

THE ROLE OF EDGE INFORMATION TO ESTIMATE THE PERCEIVED UTILITY OF NATURAL IMAGES

David M. Rouse and Sheila S. Hemami

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

ABSTRACT

In the quality assessment task, observers evaluate a natural image based on its perceptual resemblance to a reference. For the utility assessment task, observers evaluate the usefulness of a natural image as a surrogate for a reference. Humans are willing to use the information captured by an imaging system and tolerate distortions as long as the underlying task is performed reliably. Conventional notions of perceived quality cannot generally predict the perceived utility of a natural image. Many estimators have been designed to estimate perceived quality scores, and a recent estimator, referred to as the natural image contour evaluation (NICE), has been developed to estimate perceived utility scores. An analysis of edge information in natural images drives object recognition mechanisms in the human visual system, so this paper examines the role of an edge-based analysis in both popular objective quality estimators and NICE for estimating the perceived quality or utility of distorted natural images. Among the estimators evaluated, results show that estimators that emphasize an edge-based analysis provide the most accurate estimates of perceived utility scores. Estimators that augment the edge-based analysis with an energy-based analysis provide the most accurate estimates of perceived quality scores.

1. INTRODUCTION

In many imaging applications, humans are willing to use the information captured by an imaging system and tolerate distortions as long as the underlying task is performed reliably. The distortions encountered could be characterized using conventional notions of perceptual quality, which has been largely studied in the context of consumer applications. However, many applications value an assessment of distorted images according to their usefulness, or *utility*, over their perceptual *quality*. For instance, the public safety sector (e.g., law enforcement, fire control, and emergency services) and the military use imaging systems in real-time tactical scenarios to make immediate decisions on how best to respond to an incident [1, 2]. Frequently, the imaging system generates images by sensing energy at wavelengths outside of the visible spectrum of light. For example, firefighters use thermal imaging cameras to locate hot-spots in a burning structure [2]. In another example, both law enforcement and the military use infrared cameras in night-time surveillance and reconnaissance applications [2, 3]. Merely employing an existing quality estimator to predict the utility of such images is both insufficient and inappropriate, because a perceived quality score is not a proxy for a perceived utility score (cf. Figure 2). A decrease in perceived quality may not affect the perceived utility (cf. Figure 1).

Present quality estimators 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, it is often relevant to consider the *utility assessment task*. For the utility assessment task, human observers evaluate the usefulness of a natural image as a surrogate for a reference.

Many objective estimators have been designed to estimate perceived quality scores, and a recent estimator, referred to as the natural image contour evaluation (NICE), has been developed to estimate perceived utility scores. Objective estimators employ various methods to analyze image features to estimate a subjective score, and these methods reflect a tradeoff between an analysis of image information via an edge-based or an energy-based approach. Edge-based approaches conduct an analysis of high-frequency signal components, whereas energy-based approaches (e.g., mean-square error) compute the second moment of a signal. This paper examines the role of these two analyses in objective estimators for estimating the perceived quality or utility of distorted natural images.

This paper has the following organization: Section 2 discusses the differences in perceived quality and perceived utility using an image database with subjective scores for each task. Section 3 summarizes full-reference methods to objectively estimate subjective scores of distorted natural images. A summary of the capabilities of these objective estimators to predict perceived quality and perceived utility scores is presented in Section 4. This summary is followed by a discussion of the role of an energy-based and edge-based analyses by objective estimators to estimate perceived quality and perceived utility scores. Conclusions are presented in Section 5.

2. PERCEIVED QUALITY AND PERCEIVED UTILITY

The CU-Nantes image database is a collection of distorted natural images for which both perceived quality scores and perceived utility scores have been recorded [4]. The database consists of 5 reference grayscale images and 90 test images that were generated from the reference images. The test images correspond to image representations investigated in a previous study by the authors (cf. Figure 1) [4]: signal-based (SB) and boundary-preserving (BP). These image representations induce distortions that are spatially correlated with the natural image and disrupt different image characteristics to deteriorate the visual information. Inspired by low-level HVS models, the SB representation corresponds to a class of images whose distortions are induced by quantizing wavelet subband coefficients (e.g., JPEG-2000 compression). The BP representation preserves image features relevant to higher-level HVS processing and corresponds to a class of images whose texture has been removed with limited disruption to object boundaries and edges. The BP repre-

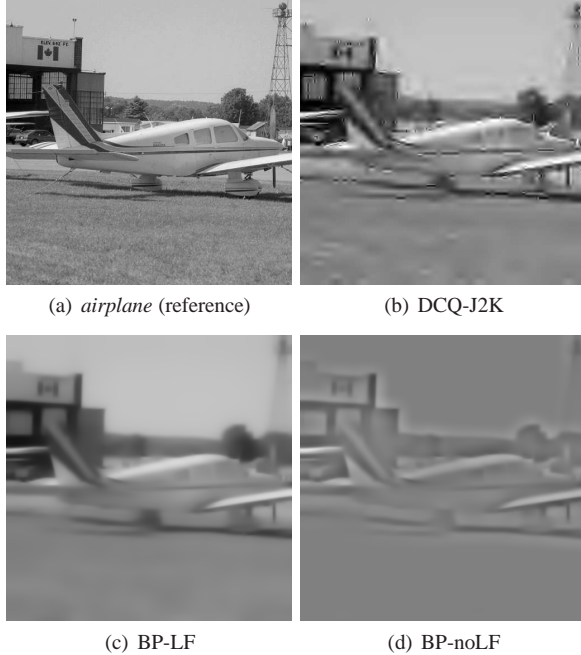


Fig. 1. Original reference *airplane* and average observer recognition thresholds for two types of boundary-preserving (BP) representations and the DCQ signal-based representation (DCQ-J2K). The BP representations shown differ with regard to the inclusion of low-frequency (LF) information.

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 BP representations that exclude LF signal information, denoted BP-noLF, were also generated.

The CU-Nantes database includes both perceived quality and perceived utility scores for the distorted images [4]. Perceived quality scores are reported as mean opinion scores (MOS) that were collected using the SAMVIQ protocol [5]. Quality scores range from 0 to 100, where a value of 100 is the highest perceived quality score.

Perceived utility scores were obtained for a content recognition task using paired comparison tests [4]. A utility score of zero corresponds to the *recognition threshold* (RT) of a reference image. A utility score of 100 defines the reference performance level (RPL) and corresponds to an image that leads to task performance as though the reference image was provided.¹ The RT specifies a collection of maximally degraded images for which an observer still understands the content.

The collected subjective data demonstrates that a perceived quality score is not a suitable 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 image representation. The quality adjectives delineating the quality rating scale have been provided. The two thresholds, the RT and the RPL, associated with the perceived utility scores are indicated on the scatter plots.

The relationship between perceived quality and perceived utility is monotonic but nonlinear, and perceived quality does not uniquely map to perceived utility. The linear relationship between quality

¹Images that are not useful surrogates for a reference (i.e., unrecognizable) have perceived utility scores below zero. An enhancement with respect to an image’s usefulness yields a utility score above 100.

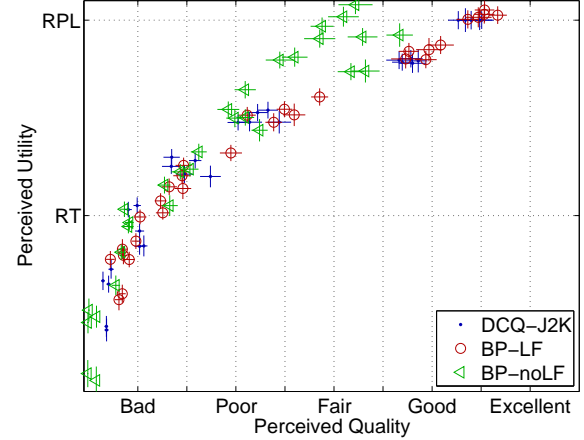


Fig. 2. Relationship between perceived utility scores and the perceived quality scores for five natural images. Quality adjectives delineating the quality rating scale and the two utility thresholds, the recognition threshold (RT) and the reference performance level (RPL), are indicated on the scatter plots. Standard error bars have been included with the subjective scores.

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. For images with perceived quality scores above 40, the BP-noLF images have nearly equal perceived utility scores to their BP-LF counterparts, yet many of the BP-noLF images have significantly lower perceived quality scores (about 20 quality points lower) than their BP-LF counterparts. The relationship between the subjective scores demonstrates that any image quality estimator that accurately estimates perceived quality scores cannot also accurately estimate perceived utility scores.

3. OBJECTIVE ESTIMATORS OF SUBJECTIVE SCORES

This section reviews several full-reference methods to objectively estimate subjective scores of distorted natural images \hat{X} that use the corresponding reference image X . In addition, this section describes how each objective estimator conducts an energy-based and/or an edge-based analysis to estimate a subjective score.

3.1. PSNR: Peak signal-to-noise ratio

The peak signal-to-noise (PSNR), which is computed using mean-square error (MSE), provides a computationally simple evaluation of signal fidelity frequently adopted to estimate the perceived quality of natural images despite its known shortcomings [6, 7, 8]. PSNR conducts a strictly energy-based analysis, since it evaluates fidelity based on the energy of the difference image $X - \hat{X}$.

3.2. SSIM: Structural Similarity Metric

The structural similarity (SSIM) [7] 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. Mathematically, SSIM computes a similarity measure between two image patches as the product of three components computed over the patches: mean,

Table 1. Results summarizing the performance of various objective estimators for the quality and utility assessment tasks. Objective scores, fitted to subjective scores with a logistic function (Eq. 2), are evaluated with respect to the subjective scores using the Spearman rank order correlation coefficient (ROCC), the Pearson (linear) correlation coefficient R , the root mean squared error (RMSE), and an F -test to individually compare the residual variance of NICE to the other estimators. F_{stat} values that indicate statistically equivalent estimation accuracy with NICE at the 95% confidence level appear in bold typeface.

<i>Estimator</i>	<i>Energy-Based or Edge-Based</i>	<i>Quality</i>				<i>Utility (Content Recognition Task)</i>				
		R	ROCC	RMSE	F_{stat}	AUC	R	ROCC	RMSE	F_{stat}
PSNR	energy-based	0.798	0.609	16.1	3.5	0.718	0.680	0.478	26.6	22
SSIM	both	0.957	0.946	7.78	0.8	0.944	0.855	0.840	18.9	11
$m \times v$	energy-based	0.898	0.905	11.8	1.9	0.923	0.754	0.751	23.9	17
r	edge-based	0.780	0.702	16.7	3.8	0.760	0.948	0.928	11.6	4.1
VIF	both (edge-based emphasized)	0.966	0.976	6.92	0.7	0.933	0.990	0.976	5.22	0.8
VIF*	both	0.990	0.987	3.82	0.2	0.971	0.955	0.933	10.8	3.6
NICE	edge-based	0.948	0.951	8.57	1	0.994	0.988	0.983	5.73	1

variance, and cross-correlation.² The two patches correspond to the same spatial window of the reference and test images. The similarity measures for the patches are averaged to estimate the degradation of the test image relative to the reference image.

SSIM employs both an energy-based and an edge-based comparison of the test and reference images to estimate the subjective score of the test image. The mean and variance components of SSIM constitute the energy-based comparison, whereas the cross-correlation component provides an edge-based comparison.

3.3. VIF: Visual Information Fidelity Criterion

The visual information fidelity (VIF) criterion [8] generates objective scores based on a measurement of the mutual information between the test and reference images. VIF uses fundamentally Gaussian models of the wavelet coefficients of the test and reference images that reduce the mutual information measurement to a local signal-to-noise ratio (SNR) in the wavelet domain. A modification of VIF, denoted VIF*, that adjusts the weights used to pool objective scores produced by VIF across image scales is also evaluated [4].³

VIF employs both an energy-based and an edge-based analysis of the reference and test images to estimate the subjective score of the test image. The local SNR calculation at multiple image scales provide an energy-based analysis. The weights applied to the individual image scale computations before linear pooling across image scale adjust the influence of VIF's edge-based analysis to the overall objective score. Greater emphasis on finer scale analysis increase the role of the edge-based analysis, while greater emphasis on the coarser image scale analysis increases the contribution of the energy-based analysis.

3.4. NICE: Natural Image Contour Evaluation

Object recognition is widely believed to rely on the perception of image details, such as sharp edges, which are conveyed by high spatial frequencies [9]. Edges or contours, defined by sudden intensity changes, can be identified by the presence of an absolute maximum magnitude in the gradient of an image. The natural image contour evaluation (NICE) strictly conducts an edge-based analysis that compares the contours of a test image to those of a reference image to

produce a numerical score indicating the estimated utility score of the test image. Numerous algorithms have been designed to detect contours in natural images. NICE has been evaluated using several different methods of contour identification, where the Sobel edge-detector was among the best methods to estimate perceived utility scores [10].

An objective score with NICE is computed by comparing the contours of the reference and test images, represented as the respective binary images B and \hat{B} . The binary images B and \hat{B} are subjected to morphological dilation with a 3×3 "plus-sign" shaped structuring element. The morphological dilation accommodates small shifts in image contours that result from distortion artifacts in a test image and should not be quantified as errors. The overall NICE score for the test image is computed as

$$\text{NICE} = \frac{1}{N_B} d_H(B, \hat{B}), \quad (1)$$

where N_B is the number of non-zero elements of B and $d_H(X, Y)$ denotes the Hamming distance between the two binary vectors X and Y . The Hamming distance counts the number of dissimilar elements between two vectors.

4. OBJECTIVE ESTIMATES OF SUBJECTIVE SCORES

This section summarizes the capabilities of the various objective estimators presented in Section 3 to estimate both perceived quality and perceived utility scores of natural images. The capability of an objective estimator to estimate perceived quality scores is evaluated using the entire CU-Nantes database of 90 test images. For the content recognition task, there are two basic operating scenarios for an objective estimator: 1) determining if a test image is recognizable, and 2) estimating the perceived utility score of a recognizable image.

Objective estimators can be used to determine if test images are recognizable by applying an appropriate threshold to the score generated by that estimator. Casting this as a two-class detection problem (i.e., an image is either recognizable or unrecognizable), the performance of an estimator can be characterized by determining the receiver operating characteristic (ROC)[11]. A ROC curve summarizes the relationship between the proportion of true positives and false positives for a given estimator using a range of threshold values. The area under the ROC curve (AUC) collapses the performance of an objective estimator to a single number. The AUC represents the probability that for a pair of test images belonging to each class (i.e., one recognizable and one unrecognizable) the recognizable image is correctly identified by a candidate detector.

²These three components have been interpreted to evaluate distortions attributed to image properties: luminance (mean), contrast (variance), and structure (cross-correlation) [7].

³VIF* multiplies the individual subband calculations corresponding to $I(\hat{C}^N; \hat{E}^N | s^N)$ and $I(\hat{C}^N; \hat{F}^N | s^N)$ in Eqs. (12) and (13) of [8] by $\frac{1}{N}$.

The perceived utility need only be estimated for recognizable images, so only those image who have perceived utility scores greater than -10 are used to evaluate an estimator's capability to estimate perceived utility scores. There are 69 images whose utility score exceed -10 .

Objective estimators frequently generate objective scores that exhibit a nonlinear relationship with subjective scores. The nonlinear mapping

$$f(a) = p_1 \times [1 + \exp(p_2(a - p_3))]^{-1} + p_4 \quad (2)$$

is used to map objective scores a to subjective scores $f(a)$. The parameters $\{p_j\}_{j=1}^4$ were fitted to the data to minimize the sum of squared error between nonlinear mapped objective scores and the subjective scores. The fitted objective scores are evaluated with respect to the subjective scores using the Spearman rank order correlation coefficient (ROCC), the Pearson (linear) correlation coefficient R , the root mean squared error (RMSE), and an F -test to individually compare the residual variance of NICE to the other estimators.

4.1. Results

Objective estimators that incorporate both an energy-based and edge-based analysis produce more accurate estimates of perceived quality than those that rely primarily on either an energy-based or an edge-based analysis. Among the objective estimators evaluated, VIF* provides significantly smaller errors (RMSE = 4.16) when predicting perceived *quality* scores than the other objective estimators. Table 1 summarizes the statistical analysis of the fitted objective scores with the perceived quality scores. An F -test that compares the residual variances of the fitted VIF* scores to the fitted scores of other estimators validated the significance of its improved performance. SSIM, which incorporates both an energy-based and an edge-based analysis, produces more accurate estimates of perceived quality than when either the cross-correlation component or the combination of the mean and variance components are used to estimate perceived quality.

Objective estimators that greater emphasize a edge-based analysis provide the most accurate estimates of perceived utility scores. Table 1 summarizes the statistical analysis of the fitted objective scores with the perceived utility scores. The ROC AUC analysis shows that NICE distinguishes recognizable and unrecognizable images better than the other objective estimators evaluated, and both NICE and VIF generate statistically equivalent estimation errors when estimating perceived utility scores. According to an F -test that compares the residual variances of the fitted objective scores, both of these objective estimators are statistically superior to the other estimators evaluated. The cross-correlation component of SSIM r , which provides an edge-based comparison, yields more accurate estimates of perceived utility scores than either the combination of the mean and variance components of SSIM, $m \times v$, or SSIM.

4.2. Discussion

The results illustrate that an edge-based analysis provides the most accurate estimate of perceived utility for the content recognition task. This finding is consistent with the theory that object recognition relies upon an interpretation of contour shape [12]. However, this theory has been developed via experiments that use either cartoons or undistorted natural images. The current results imply that this theory could extend to distorted images. More importantly, the results indicate that the degradation of image contours (as characterized by

an edge-based analysis) coincides with a decrease in a human observer's ability to interpret the content of natural images.

Estimators that augment an edge-based analysis with an energy-based analysis generate more accurate estimates of perceived quality than those that solely rely upon either of these two analyses. In Section 4.1, VIF* was shown to estimate perceived quality more accurately than VIF. VIF greatly emphasizes its edge-based analysis over its energy-based analysis, whereas VIF* equalizes the significance of the the two analyses with a slight bias towards its energy-based analysis, which is a consequence of the characteristic $1/f^\alpha$ amplitude spectrum of natural scenes [4]. The current results illustrate that a general quality estimator should not discount a coarse scale analysis of natural image characteristics.

5. CONCLUSIONS

This paper examines the role of edge-based and energy-based analyses by various objective estimators to estimate perceived quality and perceived utility scores. The capabilities of the objective estimators are evaluated using the CU-Nantes image database. Results show that estimators that emphasize an edge-based image analysis provide the most accurate estimates of perceived utility scores. Estimators that augment the edge-based analysis with an energy-based analysis provide the most accurate estimates of perceived quality scores.

6. REFERENCES

- [1] C. G. Ford, M. A. McFarland, and I. W. Stange, "Subjective video quality assessment methods for recognition tasks," in *Proc. SPIE: HVEI XIV*, B. E. Rogowitz and T. N. Pappas, Eds., vol. 7240, Jan. 2009.
- [2] Video Quality in Public Safety Conference, C. Ford, P. Raush, and K. Davis, Eds. Boulder, CO: Institute for Telecommunication Sciences, Feb.4–6, 2009.
- [3] R. Driggers, E. Jacobs, R. Vollmerhausen, B. O'Kane, M. Self, S. Moyer, J. Hixson, and G. Page, "Current infrared target acquisition approach for military sensory design and wargaming," in *Proc. SPIE: Infrared Imaging Systems*, G. C. Hoist, Ed., vol. 6207, Apr. 2006.
- [4] D. Rouse, R. Pepion, S. Hemami, and P. Le Callet, "Image utility assessment and a relationship with image quality assessment," in *Proc. SPIE: HVEI XIV*, B. E. Rogowitz and T. N. Pappas, Eds., vol. 7240, San Jose, CA, Jan. 2009.
- [5] F. Kozamernik, P. Sunna, E. Wyckens, and D. I. Pettersen, "Subjective quality of internet video codecs phase ii evaluations using SAMVIQ," EBU Technical Review, European Broadcast Union (EBU), Jan. 2005.
- [6] B. Girod, "What's wrong with mean-square error?" in *Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA: MIT Press, 1993, ch. 15, pp. 207–220.
- [7] 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.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
- [8] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.
- [9] R. L. De Valois and K. K. De Valois, *Spatial Vision*. New York: Oxford University Press, 1990.
- [10] D. M. Rouse and S. S. Hemami, "Natural image utility assessment using image contours," in *Proc. IEEE Int. Conf. on Image Process. (ICIP)*, Cairo, Egypt, Nov. 2009.
- [11] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, pp. 29–36, Apr. 1982.
- [12] D. Marr, *Vision*. W. H. Freeman and Company, 1982.