NATURAL IMAGE UTILITY ASSESSMENT USING IMAGE CONTOURS

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 generally *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. This paper examines variations to basic components of a recently introduced *utility assessment algorithm* that compares the contours of a reference and test image, referred to as the *natural image contour evaluation* (NICE), in terms of their capability to improve the prediction of perceived utility scores. Results show that classical edge-detection algorithms incorporated into NICE provide statistically equivalent performance to other, more complex edge-detection algorithms.

Index Terms— utility assessment, quality assessment, edge detection

1. INTRODUCTION

In many imaging applications, humans use the information captured by an imaging system and tolerate distortions as long as the underlying task is performed reliably. 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]. Such images should be assessed according to their usefulness, or *utility*, rather than according to the conventional notion of perceptual *quality*, which has been largely studied in the context of images used in consumer applications. Merely employing an existing quality assessment (QA) algorithm to predict the utility of an image is insufficient, 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.

Present 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, 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.

Recently, the natural image contour evaluation (NICE), was introduced as a *utility assessment algorithm* [4]. NICE conducts a comparison of the contours of a test image to those of a reference image to score the test image. This contour-based image assessment algorithm demonstrates a viable departure from traditional QA algorithms that incorporate energy-based approaches and is capable of predicting perceived utility scores. This paper examines variations to basic components of the NICE algorithm to determine their impact when predicting both perceived quality and perceived utility scores.

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 describes the NICE algorithm. An analysis with regard to the capability of variations to NICE to predict subjective scores for both quality and utility assessment tasks is provided in Section 4. Conclusions are presented in Section 5.

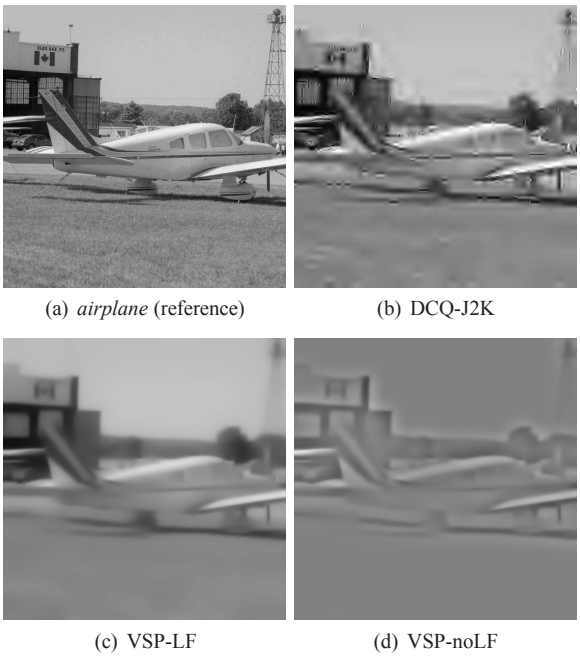
2. PERCEIVED QUALITY AND PERCEIVED UTILITY

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

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

Perceived utility scores were obtained using paired comparison tests [4]. Utility scores are reported such that a score of zero corresponds to the *recognition threshold* of a reference image and a score of 100 corresponds to an image that is visually indistinguishable from the reference image[1]. The recognition threshold specifies

[1] Images that are not useful surrogates for a reference (i.e., unrecognizable) have perceived utility scores below zero. An enhancement with respect



(a) *airplane* (reference) (b) DCQ-J2K

(c) VSP-LF (d) VSP-noLF

Fig. 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.

a collection of maximally degraded images for which an observer still understands the content [4].

The collected subjective data clearly demonstrates that 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 image representation. Test images with perceived utility scores below -10 were omitted from the set of test images, leaving 62 of the original 90 test images. Use of a linear fit to map perceived quality scores to perceived utility scores produces poor utility estimates (RMSE = 15.1).

Perceived quality does not uniquely map to perceived utility. 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. For images with perceived quality scores above 40, the VSP-noLF images have nearly equal perceived utility scores to their VSP-LF counterparts, 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 perceived quality scores cannot immediately predict perceived utility scores.

3. 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 [6, 7]. 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. This section describes the natural image contour evaluation (NICE) utility assessment algorithm [4].

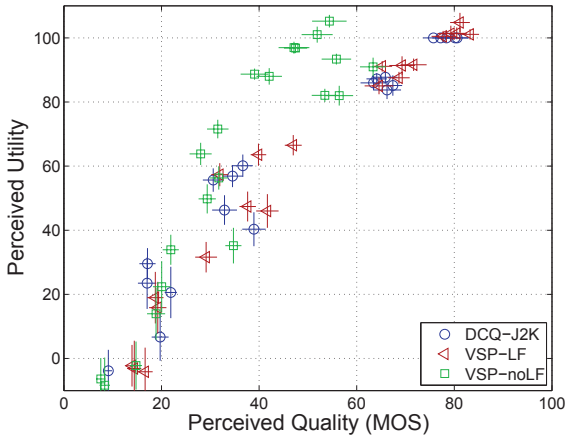


Fig. 2. Relationship between perceived utility scores and the perceived quality scores for five natural images. Standard error bars have been included for the subjective scores.

Unlike traditional QA algorithms that use energy-based computations, NICE compares the contours of a test image to that of a reference image to produce a numerical score indicating the predicted utility score of the test image. The framework has three components: 1) contour identification of reference and test images, 2) morphological dilation of binary images representing image contours, 3) element-wise exclusive-or (XOR) between dilated binary images.

The multi-channel model of the human visual system inspired the use of multi-scale contours in the original NICE design [4]. Several computationally simpler algorithms provide single-scale contours, so this section describes two methods to identify image contours: 1) using wavelet coefficients, which was considered in the original formulation of NICE [4], and 2) classical edge-detection algorithms. The section concludes by specifying the computation of the NICE score from the identified image contours.

3.1. Multi-scale Contours using Wavelet Coefficients

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. Both the reference and test image can be represented using an *undecimated* implementation of the steerable pyramid (SPYR) [8] using D orientations and S scales[2]. Let $W_{s,\theta}(i)$ and $\hat{W}_{s,\theta}(i)$ denote the i^{th} wavelet coefficient of the respective reference and test images in the subband corresponding to scale $s \in \{1, 2, \ldots, S\}$ and orientation $\theta \in \{0, \frac{\pi}{D}, \frac{2\pi}{D}, \ldots, \frac{\pi(D-1)}{D}\}$.

For each image scale s, the local modulus maxima (LMM) [9] of wavelet coefficient scales correspond to image contours for the reference and test images. The LMM are determined from gradient vectors formed from wavelet subbands corresponding to derivatives in horizontal and vertical spatial directions [9]. Define $G_s(i) = W_{s,0}(i) - \mathrm{j}W_{s,\frac{\pi}{2}}(i)$ and $\hat{G}_s(i) = \hat{W}_{s,0}(i) - \mathrm{j}\hat{W}_{s,\frac{\pi}{2}}(i)$ as the gradient of the respective reference and test images at scale s, where $\mathrm{j} = \sqrt{-1}$. 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. LMM of the reference image

to an image's usefulness has a utility score above 100.

[2]The high-pass residual generated by the steerable pyramid is not used.

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, LMM are similarly identified using $\hat{M}_s(i)$ and $\hat{A}_s(i)$. For scale s, let $\mathcal{I}_s$ and $\hat{\mathcal{I}}_s$ denote sets of indices i corresponding to LMM of the respective reference image and test images.

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 LMM that exceed a threshold β_s based on the reference image and is given as $\beta_s = \frac{1}{p}\max_i M_s(i)$ for some scalar $p > 1$. Specifically, define $B_s(i)$ and $\hat{B}_s(i)$ as

$$B_s(i) = \begin{cases} 1 & M_s(i) > \beta_s \text{ and } i \in \mathcal{I}_s \\ 0 & else \end{cases} \quad (1)$$

$\hat{B}_s(i)$ is similarly defined using $\hat{M}_s$, $\hat{\mathcal{I}}_s$, and β_s.

3.2. Single-scale Contours using Edge-Detection Algorithms

Numerous algorithms have been designed to detect edges in natural images. Three classical edge-detection algorithms are examined as alternatives to the wavelet-based contour identification approach described in Section 3.1. Although these algorithms could be extended to generate binary images corresponding to contours at several image scales, these algorithms are only used to generate the binary images B_1 and $\hat{B}_1$, corresponding to contours of one scale. Each of the algorithms incorporate a filtering operation that approximates the first-derivative of the image.

The Sobel and Prewitt operators are computationally simple operators used for edge-detection [7]. Both algorithms filter an image with two 3×3 linear filters, one that approximates a horizontally-oriented derivative and another that approximates a vertically-oriented derivative. If G_x and G_y correspond to the approximated horizontal and vertical derivatives of the original image, respectively, then an edge-intensity image, given as $G = \sqrt{G_x^2 + G_y^2}$, is subjected to hard-thresholding, using a threshold related to the average value of G, to produce a binary image identifying image contours.

The Canny edge-detector filters the image with the derivative of a Gaussian specified for a particular $\sigma > 0$ and applies thresholding to generate a binary image [10]. The parameter σ in the Canny filter controls the suppression of high frequencies before detecting edges[3].

3.3. Generating NICE Score from Identified Contours

An objective utility score 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 and $\hat{B}_s$ defined according to a method presented in Sections 3.1 or 3.2. The binary images B_s and $\hat{B}_s$ are subjected to morphological dilation [11] with a 3×3 "plus-sign" shaped structuring element, and the point-wise exclusive-or (XOR) operation of the dilated binary images produces the binary image $E_s(i)$. 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

$$\text{NICE} = \sum_{s=1}^{S} a_s N(s), \quad (2)$$

where $N(s)$ is the number of non-zero elements of E_s and the $\{a_s\}_{s=1}^{S}$ are nonnegative scalars. When using the classical edge-detection algorithms from Section 3.2, $S = 1$.

[3]This paper implements the Canny edge-detector for $\sigma = 1$

4. PREDICTING QUALITY AND UTILITY WITH NICE

This section examines the capability of NICE to predict both perceived quality and perceived utility scores for the images in the CU-Nantes database [4]. The performance of NICE is evaluated when using different contour identification algorithms and with or without the morphological dilation operation as described in Section 3. The performance of the visual information fidelity (VIF) criterion [12] and a modified implementation, denoted VIF*, that adjusts the weights used to pool the objective scores produced by VIF across image scales, are evaluated to demonstrate the comparative performance of energy-based QA algorithms [4] [4].

The original implementation of NICE identified contours using a four-scale steerable pyramid (SPYR) with $D = 6$ six orientations, but only subbands corresponding to horizontal and vertical frequencies are used [4]. To reduce the computational complexity underlying the original implementation of NICE, fewer orientations (i.e., $D = 2$ and $D = 4$) are considered. Implementations using the steerable pyramid to identify image contours compute the threshold β_s for $p = 20$ and are identified with the acronym SPYR-D, where D is the number of orientations.

Objective scores generated by QA algorithms frequently exhibit a nonlinear relationship with subjective scores. The nonlinear mapping of objective scores a to subjective scores $f(a)$ is given as

$$f(a) = p_1 \times [1 + \exp(p_2(a - p_3))]^{-1} + p_4. \quad (3)$$

The parameters $\{p_j\}_{j=1}^{4}$ were fitted to the data to minimize the sum-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 squared Pearson (linear) correlation coefficient r^2, the root mean squared error (RMSE), and an F-test to individually compare the residual variance of NICE using SPYR-6 contour identification to the other algorithms.

An F-test determines whether the residual variance of algorithm is statistically larger or smaller than the other [13]. Let σ^2_{SPYR-6} and σ^2_B denote the variance of the residuals corresponding to the SPYR-6 NICE implementation and some alternative algorithm B when used to predict subjective scores. Using the statistic $F_{stat} = \sigma^2_{SPYR-6}/\sigma^2_B$, algorithms whose value of F_{stat} lie inside the interval $[1/\mathrm{F}_{crit}, F_{crit}]$ exhibit statistically equivalent prediction performance with the SPYR-6 NICE implementation. For a 95% confidence level and 62 test images, $F_{crit} = 1.53$.

4.1. Results

Among the algorithms evaluated, VIF* provides significantly smaller errors (RMSE = 4.16) when predicting perceived *quality* scores than VIF and the various implementations of NICE. When using classical edge-detection algorithms to identify contours, morphological dilation has no statistically significant effect with regard to the performance capabilities when predicting perceived quality scores. Table 1 summarizes the results from a 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 algorithms validates the significance of its improved performance.

Most implementations of NICE that include morphological dilation and VIF provide significantly equivalent errors, based on an

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

Table 1. Results summarizing the performance of NICE using various contour identification algorithms for the quality and utility assessment tasks. Objective scores, fitted to subjective scores with a logistic function (Eq. 3), 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), and an F-test to individually compare the residual variance of NICE using SPYR-6 contour identification to the other algorithms. Algorithms whose value of F_{stat} lie inside the interval $[0.65, 1.53]$ exhibit statistically equivalent prediction performance with the SPYR-6 NICE implementation at the 95% confidence level. Values of F_{stat} inside the interval $[0.65, 1.53]$ appear in bold typeface.

		Quality Task				*Utility Task*			
Algorithm	*Contour Identification*	r^2	ROCC	RMSE	F_{stat}	r^2	ROCC	RMSE	F_{stat}
NICE	SPYR-6 (from [4])	0.917	0.952	6.66	**1.00**	0.940	0.944	8.92	**1.00**
NICE	SPYR-4	0.877	0.936	8.11	**0.67**	0.945	0.953	8.49	**1.10**
NICE	SPYR-2	0.885	0.939	7.83	**0.72**	0.950	0.958	8.11	**1.21**
NICE	Sobel (no dilation)	0.769	0.868	11.1	0.36	0.901	0.954	11.4	0.61
NICE	Sobel	0.800	0.897	10.3	0.42	0.950	0.970	8.11	**1.21**
NICE	Prewitt (no dilation)	0.775	0.879	11.0	0.37	0.905	0.956	11.2	0.63
NICE	Prewitt	0.808	0.903	10.1	0.43	0.953	0.971	7.83	**1.30**
NICE	Canny (no dilation)	0.775	0.873	11.0	0.37	0.857	0.940	13.7	0.42
NICE	Canny	0.790	0.890	10.6	0.39	0.884	0.940	12.3	0.53
VIF	n/a	0.882	0.942	7.93	**0.81**	0.959	0.969	7.33	**1.48**
VIF*	n/a	0.967	0.974	4.16	2.56	0.912	0.886	12.2	0.53

F-test, when predicting perceived *utility* scores. Table 1 summarizes the statistical analysis of the fitted objective scores with the perceived utility scores. NICE implemented with the Canny edge-detection algorithm to identify image contours exhibits poorer prediction accuracy (RMSE > 10) regardless of the inclusion of morphological dilation when computing the NICE score. However, when implemented with either the Sobel and Prewitt operators and morphological dilation, NICE performs statistically the same as the when the steerable pyramid (SPYR) is used to identify contours.

4.2. Discussion

The results demonstrate that the NICE algorithm is capable of predicting perceived *utility* scores for natural images. The results for various methods tested to identify image contours reveal that simpler, rather than more complex algorithms such as the Canny edge-detector, provide better results when coupled with morphological dilation. The Canny edge-detector was originally designed to be more robust to noise in natural images when detecting edges, but this feature is counter to the desired goal of NICE, which seeks a contour identification scheme that is less robust to distortions so differences between the contours of reference and test images can be quantified.

The results raise questions about the benefits of using contours from multiple image scales versus single-scale image contours to predict perceived utility scores, since each approach produces statistically equivalent results when coupled with morphological dilation. In general, contours from coarser image scales appear in finer image scales, making contour information from multiple scales largely redundant. This redundancy could be leveraged to identify salient contours in images. Disruptions to salient contours could be emphasized over other contours. Such variations are left for future work.

5. CONCLUSIONS

This paper examines the performance of the natural image contour evaluation (NICE) algorithm using different contour identification schemes. This edge-based image assessment algorithm is shown to be a viable alternative to traditional energy-based quality assessment algorithms and is capable of predicting perceived utility scores.

Within the NICE algorithm is a morphological dilation operation intended to accommodate small shifts in the image contours that would other have been quantified as errors. Including morphological dilation improves the prediction accuracy when using classical edge-detection algorithms. Furthermore, the Sobel and Prewitt edge-detection algorithms provide statistically similar performance to a more complex wavelet-based, multi-scale edge detection algorithm when incorporating the morphological dilation operation.

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

[7] W. K. Pratt, *Digital Image Processing: PIKS Inside*, 3rd ed. New York: Wiley-Interscience, 2001.

[8] 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.

[9] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. Image Process.*, vol. 14, no. 7, pp. 710–732, 1992.

[10] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, Nov. 1986.

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

[12] 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.

[13] J. L. Devore, *Probability and Statistics for Engineering and the Sciences*, 5th ed. Duxbury, 2000.