Statistical evaluation of quality measures. J Electron Imaging

Article *in* Journal of Electronic Imaging · April 2002

DOI: 10.1117/1.1455011 · Source: DBLP

3 authors:

Ismail Avcibas

42 PUBLICATIONS **2,757** CITATIONS

SEE PROFILE

Bulent Sankur
Bogazici University

313 PUBLICATIONS **13,874** CITATIONS

SEE PROFILE

K. Sayood
University of Nebraska at Lincoln

214 PUBLICATIONS **7,017** CITATIONS

SEE PROFILE

Statistical Evaluation of Image Quality Measures

*

*İsmail Avcıbaş[1], Bülent Sankur[2], Khalid Sayood[3]

[1]Department of Electronic Engineering, Uludağ University, Bursa, Turkey

[2]Department of Electrical and Electronic Engineering, Boğaziçi University, İstanbul, Turkey

[3]Department of Electrical Engineering, University of Nebraska at Lincoln, NE, USA

Corresponding author E-mail: sankur@boun.edu.tr

ABSTRACT

In this work we categorize comprehensively 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 HVS-based (Human Visual System-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 have been 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 has been found that measures based on phase spectrum, on multiresolution distance or HVS filtered mean square error are computationally simple and are more responsive to coding artifacts.

Keywords: Image quality measures; ANOVA analysis; Self-Organizing Map.

1. Introduction

Image quality measures (IQM) 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 well reflect the distortion on the image due to, for example, blurring, noise, compression, 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 focused on subjective image quality criteria, such as in [1], [2], [3]. In the subjective assessment of measures characteristics of the human perception become paramount, and image quality is

* The work was sponsored by NSF-INT 9996097 and The Scientific Council of Turkey TUBITAK, BDP Program.

correlated with the preference of an observer or the performance of an operator on some specific task.

In the image coding and computer vision literature, the most frequently used measures are deviations between the original and the coded images ([4], [5], [6]), with Mean Square Error (MSE) or alternatively Signal to Noise Ratio (SNR) varieties 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 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 very low bit rate coding, there has been an increase in the use of quality measures based on human perception [9], [10], [11], [12], [13], [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 measures in general gives a numerical value that will quantify the dissatisfaction of the viewer in observing the reproduced image in place of the original (though Daly's VPD map [13] is a counter example to this). The alternative is the use of subjective tests where subjects view a series of reproduced images and rate them based on the visibility of artifacts [15], [16]. Subjective tests are tedious, time consuming and expensive, and the results depend on various factors such as observer's background, motivation, etc., and furthermore actually only the display quality is being assessed. Therefore an objective measure that accurately predicts the subjective rating would be a useful guide when optimizing image compression algorithms.

Recently there has been ITU (International Telecommunications Union) efforts to establish objective measurement of video quality. Thus in the context of distribution of multimedia documents, video programming in particular, in-service continuous evaluation of video quality is needed. This continuous video quality indicator would be an input to the network management, which must guarantee a negotiated level of service quality. 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, utilizing the knowledge of the human visual system.

Similarly for computer vision tasks, prediction of 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], quantification of target detection performance and ideal observer performance [21], [22], [23] have been studied under additive noise conditions. It is of great interest to correlate coding and sensor artifacts with such algorithmic performance. More specifically one would like to identify image quality metrics that can accurately and consistently predict the performance of computer vision algorithms operating on distorted image records, the distortions being due to compression, sensor inadequacy etc. An alternative use of image quality metrics is in the 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., blur, noise, etc. amount in a distortion space) 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 Analysis of Variance method. The measures are then investigated in terms of their mutual correlation or similarity, this being put into evidence by means 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 are using. 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 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-based measures, 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

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

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

2. **Goals and Methods**

2.1 Quality Attributes

Objective video quality model attributes have been studied in [17], [18]. These attributes can be directly translated to the still image quality measures as "IQM desiderata" in the multimedia and computer vision applications:

- ♦ Prediction Accuracy: The accurate prediction of distortion, whether for algorithmic performance and subjective assessment. For example, when quality metrics are shown in box plots as 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 ability to provide consistently accurate predictions for all types of images and not to fail excessively for a subset of images.

These desired characteristics reflect on the box plots and the F scores of the quality metrics, as detailed in the sequel.

2.2 Test Image Sets and Rates

All the image quality measures are calculated in their multiband version. 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) due to Said and Pearlman [26]. The other types of image distortions are generated by the use of blurring filters with 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 well-known that the JPEG quality factor Q between 80-100 corresponds to visually imperceptible impairment, Q between 60-80 corresponds to perceptible but not annoying distortion, for Q between 40-60 the impairment becomes slightly annoying, for Q between 20-40 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 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 obtained from the JPEG experiment was also used in the SPIHT algorithm.

The test material consisted of the following image sets: 1) Ten three-band remote sensing images, which contained a fair amount of variety, i.e., edges, textures, plateaus and contrast range, 2) Ten color face images from Purdue University Face Images database [27] at rvl1.ecn.purdue.edu/aleix/Aleix_face_DB.html, 3) Ten texture images from 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 the 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 discriminative 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 5 bit rates tested. Each group had 30 sample vectors since there were 30 multispectral test images (10 remote sensing, 10 faces, 10 textures). Similarly 3 "blur groups" were created by low-pass filtering the images with 2-D Gaussian-shaped filters with increasing support. Finally 3 "noise groups" were created by contaminating the images with Gaussian noise with increasing variance, (σ^2 = 200, 600, 1700). This range of noise values spans the noisy image quality from the 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: $\mu_1 = \mu_2 = ... = \mu_k$ means of all the groups are equal,

H_A: $\mu_i \neq \mu_j$ means of the 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 evidence to reject the null hypothesis in favor of the alternative. Recall that the null hypothesis corresponds to the fact all samples are drawn from the same set and that 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 are not equal. We have opted to carry out the multiple comparison tests at the 0.05 significance level. Thus any test resulting in 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 to assert that the difference in the quality metric arises from the image coding artifacts and not from random fluctuations in the image content.

To find out whether the variability of the metric scores arises predominantly from the image quality, and not from the image set, we considered the interaction between image set and the 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 discussed in Section 3 and shown in Tables 2-3, metrics that were sensitive to distortion artifacts were naturally sensitive to image set variation as well. However for the “good” measures identified the sensitivity to image set variation was always inferior 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 the two or more groups. If the *F*-value is not high, there will be a fair amount of overlap between all of the groups. In the box plots, a steep slope and little overlap between boxes, as illustrated in Figure 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 of two successive group means by their respective variances, i.e.,

$$Q_{r,r+1} = \frac{\mu_r - \mu_{r+1}}{\sqrt{\sigma_r \sigma_{r+1}}} \tag{1}$$

$$Q = Ave\{Q_{r,r+1}\} \qquad r = 1,...,k-1$$

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 high *Q* value, which implies little overlap between groups and/or large jumps between them hence high discriminative power of the quality measure. It should be noted that the Q values and the F-scores yielded totally parallel results in our experiments.

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

Figure 1. Box plots of quality measure scores. a) Good measure, b) Moderate measure, c) Poor measure.

2.4 Visualization of Quality Metrics

The visualization of the IQMs in a 2-D 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 the SOM (Self-Organizing Map) visualization is a set of qualitative arguments for their similarity or dissimilarity. To this effect 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 as:

D1(JPEG, bitrate) = [D1(1| JPEG, bitrate) .... D1(M| JPEG, bitrate]^T

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

Recall that the self-organizing map (SOM) is a tool for visualization of high dimensional data. It maps complex, non-linear 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 the case in point, and discovery of categories and abstractions from raw data.

Let the data vectors be denoted as $\mathbf{X} = [\mathrm{x}_1,...,\mathrm{x}_M]^T \in \mathrm{R}^M$, where *M* is the number of images considered (*M* = 30 in our case). With each element in the SOM array, a parametric real vector $\mathbf{m}_i = [\mu_{i1},...,\mu_{iM}]^T \in R^M$ is associated. The location of an input vector $\mathbf{X}$ on the SOM array is defined by the decoder function $\mathrm{d}(\mathbf{X},\mathbf{m}_\mathrm{i})$, where d(.,.) is a general distance measure. The location of the input vector will have the array index *c* defined as $\mathrm{c} = \arg\min_\mathrm{i} \mathrm{d}(\mathbf{X},\mathbf{m}_\mathrm{i})$. A critical part of the algorithm is to define the $\mathbf{m}_i$ in such a way that the mapping is ordered and descriptive of distribution of $\mathbf{X}$. Finding such a set of values that minimize the distance measure resembles the standard VQ problem. In contrast, the indexing of these values is arbitrary, whereby the mapping is unordered. However, if the minimization of the objective functional based on the distance function is implemented under the conditions described in [29], then one can obtain ordered values of $\mathbf{m}_i$, almost as if the $\mathbf{m}_i$ were lying at the nodes of an elastic net. With the elastic net analogy in mind, SOM algorithm can be constructed as

$$\mathbf{m}_\mathrm{i}(\mathrm{t}+1) = \mathbf{m}_\mathrm{i}(\mathrm{t}) + \alpha(\mathrm{t})[\mathbf{X}(\mathrm{t}) - \mathbf{m}_\mathrm{i}(\mathrm{t})]$$

where $\alpha(\mathrm{t})$ 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 $\alpha(\mathrm{t}) = 0$ otherwise. During the course of ordering process, $\alpha(t)$ is decreased from 0.05 to 0.02, while radius of neighborhood is decreased from 10 to 3. Furthermore scores are normalized with respect to the range.

The component planes *j* of the SOM, i.e., the array of scalar values μ_{ij} representing the *j*'th components of the weight vectors $\mathbf{m}_i$ and having the same format as the SOM array is 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 arising from image compression schemes, in other words to find out 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 distortion causes 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 IQM desiderata.

3.1 **ANOVA Results**

The two-way ANOVA results of the image quality measures for the data obtained from all image classes (Fabrics, Faces, Remotes) are listed in Table 2. In these tables the symbols of quality measures *D1, D2...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 succeeding 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 can notice that:

- The fundamental metrics for JPEG compression are H2, H1, S2, E2, which is, HVS L2 norm, HVS absolute norm, spectral phase-magnitude, and edge stability measures. These measures are listed in decreasing order of 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 HVS L2 norm.
- The fundamental metrics for the BLUR effect are S1, E2, S2, H1, that is, spectral phase, edge stability, spectral phase-magnitude, and HVS absolute norm. Notice the similarity of 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 order, E2, H1, S2, H2, S5, that is, edge stability, HVS absolute norm, spectral phase-magnitude, HVS L2 norm, block spectral phase-magnitude. To establish the global metrics, we gave rank numbers from 1 to 26 to each one metric under the four types of distortion as 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 as the global metric, that is the ones that qualify well in all discrimination tests. These results must still be taken with some caution as, for example, none of the 5 winner 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 show also 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 change. This can be observed based on the higher F-scores for distortion effects as 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 bottom 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 into Table 3 were the outcome of the F scores.

We also investigated the metrics with respect to their ability to respond to bit rate and coder type. For this analysis the scores of the JPEG and SPIHT compressors were combined. It was observed in Table 4 that:

- 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, S5 (HVS

L2 norm, 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, HVS L2 norm, DCTune, 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..., (Mean angle-magnitude similarity, Mean square error, Modified infinity norm, Block spectral magnitude error, Mean absolute error, Mean angle similarity...). These metrics were identified by summing the two rank scores of metrics, the first being the ranks in ascending order of distortion sensitivity, the second being in descending order the image set sensitivity. Interestingly enough almost all of them are related to the mean square error varieties. Despite its many criticisms, this may explain why mean square error or signal-to-noise ratio measures have proven 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.

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

Table 3. One-way ANOVA results for each image class and two-way ANOVA results for the distortions on the combined and image set independence.

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

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 the image set variation are also not very discriminating with respect to the distortion types. The fact that the metrics are sensitive, as should be expected, to both the image content and distortion artifacts does not eclipse their potential as quality metrics. Indeed when the metrics were tested within 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 on the mutual relationship between measures. It is obvious that the quality measures must be correlated with each other as 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 some others reflect additive noise.

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

Self Organizing Map (SOM) [29] is a pictorial method 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’s 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 one of the bit rates.

The SOM organization of the measures in the 2-D space for pooled data from JPEG and SPIHT coders is shown in Figure 2. These maps are useful for the visual assessment of 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 the observation of the SOM and the correlation matrix are as follows:

- Clustering tendency of pixel difference based measures (D1, D2, D4, D5) and 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 measure (D6), E2 (edge stability measure) and C5 (mean angle-magnitude measure) are clustered in 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 image comparison by eye more closely, by assigning larger weight to low resolution components and less 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 (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 is used as an additional criterion for the selection of “good” measures subset.

3.3 Combination of Quality Measures: Super-metrics

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 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 aims to detect the presence of watermarks and to differentiate stego-documents. To this effect digital watermarking is used, which consists of an imperceptible and cryptographically secured message added to 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 stego-message. 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 (IQM) [57, 58].

In order to identify specific quality measures that prove useful in steganalysis, that in distinguishing the watermarked images from the non-watermarked ones, we again use ANOVA test. The 26 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 the 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 DFT coefficients, block 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 as: 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, Normalized Mean Square HVS Error H_2. The implication here is two-fold: One is that, using these features a steganalyzer can be designed to detect the watermarked or stegoed images using multivariate regression analysis, as we show in [57, 58, 59]. This linear combination of the IQMs for steganalysis purposes is denoted as the "supermetric" for steganalysis. It is shown in [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 on those statistically significant image features to eschew detection [59].

For the second "supermetric" we searched for 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 took a first 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. Time of observation was unlimited. The images used were all 512 X 512 RGB color images from Purdue University face database, and were viewed at 4x the image height. The results reported are based on the 850 quality evaluations of 50 encoded images (10 images compressed with JPEG at five different quality scales, Q=10, 30, 50, 70, 90) by the pool of 17 subjects. The supermetric of image quality for compression artifacts was build using the global metrics E2, H1, S2, H2, S5, that is, edge stability, HVS absolute norm, spectral phase-magnitude, HVS L2 norm, block spectral phase-magnitude) for the image distortions due to compression. The supermetric was built by regressing them against the MOS scores. The plot of this supermetric and MOS data is given in Fig. 4, where a high value of the correlation coefficient has been determined: 0.987. The correlation coefficients of the individual metrics, shown in Table 5, were all lower, as expected.

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	Super-metric	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		

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 of 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, H2)*, edge stability measure *(E2)* are most sensitive to coding and blur artifacts, while the mean square error (D1) remains as the best measure for additive noise. These "winner" metrics were elected on the basis of the summed rank scores across four artifacts; JPEG-compression' SPIHT-compression, blur and noise. This preselection of the E2, S2, S5, H1, 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 stated in Section 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 consistencies were 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 performance levels of a given compression scheme, and across different compression schemes. It appears that, as shown in Section 3.3, a combination of spectral phase-and-magnitude measure and of HVS-filtered error norm comes closest to satisfy such a measure, as it is adequately 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 has been put into evidence via Kohonen's Self-Organizing Map.

In conclusion, the subset of the E2, S2, S5, H1, H2 metrics are the prominent image quality measures as shown by the ANOVA analysis on the one hand, and by the good regression property to MOS data, on the other hand. 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.

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

APPENDIX A

In this Appendix we define and describe the multitude of image quality measures considered. In these definitions the pixel lattices of images A, B will be referred to as $A(i, j)$ and $B(i, j)$, i, j = 1...N, as the lattices are assumed to have dimensions NxN. The pixels can take values from the set $\{0,...,G\}$ 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 the pixel position *i, j,* and in band *k* as $C_k(i, j)$, where $k = 1,...,K$. 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 **C**(i,j) = [R(i,j) G(i,j) B(i,j)]^T . All these definitions are summarized in the following table :

$C_k(i, j)$	(i, j)th pixel of the kth band of image C
$\mathbf{C}(i, j)$	(i,j)th multispectral (with K bands) pixel vector
C	A multispectral image
C_k	The 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 k'th band can be calculated as $\sigma_k^2 = \sum_{i,j=0}^{N-1} C_k(i, j)^2$. All these quantities with an additional hat, i.e., $\hat{C}_k(i, j)$, $\hat{\mathbf{C}}$ etc., will correspond to the distorted versions of the same original image. As a case in point, the expression $\left\|\mathbf{C}(i, j) - \hat{\mathbf{C}}(i, j)\right\|^2$ = $\sum_{k=1}^{K} \left[C_k(i, j) - \hat{C}_k(i, j)\right]^2$ will denote the sum of errors in the spectral components at a given pixel position *i, j*. Similarly the error expression in the last row of the above table expands as $\varepsilon_k^2 = \sum_{i=1}^{N}\sum_{j=1}^{N}\left[C_k(i, j) - \hat{C}_k(i, j)\right]^2$. In the specific case of RGB color images we will occasionally revert to the notations $\{R, G, B\}$ and $\{\hat{R}, \hat{G}, \hat{B}\}$.

Measures Based on Pixel Differences

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

A.1 Minkowsky Metrics:

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

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

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

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

In what follows we have used the pixel-wise difference in the Minkowsky sum as given in Eq. (A1). For γ = 2, one obtains the well-known Mean Square Error (MSE) expression, denoted as D1:

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

An overwhelming number of quality results in the literature is in fact given in terms of the Signal to Noise Ratio (*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. Though 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 the MSE-based quality measures there has been a recent resurgence of the 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 γ = 1 one obtains the absolute difference denoted as *D2*. For $\gamma = \infty$ power in the Minkowski average the maximum difference measure

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

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 $\Delta_l\left(\mathbf{C} - \hat{\mathbf{C}}\right)$, l = 1 … N^2, resulting in a modified Minkowski infinity metric, called D3. Here $\Delta_l\left(\mathbf{C} - \hat{\mathbf{C}}\right)$ denotes the l^{th} largest deviation among all pixels [31]. Thus $\Delta_1\left(\mathbf{C} - \hat{\mathbf{C}}\right)$ is simply the error expression ε^{∞} above. Similarly Δ_2 correspond 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, Δ_1 … Δ_r .

$$\mathrm{D3} = \sqrt{\frac{1}{\mathrm{r}} \sum_{\mathrm{m}=1}^{\mathrm{r}} \Delta^2{}_{\mathrm{m}} \left(\mathbf{C} - \hat{\mathbf{C}}\right)} \qquad \text{(A3)}$$

A.2 MSE in Lab Space

The choice of the color-space for an image similarity metric is important, because the color-space must be uniform, so that the intensity difference between two colors must be consistent with the color difference estimated by a human observer. Since the RGB color-space is not well-suited for this task two color spaces are defined: 1976 CIE $\mathrm{L}^*\mathrm{u}^*\mathrm{v}^*$ and the 1976 CIE $\mathrm{L}^*\mathrm{a}^*\mathrm{b}^*$ 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 $\mathrm{L}^*\mathrm{a}^*\mathrm{b}^*$ space. Then the Euclidean distance is:

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

Note that (A4) is intended to yield perceptually uniform spacing of colors that exhibit color differences greater than JND threshold but smaller than those in 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, and not in satellite images.

A.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 the 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]:

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

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

Thus for any given pixel $\mathbf{C}(i,j)$, we search for a best matching pixel in the d-distance sense in the *wxw* neighborhood of the pixel $\hat{\mathbf{C}}(i,j)$, denoted as $\hat{\mathbf{C}}_w(i,j)$. The size of the neighborhood is typically small e.g., 3x3 or 5x5, and one can consider a square or a crossshaped support. Similarly one calculates the distance from $\hat{\mathrm{C}}(\mathrm{i},\mathrm{j})$ to $\mathrm{C}_{\mathrm{w}}(\mathrm{i},\mathrm{j})$ where again $\mathrm{C}_{\mathrm{w}}(\mathrm{i},\mathrm{j})$ denotes the pixels in the *wxw* neighborhood of coordinates (i,j) of $\mathrm{C}(\mathrm{i},\mathrm{j})$. Note that in general $d\left(\mathbf{C}(i,j),\hat{\mathbf{C}}_w(i,j)\right)$ is not equal to $d\left(\hat{\mathbf{C}}(i,j),\mathbf{C}_w(i,j)\right)$. As for the distance measure $\mathrm{d}(\cdot,\cdot)$, the city metric or the chessboard metric can be used. For example city block metric becomes

$$\mathrm{d}^{\mathrm{city}}\left(\mathbf{C}(\mathrm{i},\mathrm{j}),\hat{\mathbf{C}}(\mathrm{l},\mathrm{m})\right)=\frac{\left(|\mathrm{i}-\mathrm{l}|+|\mathrm{j}-\mathrm{m}|\right)}{\mathrm{N}}+\frac{\left\|\mathbf{C}(\mathrm{i},\mathrm{j})-\hat{\mathbf{C}}(\mathrm{l},\mathrm{m})\right\|}{\mathrm{G}}$$

where ||.|| denotes the norm of the difference between $\mathbf{C}(\mathrm{i},\mathrm{j})$ and $\hat{\mathbf{C}}(\mathrm{i},\mathrm{j})$ 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 to tune the penalties due to pixel shifts and pixel spectral differences, respectively. In our measurements we have used the city block distance with 3x3 neighborhood size.

A.4 Multiresolution Distance Measure

One limitation of standard objective measures of distance between images is that the comparison is conducted at the 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 in machine vision tasks that often use local information only.

Consider the various levels of resolution denoted by $r \geq 1$. 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 of size ($\frac{N}{2}$ x $\frac{N}{2}$) with average gray levels g_{11}, g_{12}, g_{21} and g_{22}. For the *r*th resolution level one would have than 2^{2r-2} blocks of size ($\frac{N}{2^{r-1}}$ x $\frac{N}{2^{r-1}}$), characterized by the block average gray levels g_{ij}, $i,j=1,\ldots,2^{2r-2}$. Thus for each block b_{ij} of the image C, take g_{ij} as its average gray level and $\hat{g}_{ij}$ to corresponds to its component in the image $\hat{C}$ (For simplicity a third index denoting the resolution level has been omitted). The average difference in gray level at the resolution *r* has weight $\frac{1}{2^r}$. Therefore the distortion at this level is

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

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 simply be found by summing over all resolution levels, r = 1,.., R, that is $D\left(C,\hat{C}\right)=\sum_{r=1}^{R} d_r$. 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 512x512 images one has R = 9. Finally for multispectral images one can extend this definition in two ways. In the straightforward extension, one sums the multiresolution distances d_r^k over the bands:

$$D6=\frac{1}{K}\sum_{k=1}^{K}\sum_{r=1}^{R} d_r^k \qquad \text{(A6)}$$

where d_r^k is the multiresolution distance in the *k*th band. This is the multiresolution distance definition that we used in the experiments. Alternatively the Burt pyramid was constructed to obtain the multiresolution representation. However in the tests it did not perform as well as the pyramid described in [36.

A different definition of the multiresolution distance would be to consider the vector difference of pixels:

$$\mathrm{D}\left(\mathrm{C},\hat{\mathrm{C}}\right)=\sum_{\mathrm{r}=1}^{\mathrm{R}} \mathrm{d}'_{\mathrm{r}} \quad \text{with } \mathrm{d}_{\mathrm{r}}=\frac{1}{2^{\mathrm{r}}}\frac{1}{2^{2\mathrm{r}-2}}\sum_{\mathrm{i},\mathrm{j}=1}^{2^{\mathrm{r}-1}}\left[\left(\mathrm{g}_{\mathrm{ij}}^{\mathrm{R}}-\hat{\mathrm{g}}_{\mathrm{ij}}^{\mathrm{R}}\right)^2+\left(\mathrm{g}_{\mathrm{ij}}^{\mathrm{G}}-\hat{\mathrm{g}}_{\mathrm{ij}}^{\mathrm{G}}\right)^2+\left(\mathrm{g}_{\mathrm{ij}}^{\mathrm{B}}-\hat{\mathrm{g}}_{\mathrm{ij}}^{\mathrm{B}}\right)^2\right]^{1\!/\!2}$$

where, for example, $\mathrm{g}_{\mathrm{ij}}^{\mathrm{R}}$ is the average gray level of the ij'th block in the "red" component of the image at the (implicit) resolution level r. Notice that in the latter equation the Euclidean norm of the differences of the block average color components *R*, *G*, *B* have been utilized.

Notice 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 closeness between two digital images can also be quantified in terms of 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 as follows:

Structural Content:

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

Normalized Cross-Correlation measure:

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

Czenakowski Distance:
A metric useful to compare vectors with strictly non-negative components, as in the case of color images, is given by the Czenakowski distance [37]:

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

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

Obviously as the difference between two images tends to zero $\varepsilon = \mathrm{C}-\hat{\mathrm{C}} \to 0$, all the correlation-based measures tend to 1, while as $\varepsilon^2 \to G^2$ they tend to 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.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 the **C** space. Since we deal with positive vectors $\mathbf{C},\hat{\mathbf{C}}$, we are constrained to one quadrant of the Cartesian space. Thus the normalization factor of 2/π is related to the fact that the maximum difference attained will be π/2. The combined angular correlation and magnitude difference between two vectors can be defined as follows [40, 37]:

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

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

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

$$\mathrm{C5} = \frac{1}{\mathrm{N}^2}\sum_{\mathrm{i,j=1}}^{\mathbf{N}}\chi_{\mathrm{ij}}, \qquad \text{(A11)}$$

where μ_χ is the mean of the angular differences. These moments have been previously used for the assessment of directional correlation between color vectors.

Edge Quality Measures

According to the contour-texture paradigm of images, the edges form the most informative part in images. For example, in the perception of scene content by human visual system, edges play the major role. Similarly 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 degradations are: Discontinuities in the edge, decrease of edge sharpness by smoothing effects, offset of edge position, missing edge points, falsely detected edge points etc [39]. Notice however that all the above degradations are not necessarily observed as edge and corner information in images is rather well preserved by most compression algorithms.

Since we do not possess the ground-truthed edge map, we have used the edge map obtained from the original uncompressed images as the reference. Thus to obtain edge-based quality measures we have generated edge fields from both the original and compressed images using the 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 Gaussians of each band are averaged, and the resulting average output is interpolated, thresholded and thinned in a manner similar to that in [12]. The parameters are set as in [43] at robotics.eecs.berkeley.edu/~sastry/ee20/cacode.html.

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

C1. Pratt Measure

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

$$E1 = \frac{1}{\max\{n_d, n_t\}} \sum_{i=1}^{n_d} \frac{1}{1 + ad_i^2} \qquad \text{(A12)}$$

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 candidate for the i^{th} detected edge pixel. In our study the binary edge field obtained from the uncompressed image is considered as 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 a relative weighting between smeared edges and thin but offset edges. The sum terms penalize possible shifts from the correct edge positions. In summary the smearing and offset effects are all 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 evidences across different scales in both the original and coded images [44]. Edge maps at different scales have been obtained from the images by using the Canny [43] operator at different scale parameters (with the standard deviation of the Gaussian filter assuming values σ_{m} = 1.19, 1.44, 1.68, 2.0, 2.38 (The output of this operator at scale m is thresholded with T^m , where $\text{T}^{\text{m}} = 0.1(\text{C}_{\max} - \text{C}_{\min}) + \text{C}_{\min}$. In this expression $\text{C}_{\max}$ and $\text{C}_{\min}$ denote the maximum and minimum values of the norm of the gradient output in that band. Thus the edge map at scale σ_m of the image C is obtained as

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

where $C^m(i,j)$ is the output of the Derivative of Gaussian operator at the m^{th} 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

$$Q(i,j) \quad = l \quad \text{where} \quad l = \arg\max_{l} \bigcap_{\sigma_m \le \sigma_k \le \sigma_{m+l-1}} E(i,j,\sigma_k) = 1 .$$

The edge stability index calculated from 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. Finally a fidelity measure called Edge Stability Mean Square Error (ESMSE) can be calculated by summing the differences in edge stability indexes over all edge pixel positions, n_d, that is the edge pixels of the ground-truth (undistorted) image at full resolution.

$$\mathrm{E2} = \frac{1}{n_d} \sum_{i,j=0}^{n_d} \left(Q(i,j) - \hat{Q}(i,j) \right)^2 \qquad \text{(A13)}$$

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

A property complementary to edge information could be the 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 3-D 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 has been used in [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 didn't perform particularly well in preliminary tests.

Spectral Distance Measures

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

D.1 Magnitude and Phase Spectrum

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

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

Spectral distortion measures, using difference metrics as for example given in A.1-A.3 can be extended to multispectral images. To this effect considering the phase and magnitude spectra, that is

$$\varphi(u,v) = \arctan(\Gamma(u,v))$$
$$M(u,v) = \left| \Gamma(u,v) \right|,$$

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

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

the spectral phase distortion

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

and the weighted spectral distortion

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

where λ is to be judiciously chosen e.g., to reflect quality judgement. 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, that $\lambda = 2.5\text{x}10^{-5}$.

Due to the localized nature of distortion and/or the non-stationary image field, Minkowsky averaging of block spectral distortions may be more advantageous. Thus an image can be divided into non-overlapping or overlapping *L* blocks of size *b* x *b*, say 16x16, and blockwise spectral distortions as in (A14-A15) can be computed. Let the DFT of the *l*th block of the *k*th 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 $u,v = -\frac{b}{2}...\frac{b}{2}$ and $l = 1,...,L$,or in the magnitude-phase form

$$\Gamma_k^l(u,v) = \left|\Gamma_k^l(u,v)\right|e^{j\phi_k^l(u,v)} = M_k^l(u,v)e^{\phi_k^l(u,v)}.$$

Then the following measures can be defined in the transform domain over the *l*th block.

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

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

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

with λ the relative weighting factor of the magnitude and phase spectra. Obviously measures A.16-A.18 are special cases of the above definitions for block size b covering the whole image. Various rank order operations on the block spectral differences J_M and / or J_φ can prove useful. Thus let $J^{(1)},...,J^{(L)}$ 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 $\frac{1}{2}\left(J^{\left(\frac{L}{2}\right)} + J^{\left(\frac{L+1}{2}\right)}\right)$, Maximum block distortion, $J^{(L)}$; and Average block distortion: $\frac{1}{L}\sum_{i=1}^{L}J^{(i)}$. We have found that median of the block distortions is the most effective averaging of rank ordered block spectral distortions and we have thus used:

$$S3 = \underset{l}{median}\ J_M^l \tag{A.16}$$

$$S4 = \underset{l}{median}\ J_\phi^l \tag{A.17}$$

$$S5 = \underset{l}{median}\ J^l \tag{A.18}$$

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

Context Measures

Most of the compression algorithms and computer vision tasks are based on the neighborhood information of the pixels. In this sense any loss of information in the pixel neighborhoods, that is, damage to pixel contexts 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 (p.m.f.) of pixel neighborhoods, changes in the context probabilities should be indicative of image distortions.

A major hurdle in the computation of context distortions is the requirement to calculate the high dimensional joint probability mass function. Typical p.m.f. dimensions would be of the order of *s* = 10 at least. Consequently one incurs "curse of dimensionality problems". However as detailed in [50], [51], this problem can be solved by judicious usage of kernel estimation and cluster analysis. A 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 well represents both mode and tail regions of p.m.f's, while combining the summarizing strength of histograms with generalizing strength of kernel estimates.

In what follows we have used the 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 p.m.f's obtained from 4-pixel neighborhoods in the 3-bands.

E.1 Rate-Distortion Based Distortion Measure

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

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

where X^s denotes a s-pixel neighborhood and $\mathbf{x} = [\mathrm{x}_1,...,\mathrm{x}_s]$ a random vector. Furthermore p and $\hat{p}$ are the p.m.f's of the original image contexts and that of the distorted (e.g., blurred, noisy, compressed etc.) image. The relative entropy is directly related to efficiency in compression and error rate in classification. Recall also that the optimal average bit rate is the entropy of x

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

If instead of the true probability, a perturbed version $\hat{p}$, that is the p.m.f 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 p.m.f. *p*, 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 p.m.f.'s resulting from the two compressed images, then their difference in relative entropy

$$\mathrm{Z1} = \mathrm{D}(\mathrm{p}\|\hat{\mathrm{p}}_1) - \mathrm{D}(\mathrm{p}\|\hat{\mathrm{p}}_2) = \mathrm{R}(\hat{\mathrm{p}}_1) - \mathrm{R}(\hat{\mathrm{p}}_2) \tag{A.19}$$

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 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 of quality factor 90 and $R(\hat{p}_2)$ at the bit rate of quality factor 70, etc.. Alternatively the distortion was calculated for an original image and its blurred or noise contaminated version.

E.2. f-divergences

Once the joint p.m.f 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 $\frac{\hat{\mathrm{p}}}{\mathrm{p}}$ is the likelihood ratio between, $\hat{\mathrm{p}}$, the context p.m.f. of the distorted image, p the p.m.f. function of the original image, and E_p is the expectation with respect to *p*. Some examples are as follows:

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

$$Z2 = \frac{1}{2}\int\left(\sqrt{\hat{p}} - \sqrt{p}\right)^2 d\lambda \tag{A.20}$$

Generalized Matusita Distance: $f(x) = \left|1 - x^{1/r}\right|^r,\ g(x) = x^{1/r}$

$$Z3 = \sqrt{\int\left|p^{1/r} - \hat{p}^{1/r}\right|^r d\lambda} \qquad r \geq 1 \tag{A.21}$$

Notice that integration in (A.20-A.21) are 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 p.m.f.'s do not reflect directly 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 a local level, we calculated the histograms of the original and distorted images on the basis of 16x16 blocks. To this effect 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 16x16 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 $\sum_{u=1}^{b} \max_{k=1..K}\left\{KS_u^k\right\}$ where $\mathrm{KS}_\mathrm{b}^\mathrm{k}$ denotes the Kolmogorov-Smirnov distance of the block number u and of the k'th spectral component. However the KS distance did not turn out to be effective in the ANOVA tests. Instead the SRC measure had a better performance. We again considered the SRC on a 16x16 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 their largest for each pixel neighborhood. Finally the correlation of the block ranks of the original and distorted images is calculated:

$$Z4 = \sum_{u=1}^{b} \max_{k=1..K}\left\{SRC_u^k\right\} \qquad \text{(A.22)}$$

where SRC_u^k denotes the Spearman Rank Correlation for the u'th block number and k'th spectral band.

Human Visual System Based Measures

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

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]:

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

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

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

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

Normalized Absolute Error:

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

L2 Norm:

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

F.2. A Distance Metric for Database Browsing

The metric proposed in [14], [55] based on a multiscale model of the human visual system, has actually the function of bringing forth the 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 formed. The choice of features and weights is made to maximize the consistency with similarity.

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 [14] uses 102 feature vectors extracted at different scales and orientations both in luminance and color channels. The final (dis)similarity metric is

$$H3=\sum_{i=1}^{102}\omega_i d_i \qquad \text{(A25)}$$

where ω_i are their weights as attributed in [55] and d_i are the individual feature discrepancies. We call this metric "browsing metric" for the lack of a better name. For example the color contrast distortion at scale *l* is given by

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

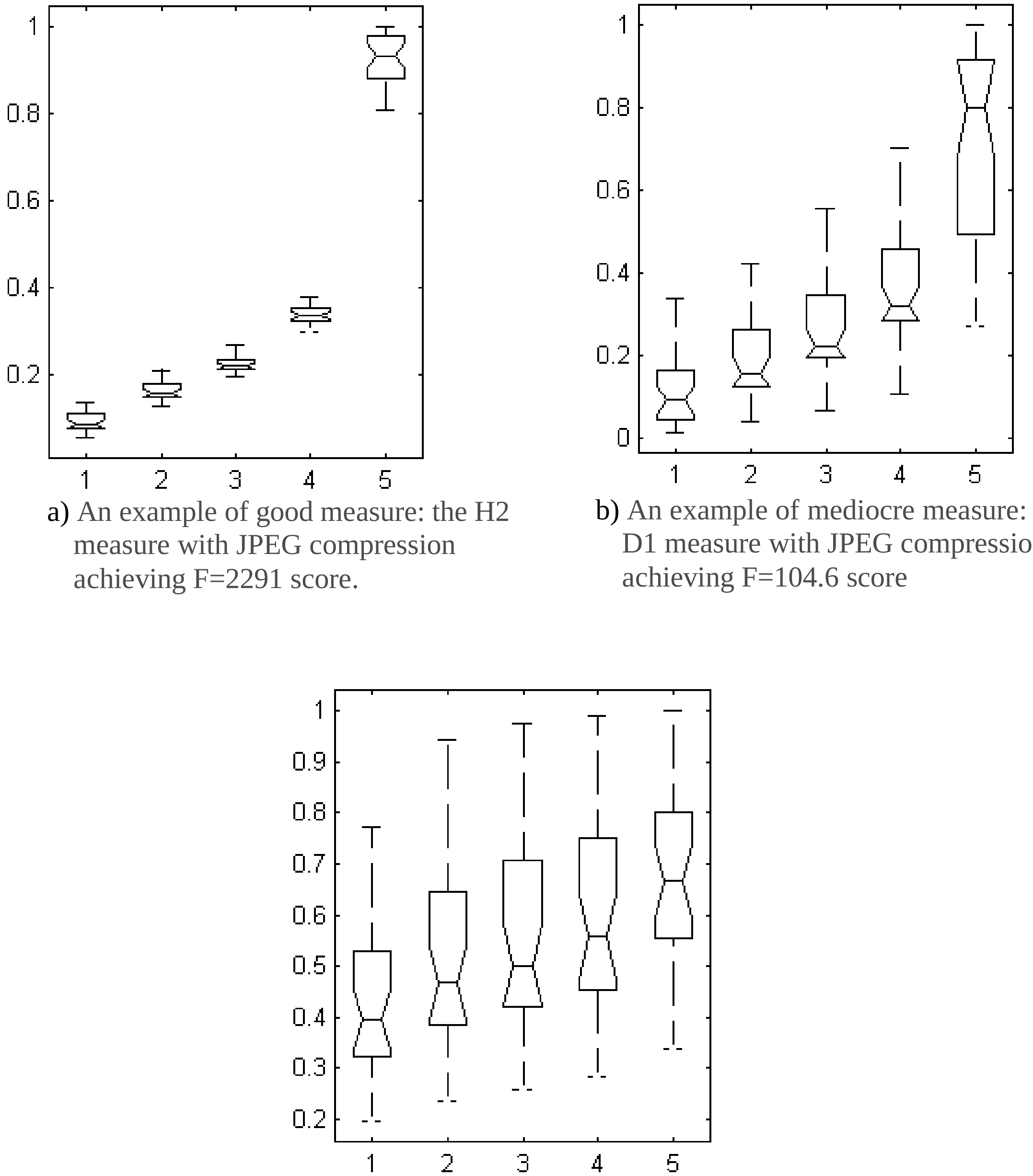
where $N_l \mathrm{x} N_l$ is the size of the image at scale *l*. $K(i,j)$ and $\hat{K}(i,j)$ denote any color or contrast channel of the original image and of the coded image at a certain level *l*. The lengthy details of the algorithm and its adaptation to our problem are summarized in [14], [55]. 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 perceptual error between two images in units of jnd (just-noticeable differences) between a reference image and a test image (http://vision.arc.nasa.gov/dctune/dctune2.0.html). This figure was used as the last metric (H4) in Table 1.

LIST OF FIGURES

LIST OF TABLES

Acknowledgements: We would like to thank H. Brettel (ENST, France), A. Eskicioglu (Thompson Comm., USA) as well as an anonymous reviewer for their invaluable help in improving the paper.



a) An example of good measure: the H2 measure with JPEG compression achieving F=2291 score.

b) An example of mediocre measure: the D1 measure with JPEG compression achieving F=104.6 score

c) An example of poor measure: the C4 measure with SPIHT compression achieving F=7.91 score.

Figure 1. Box plots of quality measure scores. a) Good measure, b) Moderate measure, c) Poor measure. The F-scores are also given.

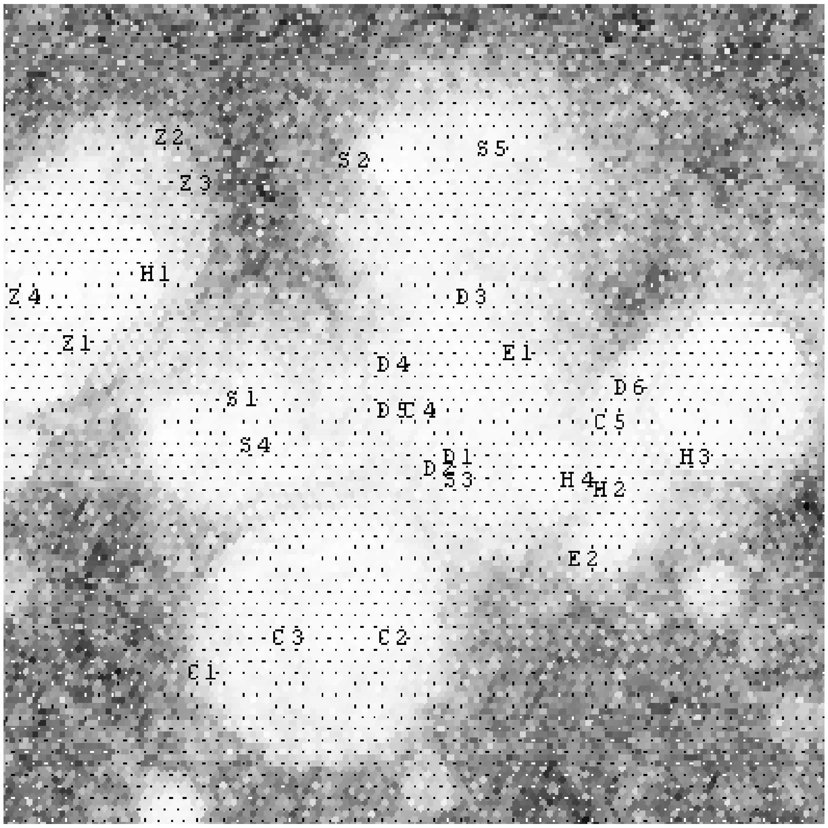


Figure 2. SOM of distortion measures for JPEG and SPIHT.

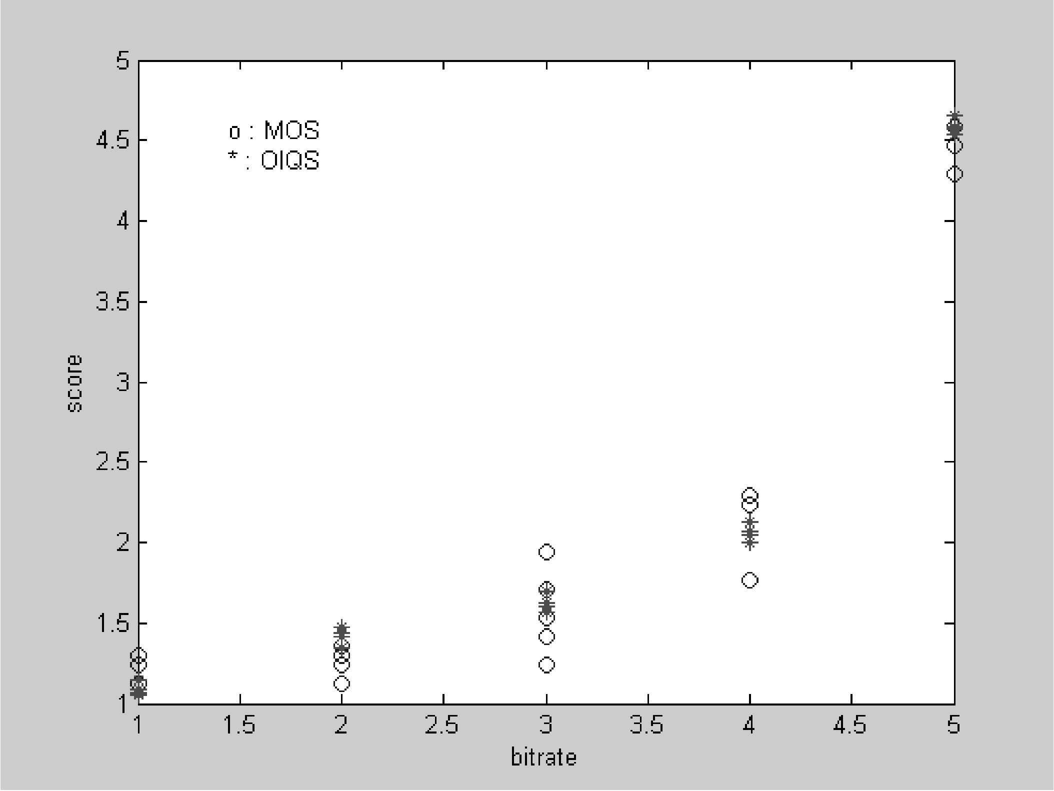


Figure 3. The plot of Mean Opinion Score (MOS) and image quality supermetric data against bit rate.

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

SYMBOL	DESCRIPTION	EQUATION
D1	Mean Square Error	A.1
D2	Mean Absolute Error	A.2
D3	Modified Infinity Norm	A.3
D4	L*a*b* Perceptual Error	A.4
D5	Neighborhood Error	A.5
D6	Multiresolution Error	A.6
C1	Normalized Cross-Correlation	A.7
C2	Image Fidelity	A.8
C3	Czekonowski Correlation	A.9
C4	Mean Angle Similarity	A.10
C5	Mean Angle-Magnitude Similarity	A.11
E1	Pratt Edge Measure	A.12
E2	Edge Stability Measure	A.13
S1	Spectral Phase Error	A.14
S2	Spectral Phase-Magnitude Error	A.15
S3	Block Spectral Magnitude Error	A.16
S4	Block Spectral Phase Error	A.17
S5	Block Spectral Phase-Magnitude Error	A.18
Z1	Rate Distortion Measure	A.19
Z2	Hellinger distance	A.20
Z3	Generalized Matusita distance	A.21
Z4	Spearman Rank Correlation	A.22
H1	HVS Absolute Norm	A.23
H2	HVS L2 Norm	A.24
H3	Browsing Similarity	A.25
H4	DCTune	

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

	JPEG		SPIHT		BLUR		NOISE	
ANOVA2	Bit rate	Image set	Bit rate	Image set	Blur	Image set	Noise	Image set
D1	104.6	42.59	39.23	13.28	43.69	2.06	9880	17.32
D2	108.5	67.45	29.56	15.93	33.94	17.76	6239	20.4
D3	63.35	29.37	53.31	48.53	38.55	24.13	1625	11.15
D4	89.93	1.99	13.75	3.71	27.87	0.96	166.4	9.88
D5	20.26	80.71	14.09	68.22	6.32	55.11	1981	43.51
D6	76.73	5.94	37.52	11.22	412.9	45.53	44.61	4.38
C1	1.35	124.6	12.05	325.5	5.61	107.2	3.82	6.17
C2	12.26	93.83	15.18	82.87	11.19	39.77	58.04	45.63
C3	82.87	83.06	24.96	22.42	30.92	1.71	567.5	52.01
C4	45.65	47.36	7.91	5.94	16.48	0.77	198.8	19.03
C5	91.42	38.17	27.51	5.28	52.57	2.44	704	10.8
E1	26.24	3.64	77.86	137	125.8	21.09	87.76	27.87
E2	176.3	92.75	212.5	200.4	768.7	23.41	158.5	24.84
S1	150.5	102.2	104	68.17	1128	60.04	47.29	38.42
S2	191.3	98.42	161	101.8	572.2	17.95	107.1	4.83
S3	145.6	56.39	38.58	26.97	24.28	6.39	2803	8.59
S4	129.1	63.26	128	46.85	215	11.17	56.04	55.1
S5	146.1	71.03	144.1	61.65	333.6	27.84	78.04	26.53
Z1	1.69	141.8	21.36	14	35.9	62.5	44.89	110.9
Z2	7.73	114.7	11.41	77.68	10.17	1.80	3.03	11.36
Z3	17.63	223	23.22	181.4	17.26	8.31	14.71	21.12
Z4	9.4	23.58	9.84	32.41	8.45	14.74	24.99	3.31
H1	371.9	0.09	107.2	40.05	525.6	69.98	230.7	19.57
H2	2291	5.46	132.9	22.82	47.28	101.7	624.3	21.32
H3	123	1.2	27.45	7.6	67.31	6.77	117.3	0.50
H4	78.83	7.14	25.2	95.72	12.55	2.11	29.06	6.69

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 image set and to the coder type.

One-way ANOVA	IMAGE SET	JPEG	SPIHT	BLUR	NOISE
	Fabrics	H4,H2,E2,S4	E1,S1,E2,S2	S1,S5,E2,S4	D1,D2,D5,D3
	Faces	H2, D1,S3,H1	H4,D3,H2,C1	S2,H1,S1,E2	D1,S3,D2,D3
	Remote Sensing	H2,H4,S4,S5	S2,S5,S4,S1	D6,S5,S4,S1	D1,D2,C3,C5
Two-way ANOVA	Combined Set	H2,H1,S2,E2	E2,S2,S5,H2	S1,E2,S2,H1	D1,D2,S3,D5
	Image Set Independence	H1,H3	D4,C5	C4,D4	H3,Z4
	Coder Type Independence	D2,D1,Z4,D3			

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

ANOVA2	JPEG+SPIHT	
Metric	Bit rate	Coder
D1	89.79	0.75
D2	74.98	2.72
D3	71.55	1.21
D4	70.52	43.85
D5	17.07	0.0005
D6	85.22	118.8
C1	2.66	45.47
C2	12.28	18.27
C3	56.48	1.56
C4	31.3	2.43
C5	78.98	2.23
E1	42.69	11.61
E2	122.4	26.28
S1	99.12	5.29
S2	140.1	12.37
S3	92.99	9.27
S4	115.5	39.1
S5	124.8	43.09
Z1	4.28	41.6
Z2	9.54	0.83
Z3	12.87	0.56
Z4	9.39	6.64
H1	278.6	52.87
H2	493	87.21
H3	97.94	16.19
H4	21.13	57.72

References

1. S.M. Perlmutter, P.C Cosman, R.M. Gray, R.A. Olshen, D.Ikeda, C.N. Adams, B.J. Betts, M.B. Williams, K.O. Perlmutter, J. Li, A. Aiyer, L. Fajardo, R. Birdwell, B.L. Daniel, "Image Quality in Lossy Compressed Digital Mammograms", *Signal Processing*, 59, 189-210 (1997).
2. C. B. Lambrecht, Ed., "Special Issue on Image and Video Quality Metrics", *Signal Processing*, vol. 70, (1998).
3. T. Lehmann, A. Sovakar, W. Schmitt, 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, 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. A. B. Watson, Ed., *Digital Images and Human Vision,* Cambridge, MA, MIT Press (1993).
8. B. Girod, "What's Wrong with Mean-squared Error", in *Digital Images and Human Vision*, A. B. Watson, ed., Chapter 15, Cambridge, MA, MIT Press (1993).
9. M. Miyahara, K. Kotani, V. R. Algazi, "Objective Picture Quality Scale (PQS) for Image Coding", *IEEE Trans. Commun*., 46(9), 1213-1226 (1998).
10. N. B. Nill, 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 Processing: Image Communication*, 13, 89-98 (1998).
12. S. Winkler: "A perceptual distortion metric for digital color images." in *Proc. 5th International Conference on Image Processing*, vol. 3, pp. 399-403, Chicago, Illinois, October 4-7, 1998.
13. S. Daly, "The visible differences predictor: An algorithm for the assessment of image fidelity", in *Digital Images and Human Vision,* A. B. Watson, ed., Cambridge, MA, MIT Press, 179-205 (1993).
14. T. Frese, C. A. Bouman and J. P. Allebach, "Methodology for Designing Image Similarity Metrics Based on Human Visual System Models", *Proceedings of SPIE/IS&T Conference on Human Vision and Electronic Imaging II*, Vol. 3016, San Jose, CA, 472-483 (1997).
15. CCIR, "Rec. 500-2 Method for the Subjective Assessment of the Quality of Television Pictures", (1986).
16. M. Van Dijk, J. B. Martens, "Subjective Quality Assessment of Compressed Images", *Signal Processing*, 58, 235-252 (1997).
17. A.M. Rohaly, P. Corriveau, J. Libert, A. Webster, V. Baroncini, J. Beerends, J.L Blin, L. Contin, T. Hamada, D. Harrison, A. Hekstra, J. Lubin, Y. Nishida, R. Nishihara, J. Pearson, A. F. Pessoa, N. Pickford, A. Schertz, M. Visca, A. B. Watson, S. Winkler:

"Video Quality Experts Group: Current results and future directions." *Proc. SPIE Visual Communications and Image Processing*, vol. 4067, Perth, Australia, June 21-23, 2000.

18. P. Corriveau, A. Webster, "VQEG Evaluation of Objective Methods of Video Quality Assessment", SMPTE Journal, 108, 645-648, 1999.
19. T. Kanugo, 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, J. Kittler, "Error-Sensitivity Assessment of Vision Algorithms", *IEE Proc.-Vis. Image Signal Proces*sing, 145(2), 124-130 (1998).
21. M. Grim, H. Szu, "Video Compression Quality Metrics Correlation with Aided Target Recognition (ATR) Applications", *J. of Electronic 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), 1261-1278 (1990).
23. H. H. Barrett, J. L. Denny, R. F. Wagner, 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, E.E. Burroughs, Developing Operational Performance Metrics Using Image Comparison Metrics and the Concept of Degradation Space, Optical Engineering, 38 (5), 836-844, 1999.
25. G. K. Wallace, "The JPEG Still Picture Compression Standard", *IEEE Trans. Consumer Electron.*, 38(1), 18-34 (1992).
26. A. Said, W. A. Pearlman, "A New Fast and Efficient Image Codec Based on Set Partitioning in Hierarchical Trees", *IEEE Trans. Circuits and Syst. Video Technol.*, 6(3), 243-250 (1996).
27. A.M.Martinez, R. Benavente, The AR Face Database, CVC Technical Report No. 24, June 1998.
28. A. C. Rencher, *Methods of Multivariate Analysis*, New York, John Wiley (1995).
29. T. Kohonen, *Self-Organizing Maps*. Springer-Verlag, Heidelberg, (1995).
30. A. W. Lohmann, D. Mendelovic, G. Shabtay, "Significance of Phase and Amplitude in the Fourier Domain", *J. Opt. Soc. of Am.*, 14, 2901 - 2904 (1997).
31. M. P. Dubuisson, A. K. Jain, "A Modified Hausdorff Distance for Object Matching", *Inter. Conf. on Pattern Recognition*, A: 566-569, Jerusalem (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), Supplement No. 2, Bureau Central de la CIE, Vienna, (1971).
33. A. K. Jain, *Fundamentals of Digital Image Processing*, New Jersey, Prentice Hall (1989).
34. V. DiGesu, V. V. Staravoitov, "Distance-based Functions for Image Comparison", *Pattern Recognition Letters*, 20(2), 207-213 (1999).
35. V. V. Starovoitov, C. Köse, B. Sankur, "Generalized Distance Based Matching of Nonbinary Images", *International Conference on Image Processing*, Chicago, (1998).
36. P. Juffs, E. Beggs, F. Deravi, "A Multiresolution Distance Measure for Images", *IEEE Signal Processing Letters*, 5(6), 138-140 (1998).
37. D. Andreutos, K. N. Plataniotis, 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*, New York, Wiley (1978).

40. P. E. Trahanias, D. Karakos, 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, B. Wegmann “The Importance of Intrinsically Two-Dimensional Image Features in Biological Vision and Picture Coding,” in *Digital Images and Human Vision,* A. B. Watson, ed., Cambridge, MA, MIT Press, 109-138 (1993).
42. P. K. Rajan, J. M. Davidson, “Evaluation of Corner Detection Algorithms”, Proc. of Twenty-First Southeastern Symposium on System Theory, 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, T. Caelli, “Region Based Coding of Color Images Using KLT”, *Graphical Models and Image Processing* 59(1), 27-38 (1997).
45. H. Tao, T. Huang, “Color Image Edge Detection using Cluster Analysis”, *IEEE International Conference On Image Processing, 834-836, IEEE Signal Processing Society,* IEEE, California, (1997)
46. P. E. Trahanias, 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 Inc., (1969).
48. M. McIvor, R. J. Valkenburg, “A Comparison of Local Surface Geometry Estimation Methods”, *Machine Vision and Applications*, 10, 17-26 (1997).
49. İ. Avcıbaş, B. Sankur, “Statistical Analysis of Image Quality Measures”, 10. European Signal Processing Conf., *EUSIPCO-2000,* 2181-2184, Tampere, Finland, 2000.
50. R. O. Duda and P. E. Hart, *Pattern Recognition and Scene Analysis*, New-York, Wiley, (1973).
51. K. Popat, R. Picard, “Cluster Based Probability Model and It’s 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*, New York, Wiley (1991).
53. M. Basseville, “Distance Measures for Signal Processing and Pattern Recognition”, *Signal Processing* 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,” *Tech. Rep. 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”, Society for Information Display Digest of Technical Papers, XXIV, 946-949, 1993.
57. İ. Avcıbaş, N. Memon and B. Sankur, “Staganalysis of Watermarking Techniques Using Image Quality Metrics”, *SPIE Conference 4314: Security and Watermarking of Multimedia Contents III*, San Jose, USA, 2001.
58. İ. Avcıbaş, N. Memon and B. Sankur, “Steganalysis Based On Image Quality Metrics”, to be presented, *MMSP’2001: IEEE Workshop on Multimedia Signal Processing*, Cannes, France, 2001.

59. İ. Avcibaş, N. Memon and B. Sankur, “Steganalysis Using Image Quality Metrics”, (under review), *IEEE Transactions on Image Processing*, 2001.