Image Utility Assessment and a Relationship with Image Quality Assessment

— Source link ⬈

David M. Rouse, Romuald Pépion, Sheila S. Hemami, Patrick Le Callet

Institutions: Cornell University

Related papers:

- Image quality assessment: from error visibility to structural similarity

- Image information and visual quality

- A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms

- Multiscale structural similarity for image quality assessment

- No-Reference Image Quality Assessment in the Spatial Domain

Image Utility Assessment and a Relationship with Image Quality Assessment

David Rouse, Romuald Pépion, Sheila Hemami, Patrick Le Callet

Image Utility Assessment and a Relationship with Image Quality Assessment

David M. Rouse⋆, Romuald Pepion†, Sheila S. Hemami⋆, and Patrick Le Callet†

⋆Visual Communications Lab, School of ECE, Cornell University, Ithaca, NY 14853
†IRCCyN, Université de Nantes, Rue Christian Pauc, 44306 Nantes, France

ABSTRACT

Present quality assessment (QA) algorithms aim to generate scores for natural images consistent with subjective scores for the *quality assessment task*. For the quality assessment task, human observers evaluate a natural image based on its perceptual resemblance to a reference. Natural images communicate useful information to humans, and this paper investigates the *utility assessment task*, where human observers evaluate the usefulness of a natural image as a surrogate for a reference. Current QA algorithms implicitly assess utility insofar as an image that exhibits strong perceptual resemblance to a reference is also of high utility. However, a perceived quality score is not a proxy for a perceived utility score: a decrease in perceived quality may not affect the perceived utility. Two experiments are conducted to investigate the relationship between the quality assessment and utility assessment tasks. The results from these experiments provide evidence that any algorithm optimized to predict perceived quality scores cannot immediately predict perceived utility scores. Several QA algorithms are evaluated in terms of their ability to predict subjective scores for the quality and utility assessment tasks. Among the QA algorithms evaluated, the visual information fidelity (VIF) criterion, which is frequently reported to provide the highest correlation with perceived quality, predicted both perceived quality and utility scores reasonably. The consistent performance of VIF for both the tasks raised suspicions in light of the evidence from the psychophysical experiments. A thorough analysis of VIF revealed that it artificially emphasizes evaluations at finer image scales (i.e., higher spatial frequencies) over those at coarser image scales (i.e., lower spatial frequencies). A modified implementation of VIF, denoted VIF*, is presented that provides statistically significant improvement over VIF for the quality assessment task and statistically worse performance for the utility assessment task. A novel *utility assessment algorithm*, referred to as the natural image contour evaluation (NICE), is introduced that conducts a comparison of the contours of a test image to those of a reference image across multiple image scales to score the test image. NICE demonstrates a viable departure from traditional QA algorithms that incorporate energy-based approaches and is capable of predicting perceived utility scores.

1. INTRODUCTION

Present quality assessment‡ (QA) algorithms aim to generate scores for natural images consistent with subjective scores for the *quality assessment task*. For the quality assessment task, human observers evaluate a natural image based on its perceptual resemblance to a reference. The reference may be either an explicit, external natural image or an internal reference, only accessible to the observer. Since natural images communicate useful information to humans, this paper investigates the *utility assessment task*. For the utility assessment task, human observers evaluate the usefulness of a natural image as a surrogate for a reference.

No algorithms exist for the *utility assessment* task for *natural images*. Driven by the quality assessment task, current QA algorithms for natural images implicitly assess utility insofar as an image that exhibits strong perceptual resemblance to the reference is also of high utility. However, applications that use other, substantially constrained classes of images (e.g., scientific and medical images) evaluate the "quality" of an image in terms of the usefulness of that image to a human observer performing a task (e.g., detecting the presence of a tumor).[1] *Model observers* that aim to imitate the performance of human observers for a particular task have been designed

---

D.M.R.: E-mail: dmr58@cornell.edu; R.P.: E-mail: romuald.pepion@univ-nantes.fr, S.S.H.: E-mail: hemami@ece.cornell.edu, P.L.C.: E-mail: patrick.lecallet@univ-nantes.fr

‡Consonant with contemporary parlance among researchers investigating natural images, this paper refers to any algorithm designed to imitate human observer performance as a quality assessment algorithm.

and studied for medical and scientific images.[2] Such model observers constitute utility assessment algorithms optimized for a specific class of images but not the unconstrained set of natural images contaminated by a variety of noise sources.

A perceived quality score is not a proxy for a perceived utility score: a decrease in perceived quality may not affect the perceived utility. In applications acquiring, processing, and/or transmitting images for use by humans, system design constraints could be relaxed to accommodate a perceived utility constraint rather than a perceived quality constraint. Consider a human that uses visual information acquired by an imaging system to maneuver a remote reconnaissance vehicle. Designing the imaging system to operate at a specified quality constraint (i.e., a minimum quality threshold) demands better resources (e.g., sensor components, communication bandwidths, memory storage, etc.) even though the perceived utility of the image may not have been compromised. Designing the system in terms of perceived utility would accommodate degradations to perceived quality without sacrificing perceived utility, demanding fewer resources and, hence, reducing costs.

Our approach to study the utility assessment task for natural images consists of two parts. In the first part, two psychophysical experiments are conducted to investigate the relationship between the quality assessment and utility assessment tasks. The first experiment collects perceived quality scores for a collection of test images, since the test images used in this study include a class of images that have not been evaluated in a formal quality assessment study. The second experiment collects perceived utility scores for the same collection of test images. The results from the experiments suggest a complex relationship between quality and utility scores dependent upon the image representation. Namely, any algorithm optimized to predict perceived quality scores cannot immediately predict perceived utility scores.

The second part of this paper evaluates the ability of several QA algorithms[3–8] to predict subjective scores for either the quality or utility assessment tasks. Although the quality assessment task inspired the design of QA algorithms, the capabilities of these algorithms for the utility assessment task is unknown. Furthermore, since these algorithms seek *objective* scores consistent with *subjective* scores, a successful algorithm could model aspects of observer performance relevant to an algorithm that predicts perceived utility scores.

Among the QA algorithms evaluated, the visual information fidelity (VIF) criterion,[6] which is frequently reported to provide the highest correlation with perceived quality, predicted both perceived quality scores and utility scores reasonably. The consistent performance of VIF for both the quality and utility tasks raised suspicions in light of the conclusions drawn from the psychophysical experiments. A thorough analysis of VIF revealed that it artificially emphasizes evaluations at finer image scales (i.e., higher spatial frequencies) over those at coarser image scales (i.e., lower spatial frequencies). A modified implementation of VIF, denoted VIF*, is described that provides statistically significant improvement over VIF for the quality assessment task and statistically worse performance than VIF for the utility assessment task. The modifications adjust the weights used to pool its objective scores across image scales.

A novel *utility assessment algorithm*, referred to as the natural image contour evaluation (NICE), is introduced. NICE conducts a comparison of the contours of a test image to those of a reference image across multiple image scales to score the test image. The proposed algorithm demonstrates a viable departure from traditional QA algorithms that incorporate energy-based approaches and is capable of predicting perceived utility scores.

This paper has the following organization: Section 2 specifies two broad classes of image stimuli used to evaluate the relationship between quality and utility. Section 3 specifies the experimental methodology. Experiment results are presented in Section 4. Section 5 reviews QA algorithms evaluated in this paper. An analysis with regard to the capability of QA algorithms to predict subjective scores is provided in Section 6. Conclusions are presented in Section 7.

2. NATURAL IMAGE REPRESENTATIONS

To investigate the relationship between the quality and utility assessment tasks, test images are necessary that produce different distortions. Two broad classes of stimuli, distinguished according to the image representation model, specify the test images used. This section describes these two image representations: signal-based representations and visual-structure-preserving representations. These image representations induce distortions

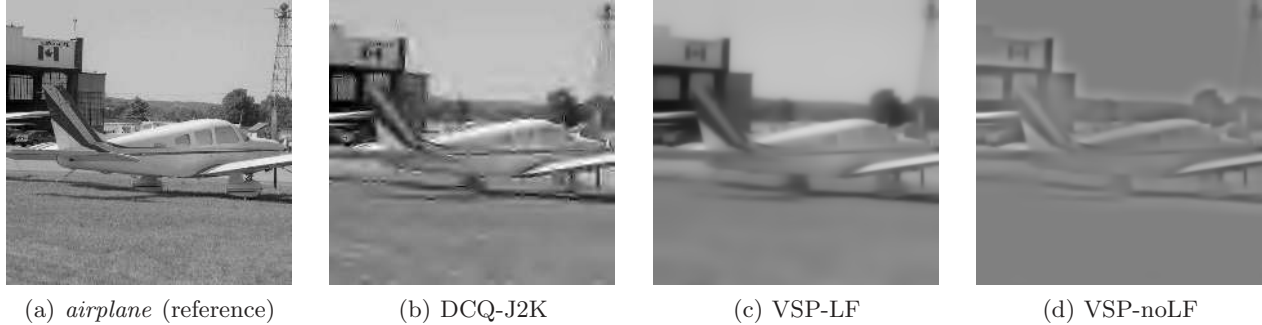|			
(a) *airplane* (reference)	(b) DCQ-J2K	(c) VSP-LF	(d) VSP-noLF

Figure 1. Original reference *airplane* and average observer recognition thresholds for two types of visual-structure-preserving (VSP) representations and the DCQ signal-based representation (DCQ-J2K). The VSP representations shown differ with regard to the inclusion of low-frequency (LF) information contained in the low-frequency residual subband.

that are spatially correlated with the natural image and disrupt different image characteristics to deteriorate the visual information.

An *image representation* specifies a sequence of images corresponding to a processing algorithm and a parameter, where subsequent images in the sequence contain additional detail or information relative to the previous images.[9] The first image recognized by human observers on average in an image representation sequence is called the *recognition threshold*. The *recognition threshold* (RT) specifies the minimum visual information of a test image that contains sufficient visual information to be a useful surrogate for the reference image. Images whose visual information exceeds that of the RT are useful to a human observer, whereas those with less visual information than the RT have little or no use. For a natural image, a recognition threshold is defined for each image representation, and the RT for each image representation conveys the same visual information from the reference image.

The signal-based (SB) representations (cf. Figure 1(b)) correspond to a class of images encountered whose distortions are induced by quantizing wavelet subband coefficients (e.g., JPEG-2000 compression). A type of signal-based representations, denoted DCQ-J2K, is used, where quantization step-sizes are assigned according to the dynamic contrast-based quantization (DCQ) algorithm.[10] DCQ computes a measure of visual distortion based on the characteristics of the image, the subband, and the display to assign subband quantization step-sizes. DCQ supports visually lossless compression. A visually lossless image is visually indistinguishable from the reference. The SB representations evolve by increasing the encoding bitrate, R.

The visual-structure-preserving (VSP) representations (cf. Figures 1(c) and 1(d)) correspond to a class of images whose texture has been removed with limited disruption to object boundaries and edges. VSP representations were obtained by implementing total variation (TV) regularization via soft-thresholding of undecimated Haar wavelet coefficients in all subbands except the low-frequency residual subband.[9, 11–13] The VSP representations may either include or exclude low-frequency (LF) signal information. Higher-frequency signal information is believed to convey salient visual information for interpretation, so VSP representations that exclude low-frequency signal information, denoted VSP-noLF, were generated that isolate object boundaries and edges. The VSP representations evolve by decreasing the soft thresholding parameter, τ.

3. METHODS: SUBJECTIVE SCORES FOR QUALITY AND UTILITY TASKS

Subjective scores for the quality assessment and utility assessment tasks were collected for test images corresponding to each image representation for five 512×512 grayscale natural images used in a previous study:[9] *airplane*, *boy & cat*, *train*, *caged birds*, *guitarist*. This section summarizes the methods used for the experiments to collect subjective responses for both assessment tasks.

Stimuli: Each image representation (i.e., DCQ-J2K, VSP-LF, and VSP-noLF) specifies a sequence of images identified by the image representation parameter value (i.e., encoding bitrate for DCQ-J2K representations and soft-thresholding parameter τ for VSP representations). A subsequence of $N = 6$ images is constructed to span the range of a given image representation (i.e., DCQ-J2K, VSP-LF, and VSP-noLF) for a reference image.

Specifically, sequences of DCQ-J2K images were generated for each natural image corresponding to bitrates logarithmically equally spaced from $R = 0.01$ to R_{VL}, where R_{VL} denotes the bitrate of a visually lossless image as specified by DCQ.[10] Sequences of VSP-LF and VSP-noLF images were generated for each natural image corresponding to $\tau = 2048, 446, 97, 21, 5, 1$.

Quality Assessment Experiment: The quality assessment experiment asks subjects to score the perceptual quality of test images from the sequences corresponding to the DCQ-J2K, VSP-LF, and VSP-noLF image representations for each reference natural image. The subjective assessment methodology for video quality (SAMVIQ) protocol provides consistent opinion scores from observers and is used for the quality assessment experiment.[14, 15] Observers provided opinion scores based on the perceived quality of the test images.

To alleviate observer fatigue due to prolonged evaluation sessions, the test images were partitioned into to two equally representative sets, creating two testing sessions. For each natural image, three anchor images, each from a different image representation and spanning the range of "quality", served as anchor images and appeared in both testing sessions to facilitate opinion score alignment across both sessions. Twenty-nine observers with normal or corrected-to-normal acuity participated in the experiment, and twenty-six opinion scores were collected for each test image (52 opinion scores were collected for the anchor images).

Utility Assessment Experiment: The utility assessment experiment asks subjects to select the appropriate image from a pair of images in response to the query "Which image tells you more about the content?" The images in the pairs are selected from the sequences of two different image representations (e.g., SB and VSP-noLF)[†]. Results from a preliminary test concluded that certain pair comparisons were unnecessary (e.g., comparing the first image of the SB sequence with the last image of the VSP-LF sequence). This reduced the number of comparisons to 54 for each natural image. Forty observers with normal or corrected-to-normal acuity participated in the experiment, and all observers provided responses for each pair of images once.

4. RESULTS: PERCEIVED QUALITY AND PERCEIVED UTILITY SCORES

This section first describes the generation of meaningful perceived quality scores and perceived utility scores from the raw experimental data collected in the experiments described in Section 3. A comparative analysis of the perceived quality scores and perceived utility scores concludes this section.

4.1 Generating Meaningful Perceived Quality Scores from the Raw Experimental Data

A meaningful perceived quality score, given as a mean opinion score (MOS), is computed for each test image from the opinion scores collected in the quality assessment experiment. A MOS is the arithmetic average of the individual observer opinion scores. A linear mapping between the MOSs of the anchor images from a specific testing session were mapped to the overall MOSs of these anchor images across both testing sessions. These mappings correct for deviations in observer opinion scores between sessions. These linear mappings were applied to the raw opinion scores for the individual sessions to generate overall MOSs for each test image.

4.2 Generating Meaningful Perceived Utility Scores from the Raw Experimental Data

The paired comparison test acquires subjective data to estimate $p_{ij} = Pr(X_i > X_j)$, the probability that stimulus X_i "tells you more about the content" than stimulus X_j. Bradley and Terry[16] specify a mathematical model that relates the p_{ij} to a continuum of raw scale values that ranks the collection of stimuli $\{X_i\}_{i=1}^n$. We obtain raw scale values for each natural using a generalized linear model, which Critchlow and Flinger[17] demonstrate is equivalent to traditional the maximum-likelihood method used by Bradley and Terry.[16]

The raw scale values order test images corresponding to a single natural image but are not aligned across natural images. Meaningful utility scores across natural images were derived from the raw scale values using a linear mapping for each natural image. The linear mapping is defined such that a utility score of 0 corresponds to a natural image's recognition threshold (RT) and a utility score of 100 corresponds to any image that is visually lossless with respect to the reference natural image. The raw scale value corresponding to the RT for each image

---

[†]It is assumed that the ordering of the image representation sequences also corresponds to the utility of the images within an image representation.

(a) Subjective scores plotted by natural image.

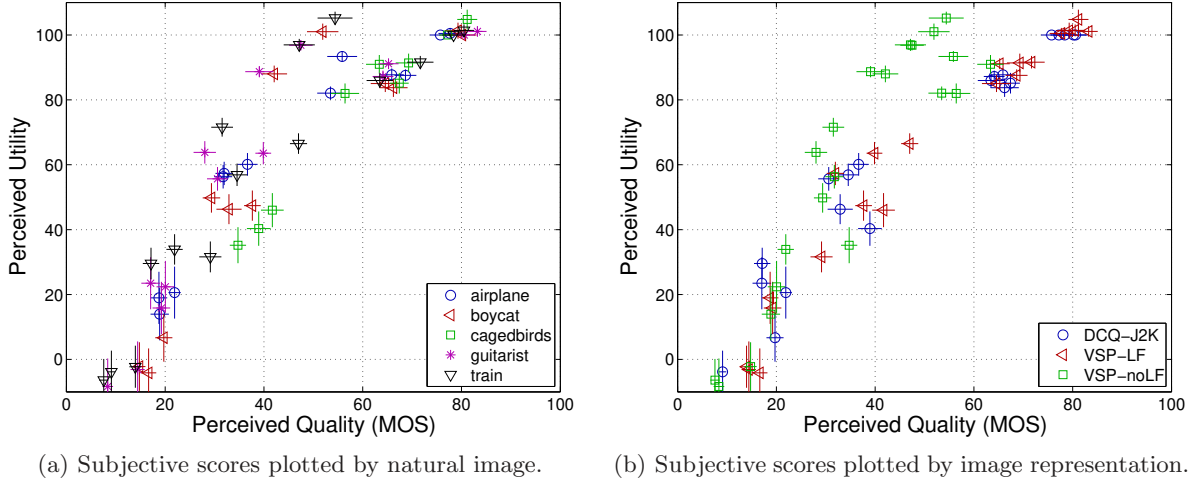(b) Subjective scores plotted by image representation.

Figure 2. Scatter plots showing the relationship between perceived utility scores and the perceived quality scores for five natural images. Standard error bars have been included for the subjective scores.

representation was linearly interpolated using the RT parameterized by the corresponding image representation parameter (i.e., R for the DCQ-J2K and τ for the VSP representations).[9] The RTs were obtained in a previous study.[9] The raw scale values corresponding to the RTs were consistent across the image representations; therefore, the interpolated raw scale values corresponding to the RT for each image representation were averaged to obtain the raw scale value corresponding to the RT of the natural image. Using the raw scale value corresponding to the RT of the natural image and the raw scale value corresponding to the DCQ-J2K visually lossless image, a linear mapping from raw scale values to meaningful utility scores is obtained. Test images with perceived utility scores below -10 were omitted from set of test images, leaving 62 test images.

4.3 Comparative Analysis Between Perceived Quality and Perceived Utility

A perceived quality score is not a proxy for a perceived utility score. Figure 2 shows the relationship between the perceived utility scores as a function of the perceived quality scores plotted by natural image and by image representation. Although the relationship appears linear, use of a linear fit to map perceived quality scores to perceived utility scores will produce poor utility estimates (RMSE = 15.1). Perceived utility exhibits an interesting relationship with perceived quality. Namely, the linear relationship between quality and utility for images with quality scores below 30 suggests that observers judge very low quality images in terms of the ability to interpret the content. Furthermore, for images with perceived quality scores above 40, the VSP-noLF images have nearly equal perceived utility scores to their VSP-LF counterparts (i.e., equal values of τ), yet many of the VSP-noLF images have significantly lower perceived quality scores (about 25 quality points lower) than their VSP-LF counterparts. In general, any algorithm optimized to predict subjective quality scores cannot immediately predict subjective utility scores.

5. FULL-REFERENCE QUALITY ASSESSMENT ALGORITHMS

This section reviews the full-reference QA algorithms investigated for their ability to predict the subjective quality scores and subjective utility scores. Full-reference quality assessment algorithms harness the availability of an explicit, external reference image $\boldsymbol{X}$ to predict the subjective score of a test image $\hat{\boldsymbol{X}}$, and serve as a tool to compare the two images. These full-reference QA algorithms can be categorized as 1) conventional distortion measures, 2) algorithms based on 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 produced objective scores consistent with perceived quality scores, this work specifically focuses on the visual information fidelity (VIF) criterion and a novel algorithm refer to as the natural image contour evaluation (NICE). A mathematical description of these two QA algorithms is included this section.

5.1 Conventional Signal Fidelity 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 fidelity. These measures evaluate fidelity solely in terms of the energy of the distortions. Root-mean-squared (RMS) distortion contrast $C_{rms}(\boldsymbol{E})$ evaluates fidelity based on the visibility of the distortions $\boldsymbol{E} = \hat{\boldsymbol{X}} - \boldsymbol{X}$ when comparing the images on a particular display device.[18]

5.2 Algorithms Based on Properties of the Human Visual System

Several quality assessment algorithms capitalize on models and principles characterizing low-level HVS properties such as contrast sensitivity,[19] contrast masking,[19–21] and perceived contrast.[22,23] These properties model the detection of a visual target (e.g., the distortions in an image) under a variety of conditions based on the contrast of the distortions. Many quality assessment algorithms have been proposed,[24–34] but this section summaries a subset that represents a variety of approaches.

The weighted SNR (WSNR) and noise quality measure (NQM) algorithms evaluate images 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.[3] The visual signal-to-noise ratio (VSNR) algorithm evaluates images according to a contrast model accounting for low-level HVS properties and the mid-level HVS property of global precedence.[7,35] The criterion 4 (C4) algorithm assesses images using elaborate models of several processing areas of the visual cortex.[8] The models in C4 describe color vision; frequency-orientation analysis; contour detection; perceptual and localization of patterns; object discrimination; and visual memory.

5.3 Algorithms based on Hypothetical Objectives of the Human Visual System

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, produces objective scores according to the resemblance of a test image's structural information (i.e., edges and object boundaries) to that of the reference image.[36] Algorithms from this family include the structural similarity (SSIM) metric,[5] the information fidelity criterion (IFC),[36] and the visual information fidelity (VIF) criterion.[6] This section first summarizes these QA algorithms. Second, this section presents the mathematical specification of VIF. This section concludes with a discussion and specification of a modified version of VIF, denoted VIF*.

5.3.1 Summary of SSIM, MS-SSIM, and VIF

The structural similarity (SSIM)[5] metric employs a local measure of spatial correlation between the pixels of the reference and test images that is modulated by distortions quantified by locally normalized first (mean) and second (variance) moments. MS-SSIM extends SSIM by evaluating this modified spatial correlation measure across several image scales.[4] Two of the authors present extended discussions of SSIM and MS-SSIM elsewhere.[9,37]

The information fidelity criterion (IFC)[36] and the visual information fidelity (VIF) criterion[6] generate objective scores based on a measurement of the mutual information between the test and reference image. Both use fundamentally Gaussian models of the wavelet coefficients of the test image and reference image which reduces the mutual information measurement to a local signal-to-noise ratio in the wavelet domain (cf. Eq. (2)).

5.3.2 Mathematical Specification of VIF

This section reviews VIF, which is an extension of IFC that incorporates a simple HVS model.[6,36] The HVS is modeled as an additive Gaussian noise source attributed to low-level HVS processing.[6] The assessment of a test image is based on spatially local SNR measurements, computed at multiple image scales, of both the reference and test images contaminated with the modeled, low-level HVS noise. Let the elements of the length N_k vectors $\boldsymbol{C}^k$ and $\boldsymbol{D}^k$ denote the wavelet coefficients of the kth subband of the reference and test images, respectively[‡]. The elements of the length N_k vectors $\boldsymbol{E}^k$ and $\boldsymbol{F}^k$ denote the wavelet coefficients of the kth subband of the respective reference and test images that have been contaminated with visual noise.

[‡]The subscript k for N_k accounts for decimated wavelet decompositions, such as the steerable pyramid, whose subbands in coarser image scales have fewer coefficients than subbands in finer image scales.

VIF parses each wavelet subband into disjoint blocks composed of P coefficients. The following discussion assumes only one subband, so the superscript k is omitted in the subsequent discussion. Let $\vec{C}_b$ and $\vec{D}_b$ correspond to the bth block of P spatially adjacent coefficients of $\boldsymbol{C}$ and $\boldsymbol{D}$, respectively. The bth block of wavelet coefficients in the subband of the reference image may be modeled as a Gaussian scale mixture[38, 39] (GSM) random vector given as $\vec{C}_b = s_b\vec{U}$, where s_b is a positive random scalar and $\vec{U}$ is a zero mean Gaussian random vector of length P with covariance $\boldsymbol{K}_{\vec{U}}$. Given s_b, the coefficient block $\vec{C}_b$ is a zero mean Gaussian random scalar with covariance $s_b^2 \boldsymbol{K}_{\vec{U}}$, and $\vec{C}_b$ is conditionally independent of $\vec{C}_m$ for all $m \neq b$. VIF relates the bth block of wavelet coefficient of the test and reference images according to the linear model $\vec{D}_b = g_b\vec{C}_b + \vec{V}_b$, where g_b is a deterministic scalar defined for each block and $\vec{V}_b$ is a zero mean Gaussian random vector of length P with covariance matrix $\sigma_{\vec{V}_b}^2 \boldsymbol{I}$ specified for each block b. Thus, given s_b, the block of coefficients $\vec{D}_b$ is also a Gaussian random vector with covariance $g_b^2 s_b^2 \boldsymbol{K}_{\vec{U}} + \sigma_{\vec{V}_b}\boldsymbol{I}$.

Independent zero-mean additive Gaussian noise sources model low-level HVS noise in VIF; coefficients of the reference and test images are contaminated with visual noise. Let $\vec{E}_b$ and $\vec{F}_b$ correspond to the bth block of P spatially adjacent coefficients of $\boldsymbol{E}$ and $\boldsymbol{F}$, respectively. The output of the HVS model for the reference image is $\vec{E}_b = \vec{C}_b + \vec{M}_b$, and the output of the HVS model for the test image is $\vec{F}_b = \vec{D}_b + \vec{N}_b$. The terms $\vec{M}_b$ and $\vec{N}_b$ are a zero mean Gaussian random vectors of length P with covariance $\sigma_M^2\boldsymbol{I} = \sigma_N^2\boldsymbol{I}$, where $\sigma_N^2 = \sigma_M^2$ is the HVS model parameter. Thus, given s_b, the block of coefficients $\vec{E}_b$ is a Gaussian random vector with covariance $s_b^2 \boldsymbol{K}_{\vec{U}} + \sigma_N^2\boldsymbol{I}$, and the block of coefficients $\vec{F}_b$ is also a Gaussian random vector with covariance $g_b^2 s_b^2 \boldsymbol{K}_{\vec{U}} + \sigma_{\vec{V}_b}^2\boldsymbol{I} + \sigma_N^2\boldsymbol{I}$

VIF combines two evaluations to yield an overall assessment of a test image. First, an evaluation comparing the reference coefficients before and after the HVS model value is computed. Second, an evaluation comparing the reference coefficients before the HVS model to the processed coefficients after the HVS model is computed. These two evaluations are computed for each wavelet subband. The ratio of the sum of these evaluations across the subbands provides an overall assessment of the test image. Let $\boldsymbol{s}$ be a length B_k vector whose b^{th} element is s_b. Given $\boldsymbol{s}$, the VIF value is given by

$$\text{VIF} = \frac{\sum_{k=1}^{K} IFC(\boldsymbol{C}^k, \boldsymbol{F}^k)}{\sum_{k=1}^{K} IFC(\boldsymbol{C}^k, \boldsymbol{E}^k)}. \tag{1}$$

The terms $IFC(\boldsymbol{C}^k, \boldsymbol{F}^k)$ and $IFC(\boldsymbol{C}^k, \boldsymbol{E}^k)$ are based on IFC[36] and are defined as

$$IFC(\boldsymbol{C}^k, \boldsymbol{F}^k) = \sum_{b=1}^{B_k} \log_2\left(\frac{|g_b^2 s_b^2 \boldsymbol{K}_{\vec{U}} + (\sigma_{\vec{V}_b}^2 + \sigma_N^2)\boldsymbol{I}|}{|(\sigma_{\vec{V}_b}^2 + \sigma_N^2)\boldsymbol{I}|}\right) \qquad IFC(\boldsymbol{C}^k, \boldsymbol{E}^k) = \sum_{b=1}^{B_k} \log_2\left(\frac{|s_b^2 \boldsymbol{K}_{\vec{U}} + \sigma_N^2\boldsymbol{I}|}{|\sigma_N^2\boldsymbol{I}|}\right), \quad (2)$$

where $|\cdot|$ denotes the matrix determinant and the terms g_b, s_b, $\boldsymbol{K}_{\vec{U}}$, and $\sigma_{\vec{V}_b}$ vary with k and are computed from $\boldsymbol{C}^k$ and $\boldsymbol{D}^k$. For subband k, the term g_b is estimated as the linear regression of block $\vec{D}_b$ on the block $\vec{C}_b$, and the variance of the additive zero mean Gaussian noise $\vec{V}_b$ is the mean squared error of the regression.

5.3.3 VIF* Specification

VIF artificially emphasizes evaluations at finer image scales (i.e., higher spatial frequencies) over those at coarser image scales (i.e., lower spatial frequencies). Thus, VIF is invariant to changes to low spatial frequency energy, which distinguishes the VSP-LF and VSP-noLF representations. Figure 3(a) demonstrates this characteristic of VIF for the *airplane* image under the DCQ-J2K (MOS = 65.9), VSP-LF (MOS = 68.6), and VSP-noLF (MOS = 53.5) image representations. In particular, Figure 3(a) shows the values of $IFC(\boldsymbol{C}^k, \boldsymbol{F}^k)$ (i.e., the numerator[§] of VIF) for the subband capturing horizontal edges as a function of image scale k. The values of $IFC(\boldsymbol{C}^k, \boldsymbol{F}^k)$ corresponding to finer scales are an order of magnitude larger than those at coarser scales. Thus, VIF cannot produce distinct scores that reflect the changes in the perceived quality scores for these images.

The problem identified with VIF (cf. Figure 3) is due to the variation in the number of coefficients blocks B_k for subbands at different image scales. Specifically, subbands corresponding to finer image scales have more

---

[§]All of these test images correspond to the same reference image, so the denominator $IFC(\boldsymbol{C}^k; \boldsymbol{E}^k)$ is irrelevant to a discussion of VIF and merely serves as a scale factor unique to each reference image.

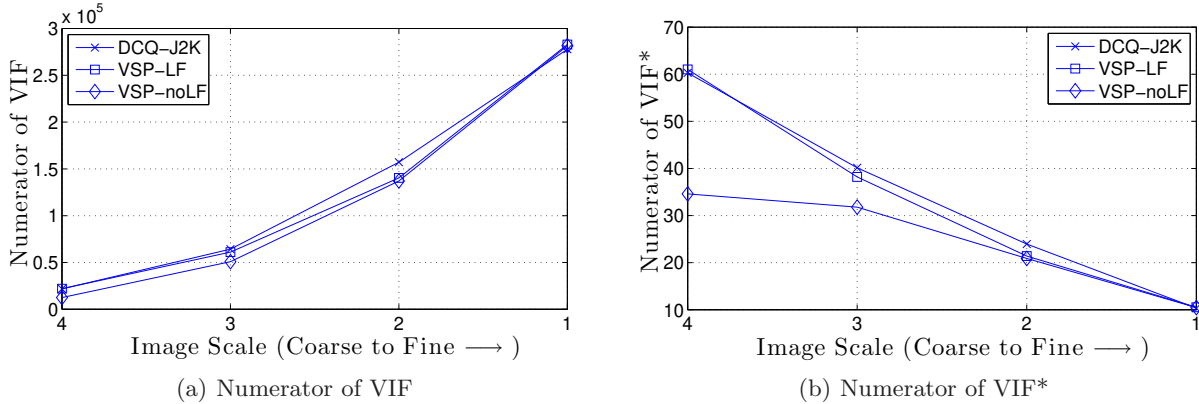(a) Numerator of VIF   (b) Numerator of VIF*

Figure 3. VIF artificially emphasizes evaluations at finer image scales (i.e., higher spatial frequencies) over those at coarser image scales (i.e., lower spatial frequencies). Thus, VIF is invariant to changes to low spatial frequency energy, which distinguishes the VSP-LF and VSP-noLF representations. Figure 3(a) demonstrates this characteristic of VIF for the *airplane* image under the DCQ-J2K (MOS = 65.9), VSP-LF (MOS = 68.6), and VSP-noLF (MOS = 53.5) image representations. The value of the numerator of VIF for the subband capturing horizontal edges is shown at different image scales. VIF* resolves this problem with VIF and is demonstrated in Figure 3(b) for the same three images.

wavelet coefficients that subbands corresponding to coarser image scales due to the use of a decimated wavelet transform; for a fixed block size P, the number of coefficient blocks is smaller for subbands corresponding to coarser image scales. A modified definition of VIF, denoted VIF*, is given that scales the calculation for each subband by the number of blocks B_k for that subband. VIF* is given as

$$\text{VIF}^* = \frac{\sum_{k=1}^{K} \frac{1}{B_k} IFC(\boldsymbol{C}^k, \boldsymbol{F}^k)}{\sum_{k=1}^{K} \frac{1}{B_k} IFC(\boldsymbol{C}^k; \boldsymbol{E}^k)}, \tag{3}$$

where $IFC(\boldsymbol{C}^k; \boldsymbol{F}^k)$ and $IFC(\boldsymbol{C}^k; \boldsymbol{E}^k)$ are defined as in Eq. (2). Figure 3(b) shows the value of $\frac{1}{B_k} IFC(\boldsymbol{C}^k; \boldsymbol{F}^k)$ (i.e., the numerator of VIF*) for the subbands capturing horizontal edges for different image scales for the same three images corresponding to the *airplane* image described in Figure 3(a). VIF* reveals greater differences between the images of the VSP-LF and VSP-noLF representations corresponding to the coarsest scale (scale 4). Thus, VIF* produces distinct scores that reflect the changes in the perceived quality scores for these images.

5.4 NICE: Natural Image Contour Evaluation

This section describes the proposed natural image contour evaluation (NICE) algorithm for utility assessment. NICE compares the contours of a test image to that of a reference image across multiple image scales to produce a numerical score of the test image. This novel algorithm demonstrates a viable departure from traditional QA algorithms that incorporate energy-based approaches.

A wavelet representation of an image provides multiscale directional derivatives of that image, which can be used to identify image contours at different image scales. For NICE, both the reference and test image are represented using an *undecimated* implementation of the steerable pyramid[40] using D orientations and S scales[¶]. Let $W_{s,\theta}(i)$ and $\hat{W}_{s,\theta}(i)$ denote the ith wavelet coefficient of the respective reference and test images in the subband corresponding to scale $s \in \{1, 2, \ldots, S\}$ and orientation $\theta \in \{0, \frac{\pi}{D}, \frac{2\pi}{D}, \ldots, \frac{\pi(D-1)}{D}\}$. The remainder of this section describes the algorithm to identify image contours and the comparison of those contours to generate a NICE score for the test image.

5.4.1 Image Contour Identification from Local Modulus Maxima

For each image scale s, the local modulus maxima[41] of wavelet coefficient scales are used to identify image contours for the reference and test images. The local modulus maxima are determined from gradient vectors

---

[¶]The high-pass residual generated by the steerable pyramid is not used to compute the NICE score.

formed from wavelet subbands corresponding to derivatives in horizontal and vertical spatial directions.[41] Define $G_s(i) = W_{s,0}(i) - jW_{s,\frac{\pi}{2}}(i)$ and $\hat{G}_s(i) = \hat{W}_{s,0}(i) - j\hat{W}_{s,\frac{\pi}{2}}(i)$ as the gradient of the respective reference and test images at scale s, where $j = \sqrt{-1}$. For image scale s, let $M_s(i) = |G_s(i)|$ and $A_s(i) = \angle G_s(i)$ denote the respective modulus and angle of the gradient of the reference image. Similarly, define $\hat{M}_s(i) = |\hat{G}_s(i)|$ and $\hat{A}_s(i) = \angle \hat{G}_s(i)$ for the test image. Local modulus maxima of the reference image correspond to points of $M_s(i)$ greater than the two adjacent neighbors in the direction indicated by $A_s(i)$, and for the test image, local modulus maxima are similarly identified using $\hat{M}_s(i)$ and $\hat{A}_s(i)$. Let $\mathcal{I}_s$ denote the set of indices i corresponding to local modulus maxima of the reference image at scale s, and let $\hat{\mathcal{I}}_s$ denote the set of indices i corresponding to local modulus maxima of the test image at scale s.

Binary images represent image contours of the reference and test images. The image contours at scale s of the reference and test images are identified as local modulus maxima that exceed a threshold β_s based on the reference image and is given as

$$\beta_s = \frac{1}{p} \max_i M_s(i),\tag{4}$$

for some scalar $p > 0$. Specifically, define $B_s(i)$ and $\hat{B}_s(i)$ as

$$B_s(i) = \begin{cases} 1 & M_s(i) > \beta_s \text{ and } i \in \mathcal{I}_s \\ 0 & else \end{cases} \qquad \hat{B}_s(i) = \begin{cases} 1 & \hat{M}_s(i) > \beta_s \text{ and } i \in \hat{\mathcal{I}}_s \\ 0 & else \end{cases}.\tag{5}$$

5.4.2 NICE Score

An objective score based for NICE is computed by comparing the contours of the reference and test images at each image scale s, represented as the respective binary images $B_s(i)$ and $\hat{B}_s(i)$ defined in Section 5.4.1. The binary images $B_s(i)$ and $\hat{B}_s(i)$ are subjected to morphological dilation[42] with a 3×3 "plus-sign" shaped structuring element, and the point-wise exclusive-or operation of the dilated binary images produces the binary image $E_s(i)$. The overall NICE score for the test image is computed as

$$\text{NICE} = \sum_{s=1}^{S} a_s N(s),\tag{6}$$

where $N(s)$ denotes the number of non-zero elements of E_s and the $\{a_s\}_{s=1}^{S}$ are nonnegative scalars.

6. ANALYSIS AND DISCUSSION

Although the psychophysical evidence presented in Section 4 establishes a complex relationship between scores for quality task and utility task, quality assessment (QA) algorithms, which are typically developed for the quality task, could be useful to predict subjective utility scores. The QA algorithms and the proposed utility assessment algorithm described in Section 5 are evaluated in terms of the their capability to predict subjective scores of test images for both the quality and utility assessment tasks. VIF, VIF*, and NICE use a four-scale steerable pyramid with six orientations, but only subbands corresponding to horizontal and vertical frequencies are used. The implementations for most of the algorithms were obtained from the respective authors of the algorithms.

Objective scores generated by QA algorithms frequently exhibit a nonlinear relationship with subjective scores. Objective scores are fitted to subjective scores with a third order polynomial subject to a monotonicity constraint. The fitted objective scores are evaluated with respect to the subjective scores using the Spearman rank order correlation coefficient (ROCC), the squared Pearson (linear) correlation coefficient r^2, the root mean squared error (RMSE), an F-test to individually compare the residual variance of NICE to the other algorithms, and the outlier ratio (OR). The OR is the proportion of fitted objectives scores that differ from the subjective scores by more than twice the subjective score's estimated standard error.

An F-test determines whether the residual variance of algorithm is statistically larger or smaller than the other.[43] Suppose σ_A^2 and σ_B^2 denote the variance of the residuals corresponding to algorithms A and B when

Table 1. Results summarizing the performance of quality assessment algorithms and proposed NICE utility assessment for the quality assessment and utility assessment tasks. NICE is designed for the utility assessment task and *not* the quality assessment task. Results on its performance for the quality assessment task demonstrate the divergence between performance for the quality and utility assessment tasks when utility is predicted well (cf. Section 4.3). Objective scores, fitted to subjective scores with a monotonic third order polynomial, are evaluated with respect to the subjective scores using the Spearman rank order correlation coefficient (ROCC), the squared Pearson (linear) correlation coefficient r^2, the root mean squared error (RMSE), an F-test to individually compare the residual variance of NICE to the other algorithms, and the outlier ratio (OR). Values of $F_{statistic} > F_{critical} = 1.53$ (or $F_{statistic} < 1/F_{critical} = 0.654$) have statistically larger (or smaller) residual variances than the proposed algorithm at the 95% confidence level.

Algorithm	Quality Task					Utility Task				
	r^2	ROCC	RMSE	$F_{statistic}$	OR	r^2	ROCC	RMSE	$F_{statistic}$	OR
PSNR	0.624	0.700	14.2	4.52	0.790	0.191	0.471	43.5	23.7	0.887
$C_{rms}(\boldsymbol{E})$	0.424	0.691	25.9	15.1	0.935	0.205	0.510	50.7	32.1	0.984
WSNR	0.508	0.663	16.2	5.92	0.871	0.187	0.425	32.8	13.5	0.839
NQM	0.415	0.644	17.7	7.05	0.871	0.318	0.467	35.4	15.7	0.774
VSNR	0.636	0.660	19.9	8.91	0.903	0.371	0.473	31.8	12.7	0.742
C4	0.788	0.848	10.6	2.56	0.597	0.526	0.676	25.0	7.82	0.742
SSIM	0.904	0.933	7.15	1.15	0.532	0.749	0.871	18.2	4.13	0.532
MS-SSIM	0.748	0.871	11.6	3.03	0.742	0.566	0.726	23.9	7.16	0.774
VIF	0.882	0.942	7.94	1.42	0.645	0.959	0.969	7.32	0.671	0.355
VIF*	0.967	0.974	4.22	0.402	0.306	0.882	0.912	12.5	1.95	0.565
NICE	0.917	0.952	6.66	1	0.516	0.939	0.944	8.94	1	0.484

used to predict subjective scores. Using the statistic $F_{statistic} = \frac{\sigma_A^2}{\sigma_B^2}$, the F-test computes a critical value $F_{critical}$ based on the number of residuals and the confidence level $(1-\alpha)\%$ of the test, and values of $F_{statistic} > F_{critical}$ (or $F_{statistic} < 1/F_{critical}$) have statistically larger (or smaller) residual variances than the proposed algorithm at the $(1-\alpha)\%$ confidence level. For a 95% confidence level and 62 test images, $F_{critical} = 1.53$.

This section first evaluates the ability of the QA algorithms and the proposed NICE utility assessment algorithm to the predict subjective scores for the quality and utility assessment tasks. NICE is designed for the utility assessment task and *not* the quality assessment task. Results on its performance for the quality assessment task demonstrate the divergence between performance for the quality and utility assessment tasks when utility is predicted well (cf. Section 4.3). Second, this section investigates the benefits of appropriate pooling across image scale evaluations with VIF and NICE to predict subjective scores for both assessment tasks.

6.1 Predicting Perceived Quality Scores and Perceived Utility Scores with Quality Assessment Algorithms

Among the QA algorithms evaluated, VIF* provides significantly smaller errors (RMSE = 4.22) for predicting perceived quality scores than the other QA algorithms. Table 1 summarizes the results from a statistical analysis of the fitted QA algorithm scores with the perceived quality scores This significance of the improvement with VIF* was established using an F-test to compare the residual variances of the fitted VIF* scores to the fitted scores of other algorithms. The performance improvement with VIF* for the quality task, validates the proposed modifications to VIF.

Overall, the QA algorithms based on hypothetical properties of the HVS and the proposed NICE utility assessment algorithm provide better predict perceived quality scores than the other classes of algorithms. An inspection of the QA algorithms classified as either conventional signal fidelity measures or based on properties of the HVS revealed that these algorithms assigned relatively low scores to images corresponding to the VSP-noLF representation, which suggests that they overemphasize the relevance of distortions to low spatial frequencies.

For the utility assessment task, both VIF and NICE provide significantly smaller errors (RMSE = 7.32 and 8.94, respectively) for predicting perceived utility scores than the other algorithms. Table 1 summarizes the statistical analysis of the fitted algorithm scores with the perceived utility scores. The variance of the residuals corresponding to fitted objective scores produced by VIF and NICE were found to be statistically equivalent

Table 2. A summary of the prediction abilities of VIF and the proposed NICE utility assessment algorithm for the quality and utility assessment tasks using individual image scale objective scores and the minimum mean squared error (MMSE) linear estimator that combines objective scores from each image scale. Refer to Table 1 for a description of the column acronyms.

Algorithm	Quality Task				Utility Task			
	r^2	ROCC	RMSE	OR	r^2	ROCC	RMSE	OR
VIF (scale 1)	0.861	0.911	8.61	0.597	0.959	0.966	7.35	0.371
VIF (scale 2)	0.895	0.950	7.48	0.532	0.959	0.961	7.32	0.387
VIF (scale 3)	0.967	0.978	4.22	0.339	0.917	0.920	10.5	0.532
VIF (scale 4)	0.802	0.876	10.3	0.613	0.602	0.718	22.9	0.710
VIF (MMSE)	0.971	0.979	4.01	0.323	0.962	0.970	7.13	0.371
NICE (scale 1)	0.788	0.882	10.6	0.855	0.844	0.930	14.3	0.581
NICE (scale 2)	0.893	0.943	7.56	0.581	0.954	0.958	7.82	0.419
NICE (scale 3)	0.854	0.922	8.81	0.597	0.805	0.847	16.0	0.629
NICE (scale 4)	0.676	0.773	13.2	0.790	0.476	0.608	26.9	0.790
NICE (MMSE)	0.923	0.958	6.67	0.581	0.963	0.966	7.04	0.387

Table 3. Weights to linearly combine objective score from each image scale found by minimizing the mean squared error between the combined objective scores and subjective scores for each task.

Algorithm	Quality Task				Utility Task			
	scale 1	scale 2	scale 3	scale 4	scale 1	scale 2	scale 3	scale 4
VIF (MMSE)	0.46	-0.39	1	0.10	0.74	1	-0.18	0.06
NICE (MMSE)	0.38	1	0.68	0.26	0.02	1	-0.25	0.02

based on an F-test. The inability of VIF*, as well as the other QA algorithms, to produce accurate objective utility scores indicates that an algorithm capable of predicting perceived utility scores should not overemphasize degradations to energy at lower spatial frequencies.

6.2 Improving Pooling Strategies Across Image Scales to Predict Subjective Scores

The HVS has been characterized as a multi-scale analyzer of visual data.[19] Thus, many QA algorithms evaluate a test image with respect to a reference image at multiple image scales. In light of the improvement of VIF* over VIF for the quality task, this section demonstrates the benefits of properly pooling multi-scale evaluations of VIF and NICE using the optimal linear estimator of the subjective scores given the objective scores for each image scale. This approach is compared to using only single image scale objective scores.

Given objective scores computed for each image scale of either VIF[†] or NICE, the optimal linear estimator of the subjective scores is found after fitting the single scale objective scores to the subjective scores with a third order polynomial subject to a monotonicity constraint. The polynomial fit resolves the nonlinear relationship between the objective scores and the subjective scores.

The optimal pooling weights improve the predictive capabilities of VIF for both the quality task (RMSE = 4.01) and the utility task (RMSE = 7.13); the optimal pooling weights improve NICE for the the utility task (RMSE = 7.04) and exhibit essentially the same performance for the quality task. For both algorithms, the optimal weights differ for each assessment task. Table 2 summarizes the performance analysis of VIF and NICE using individual image scale objective scores and the minimum mean squared error (MMSE) linear estimator that combines objective scores from each image scale. Table 3 lists the optimal weights for VIF and NICE for the quality and utility task. The weights have been normalized such that the absolute maximum weight for each algorithm and task is one.

With only 62 test images available, the number of parameters used to fit the objective scores to the subjective data restrict the inferences one can make from these results. Each nonlinear fit requires four parameters for each scale, and the addition of the four weights yields to a total of 20 parameters. Despite this shortcoming, these results advocate the potential benefits of combining objective scores across image scales.

[†]The modifications of VIF that specify VIF* do not change the results using single image scale objective scores.

7. CONCLUSIONS AND FUTURE WORK

This paper investigates the relationship between the quality assessment and utility assessment tasks for natural images. Two psychophysical experiments are conducted to collect perceived quality scores and perceived utility scores for a collection of test images corresponding to signal-based representations and visual-structure-preserving representations. The results from these experiments provide evidence that any quality assessment (QA) algorithm optimized to predict perceived quality scores cannot immediately predict perceived utility scores.

Several QA algorithms are evaluated in terms of their ability to predict subjective scores for the quality and utility assessment tasks. An analysis of VIF, instigated by its suspiciously consistent performance for both tasks, revealed that it artificially emphasized evaluations at finer image scales (i.e., higher spatial frequencies) over those at coarser image scales (i.e., lower spatial frequencies). A proposed implementation of VIF, denoted VIF*, demonstrates statistically significant improvement over VIF for the quality assessment task and statistically worse performance for the utility assessment task. NICE, a novel utility assessment algorithm is introduced that conducts a comparison of the contours of a test image to those of a reference image across multiple image scales to score the test image. NICE demonstrates a viable departure from traditional QA algorithms that incorporate energy-based approaches and is capable of predicting perceived utility scores.

This investigation established several encouraging areas of future work. First, while formulating a relationship between perceived quality scores and perceived utility scores based on image representations is useful, it is hypothesized that a more appropriate relationship between could be modeled as the perceptual change to an image when removing lower spatial frequency content. Second, though NICE establishes a novel framework that provides reasonable prediction of perceived utility scores, a rigorous evaluation of its parameters and components could yield significant improvements. Last, the demonstrated benefits of linear pooling to combine objective scores across image scales urge an extensive analysis of other pooling strategies.

8. ACKNOWLEDGEMENTS

REFERENCES

1. H. Barrett, K. Myers, N. Devaney, and C. Dainty, "Objective assessment of image quality IV: Application to adaptive optics," *J. Opt. Soc. of Am. A* **23**(12), pp. 3080–3105, 2006.
2. H. Barrett, J. Yao, J. Rolland, and K. Myers, "Model observers for assessment of image quality," in *Proc. of the National Academy of Sciences of the USA*, **90**, pp. 9758–9756, Feb. 1993.
3. N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Process.* **9**, pp. 636–650, Apr. 2000.
4. Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proc. of the 37th IEEE Asilomar Conf. on Sig., Sys. and Comp.*, (Pacific Grove, CA), Nov. 2003.
5. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.* **13**, pp. 600–612, Apr. 2004.
6. H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.* **15**, pp. 430–444, Feb. 2006.
7. D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.* **16**, pp. 2284–2298, Sept. 2007.
8. M. Carnec, P. Le Callet, and D. Barba, "Objective quality assessment of color images based on a generic perceptual reduced reference," *Signal Processing: Image Communication* , 2008. doi:10.1016/j.image.2008.02.003.
9. D. M. Rouse and S. S. Hemami, "Analyzing the role of visual structure in the recognition of natural image content with multi-scale SSIM," in *Proc. SPIE: HVEI XIII*, B. E. Rogowitz and T. N. Pappas, eds., **6806**, (San Jose, CA), Jan. 2008.
10. D. M. Chandler and S. S. Hemami, "Dynamic contrast-based quantization for lossy wavelet image compression," *IEEE Trans. Image Process.* **14**, pp. 397–410, Apr. 2005.

11. L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noised removal algorithm," *Physica D* **60**, pp. 259–268, 1992.

12. G. Steidl, J. Weickert, T. Brox, P. Mrazek, and M. Welk, "On the equivalence of soft wavelet shrinkage, total variation diffusion, total variation regularization, and SIDEs," *SIAM J. of Numerical Analysis* **42**(2), pp. 686–713, 2004.

13. J.-L. Starck, M. Elad, and D. L. Donoho, "Image decomposition via the combination of sparse representations and a variational approach," *IEEE Trans. Image Process.* **14**, pp. 1570–1582, Oct. 2005.

14. "SAMVIQ - subjective assessment methodology for video quality," Tech. Rep. BPN 056, European Broadcast Union (EBU), May 2003.

15. F. Kozamernik, P. Sunna, E. Wyckens, and D. I. Pettersen, "Subjective quality of internet video codecs phase ii evaluations using SAMVIQ." EBU Technical Review, Jan. 2005.

16. R. A. Bradley and M. E. Terry, "The rank analysis of incomplete block designs I: The method of paired comparisons," *Biometrika* **39**, pp. 324–345, 1952.

17. D. E. Critchlow and M. A. Fligner, "Paired comparisons, triple comparisons, and ranking experiments as generalized linear models, and their implementation on glim," *Psychometrika* **56**(3), pp. 517–533, 1991.

18. C. Poynton, "The rehabilitation of gamma," in *Proc. SPIE: HVEI III*, B. E. Rogowitz and T. N. Pappas, eds., (San Jose, CA), 1998.

19. R. L. De Valois and K. K. De Valois, *Spatial Vision*, Oxford University Press, New York, 1990.

20. G. Legge and J. Foley, "Contrast masking in human vision," *J. Opt. Soc. Am.* **70**, pp. 1458–1470, 1980.

21. D. M. Chandler and S. S. Hemami, "Effects of natural images on the detectability of simple and compound wavelet subband quantization distortions," *J. Optical Soc. of Am. A* **20**, July 2003.

22. M. A. Georgeson and G. D. Sullivan, "Contrast constancy: Debluring in human vision by spatial frequency channels," *Journal of Physiology* **252**, pp. 627–656, 1975.

23. N. Brady and D. J. Field, "What's constant in contrast constancy? The effects of scaling on the perceived contrast of bandpass patterns," *Vision Research* **35**(6), pp. 739–756, 1995.

24. T. Stockham, "Image processing in the context of a visual model," *Proc. IEEE* **60**(7), pp. 828–842, 1972.

25. J. L. Mannos, "The effects of a visual fidelity criterion on the encoding of images," *IEEE Trans. Inf. Theory* **20**(4), pp. 525–536, 1974.

26. W. A. Pearlman, "A visual system model and a new distortion measure in the context of image processing," *J. Optical Soc. Am.* **68**(3), pp. 374–386, 1978.

27. D. Granrath, "The role of human visual models in image processing," *Proc. IEEE* **69**(5), pp. 552–561, 1981.

28. R. J. Safranek and J. D. Johnston, "A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression," in *Proc. of the IEEE Intl. Conf. on Acoustics, Speech, and Sig. Processing*, **3**, pp. 1945–1948, (Glasgow, Scotland), May 1989.

29. S. J. Daly, "The visible difference predictor: An algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, A. B. Watson, ed., ch. 14, pp. 179–206, MIT Press, Cambridge, MA, 1993.

30. J. Lubin, "The use of psychophysical data and models in the analysis of display system performance," in *Digital Images and Human Vision*, A. B. Watson, ed., ch. 13, pp. 163–178, MIT Press, Cambridge, MA, 1993.

31. A. B. Watson, "DCT quantization matrices visually optimized for individual images," in *Proc. SPIE: Human Vision, Visual Process., and Dig. Display IV*, B. E. Rogowitz and J. P. Allebach, eds., **1913**, pp. 202–216, (San Jose, CA), Feb. 1993.

32. P. Teo and D. Heeger, "Perceptual image distortion," in *Proc. SPIE: Human Vision, Visual Process., and Digital Display V*, B. E. Rogowitz and J. P. Allebach, eds., **2179**, pp. 127–141, (San Jose, CA), Feb. 1994.

33. A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Process.* **6**(8), pp. 1164–1175, 1997.

34. T. N. Pappas and R. J. Safranek, "Perceptual criteria for image quality evaluation," in *Handbook of Image and Video Processing*, A. C. Bovik, ed., Academic, New York, 2000.

35. D. Navon, "Forest before trees: The precedence of global features in visual perception," *Cognitive Psychology* **9**, pp. 353–383, 1977.

36. H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.* **14**(12), pp. 2117–2128, 2005.

37. D. M. Rouse and S. S. Hemami, "Understanding and simplifying the structural similarity metric," in *Proc. IEEE Int. Conf. on Image Process. (ICIP)*, (San Diego, CA), Oct. 2008.

38. M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of gaussians and the statistics of natural images," in *Advances in Neural Information Processing Systems*, S. A. Solla, T. K. Leen, and K.-R. Miller, eds., pp. 855–861, MIT Press, 2000.

39. M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky, "Random cascades on wavelet trees and their use in analyzing and modeling natural images," *Applied and Comp. Harmonic Analysis* **11**, pp. 89–123, 2001.

40. E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in *Proc. IEEE Intl. Conf. on Image Process.*, (Washington, D.C.), Oct. 1995.

41. S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. Image Process.* **14**(7), pp. 710–732, 1992.

42. C. Giardina and E. Dougherty, *Morphological Methods in Image and Signal Process.*, Prentice Hall, 1998.

43. J. L. Devore, *Probability and Statistics for Engineering and the Sciences*, Duxbury, fifth ed., 2000.