Journal of Electronic Imaging 11(2), 206–223 (April 2002).

Statistical evaluation of image quality measures

İsmail Avcıbaş
Uludağ University
Department of Electronic Engineering
Bursa, Turkey

Bülent Sankur
Boğaziçi University
Department of Electrical and Electronic Engineering
Istanbul, Turkey
E-mail: sankur@boun.edu.tr

Khalid Sayood
University of Nebraska at Lincoln
Department of Electrical Engineering
Lincoln, Nebraska

Abstract. *In this work we comprehensively categorize image quality measures, extend measures defined for gray scale images to their multispectral case, and propose novel image quality measures. They are categorized into pixel difference-based, correlation-based, edge-based, spectral-based, context-based and human visual system (HVS)-based measures. Furthermore we compare these measures statistically for still image compression applications. The statistical behavior of the measures and their sensitivity to coding artifacts are investigated via analysis of variance techniques. Their similarities or differences are illustrated by plotting their Kohonen maps. Measures that give consistent scores across an image class and that are sensitive to coding artifacts are pointed out. It was found that measures based on the phase spectrum, the multiresolution distance or the HVS filtered mean square error are computationally simple and are more responsive to coding artifacts. We also demonstrate the utility of combining selected quality metrics in building a steganalysis tool.* © *2002 SPIE and IS&T.*
[DOI: 10.1117/1.1455011]

1 Introduction

Image quality measures (IQMs) are figures of merit used for the evaluation of imaging systems or of coding/processing techniques. In this study we consider several image quality metrics and study their statistical behavior when measuring various compression and/or sensor artifacts.

A good objective quality measure should reflect the distortion on the image well due to, for example, blurring, noise, compression, and sensor inadequacy. One expects that such measures could be instrumental in predicting the performance of vision-based algorithms such as feature extraction, image-based measurements, detection, tracking, and segmentation, etc., tasks. Our approach is different from companion studies in the literature that focused on subjective image quality criteria, such as those in Refs. 1–3. In a subjective assessment measures characteristics of human perception become paramount, and image quality is correlated with the preference of an observer or the performance of an operator for some specific task.

In the image coding and computer vision literature, the most frequently used measures are deviations between the original and coded images,[4–6] with varieties of the mean square error (MSE) or signal to noise ratio (SNR) being the most common measures. The reasons for their widespread popularity are their mathematical tractability and the fact that it is often straightforward to design systems that minimize the MSE. Raw error measures such as the MSE work best when the distortion is due to additive noise contamination. However they do not necessarily correspond to all aspects of the observer's visual perception of the errors,[7,8] nor do they correctly reflect structural coding artifacts.

For multimedia applications and for very low bit rate coding, there has been an increase in the use of quality measures based on human perception.[9–14] Since a human observer is the end user in multimedia applications, an image quality measure that is based on a human vision model seems to be more appropriate for predicting user acceptance and for system optimization. This class of distortion measure in general gives a numerical value that will quantify the dissatisfaction of the viewer in observing the reproduced image in place of the original (although Daly's VPD map[13] is an example opposite to this). The alternative is the use of subjective tests in which subjects view a series of reproduced images and rate them based on the visibility of the artifacts.[15,16] Subjective tests are tedious, time consuming and expensive, and the results depend on various factors such as the observer's background, motivation, etc., and really actually only the display quality is being assessed. Therefore an objective measure that accurately pre-

Paper JEI 99022 received Apr. 28, 1999; revised manuscripts received Nov. 16, 2000 and Aug. 27, 2001; accepted for publication Oct. 19, 2001.
1017-9909/2002/$15.00 © 2002 SPIE and IS&T.

dicts the subjective rating would be a useful guide when optimizing image compression algorithms.

Recently there have been efforts by the International Telecommunications Union (ITU) to establish an objective measurement of video quality. Thus within the context of the distribution of multimedia documents, video programming, in particular, in-service continuous evaluation of video quality, is needed. This continuous video quality indicator would be input to the network management, which must guarantee a negotiated level of quality of service. Obviously such quality monitoring can only be realized with objective methods.[17,18] It must be pointed out, however, that subjective assessment, albeit costly and time consuming, if not impractical, is accurate. Objective methods, on the other hand, can at best try to emulate the performance of subjective methods, and utilize knowledge of the human visual system.

Similarly for computer vision tasks, prediction of the algorithmic performance in terms of imaging distortions is of great significance.[19,20] In the literature the performance of feature extraction algorithms, like lines and corners,[19] propagation of covariance matrices,[20] and quantification of target detection performance and ideal observer performance,[21–23] have been studied under additive noise conditions. It is of great interest to correlate coding and sensor artifacts with this kind of algorithmic performance. More specifically, one would like to identify image quality metrics that can accurately and consistently predict the performance of computer vision algorithms that operate on distorted image records, the distortions being due to compression, sensor inadequacy, etc. An alternative use of image quality metrics is in inverse mapping from metrics to the nature of distortions.[24] In other words, given the image quality metrics, one tries to reconstruct the distortions (e.g., the amount of blur, noise, etc., in distortion coordinates) that could have resulted in the measured metric values.

In this paper we study objective measures of image quality and investigate their statistical performance. Their statistical behavior is evaluated first, in terms of how discriminating they are to distortion artifacts when tested on a variety of images using the analysis of variance method. The measures are then investigated in terms of their mutual correlation or similarity in the form of Kohonen maps.

Twenty-six image quality metrics are listed and described in Appendix A and summarized in Table 1. These quality metrics are categorized into six groups according to the type of information they use. The categories used are:

1. pixel difference-based measures such as mean square distortion;
2. correlation-based measures, that is, correlation of pixels, or of the vector angular directions;
3. edge-based measures, that is, displacement of edge positions or their consistency across resolution levels;
4. spectral distance-based measures, that is, the Fourier magnitude and/or phase spectral discrepancy on a block basis;
5. context-based measures, that is, penalties based on various functionals of the multidimensional context probability;
6. human visual system (HVS)-based measures, that is, measures either based on the HVS-weighted spectral distortion measures or (dis)similarity criteria used in image base browsing functions.

Table 1 List of symbols and equation numbers of the quality metrics.

Symbol	Description	Equation
$D1$	Mean square error	(A1)
$D2$	Mean absolute error	(A2)
$D3$	Modified infinity norm	(A3)
$D4$	$L^*a^*b^*$ perceptual error	(A4)
$D5$	Neighborhood error	(A5)
$D6$	Multiresolution error	(A6)
$C1$	Normalized cross correlation	(A7)
$C2$	Image fidelity	(A8)
$C3$	Czekonowski correlation	(A9)
$C4$	Mean angle similarity	(A10)
$C5$	Mean angle-magnitude similarity	(A11)
$E1$	Pratt edge measure	(A12)
$E2$	Edge stability measure	(A13)
$S1$	Spectral phase error	(A14)
$S2$	Spectral phase-magnitude error	(A15)
$S3$	Block spectral magnitude error	(A16)
$S4$	Block spectral phase error	(A17)
$S5$	Block spectral phase-magnitude error	(A18)
$Z1$	Rate distortion measure	(A19)
$Z2$	Hellinger distance	(A20)
$Z3$	Generalized Matusita distance	(A21)
$Z4$	Spearman rank correlation	(A22)
$H1$	HVS absolute norm	(A23)
$H2$	HVS $L2$ norm	(A24)
$H3$	Browsing similarity	(A25)
$H4$	DCTune	

We define several distortion measures in each category. The specific measures are denoted by $D1$, $D2$, etc. in Appendix A, in the pixel difference category, as $C1$, $C2$, etc. in Appendix B, in the correlation category and so on for ease of reference in the results and discussion sections.

The paper is organized as follows: The methodology and data sets are given in Sec. 2. The descriptions of the specific measures used are relegated to the Appendix and its six subsections. The results of the experiments and statistical analyses are presented in Sec. 3. We discuss the main conclusions and related future work in Sec. 4.

2 Goals and Methods

2.1 *Quality Attributes*

Objective video quality model attributes were reported in Refs. 17 and 18. These attributes can be directly translated to the still image quality measures as ''IQM desiderata'' in multimedia and computer vision applications.

- Prediction accuracy: The accurate prediction of distor-

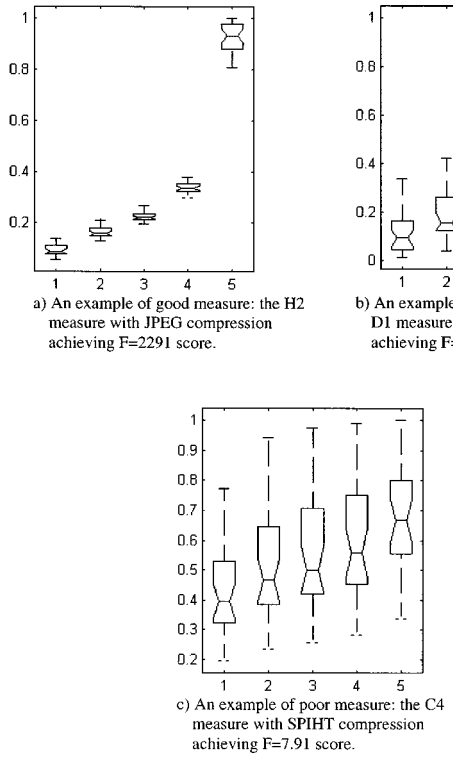


Fig. 1 Box plots of quality measure scores: (a) good measure, (b) moderately good measure, (c) poor measure.

tion, whether for algorithmic performance and subjective assessment. For example, when quality metrics are shown in box plots, like in Fig. 1, an accurate metric will possess a small scatter plot.

- Prediction monotonicity: The objective image quality measure's scores should be monotonic in their relationship to the performance scores.
- Prediction consistency: This attribute relates to the objective quality measure's capability to provide consistently accurate predictions for all types of images and not to fail badly for a subset of images.

These desired characteristics are captured in the statistical measures such as the F scores of the quality metrics, as detailed in Tables 1–3.

2.2 *Test Image Sets and Rates*

All the image quality measures are calculated in their multiband versions. In the current study of the quality measures in image compression, we used two well-known compression algorithms: the popular DCT based JPEG[25] and wavelet zero-tree method ''set partitioning in hierarchical trees'' (SPIHT) formulated by Said and Pearlman.[26] The other types of image distortions are generated by the use of blurring filters of various support sizes and by the addition of white Gaussian noise at various levels.

The rate selection scheme was based on the accepted rate ranges of JPEG. It is known that the JPEG quality factor Q between 80 and 100 corresponds to visually imperceptible impairment, Q between 60 and 80 corresponds to perceptible but not annoying distortion, for Q between 40 and 60 the impairment becomes slightly annoying, for Q between 20 and 40 the impairment is annoying, and, finally, for Q less than 20 the degradation is very annoying. Thus each image class was compressed with 5 JPEG Q factors of 90, 70, 50, 30, and 10. For each quality class the average length of compressed files was calculated and the corresponding bit rate (bit/pixel) was accepted as the class' rate. The same rate as that obtained from the JPEG experiment was also used in the SPIHT algorithm.

The test material consisted of the following image sets: (1) 10 three-band remote sensing images, which contained a fair amount of variety, i.e., edges, textures, plateaus, and contrast range, (2) 10 color face images from the Purdue University Face Image database[27] at rvl1.ecn.purdue.edu/aleix/Aleix_face_DB.html, and (3) 10 texture images from the MIT Texture Database (VISTEX) at www-white.media.edu/vismod/imagery/VisionTexture/vistex.html.

2.3 *Analysis of Variance*

Analysis of variance (ANOVA)[28] was used as a statistical tool to evaluate the merits of the quality measures. In other words, ANOVA was used to show whether variation in the data could be accounted for by the hypothesized factor, for example, the factor of image compression type, the factor of image class, etc. The output of the ANOVA is the identification of those image quality measures that are most consistent and discerning of the distortion artifacts due to compression, blur, and noise.

Recall that ANOVA is used to compare the means of more than two independent Gaussian distributed groups. In our case each ''compression group'' consists of quality scores from various images at a certain bit rate, and there are $k=5$ groups corresponding to the five bit rates tested. Each group had 30 sample vectors since there were 30 multispectral test images (10 remote sensing, 10 faces, 10 textures). In a similarly way three ''blur groups'' were created by low-pass filtering the images with two-dimensional (2D) Gaussian-shaped filters with increasing support. Finally three ''noise groups'' were created by contaminating the images with Gaussian noise with variance set at ($\sigma^2 = 200$, 600, and 1700). This range of noise values spans the noisy image quality from just noticeable distortion to annoying degradation. In a concomitant experiment[57] images were watermarked at four different insertion strengths.

Since we have two coders (i.e., JPEG and SPIHT algorithms) two-way ANOVA is appropriate. The hypotheses for the comparison of independent groups are

H_0: $\mu_1=\mu_2=\ldots=\mu_k$
mean values of all groups are equal,

H_A: $\mu_i \neq \mu_j$
mean values of two or more groups are not equal.

It should be noted that the test statistic is an F test with $k-1$ and $N-k$ degrees of freedom, where N is the total number of compressed images. A low p-value (high F value) for this test indicates one should reject the null hypothesis in

favor of the alternative. Recall that the null hypothesis corresponds to a situation in which all samples are drawn from the same set and there is no significant difference between their means. A low value of p (correspondingly, a high value of F) casts doubt on the null hypothesis and provides strong evidence that at least one of the means is significantly different. In other words, there is evidence that at least one pair of means is not equal. We have opted to carry out multiple comparison tests at the 0.05 significance level. Thus any test resulting of a p value under 0.05 would be significant, and, therefore, one would reject the null hypothesis in favor of the alternative hypothesis. This is done to assert that the difference in quality metric arises from image coding artifacts and not from random fluctuations in the image content.

To find out whether the variation of the metric scores arises predominantly from image quality, and not from the image set, we considered the interaction between the image set and distortion artifacts (i.e., compression bit rate, blur, etc.). To this end we considered the F scores with respect to the image set as well. As will be discussed in Sec. 3 and shown in Tables 2 and 3, metrics that were sensitive to distortion artifacts were naturally sensitive to variations in the image set as well. However for the ''good'' measures identified, the sensitivity to image set variation was always less to the distortion sensitivity.

A graphical comparison based on box plots, where each box is centered on the group median and sized to the upper and lower 50 percentiles, allows one to see the distribution of the groups. If the F value is high, there will be little overlap between two or more groups. If the F value is not high, there will be a fair amount of overlap among all of the groups. In the box plots, a steep slope and little overlap between boxes, as illustrated in Fig. 1, are both indicators of a good quality measure. In order to quantify the discriminative power of a quality measure, we have normalized the difference between two successive group means by their respective variances, i.e.,

$$Q_{r,r+1}=\frac{\mu_r-\mu_{r+1}}{\sqrt{\sigma_r\sigma_{r+1}}},$$
$$Q=\mathrm{ave}\{Q_{r,r+1}\} \quad r=1,\ldots k-1, \tag{1}$$

where μ_r denotes the mean value of the image quality measure for the images compressed at rate r and σ_r is the standard deviation; k is the number of different bit rates at which quality measures are calculated. A good image quality measure should have a high Q value, which implies little overlap between groups and/or large jumps between them hence a highly discriminative power of the quality measure. It should be noted that the Q values and the F scores yielded identical results in our experiments.

In Fig. 1 we give box plot examples of a good, a moderate, and a poor measure. For the box plot visualization the data have been appropriately scaled without any loss of information.

2.4 *Visualization of Quality Metrics*

The visualization of the IQMs in a 2D display is potentially helpful to observe the clustering behavior of the quality metrics, and conversely to deduce how differently they respond to distortion artifacts arising from compression, blur and noise. The output of self-organizing map (SOM) visualization is a set of qualitative arguments showing their similarity or dissimilarity. To see this we organized them as vectors and fed them to a SOM algorithm. The elements of the vectors are simply the measured quality scores. For example, consider the MSE error ($D1$) for a specific compression algorithm (e.g., JPEG) at a specific rate. The corresponding vector **D1** is M dimensional, where M is the number of images, and it reads

$$\mathbf{D1}(\mathrm{JPEG},\mathrm{bitrate})$$
$$=[D1(1|\mathrm{JPEG},\mathrm{bitrate}),\ldots,D1(M)|\ \mathrm{JPEG},\mathrm{bitrate}]^T.$$

There will be five such vectors, one for each bit rate considered. We used a total of 30 images$\times$5 bit rates $\times$2 compressors$\times$26 metrics$=$7800 vectors to train the SOM.

Recall that the SOM is a tool for visualization of high dimensional data. It maps complex, nonlinear high dimensional data into simple geometric relationships on a low dimensional array and thus serves to produce abstractions. Among the important applications of the SOM one can cite the visualization of high dimensional data, as a case in point, and the discovery of categories and abstractions from raw data.

Let the data vectors be denoted as $\mathbf{X}=[x_1,\ldots,x_M]^T \in R^M$, where M is the number of images considered ($M=30$ in our case). With each element in the SOM array, a parametric real vector $\mathbf{m}_i=[\mu_{i1},\ldots,\mu_{iM}]^T\in R^M$ that is associated. The location of an input vector $\mathbf{X}$ in the SOM array is defined by the decoder function $d(\mathbf{X},\mathbf{m}_i)$, where $d(.,.)$ is a general measure of distance. The location of the input vector will have the array index c defined as $c=\underset{i}{\mathrm{argmin}}\, d(\mathbf{X},\mathbf{m}_i)$. A critical part of the algorithm is defining $\mathbf{m}_i$ in such a way that the mapping is ordered and descriptive of the distribution of $\mathbf{X}$. Finding such a set of values that minimizes the distance measure resembles the standard vector quantization (VQ) problem. In contrast, the indexing of these values is arbitrary, whereby the mapping is unordered. However, if minimization of the objective functional based on the distance function is implemented under the conditions described in Ref. 29, then one can obtain ordered values of $\mathbf{m}_i$, almost as if $\mathbf{m}_i$ were lying at the nodes of an elastic net. With the elastic net analogy in mind, the SOM algorithm can be constructed as

$$\mathbf{m}_i(t+1)=\mathbf{m}_i(t)+\alpha(t)[\mathbf{X}(t)-\mathbf{m}_i(t)],$$

where $\alpha(t)$ is a small scalar, if the distance between units c and i in the array is smaller than or equal to a specified limit (radius), and $\alpha(t)=0$ otherwise. During the course of the ordering process, $\alpha(t)$ is decreased from 0.05 to 0.02, while the radius of the neighborhood is decreased from 10 to 3. Furthermore scores are normalized with respect to the range.

Table 2 ANOVA results (*F* scores) for the JPEG and SPIHT compression distortions as well as for additive noise and blur artifacts. For each distortion type the variation due to the image set is also established. For compression the degrees of freedom are 4 (bit rate) and 2 (image class) while they are 2 for both the blur and noise experiments.

	JPEG		SPIHT		Blur		Noise	
ANOVA2	Bit rate	Image set	Bit rate	Image set	Blur	Image set	Noise	Image set
*D*1	104.6	42.59	39.23	13.28	43.69	2.06	9880	17.32
*D*2	108.5	67.45	29.56	15.93	33.94	17.76	6239	20.4
*D*3	63.35	29.37	53.31	48.53	38.55	24.13	1625	11.15
*D*4	89.93	1.99	13.75	3.71	27.87	0.96	166.4	9.88
*D*5	20.26	80.71	14.09	68.22	6.32	55.11	1981	43.51
*D*6	76.73	5.94	37.52	11.22	412.9	45.53	44.61	4.38
*C*1	1.35	124.6	12.05	325.5	5.61	107.2	3.82	6.17
*C*2	12.26	93.83	15.18	82.87	11.19	39.77	58.04	45.63
*C*3	82.87	83.06	24.96	22.42	30.92	1.71	567.5	52.01
*C*4	45.65	47.36	7.91	5.94	16.48	0.77	198.8	19.03
*C*5	91.42	38.17	27.51	5.28	52.57	2.44	704	10.8
*E*1	26.24	3.64	77.86	137	125.8	21.09	87.76	27.87
*E*2	176.3	92.75	212.5	200.4	768.7	23.41	158.5	24.84
*S*1	150.5	102.2	104	68.17	1128	60.04	47.29	38.42
*S*2	191.3	98.42	161	101.8	572.2	17.95	107.1	4.83
*S*3	145.6	56.39	38.58	26.97	24.28	6.39	2803	8.59
*S*4	129.1	63.26	128	46.85	215	11.17	56.04	55.1
*S*5	146.1	71.03	144.1	61.65	333.6	27.84	78.04	26.53
*Z*1	1.69	141.8	21.36	14	35.9	62.5	44.89	110.9
*Z*2	7.73	114.7	11.41	77.68	10.17	1.80	3.03	11.36
*Z*3	17.63	223	23.22	181.4	17.26	8.31	14.71	21.12
*Z*4	9.4	23.58	9.84	32.41	8.45	14.74	24.99	3.31
*H*1	371.9	0.09	107.2	40.05	525.6	69.98	230.7	19.57
*H*2	2291	5.46	132.9	22.82	47.28	101.7	624.3	21.32
*H*3	123	1.2	27.45	7.6	67.31	6.77	117.3	0.50
*H*4	78.83	7.14	25.2	95.72	12.55	2.11	29.06	6.69

The component planes j of the SOM, i.e., the array of scalar values μ_{ij} representing the jth components of the weight vectors $\mathbf{m}_i$ and having the same format as the SOM array, are displayed as shades of gray.

3 Statistical Analysis of Image Quality Measures

Our first goal is to investigate the sensitivity of quality measures to distortions that arise from image compression schemes, in other words, to find the degree to which a quality measure can discriminate the coding artifacts and translate it into a meaningful score. We similarly establish the response sensitivity of the measures to other causes of distortion such as blur and noise. Our second goal is to establish how various quality measures are related to each other and to show the degree to which measures respond (dis)similarly to coding and sensor artifacts. As the outcome of these investigations we intend to extract a subset of measures that satisfies the image quality measure desiderata.

3.1 *ANOVA Results*

The two-way ANOVA results of the image quality measures of the data obtained from all image classes (fabrics, faces, remotes) are listed in Table 2. In Table 2 the symbols of quality measures $D1, D2, \ldots, H3, H4$ are listed in the first column while the F scores of JPEG compression, of SPIHT compression, of blur and of noise distortions are given, respectively, in the remaining four columns.

The metric that responds most strongly to one distortion type is called the ''fundamental metric'' of that distortion type.[24] Note that there could be more than one fundamental metric. Similarly, the metric that responds adequately to all sorts of distortion effects is denoted as the ''global metric.'' One notices the following.

- The fundamental metrics for JPEG compression are $H2$, $H1$, $S2$, and $E2$, which is the human visual system (HVS) $L2$ norm, the HVS absolute norm, the

Table 3 Classification of metrics according to their sensitivity for different types of distortion on individual and combined image sets. The bottom two rows indicate the metrics that are least sensitive to the image set and to the coder type.

	Image set	JPEG	SPIHT	Blur	Noise
One-way ANOVA	Fabrics	*H*4, *H*2, *E*2, *S*4	*E*1, *S*1, *E*2, *S*2	*S*1, *S*5, *E*2, *S*4	*D*1, *D*2, *D*5, *D*3
	Faces	*H*2, *D*1, *S*3, *H*1	*H*4, *D*3, *H*2, *C*1	*S*2, *H*1, *S*1, *E*2	*D*1, *S*3, *D*2, *D*3
	Remote sensing	*H*2, *H*4, *S*4, *S*5	*S*2, *S*5, *S*4, *S*1	*D*6, *S*5, *S*4, *S*1	*D*1, *D*2, *C*3, *C*5
Two-way ANOVA	Combined set	*H*2, *H*1, *S*2, *E*2	*E*2, *S*2, *S*5, *H*2	*S*1, *E*2, *S*2, *H*1	*D*1, *D*2, *S*3, *D*5
	Image set independence	*H*1, *H*3	*D*4, *C*5	*C*4, *D*4	*H*3, *Z*4
	Coder type independence	*D*2, *D*1, *Z*4, *D*3			

spectral phase magnitude, and edge stability measures. These measures are listed in decreasing order of the F score.

- The fundamental metrics for SPIHT compression are $E2$, $S2$, $S5$, and $H2$, that is, edge stability, spectral phase magnitude, block spectral phase magnitude, and the HVS $L2$ norm.
- The fundamental metrics for the blur effect are $S1$, $E2$, $S2$, and $H1$, that is, spectral phase, edge stability, spectral phase magnitude, and the HVS absolute norm. Notice the similarity of the metrics between SPIHT and blur. This is due to the fact that we primarily encounter blur artifacts in wavelet-based compression.
- The fundamental metric for the noise effect is, as expected, $D1$, the mean square error.
- Finally the image quality metrics that are sensitive to all distortion artifacts are, in rank of order, $E2$, $H1$, $S2$, $H2$, and $S5$, that is, edge stability, the HVS absolute norm, spectral phase magnitude, the HVS $L2$ norm, and block spectral phase magnitude. To establish the global metrics, we gave rank numbers from 1 to 26 to each metric under the four types of distortion in Table 2. For example, for JPEG the metrics are ordered as $H2$, $H1$, $S2$, $E2$, etc., if we take into consideration their F scores. Then we summed their rank numbers, and the metrics for which the sum of the scores were the smallest were declared the global metric, that is, the ones that qualify well in all discrimination tests. These results must still be taken with some caution since, for example, none of the five winning scores is as sensitive to additive noise as the $D1$ and $D2$ scores.
- The metrics that were the least sensitive to image set variation are $D4$, $H3$, $C4$, $C5$, $D6$, etc. It can be observed that these metrics in general also show poor performance in discriminating distortion effects. On the other hand, for the distortion sensitive metrics, even though their image set dependence is higher than the so-called ''image independent'' metrics, more of the score variability is due to distortion than to image set changes. This can be observed based on the higher F scores for distortion effects compared to image set related F scores.

These observations are summarized in Table 3 where one-way results are given for each image class (fabrics, faces, remote sensing) separately, and two-way ANOVA results are presented for the combined set. In the two bottommost

Table 4 ANOVA results for the effect of bit rate (pooled data from JPEG and SPIHT) and of coder type. The degrees of freedom are 4 (bit rate) and 1 (coder type).

ANOVA2	JPEG+SPIHT	
Metric	Bit rate	Coder
*D*1	89.79	0.75
*D*2	74.98	2.72
*D*3	71.55	1.21
*D*4	70.52	43.85
*D*5	17.07	0.0005
*D*6	85.22	118.8
*C*1	2.66	45.47
*C*2	12.28	18.27
*C*3	56.48	1.56
*C*4	31.3	2.43
*C*5	78.98	2.23
*E*1	42.69	11.61
*E*2	122.4	26.28
*S*1	99.12	5.29
*S*2	140.1	12.37
*S*3	92.99	9.27
*S*4	115.5	39.1
*S*5	124.8	43.09
*Z*1	4.28	41.6
*Z*2	9.54	0.83
*Z*3	12.87	0.56
*Z*4	9.39	6.64
*H*1	278.6	52.87
*H*2	493	87.21
*H*3	97.94	16.19
*H*4	21.13	57.72

rows of Table 3 the metrics that are least sensitive to the coder type and to the image set are given. The criteria for omitting and entering the metrics in Table 3 were the outcome of the F scores.

We also investigated the metrics with respect to their capability to respond to bit rate and coder type. For this analysis the scores of the JPEG and SPIHT compressors were combined. The following can be observed in Table 4.

- The metrics that were best in discriminating compression distortion as parameterized by the bit rate, whatever the coder type, that is JPEG or SPIHT, were $H2$, $H1$, $S2$, and $S5$ (the HVS $L2$ norm, the HVS absolute norm, spectral phase magnitude, block spectral phase magnitude, etc.
- The metrics that were capable of discriminating the coder type (JPEG versus SPIHT) were similar in the sense that they all belong to the human vision system inspired types, namely, $D6$, $H2$, $H4$, and $H1$ (multiresolution error, the HVS $L2$ norm, DCTune, and the HVS $L1$ norm).
- Finally, the metrics that were most sensitive to distortion artifacts, but at the same time least sensitive to image set variation, were $C5$, $D1$, $D3$, $S3$, $D2$, $C4$, etc. (mean angle-magnitude similarity, mean square error, modified infinity norm, block spectral magnitude error, mean absolute error, mean angle similarity, etc.). These metrics were identified by summing the two rank scores of the metrics, the first being the ranks in ascending order of distortion sensitivity, the second in descending order of the image set sensitivity. Interestingly enough almost all of them are related to a variety of mean square error. Despite its many criticisms, this may explain why mean square error or signal-to-noise ratio measures have proven to be so resilient over time. Again this conclusion should be accepted with some caution. For example, common experience indicates that MSE measures do not necessarily reflect all the objectionable coding artifacts especially at low bit rates.

As expected the metrics that are responsive to distortions are also almost always responsive to the image set. Conversely, the metrics that do not respond to variation of the image set are also not very discriminating with respect to distortion types. The fact that the metrics are sensitive, as would be expected, to both the image content and distortion artifacts does not eclipse their potential as quality metrics. Indeed, when the metrics were tested in more homogeneous image sets (that is, only within face images or remote sensing images, etc.) the same high-performance metrics scored consistently higher. Furthermore, when one compares the F scores of the metrics with respect to bit rate variation and image set variation, even though there is a non-negligible interaction factor, the F score due to bit rate is always larger than the F score due to image sets.

3.2 *Self-Organizing Map of Quality Measures*

Our second investigation was of the mutual relationship between measures. It is obvious that the quality measures must be correlated with each other since most of them must respond to compression artifacts in similar ways. On the other hand, one can conjecture that some measures must be more sensitive to blurring effects, while others respond to blocking effects, while still others reflect additive noise.

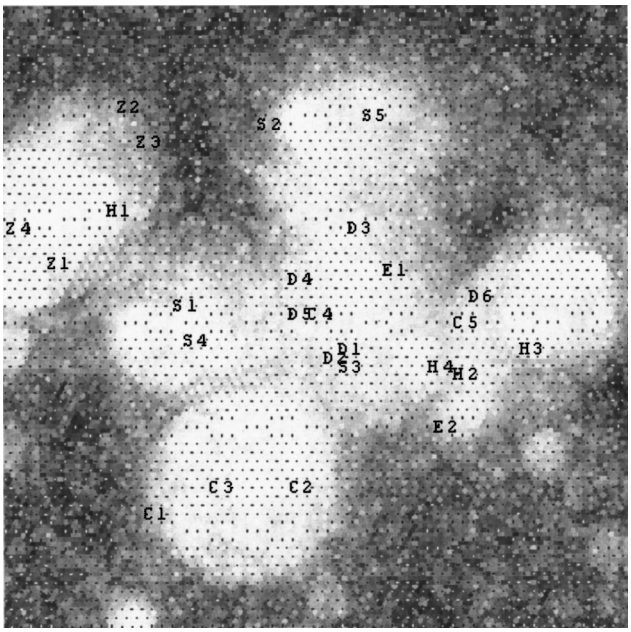


Fig. 2 SOM map of distortion measures for JPEG and SPIHT.

The SOM[29] is a pictorial method by which to display similarities and differences between statistical variables, such as quality measures. We have therefore obtained spatial organization of these measures via Kohonen's self-organizing map algorithm. The input to the SOM algorithm was vectors whose elements are the scores of the measure resulting from different images. More explicitly, consider one of the measures, let us say, $D1$, and a certain compression algorithm, e.g., JPEG. The instances of this vector will be 60 dimensional, one for each of the images in the set. The first 30 components consist of 30 images compressed with JPEG, the next 30 juxtaposed components of the same images compressed with SPIHT. Furthermore there will be five such vectors, one for each of the bit rates.

The SOM organization of the measures in 2D space for pooled data from JPEG and SPIHT coders is shown in Fig. 2. These maps are useful for visual assessment of any possible correlation present in the measures. One would expect that measures with similar trends and which respond in similar ways to artifacts would cluster together spatially. The main conclusions from observation of the SOM and the correlation matrix are the following.

- The clustering tendency of pixel difference based measures ($D1$,$D2$,$D4$,$D5$) and the spectral magnitude based method ($S3$) is obvious in the center portion of the map, a reflection of the Parseval relationship. However notice that spectral phase-magnitude measures ($S2$,$S5$) stay distinctly apart from these measures. In a similar vein purely spectral phase measures also form a separate cluster.
- The human visual system based measures ($H2$,$H3$,$H4$), multiresolution pixel-difference mea-

sure ($D6$), $E2$ (edge stability measure), and $C5$ (mean angle-magnitude measure) are clustered on the right side of the map. The correlation of the multiresolution distance measure, $D6$, with HVS based measures ($H2$,$H3$,$H4$) is not surprising since the idea behind this measure is to mimic an image comparison by eye more closely by assigning a larger weight to low resolution components and a lesser weight to the detailed high frequency components.

- The three correlation based measures ($C1$,$C2$,$C3$) are together in the lower part of the map while the two spectral phase error measures ($S2$,$S5$) are concentrated separately in the upper part of the map.
- It is interesting to note that all the context-based measures ($Z1$,$Z2$,$Z3$,$Z4$) are grouped in the upper left region of the map together with $H1$ (the HVS filtered absolute error).
- The proximity of the Pratt measure ($E1$) and the maximum difference measures ($D3$) is meaningful, since the maximum distortions in reconstructed images are near the edges. The constrained maximum distance or sorted maximum distance measures can be used in codec designs to preserve the two-dimensional features, such as edges, in reconstructed images.

In conclusion the relative positioning of measures in the two-dimensional map was in agreement with one's intuitive grouping and with the ANOVA results. We would like to emphasize here that in the above SOM discussions it is only the relative position of the measures that is significant, while their absolute positioning is arbitrary. Furthermore, the metrics that behave in an uncorrelated way in the SOM display are conjectured to respond to different distortion artifacts and are used as an additional criterion for the selection of ''good'' measure subsets.

3.3 *Combination of Quality Measures: Supermetrics*

It was conjectured that a judicious combination of image quality metrics could be more useful in image processing tasks. We present two instances of the application of an IQM combination, namely, in steganalysis and in predicting subjective quality measures.

Steganography refers to the art of secret communication while steganalysis is the ensemble of techniques that can detect the presence of watermarks and differentiate stegodocuments. For this digital watermarking is used, which consists of an imperceptible and cryptographically secure message added to the digital content, to be extracted only by the recipient. However, if digital watermarks are to be used in steganography applications, detection of their presence by an unauthorized agent defeats their very purpose. Even in applications that do not require hidden communication, but only watermarking robustness, we note that it would be desirable to first detect the possible presence of a watermark before trying to remove or manipulate it.

The underlying idea of watermarking is to create a new document, e.g., an image, which is *perceptually identical but statistically different* from the host signal. Watermark decoding uses this statistical difference in order to extract the stegomessage. However, the very same statistical difference that is created could potentially be exploited to determine if a given image is watermarked or not. The answer to this conjecture is positive in that we show that watermarking leaves unique artifacts, which can be detected using image quality measures (IQMs).[57,58]

In order to identify specific quality measures that prove useful in steganalysis, that is, distinguishing the watermarked images from the nonwatermarked ones, we again use the ANOVA test. Twenty-six quality measures are subjected to a statistical test to determine if the fluctuations of the measures result from image variety or whether they arise due to treatment effects, that is, watermarking and stego-message embedding. Thus any test resulting in a p value under 0.05 would be significant, and, therefore, one would accept the assertion that the difference in quality metric arises from the ''strength'' parameter of the watermarking or steganography artifacts, and not from variations in the image content. The idea of employing more than one IQM in the steganalyzer is justified since different watermarking algorithms mark different features of the image, such as global discrete Fourier transform (DFT) coefficients, block discrete cosine transform (DCT) coefficients, pixels directly, etc.

We performed ANOVA tests for several watermarking and steganography algorithms. For example, the most discriminating IQMs for the pooled steganography and watermarking algorithms were found to be the mean absolute error D_2, mean square error D_1, Czekonowsky correlation measure C_3, angle mean C_4, spectral magnitude distance S_2, median block spectral phase distance S_4, median block weighted spectral distance S_5, and normalized mean square HVS error H_2. The implication here is twofold: One is that, by using these features, a steganalyzer can be designed to detect the watermarked or stegoed images using multivariate regression analysis, as we showed in Refs. 57–59. This linear combination of IQMs for steganalysis purposes is denoted as the ''supermetric'' for steganalysis. It was shown in Ref. 57 that the steganalysis supermetric can detect the presence of watermarking with 85% accuracy and can even predict whose watermark it is.[58] The other implication is that current watermarking or steganographic algorithms should exercise more care in those statistically sig-

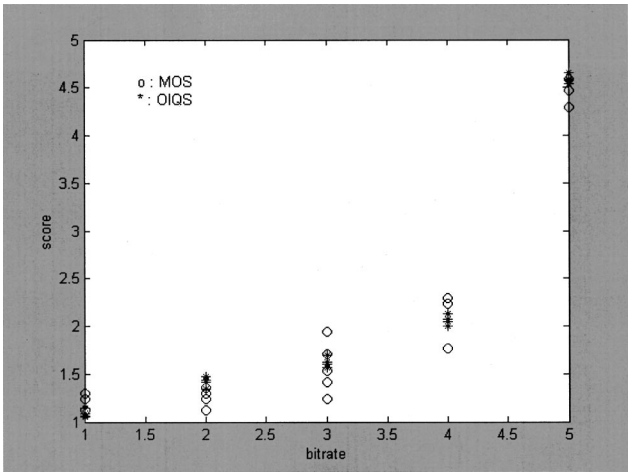


Fig. 3 Plot of the mean opinion score and image quality supermetric data.

Table 5 Image quality metrics and their correlation coefficients with MOS data.

$D1$	0.893	$C1$	0.501	$E2$	0.890	$Z1$	0.502	$H3$	0.936
$D2$	0.895	$C2$	0.810	$S1$	0.929	$Z2$	0.543	$H4$	0.982
$D3$	0.720	$C3$	0.926	$S2$	0.903	$Z3$	0.609	Supermetric	0.987
$D4$	0.901	$C4$	0.912	$S3$	0.930	$Z4$	0.517		
$D5$	0.381	$C5$	0.917	$S4$	0.883	$H1$	0.890		
$D6$	0.904	$E1$	0.833	$S5$	0.865	$H2$	0.938		

nificant image features to eschew detection.[59]

For the second supermetric we searched for a correlation between the subjective opinions and an objective measure derived from a combination of our IQMs. The subjective image quality experiment was conducted with a group of 17 subjects (students that first took a course in image processing) who noted their image quality opinion scores in the 1–5 range, 1 being no distortion could be observed and 5 meaning very annoying quality. The time of observation was unlimited. The images used were all 512×512 red–green–blue (RGB) color images from the Purdue University face database, and were viewed at $4\times$ the image height. The results reported are based on 850 quality evaluations of 50 encoded images (10 images compressed with JPEG at five different quality scales, $Q=10$, 30, 50, 70, and 90) by the pool of 17 subjects. The supermetric of image quality for compression artifacts was built using global metrics $E2$, $H1$, $S2$, $H2$, and $S5$, that is, the edge stability, HVS absolute norm, spectral phase magnitude, HVS $L2$ norm, and block spectral phase magnitude) for the image distortions due to compression. The supermetric was built by regressing them against the mean opinion scores (MOS). The plot of this supermetric and MOS data are given in Fig. 3, where a high value of the correlation coefficient was determined: 0.987. The correlation coefficients of the individual metrics, shown in Table 5, were all lower, as expected.

4 Conclusions

In this work we have presented collectively a comprehensive set of image quality measures and categorized them. Using statistical tools we were able to classify more than two dozen measures based on their sensitivity to different types of distortions.

Statistical investigation of 26 different measures using ANOVA analyses has revealed that local phase-magnitude measures ($S2$ or $S5$), HVS-filtered $L1$ and $L2$ norms ($H1$ and $H2$), and the edge stability measure ($E2$) are most sensitive to coding and blur artifacts, while the mean square error ($D1$) remains the best measure for additive noise. These ''winning'' metrics were selected on the basis of the sum of the rank scores over four artifacts: JPEG-compression/SPIH-compression, blur, and noise. This preselection of the $E2$, $S2$, $S5$, $H1$, and $H2$ subset was based, on the one hand, on their superior F scores and, on the other hand, on the fact they appeared to behave in an uncorrelated way in their SOM maps.

These metrics satisfied, in their category of distortion, the IQM desiderata given in Sec. 2.1, namely, accuracy, monotonicity, and consistency. The $H1$, $H2$, $S2$, $S5$, and $D1$ metrics were accurate in that they responded predominantly to the type of distortion stated than to any other factor. They responded monotonically to the level of distortion, that is, the metric versus distortion parameter plotted monotonically (graph not shown). Finally their consistency was shown when they were tested on widely differing image classes (faces, textures, remote sensing).

Ideally speaking, one would like to have a quality measure that is able to give accurate results for different levels of performance of a given compression scheme, and across different compression schemes. It appears that, as shown in Sec. 3.3, a combination of spectral phase-and-magnitude measures and of the HVS-filtered error norm comes closest to satisfying such a measure, because it is sufficiently sensitive to a variety of artifacts. The Kohonen map of the measures has been useful in depicting measures that behave similarly or in an uncorrelated way. The correlation between various measures as are depicted in Kohonen's self-organizing map.

In conclusion, the subsets of the $E2$, $S2$, $S5$, $H1$, and $H2$ metrics are the prominent image quality measures, as shown from both ANOVA analysis and MOS scores points of view. The implication is that more attention should be paid to the spectral phase and HVS-filtered quality metrics in the design of coding algorithms and sensor evaluation. We have also shown the validity of the ANOVA methodology in an alternate application, that is, when we applied it to the selection of IQMs for the construction of a steganalyzer.

In future work we will address extension of the subjective experiments. Note that we have only shown in one experiment that the IQMs selected regress well in the mean opinion scores. However this experiment must be repeated on yet unseen data to understand how well it predicts a subjective opinion. In a similar vein the database for detection experiments will be extended to cover a larger variety of watermarking and steganography tools.

Acknowledgments

The authors would like to thank H. Brettel (ENST, France), A. Eskicioglu (Thompson Communication, Indianapolis, USA), as well as an anonymous reviewer for their invaluable help in improving this paper.

Appendix

Here in the Appendix we define and describe the multitude of image quality measures considered. In these definitions the pixel lattices of images A and B will be referred to as $A(i,j)$ and $B(i,j)$, $i, j=1,\ldots,N$, since the lattices are assumed to have dimensions of $N\times N$. The pixels can take

values from the set $\{0,\ldots,G\}$ in any spectral band. The actual color images we considered had G=255 in each band. Similarly we will denote the multispectral components of an image at pixel positions *i* and *j*, and in band *k* as $C_k(i,j)$, where $k=1,\ldots,K$. The boldface symbols $\mathbf{C}(i,j)$ and $\hat{\mathbf{C}}(i,j)$ will indicate the multispectral pixel vectors at position (*i,j*). For example, for the color images in the RGB representation one has $\mathbf{C}(i,j)=[\mathrm{R}(i,j)\ \mathrm{G}(i,j)\ \mathrm{B}(i,j)]^T$. All these definitions are summarized in the following:

$C_k(i,\ j)$ $(i,\ j)$th pixel of the kth band of image C

$\mathbf{C}(i,\ j)$ $(i,\ j)$th multispectral (with K bands) pixel vector

$\mathbf{C}$ multispectral image

C_k kth band of a multispectral image C

$\varepsilon_k=C_k-\hat{C}_k$ error over all the pixels in the kth band of a multispectral image C

Thus, for example, the power in the kth band can be calculated as $\sigma_k^2=\Sigma_{i,j=0}^{N-1}C_k(i,j)^2$. All the quantities with a caret, i.e., $\hat{C}_k(i,j)$, $\hat{\mathbf{C}}$, etc., will correspond to distorted versions of the same original image. As a case in point, the expression $\|\mathbf{C}(i,j)-\hat{\mathbf{C}}(i,j)\|^2=\Sigma_{k=1}^{K}[C_k(i,j)-\hat{C}_k(i,j)]^2$ will denote the sum of errors in the spectral components at given pixel positions *i,j*. In a similar way the error in the last row of the above minitable expands as $\varepsilon_k^2=\Sigma_{i=1}^{N}\Sigma_{j=1}^{N}[C_k(i,j)-\hat{C}_k(i,j)]^2$. In the specific case of RGB color images we will occasionally revert back to notations {R, G, B} and $\{\hat{\mathrm{R}},\hat{\mathrm{G}},\hat{\mathrm{B}}\}$.

A *Measures Based on Pixel Differences*

The measures here calculate the distortion between two images on the basis of their pixelwise differences or certain moments of the difference (error) image.

A.1.1 *Minkowsky metrics*

The L_γ norm of the dissimilarity of two images can be calculated by taking the Minkowsky average of pixel differences spatially and then chromatically (that is, over the bands):

$$\varepsilon^\gamma=\frac{1}{K}\sum_{k=1}^{K}\left\{\frac{1}{N^2}\sum_{i,j=0}^{N-1}|C_k(i,j)-\hat{C}_k(i,j)|^\gamma\right\}^{1/\gamma}. \tag{A1}$$

Or, the Minkowsky average can first be carried over the bands and then spatially, as in the following expression:

$$\varepsilon^\gamma=\frac{1}{N^2}\left\{\sum_{i,j=0}^{N-1}\left[\frac{1}{K}\sum_{k=1}^{K}|C_k(i,j)-\hat{C}_k(i,j)|\right]^\gamma\right\}^{1/\gamma}.$$

In what follows we have used the pixelwise difference in the Minkowsky sum given in Eq. (A1). For $\gamma=2$, one obtains the well-known mean square error expression, denoted as $D1$:

$$D1=\frac{1}{K}\frac{1}{N^2}\sum_{i,j=0}^{N-1}\|\mathbf{C}(i,j)-\hat{\mathbf{C}}(i,j)\|^2=\frac{1}{K}\sum_{k=1}^{K}\varepsilon_k^2. \tag{A2}$$

An overwhelming number of quality results in the literature is in fact given in terms of the SNR or the peak SNR (PSNR), which are obtained, respectively, by dividing the image power by $D1$ and by dividing the peak power G^2 by $D1$. Although the SNR and the PSNR are very frequently used in quantifying coding distortions, their shortcomings have been pointed out in various studies.[13] However, despite these oft cited criticisms of MSE-based quality measures there has been a recent resurgence of SNR/PSNR metrics.[17,18] For example, studies of the video quality expert Group (VQEG)[17] have shown that the PSNR measure is a very good indicator of subjective preference in video coding.

For $\gamma=1$ one obtains the absolute difference, denoted as $D2$. For $\gamma=\infty$ power in the Minkowski average the maximum difference measure,

$$\varepsilon^\infty=\max_{i,j}\sum_{k=1}^{K}\frac{1}{K}|C_k(i,j)-\hat{C}(i,j)|=\max_{i,j}\|\mathbf{C}(i,j)-\hat{\mathbf{C}}(i,j)\|,$$

is obtained. Recall that in signal and image processing the maximum difference or the infinity norm is very commonly used.[6] However given the noise-prone nature of the maximum difference, this metric can be made more robust by considering the ranked list of pixel differences $\Delta_l(\mathbf{C}-\hat{\mathbf{C}})$, $l=1,\ldots N^2$, resulting in a modified Minkowski infinity metric, called $D3$. Here $\Delta_l(\mathbf{C}-\hat{\mathbf{C}})$ denotes the lth largest deviation among all pixels.[31] Thus $\Delta_l(\mathbf{C}-\hat{\mathbf{C}})$ is simply the error expression ε^∞ above. Similarly, Δ_2 corresponds to the second largest term, etc. Finally a modified maximum difference measure using the first r of Δ_m terms can be constructed by computing the root mean square value of the ranked largest differences, $\Delta_l,\ldots,\Delta_r$.

$$D3=\sqrt{\frac{1}{r}\sum_{m=1}^{r}\Delta_m^2(\mathbf{C}-\hat{\mathbf{C}})}. \tag{A3}$$

A.1.2 *MSE in L*a*b* space*

The choice of color space for measuring image similarity is important, because the color space must be uniform, so the intensity difference between the two colors must be consistent with the color difference estimated by a human observer. Since the RGB color space is not well suited to this task two color spaces are defined: 1976 CIE $L^*u^*v^*$ and 1976 CIE $L^*a^*b^*$ color spaces.[32] One recommended color-difference equation for the Lab color space is given by the Euclidean distance.[33] Let

$$\Delta L^*(i,j)=L^*(i,j)-\hat{L}^*(i,j),$$

$$\Delta a^*(i,j)=a^*(i,j)-\hat{a}^*(i,j),$$

$$\Delta b^*(i,j)=b^*(i,j)-\hat{b}^*(i,j),$$

denote the color component differences in $L^*a^*b^*$ space. Then the Euclidean distance is

$$D4=\frac{1}{N^2}\sum_{i,j=0}^{N-1}[\Delta L^*(i,j)^2+\Delta a^*(i,j)^2+\Delta b^*(i,j)^2]. \tag{A4}$$

Note that Eq. (A4) is intended to yield a perceptually uniform spacing of colors that exhibit color differences greater than the just-noticeable difference (JND) threshold but smaller than those in the Munsell book of color.[33] This measure applies obviously to color images only and cannot be generalized to arbitrary multispectral images. Therefore it has been used only for the face images and texture images, not the satellite images.

A.1.3 *Difference over a neighborhood*

Image distortion on a pixel level can arise from differences in the gray level of the pixels and/or from displacements of the pixel. A distortion measure that penalizes in a graduated way spatial displacements in addition to gray level differences, and that allows therefore some tolerance for pixel shifts can be defined as follows:[34,35]

$$D5=\sqrt{\frac{1}{2(N-w)^2}\sum_{i,j=w/2}^{N-w/2}\left(\min_{l,m\in w_{i,j}}\{d[\mathbf{C}(i,j),\hat{\mathbf{C}}(l,m)]\}\right)^2+\left(\min_{l,m\in w_{i,j}}\{d[\hat{\mathbf{C}}(i,j),\mathbf{C}(l,m)]\}\right)^2}, \tag{A5}$$

where $d(\cdot,\cdot)$ is some appropriate distance metric. Notice that for $w=1$ this metric reduces to the mean square error like in $D1$.

Thus for any given pixel $\mathbf{C}(i,j)$, we search for the best matching pixel in the d distance sense in the wxw neighborhood of pixel $\hat{\mathbf{C}}(i,j)$, denoted as $\hat{\mathbf{C}}_w(i,j)$. The size of the neighborhood is typically small, e.g., 3×3 or 5×5, and one can consider a square or a cross-shaped support. Similarly, one calculates the distance from $\hat{C}(i,j)$ to $C_w(i,j)$ where again $C_w(i,j)$ denotes the pixels in the wxw neighborhood of coordinates (i,j) of $C(i,j)$. Note that in general $d[\mathbf{C}(i,j),\hat{\mathbf{C}}_w(i,j)]$ is not equal to $d[\hat{\mathbf{C}}(i,j),\mathbf{C}_w(i,j)]$. As for the distance measure $d(\cdot,\cdot)$, a city metric or a chessboard metric can be used. For example, a city block metric becomes

$$d^{\text{city}}[\mathbf{C}(i,j),\hat{\mathbf{C}}(l,m)]=\frac{(|i-1|+|j-m|)}{N}+\frac{\|\mathbf{C}(i,j)-\hat{\mathbf{C}}(l,m)\|}{G},$$

where $\|\cdot\|$ denotes the norm of the difference between $\mathbf{C}(i,j)$ and $\hat{\mathbf{C}}(i,j)$ vectors. Thus both the pixel color difference and search displacement are considered. In this expression N and G are one possible set of normalization factors with which to tune deviations due to pixel shifts and pixel spectral differences, respectively. In our measurements we have used the city block distance with a 3×3 neighborhood size.

A.1.4 *Multiresolution distance measure*

One limitation of standard objective measures of distance between images is that the comparison is conducted at full image resolution. Alternative measures can be defined that resemble image perception in the human visual system more closely by assigning larger weights to low resolutions and smaller weights to the detail image.[36] Such measures are also more realistic for machine vision tasks that often use local information only.

Consider the various levels of resolution denoted by $r\geqslant1$. For each value of r the image is split into blocks b_1 to b_n where n depends on scale r. For example, for $r=1$, at the lowest resolution, only one block covers the whole image characterized by its average gray level g. For $r=2$ one has four blocks each $N/2\times N/2$ with average gray levels of g_{11}, g_{12}, g_{21}, and g_{22}. For the rth resolution level one would then have 2^{2r-2} blocks of size $N/2^{r-1}\times N/2^{r-1}$, characterized by the block average gray levels g_{ij}, $i,j=1,\ldots,2^{2r-2}$. Thus for each block b_{ij} of image C, take g_{ij} as its average gray level and $\hat{g}_{ij}$ to correspond to its component in image $\hat{C}$ (for simplicity a third index that denotes the resolution level was omitted). The average difference in gray level at resolution r has weight of $1/2^r$. Therefore the distortion at this level is

$$d_r=\frac{1}{2^r}\frac{1}{2^{2r-2}}\sum_{i,j=1}^{2^{r-1}}|g_{ij}-\hat{g}_{ij}|,$$

where 2^{r-1} is the number of blocks along either the i and j indices. If one considers a total of R resolution levels, then a distance function can be found simply by summing over all the resolution levels, $r=1,\ldots,R$, that is, $D(C,\hat{C})=\Sigma_{r=1}^{R}d_r$. The actual value of R (the number of resolution levels) will be set by the initial resolution of the digital image. For example, for a 512×512 image one has $R=9$. Finally, for multispectral images one can extend this definition in two ways. In a straightforward extension, one sums the multiresolution distances d_r^k over the bands,

$$D6=\frac{1}{K}\sum_{k=1}^{K}\sum_{r=1}^{R}d_r^k, \tag{A6}$$

where d_r^k is the multiresolution distance in the kth band. This is the multiresolution distance definition that we used in the experiments. As an alternative, a Burt pyramid was constructed to obtain a multiresolution representation. However in the tests it did not perform as well as the pyramid described in Ref. 36.

A different way in which to define the multiresolution distance would be to consider the vector difference of pixels:

$$D(C,\hat{C})=\sum_{r=1}^{R} d_r', \text{ with}$$

$$d_r=\frac{1}{2^r}\frac{1}{2^{2r-2}}\sum_{i,j=1}^{2^{r-1}}[(g_{ij}^R-\hat{g}_{ij}^R)^2+(g_{ij}^G-\hat{g}_{ij}^G)^2 +(g_{ij}^B-\hat{g}_{ij}^B)^2]^{1/2},$$

where, for example, g_{ij}^R is the average gray level of the i,jth block in the ''red'' component of the image at (implicit) resolution level r. Notice that in the latter equation the Euclidean norm of the differences of the block average color components R, G, and B have been utilized.

Notice also that the last two measures, that is, the neighborhood distance measure and the multiresolution distance measure, have not been previously used in evaluating compressed images.

B Correlation-Based Measures

B.1 *Image Correlation Measures*

The similarity between two digital images can also be quantified in terms of the correlation function.[5] These measures measure the similarity between two images, hence in this sense they are complementary to the difference-based measures: Some correlation based measures are the following.

Structural content:

$$C1=\frac{1}{K}\sum_{k=1}^{K}\frac{\Sigma_{i,j=0}^{N-1}C_k(i,j)^2}{\Sigma_{i,j=0}^{N-1}\hat{C}_k(i,j)^2}. \tag{A7}$$

normalized cross-correlation measure:

$$C2=\frac{1}{K}\sum_{k=1}^{K}\frac{\Sigma_{i,j=0}^{N-1}C_k(i,j)\hat{C}_k(i,j)}{\Sigma_{i,j=0}^{N-1}C_k(i,j)^2}. \tag{A8}$$

Czenakowski distance: A metric that is useful for comparing vectors with strictly non-negative components, like in the case of color images, is given by the Czenakowski distance:[37]

$$C3=\frac{1}{N^2}\sum_{i,j=0}^{N-1}\left(1-\frac{2\Sigma_{k=1}^{K}\min[C_k(i,j),\hat{C}_k(i,j)]}{\Sigma_{k=1}^{K}[C_k(i,j)+\hat{C}_k(i,j)]}\right). \tag{A9}$$

The Czenakowski coefficient[38] (also called the percentage of similarity) measures the similarity among different samples, communities, and quadrates.

Obviously as the difference between two images tends towards zero $\varepsilon=C-\hat{C}\to 0$, all the correlation-based measures tend towards 1, while as $\varepsilon^2\to G^2$ they tend towards 0. Recall also that distance measures and correlation measures are complementary, so that under certain conditions, minimizing distance measures is tantamount to maximizing the correlation measure.[39]

B.1.2 *Moments of the angles*

A variant of correlation-based measures can be obtained by considering the statistics of the angles between the pixel vectors of the original and coded images. Similar ''colors'' will result in vectors pointing in the same direction, while significantly different colors will point in different directions in **C** space. Since we deal with positive vectors **C**, $\hat{\mathbf{C}}$, we are constrained to one quadrant of Cartesian space. Thus the normalization factor of $2/\pi$ is related to the fact that the maximum difference attained will be $\pi/2$. The combined angular correlation and magnitude difference between two vectors can be defined as[37,40]

$$\chi_{ij}=1-\left[1-\frac{2}{\pi}\cos^{-1}\frac{\langle\mathbf{C}(i,j),\hat{\mathbf{C}}(i,j)\rangle}{\|\mathbf{C}(i,j)\|\|\hat{\mathbf{C}}(i,j)\|}\right]\times\left[1-\frac{\|\mathbf{C}(i,j)-\hat{\mathbf{C}}(i,j)\|}{\sqrt{3\times 255^2}}\right].$$

We can use the moments of the spectral (chromatic) vector differences as distortion measures. To this end we have used the mean of the angle difference ($C4$) and the mean of the combined angle-magnitude difference ($C5$) in the following two measures:

$$C4=\mu_\chi=1-\frac{1}{N^2}\sum_{i,j=1}^{N}\left(\frac{2}{\pi}\cos^{-1}\frac{\langle\mathbf{C}(i,j),\hat{\mathbf{C}}(i,j)\rangle}{\|\mathbf{C}(i,j)\|\|\hat{\mathbf{C}}(i,j)\|}\right), \tag{A10}$$

$$C5=\frac{1}{N^2}\sum_{i,j=1}^{N}\chi_{ij}, \tag{A11}$$

where μ_χ is the mean of the angular differences. These moments have previously been used to assess the directional correlation among color vectors.

C Edge Quality Measures

According to the contour-texture paradigm of images, the edges form the most informative part of the image. For example, in the perception of scene content by the human visual system, edges play a major role. In a similar way, machine vision algorithms often rely on feature maps obtained from the edges. Thus, task performance in vision, whether by humans or machines, is highly dependent on the quality of the edges and other two-dimensional features such as corners.[9,41,42] Some examples of edge degradation are discontinuities at the edge, a decrease in edge sharpness by smoothing effects, offset of the edge position, missing edge points, falsely detected edge points, etc.[39] Notice, however, that all the above degradations are not necessarily observed since edge and corner information in images is rather well preserved by most compression algorithms.

Since we do not possess the ground-truth edge map, we have used the edge map obtained from the original uncom-

pressed images as a reference. Thus to obtain edge-based quality measures we have generated edge fields from both the original and compressed images using a Canny detector.[43] We have not used any multiband edge detector; instead a separate edge map from each band has been obtained. The outputs of the derivative of the Gaussians of each band are averaged, and the resulting average output is interpolated, thresholded, and thinned in a manner similar to that in Ref. 12. The parameters are set like those in Ref. 43 at robotics.eecs.berkeley.edu/~sastry/ee20/cacode.html.

In summary, for each band $k=1,\ldots K$, horizontal and vertical gradients and their norms, G_x^k, G_y^k and $N^k=\sqrt{G_x^{k^2}+G_y^{k^2}}$ are found. Their average over bands is calculated and thresholded with $T=\alpha(T_{\max}-T_{\min})+T_{\min}$, where $T_{\max}=1/K\sum_k \max(N^k)$ and $T_{\min}=1/K\sum_k \min(N^k)$, $\alpha=0.1$. Finally they are thinned by interpolation to find the pixels in which the norms of gradient constitute the local maxima.

C.1 *Pratt Measure*

A measure introduced by Pratt[39] considered both the accuracy of the edge location and missing/false alarm edge elements. This measure is based on knowledge of an ideal reference edge map, in which the reference edges should preferably have a width of one pixel. The figure of merit is defined as

$$E1=\frac{1}{\max\{n_d,n_t\}}\sum_{i=1}^{n_d}\frac{1}{1+ad_i^2}, \tag{A12}$$

where n_d and n_t are the number of detected and ground-truth edge points, respectively, and d_i is the distance to the closest edge possible for the ith edge pixel detected. In our study the binary edge field obtained from the uncompressed image is considered the "ground truth," or the reference edge field. The factor $\max\{n_d,n_t\}$ penalizes the number of false alarm edges or, conversely, missing edges.

This scaling factor provides the relative weighting between smeared edges and thin but offset edges. The terms in the sum penalize possible shifts from the correct edge positions. In summary the smearing and offset effects are both included in the Pratt measure, which provides an impression of overall quality.

C.2 *Edge Stability Measure*

Edge stability is defined as the consistency of edge that is evident across different scales in both the original and coded images.[44] Edge maps at different scales have been obtained from the images using the Canny[43] operator for different scale parameters (with standard deviation of the Gaussian filter assuming values of $\sigma_m=1.19$, 1.44, 1.68, 2.0, and 2.38). The output of this operator at scale m is decided at the threshold T^m, where $T^m=0.1(C_{\max}-C_{\min})+C_{\min}$. In this expression $C_{\max}$ and $C_{\min}$ denote, respectively, the maximum and minimum values of the norm of the gradient output in that band. Thus the edge map at scale σ_m of image C is obtained as

$$E(i,j,\sigma_m)=\begin{cases}1 & C^m(i,j)>T^m \text{ at } (i,j),\\ 0 & \text{otherwise,}\end{cases}$$

where $C^m(i,j)$ is the output of the derivative of the Gaussian operator at the mth scale. In other words, using a continuous function notation one has $C^m(x,y)=C(x,y)^{**}G_m(x,y)$ where

$$G_m(x,y)=\frac{1}{2\pi\sigma_m^4}xy\exp\left\{-\frac{x^2+y^2}{2\sigma_m^2}\right\}.$$

An edge stability map $Q(i,j)$ is obtained by considering the longest subsequence $E(i,j,\sigma_m),\ldots,E(i,j,\sigma_{m+l-1})$ of edge images such that

$$Q(i,j)=l,$$

where

$$l=\arg\max_l \bigcap_{\sigma_m\leqslant\sigma_k\leqslant\sigma_{m+l-1}} E(i,j,\sigma_k)=1.$$

The edge stability index calculated from the distorted image at pixel position i,j will be denoted by $\hat{Q}(i,j)$. We have used five scales to obtain the edge maps of five band-pass filtered images. Then a fidelity measure called the edge stability mean square error (ESMSE) can be calculated by summing the differences in the edge stability indices over all edge pixel positions, n_d, that is, the edge pixels of the ground-truth (undistorted) image at full resolution.

$$E2=\frac{1}{n_d}\sum_{i,j=0}^{n_d}[Q(i,j)-\hat{Q}(i,j)]^2. \tag{A13}$$

For multispectral images the index in Eq. (A13) can simply be averaged over the bands. Alternatively, a single edge field from multiband images[45,46] can be obtained and the resulting edge discrepancies measured like in Eq. (A13).

A property that is complementary to edge information could be surface curvature,[47] which is a useful feature for scene analysis, feature extraction, and object recognition. Estimates of local surface types,[48] based on the signs of the mean and Gaussian curvatures, have been widely used for image segmentation and classification algorithms. If one models a gray level image as a three-dimensional (3D) topological surface, then one can analyze this surface locally using differential geometry. A measure based on the discrepancy of mean and Gaussian curvatures between an image and its distorted version was used in Ref. 49. However this measure was not pursued further due to the subjective assignment of weights to the surface types and the fact that this measure did not perform particularly well in preliminary tests.

D Spectral Distance Measures

In this category we consider the distortion penalty functions obtained from the complex Fourier spectrum of images.[10,30]

D.1 *Magnitude and Phase Spectrum*

Let the DFT of the kth band of the original and coded images be denoted by $\Gamma_k(u,v)$ and $\hat{\Gamma}_k(u,v)$, respectively. The spectra are defined as

$$\Gamma_k(u,v)=\sum_{m,n=0}^{N-1} C_k(m,n)\exp\left[-2\pi i m\frac{u}{N}\right]\times\exp\left[-2\pi i n\frac{v}{N}\right],\quad k=1,\ldots,K.$$

Spectral distortion measures, using difference metrics like, for example, those given in Eqs. (A1)–(A3), can be extended to multispectral images. To this end considering the phase and magnitude spectra, that is,

$$\varphi(u,v)=\arctan[\Gamma(u,v)],$$

$$M(u,v)=|\Gamma(u,v)|,$$

the distortion that occurs in the phase and magnitude spectra can be calculated and weighted separately. Thus one can define the spectral magnitude distortion,

$$S=\frac{1}{N^2}\sum_{u,v=0}^{N-1}|M(u,v)-\hat{M}(u,v)|^2,$$

the spectral phase distortion,

$$S1=\frac{1}{N^2}\sum_{u,v=0}^{N-1}|\varphi(u,v)-\hat{\varphi}(u,v)|^2, \tag{A14}$$

and the weighted spectral distortion,

$$S2=\frac{1}{N^2}\left(\lambda\sum_{u,v=0}^{N-1}|\varphi(u,v)-\hat{\varphi}(u,v)|^2+(1-\lambda)\times\sum_{u,v=0}^{N-1}|M(u,v)-\hat{M}(u,v)|^2\right), \tag{A15}$$

where λ is chosen to attach commensurate weights to the phase and magnitude terms. These ideas can be extended in a straightforward manner to multiple band images by summing over all band distortions. In the following computations, λ is chosen so as to render the contributions of the magnitude and phase terms commensurate, so that $\lambda=2.5\times10^{-5}$.

Due to the localized nature of distortion and/or the nonstationary image field, Minkowsky averaging of block spectral distortions may be more advantageous. An image is divided into nonoverlapping or overlapping L blocks of $b\times b$, say, 16×16, and blockwise spectral distortions like those in Eqs. (A14) and (A15) can be computed. Let the DFT of the lth block of the kth band image $C_k^l(m,n)$ be $\Gamma_k^l(u,v)$:

$$\Gamma_k^l(u,v)=\sum_{m,n=0}^{b-1} C_k^l(m,n)\exp\left[-2\pi i m\frac{u}{b}\right]\exp\left[-2\pi i n\frac{v}{b}\right],$$

where $u,\ v=-b/2,\ldots,b/2$ and $l=1,\ldots,L$, or in magnitude-phase form

$$\Gamma_k^l(u,v)=|\Gamma_k^l(u,v)|e^{i\phi_k^l(u,v)}=m_k^l(u,v)e^{\phi_k^l(u,v)}.$$

Then the following measures can be defined in the transform domain over the lth block:

$$J_M^l=\frac{1}{K}\sum_{k=1}^{K}\left(\sum_{u,v=0}^{b-1}[|\Gamma_k^l(u,v)|-|\hat{\Gamma}_k^l(u,v)|]^\gamma\right)^{1/\gamma},$$

$$J_\varphi^l=\frac{1}{K}\sum_{k=1}^{K}\left(\sum_{u,v=0}^{b-1}[|\phi_k^l(u,v)|-|\hat{\phi}_k^l(u,v)|]^\gamma\right)^{1/\gamma},$$

$$J^l=\lambda J_M^l+(1-\lambda)J_\varphi^l,$$

with λ the relative weighting factor of the magnitude and phase spectra. Obviously the measures of Eqs. (A16)–(A18) are special cases of the above definitions for block size b that cover the whole image. Various rank order operations of the block spectral differences J_M and/or J_φ can prove useful. Thus let $J^{(1)},\ldots,J^{(L)}$ be the rank ordered block distortions, such that, for example, $J^{(L)}=\max_l\{J^l\}$. Then one can consider the following rank order averages: median block distortion, $\frac{1}{2}(J^{L/2}+J^{(L+1/2)})$, maximum block distortion, $J^{(L)}$, and average block distortion, $1/L\,\Sigma_{i=1}^{L}J^{(i)}$. We have found that the median of the block distortions is the most effective averaging of rank ordered block spectral distortions and we have thus used

$$S3=\operatorname*{median}_l\ J_m^l, \tag{A16}$$

$$S4=\operatorname*{median}_l\ J_\phi^l, \tag{A17}$$

$$S5=\operatorname*{median}_l\ J^l. \tag{A18}$$

In this study we have averaged the block spectra with $\gamma=2$ and for the choice of block size we have found that block sizes of 32 and 64 yield better results than sizes in the lower or higher range.

E Context Measures

Most of the compression algorithms and computer vision tasks are based on neighborhood information of the pixels. In this sense any loss of information in the pixel neighborhoods, that is, damage to the pixel context, could be a good measure of overall image distortion. Since such statistical information lies in the context probabilities, that is, the joint probability mass function (PMF) of pixel neighborhoods, changes in the context probabilities should be indicative of image distortion.

A major hurdle in the computation of context distortion is the requirement to calculate the high dimensional joint probability mass function. Typical PMF dimensions would be of the order of $s=10$ at least. Consequently one incurs the ''curse of dimensionality problems.'' However, as de-

tailed in Refs. 50 and 51, this problem can be solved by judicious usage of kernel estimation and cluster analysis. One modification of the kernel method is to identify the important regions in a s-dimensional space X^s by cluster analysis and to fit region-specific kernels to these locations. The result is a model that represents both mode and tail regions of PMFs well, while combining the summarizing strength of histograms with the generalizing property of kernel estimates.

In what follows we have used a causal neighborhood of pixels, i.e., $C_k(i,j)$, $C_k(i-1,j)$, $C_k(i,j-1)$, $C_k(i-1,j-1)$, $k=1, 2, 3$. Hence we have derived $s=12$ dimensional PMF's obtained from four-pixel neighborhoods in the three bands.

E.1 *Rate-Distortion Based Distortion Measure*

A method by which to quantify the changes in context probabilities is the relative entropy,[52] defined as

$$D(p\|\hat{p})=\sum_{\mathbf{x}\in X^s} p(\mathbf{x})\log\frac{p(\mathbf{x})}{\hat{p}(\mathbf{x})},$$

where X^s denotes an s-pixel neighborhood and $\mathbf{x}=[x_1,\ldots,x_s]$ a random vector. Furthermore, p and $\hat{p}$ are the PMFs of the original image context and that of the distorted (e.g., blurred, noisy, compressed, etc.) image. The relative entropy is directly related to the efficiency in compression and the error rate in classification. Recall also that the optimal average bit rate is the entropy of x,

$$H(X)=-\sum_{X\in X^s} p(X)\log p(X)=R(p).$$

If, instead of the true probability, a perturbed version $\hat{p}$, that is, the PMF of the distorted image, is used, then the average bit rate $R(\hat{p})$ becomes

$$R(\hat{p})=-\sum_{\mathbf{X}\in X^s} p(\mathbf{X})\log_2\hat{p}(\mathbf{X})=H(\mathbf{X})+D(p\|\hat{p}).$$

The increase in the entropy rate is also indicative of how much the context probability differs from the original due to coding artifacts. However we do not know the true PMF p nor, hence, its rate. We can bypass this problem by comparing two competing compression algorithms in terms of the resulting context probabilities $\hat{p}_1$ and $\hat{p}_2$. If $\hat{p}_1$ and $\hat{p}_2$ are the PMFs that result from the two compressed images, then their difference in relative entropy,

$$Z1=D(p\|\hat{p}_1)-D(p\|\hat{p}_2)=R(\hat{p}_1)-R(\hat{p}_2), \tag{A19}$$

is easily and reliably estimated from a moderate-size sample by subtracting the sample average of $-\log\hat{p}_2$ from that of $-\log\hat{p}_1$.[51] The comparison can be carried out for more than two images compressed to different bit rates in a similar way, that is, by comparing them two by two since the unknown entropy term is common to all of them.

As a quality measure for images we have calculated $Z1$ for each image when they were compressed at two consecutive bit rates, for example, $R(\hat{p}_1)$ at the bit rate of quality factor 90 and $R(\hat{p}_2)$ at the bit rate of quality factor 70, etc. As an alternative, the distortion was calculated for an original image and its blurred or noise contaminated version.

E.2 *f Divergences*

Once the joint PMF of a pixel context is obtained, several information theoretic distortion measures[53] can be used. Most of these measures can be expressed in the following general form:

$$d(p,\hat{p})=g\left\{E_p\left[f\left(\frac{\hat{p}}{p}\right)\right]\right\},$$

where $\hat{p}/p$ is the likelihood of the ratio of $\hat{p}$, the context PMF of the distorted image, and of p the PMF function of the original image, and E_p is the expectation with respect to p. Some examples follows.

Hellinger distance: $f(x)=(\sqrt{x}-1)^2$, $g(x)=\frac{1}{2}x$,

$$Z2=\frac{1}{2}\int(\sqrt{\hat{p}}-\sqrt{p})^2 d\lambda. \tag{A20}$$

Generalized Matusita distance: $f(x)=|1-x^{1/r}|^r$, $g(x)=x^{1/r}$,

$$Z3=\sqrt{\int|p^{1/r}-\hat{p}^{1/r}|^r d\lambda}, \quad r\geqslant 1. \tag{A21}$$

Notice that integration in Eqs. (A20) and (A21) is carried out in s-dimensional space. Also, we have found according to ANOVA analysis that the choice of $r=5$ in the Matusita distance yields good results. Despite the fact that the PMFs do not directly reflect the structural content or the geometrical features in an image, they perform sufficiently well to differentiate artifacts between the original and test images.

E.3 *Local Histogram Distances*

In order to reflect the differences between two images at the local level, we calculated the histograms of the original and distorted images on the basis of 16×16 blocks. To this end we considered both the Kolmogorov–Smirnov (KS) distance and the Spearman rank correlation (SRC).

For the KS distance we calculated the maximum deviation between the respective cumulatives. For each of the 16×16 blocks of the image, the maximum of the KS distances over the K spectral components was found and these local figures were summed over all the blocks to yield $\Sigma_{u=1}^{b}\max_{k=1,\ldots,K}\{KS_u^k\}$ where KS_b^k denotes the Kolmogorov–Smirnov distance of block number u and of the kth spectral component. However the KS distance did not turn out to be effective in the ANOVA tests. Instead the SRC measure had better performance. We again considered the SRC on a 16×16 block basis and we took the maximum over the three spectral bands. The block SRC measure was computed by computing the rank scores of the ''gray'' levels in the bands and for each pixel the largest of the three scores was selected. Then the correlation of the block ranks of the original and distorted images was calculated:

$$Z4=\sum_{u=1}^{b}\max_{k=1,\ldots,K}\{\mathrm{SRC}_u^k\}, \tag{A22}$$

where SRC_u^k denotes the Spearman rank correlation for the uth block number and the kth spectral band.

F Human Visual System Based Measures

Despite the search for an objective image distortion measure it is intriguing to learn the role of HVS-based measures. The HVS is too complex to be fully understood with the present psychophysical means, but the incorporation of even a simplified HVS model into objective measures reportedly[7,10,14,54] leads to a better correlation with the subjective ratings. It is conjectured therefore that in machine vision tasks HVS-based measures may have some relevance as well.

F.1 *HVS Modified Spectral Distortion*

In order to obtain a closer relation with the assessment by the human visual system, both the original and coded images can be preprocessed via filters that simulate the HVS. One of the models for the human visual system is given as a band-pass filter with a transfer function in polar coordinates:[54]

$$H(\rho)=\begin{cases}0.05e^{\rho^{0.554}}, & \rho<7,\\ e^{-9[|\log_{10}\rho-\log_{10}9|]^{2.3}}, & \rho\geqslant 7,\end{cases}$$

where $\rho=(u^2+v^2)^{1/2}$. An image processed through such a spectral mask and then inverse discrete cosine transformed can be expressed via the $U\{\cdot\}$ operator, i.e.,

$$U\{C(i,j)\}=\mathrm{DCT}^{-1}\{H(\sqrt{u^2+v^2})\Omega(u,v)\},$$

where $\Omega(u,v)$ denotes the 2D DCT of the image and DCT^{-1} is the 2D inverse DCT. Some possible measures[5,49] for the K component multispectral image are normalized absolute error:

$$H1=\frac{1}{K}\sum_{k=1}^{K}\frac{\Sigma_{i,j=0}^{N-1}|U\{C_k(i,j)\}-U\{\hat{C}_k(i,j)\}|}{\Sigma_{i,j=0}^{N-1}|U\{C_k(i,j)\}|}, \tag{A23}$$

L2 norm:

$$H2=\frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{N^2}\sum_{i,j=0}^{N-1}|U\{C_k(i,j)\}-U\{\hat{C}_k(i,j)\}|^2\right]^{1/2}. \tag{A24}$$

F.2 *Distance Metric for Database Browsing*

The metric proposed in Refs. 14 and 55 based on a multiscale model of the human visual system actually brings forth similarities between image objects for database search and browsing purposes. This multiscale model includes channels, which account for perceptual phenomena such as color, contrast, color contrast, and orientation selectivity. From these channels, features are extracted and then an aggregate measure of similarity using a weighted linear combination of the feature differences is found. The choice of features and weights is made to reflect objects similarly.

We have adopted this database search algorithm to measure discrepancies between an original image and its distorted version. In other words, an image similarity metric that was conceived for browsing and searching in image databases was adapted to measure the similarity (or the difference) between an image and its distorted version.

More specifically, we exploit a vision system designed for image database browsing and object identification to measure image distortion. The image similarity metric in Ref. 14 used 102-dimension feature vectors extracted at different scales and orientations both in luminance and in color channels. The final (dis)similarity metric is

$$H3=\sum_{i=1}^{102}\omega_i d_i, \tag{A25}$$

where ω_i are the weights of the component features stated in Ref. 55 and d_i are the individual feature discrepancies. We call this metric a ''browsing metric'' for lack of a better name. For example, the color contrast distortion at scale l is given by

$$d_\mu=\frac{1}{N_l N_l}\sum_{i,j=0}^{N_l}[K(i,j)-\hat{K}(i,j)]^2,$$

where $N_l\times N_l$ is the size of the image at scale l. $K(i,j)$ and $\hat{K}(i,j)$ denote any color or contrast channel of the original image and of the coded image at a certain level l. The lengthy details of the algorithm and its adaptation to our problem are summarized in Refs. 14 and 55. Finally, note that this measure was used only for color images, and not in the case of satellite three-band images.

The last quality measure we used that reflects the properties of the human visual system was the DCTune algorithm.[56] DCTune is in fact a technique for optimizing JPEG still image compression. DCTune calculates the best JPEG quantization matrices to achieve the maximum possible compression for a specified perceptual error, given a particular image and a particular set of viewing conditions. DCTune also allows the user to compute the percepted error between two images in units of JNDs between a reference image and a test image (http://vision.arc.nasa.gov/dctune/dctune2.0.html). This JND measure was used as the last metric ($H4$) in Table 1.

Acknowledgments

This work was sponsored by NSF INT 9996097, the Scientific Council of Turkey: TUBITAK BDP Program. K. Sayood was supported in part by NASA GSFC.

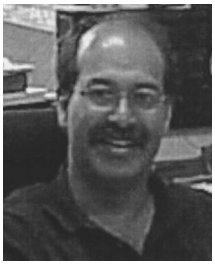
References

1. S. M. Perlmutter *et al.*, ''Image quality in lossy compressed digital mammograms,'' *Signal Process.* **59**, 189–210 (1997).
2. ''Special issue on image and video quality metrics,'' *Signal Process. edited by C. B. Lambrecht* **70**, 153–297 (1998).
3. T. Lehmann, A. Sovakar, W. Schmitt, and R. Repges, ''A comparison of similarity measures for digital subtraction radiography,'' *Comput. Biol. Med.* **27**(2), 151–167 (1997).

4. A. M. Eskicioğlu, ''Application of multidimensional quality measures to reconstructed medical images,'' *Opt. Eng.* **35**(3), 778–785 (1996).
5. A. M. Eskicioğlu and P. S. Fisher, ''Image quality measures and their performance,'' *IEEE Trans. Commun.* **43**(12), 2959–2965 (1995).
6. H. de Ridder, ''Minkowsky metrics as a combination rule for digital image coding impairments,'' in *Human Vision, Visual Processing, and Digital Display III, Proc. SPIE* **1666**, 17–27 (1992).
7. *Digital Images and Human Vision*, edited by A. B. Watson, MIT Press, Cambridge, MA (1993).
8. B. Girod, ''What's wrong with mean-squared error,'' in *Digital Images and Human Vision*, edited by A. B. Watson, Chap. 15, MIT Press, Cambridge, MA (1993).
9. M. Miyahara, K. Kotani, and V. R. Algazi, ''Objective picture quality scale (PQS) for image coding,'' *IEEE Trans. Commun.* **46**(9), 1213–1226 (1998).
10. N. B. Nill and B. H. Bouzas, ''Objective image quality measure derived from digital image power spectra,'' *Opt. Eng.* **31**(4), 813–825 (1992).
11. P. Franti, ''Blockwise distortion measure for statistical and structural errors in digital images,'' *Signal Process. Image Commun.* **13**, 89–98 (1998).
12. S. Winkler, ''A perceptual distortion metric for digital color images,'' in *Proc. 5th Int. Conf. on Image Processing*, Vol. 3, pp. 399–403, Chicago, IL (1998).
13. S. Daly, ''The visible differences predictor: An algorithm for the assessment of image fidelity,'' in *Digital Images and Human Vision*, edited by A. B. Watson, pp. 179–205, MIT Press, Cambridge, MA (1993).
14. T. Frese, C. A. Bouman, and J. P. Allebach, ''Methodology for designing image similarity metrics based on human visual system models,'' *Proc. SPIE /IS&T Conf. on Human Vision and Electronic Imaging II*, **3016**, 472–483 (1997).
15. CCIR, ''Reccomendation 500-2 method for the subjective assessment of the quality of television pictures'' (1986).
16. M. Van Dijk and J. B. Martens, ''Subjective quality assessment of compressed images,'' *Signal Process.* **58**, 235–252 (1997).
17. A. M. Rohaly *et al.*, ''Video Quality Experts Group: Current results and future directions,'' *Visual Communications and Image Processing, Proc. SPIE* **4067**, 742–753 (2000).
18. P. Corriveau and A. Webster, ''VQEG evaluation of objective methods of video quality assessment,'' *SMPTE J.* **108**, 645–648 (1999).
19. T. Kanugo and R. M. Haralick, ''A methodology for quantitative performance evolution of detection algorithms,'' *IEEE Trans. Image Process.* **4**(12), 1667–1673 (1995).
20. R. Matrik, M. Petrou, and J. Kittler, ''Error-sensitivity assessment of vision algorithms,'' *IEE Proc. Vision Image Signal Process.* **145**(2), 124–130 (1998).
21. M. Grim and H. Szu, ''Video compression quality metrics correlation with aided target recognition (ATR) applications,'' *J. Electron. Imaging* **7**(4), 740–745 (1998).
22. H. H. Barrett, ''Objective assessment of image quality: Effects of quantum noise and object variability,'' *J. Opt. Soc. Am. A* **7**, 1266–1278 (1990).
23. H. H. Barrett, J. L. Denny, R. F. Wagner, and K. J. Myers, ''Objective assessment of image quality II: Fisher information, Fourier-crosstalk, and figures of merit for task performance,'' *J. Opt. Soc. Am. A* **12**, 834–852 (1995).
24. C. E. Halford, K. A. Krapels, R. G. Driggers, and E. E. Burroughs, ''Developing operational performance metrics using image comparison metrics and the concept of degradation space,'' *Opt. Eng.* **38**(5), 836–844 (1999).
25. G. K. Wallace, ''The JPEG still picture compression standard,'' *IEEE Trans. Consum. Electron.* **38**(1), 18–34 (1992).
26. A. Said and W. A. Pearlman, ''A new fast and efficient image codec based on set partitioning in hierarchical trees,'' *IEEE Trans. Circuits Syst. Video Technol.* **6**(3), 243–250 (1996).
27. A. M. Martinez and R. Benavente, ''The AR Face Database,'' CVC Technical Report No. 24 (June 1998).
28. A. C. Rencher, *Methods of Multivariate Analysis*, Wiley, New York (1995).
29. T. Kohonen, *Self-Organizing Maps*, Springer, Heidelberg (1995).
30. A. W. Lohmann, D. Mendelovic, and G. Shabtay, ''Significance of phase and amplitude in the Fourier domain,'' *J. Opt. Soc. Am. A* **14**, 2901–2904 (1997).
31. M. P. Dubuisson and A. K. Jain, ''A modified Hausdorff distance for object matching,'' *Int. Conf. on Pattern Recognition*, A: pp. 566–569, (1994).
32. International Commission of Illumination (CIE), ''Recommendations on uniform color spaces, color difference equations, psychometric color terms,'' Publication CIE 15 (E-1.3.1), Supp. 2, Bureau Central de la CIE, Vienna (1971).
33. A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice–Hall, Englewood Cliffs, NJ (1989).
34. V. DiGesu and V. V. Staravoitov, ''Distance-based functions for image comparison,'' *Pattern Recogn. Lett.* **20**(2), 207–213 (1999).
35. V. V. Starovoitov, C. Köse, and B. Sankur, ''Generalized distance based matching of nonbinary images,'' *Int. Conf. on Image Processing*, Chicago (1998).
36. P. Juffs, E. Beggs, and F. Deravi, ''A multiresolution distance measure for images,'' *IEEE Signal Process. Lett.* **5**(6), 138–140 (1998).
37. D. Andreutos, K. N. Plataniotis, and A. N. Venetsanopoulos, ''Distance measures for color image retrieval,'' *IEEE International Conference On Image Processing, IEEE Signal Processing Society*, IEEE, Chicago (1998).
38. http://ag.arizona.edu/classes/rnr555/lecnotes/10.html.
39. W. K. Pratt, *Digital Image Processing*, Wiley, New York (1978).
40. P. E. Trahanias, D. Karakos, and A. N. Venetsanopoulos, ''Directional processing of color images: Theory and experimental results,'' *IEEE Trans. Image Process.* **5**(6), 868–880 (1996).
41. C. Zetsche, E. Barth, and B. Wegmann, ''The importance of intrinsically two-dimensional image features in biological vision and picture coding,'' in *Digital Images and Human Vision*, edited by A. B. Watson, MIT Press, Cambridge, MA pp. 109–138 (1993).
42. P. K. Rajan and J. M. Davidson, ''Evaluation of corner detection algorithms,'' *Proc. Twenty-First Southeastern Symp. on System Theory*, pp. 29–33 (1989).
43. J. Canny, ''A computational approach to edge detection,'' *IEEE Trans. Pattern Anal. Mach. Intell.* **8**(6), 679–698 (1986).
44. D. Carevic and T. Caelli, ''Region based coding of color images using KLT,'' *Graph. Models Image Process.* **59**(1), 27–38 (1997).
45. H. Tao and T. Huang, ''Color image edge detection using cluster analysis,'' *IEEE Int. Conf. on Image Processing*, pp. 834–836, IEEE Signal Processing Society, IEEE, Santa Barbara, California (1997).
46. P. E. Trahanias and A. N. Venetsanopoulos, ''Vector order statistics operators as color edge detectors,'' *IEEE Trans. Syst. Man Cybern.* **26**(1), 135–143 (1996).
47. M. M. Lipschutz, *Theory and Problems of Differential Geometry*, McGraw–Hill, New York (1969).
48. M. McIvor and R. J. Valkenburg, ''A comparison of local surface geometry estimation methods,'' *Mach. Vision Appl.* **10**, 17–26 (1997).
49. I. Avcibaş and B. Sankur, ''Statistical analysis of image quality measures,'' *European Signal Processing Conf., EUSIPCO-2000*, Tampere, Finland, pp. 2181–2184 (2000).
50. R. O. Duda and P. E. Hart, *Pattern Recognition and Scene Analysis*, Wiley, New York (1973).
51. K. Popat and R. Picard, ''Cluster based probability model and its application to image and texture processing,'' *IEEE Trans. Image Process.* **6**(2), 268–284 (1997).
52. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York (1991).
53. M. Basseville, ''Distance measures for signal processing and pattern recognition,'' *Signal Process.* **18**, 349–369 (1989).
54. N. B. Nill, ''A visual model weighted cosine transform for image compression and quality assessment,'' *IEEE Trans. Commun.* **33**(6), 551–557 (1985).
55. T. Frese, C. A. Bouman and J. P. Allebach, ''A methodology for designing image similarity metrics based on human visual system models,'' Technical Report TR-ECE 97-2, Purdue University, West Lafayette, IN (1997).
56. A. B. Watson, ''DCTune: A technique for visual optimization of DCT quantization matrices for individual images,'' *SID Dig.* **XXIV**, 946–949 (1993).
57. I. Avcibaş, N. Memon, and B. Sankur, ''Steganalysis of watermarking techniques using image quality metrics,'' *Security and Watermarking of Multimedia Contents III, Proc. SPIE* **4314**, 523–531 (2001).
58. I. Avcibaş, N. Memon, and B. Sankur, ''Steganalysis based on image quality metrics,'' *IEEE Workshop on Multimedia Signal Processing, MMSP'2001*, Session 15, Cannes, France (2001).
59. I. Avcibaş, N. Memon, and B. Sankur, ''Steganalysis using image quality metrics,'' *IEEE Trans. Image Process.* (submitted).

İsmail Avcıbaş received BS and MS degrees in electronic engineering from Uludağ University, Turkey, in 1992 and 1994, and a PhD degree in electrical and electronic engineering from Boğaziçi University, Turkey, in 2001. He is currently with the Electronic Engineering Department at Uludağ University as a lecturer. His research interests include image processing, data compression, information hiding, and multimedia communications.

Bülent Sankur received his BS degree in electrical engineering at Robert College, Istanbul, and completed his MS and PhD degrees at Rensselaer Polytechnic Institute, Troy, New York. He is currently at Boğaziçi (Bosphorus) University in the Department of Electric and Electronic Engineering. His research interests are in the areas of digital signal processing, image and video compression, and multimedia systems. Dr. Sankur has held visiting professor positions at University of Ottawa, Technical University of Delft, and ENST, France.



Khalid Sayood received his BS and MS in electrical engineering from the University of Rochester in 1977 and 1979, respectively, and his PhD in electrical engineering from Texas A&M University in 1982. He joined the University of Nebraska in 1982 where he currently serves as the Henson Professor of Engineering. From 1995 to 1996 he served as the founding head of the computer vision and image processing group at the Turkish National Research Council Informatics Institute (TUBITAK–MAM) and spent the 1996–1997 academic year as visiting professor at Bogazici University, Istanbul, Turkey. He is the author of *Introduction to Data Compression* (second edition) published by Morgan Kaufmann and the editor of the upcoming *Handbook of Lossless Compression* to be published by Academic Press. His current research includes joint source/channel coding, biological sequence analysis, and data compression.