Evaluating Quality of Compressed Medical Images: SNR, Subjective Rating, and Diagnostic Accuracy

PAMELA C. COSMAN, MEMBER, IEEE, ROBERT M. GRAY, FELLOW, IEEE, AND RICHARD A. OLSHEN

Invited Paper

Compressing a digital image can facilitate its transmission, storage, and processing. As radiology departments become increasingly digital, the quantities of their imaging data are forcing consideration of compression in picture archiving and communication systems (PACS) and evolving teleradiology systems. Significant compression is achievable only by lossy algorithms, which do not permit the exact recovery of the original image. This loss of information renders compression and other image processing algorithms controversial because of the potential loss of quality and consequent problems regarding liability, but the technology must be considered because the alternative is delay, damage, and loss in the communication and recall of the images. How does one decide if an image is good enough for a specific application, such as diagnosis, recall, archival, or educational use? We describe three approaches to the measurement of medical image quality: signal-to-noise ratio (SNR), subjective rating, and diagnostic accuracy. We compare and contrast these measures in a particular application, consider in some depth recently developed methods for determining diagnostic accuracy of lossy compressed medical images, and examine how good the easily obtainable distortion measures like SNR are at predicting the more expensive subjective and diagnostic ratings. The examples are of medical images compressed using predictive pruned tree-structured vector quantization, but the methods can be used for any digital image processing that produces images different from the original for evaluation.

I. Compression and Quality Measurement

The overall goal of compression is to represent an image with the smallest possible number of bits, thereby speeding transmission and minimizing storage requirements. Alternatively, the goal is to achieve the best possible fidelity for an available communcation or storage bit rate capacity. A compression system typically consists of one or more of the following operations, which may be combined with each other or with additional signal processing:

- *Sampling*: The intensity of an analog image is measured on a regular grid of points called *picture elements* or *pixels*.
- *Signal decomposition*: The image is decomposed into several images for separate processing, typically by linear transformation by a Fourier or discrete cosine transform or by filtering with a subband or wavelet filter bank. The goal is to concentrate energy in a few coefficients, to reduce correlation, or to provide a useful data structure.
- *Quantization*: Analog or high-rate digital pixels are converted into a relatively small number of bits. This operation is nonlinear and noninvertible; it is "lossy." The conversion can operate on individual pixels (scalar quantization) or groups of pixels (vector quantization). Quantization can include throwing away some of the components of the signal decomposition step.
- *Lossless compression*: Further compression is achieved by an invertible (lossless, entropy) code such as Huffman, Lempel–Ziv, or arithmetic code. The idea here is to assign codewords with a few bits to likely symbols and codewords with more bits to unlikely symbols so that the average number of bits is minimized.

Decompression reverses the above process to the extent possible. Shannon's data transmission theorem [1]–[3] and experience argue that a good compression system can be designed by focusing separately on each individual operation, although simpler or better implementations may be obtained by combining some operations. The *code rate* or *bit rate* of the compression system is defined as the average number of

Manuscript received November 1, 1993; revised January 15, 1994. This work was supported by the National Institutes for Health under Grants CA49697-02 and 5 RO1 CA55325, and by the National Science Foundation under Grant DMS-9101528.

P. C. Cosman and R. M. Gray are with the Information Systems Laboratory, Department of Electrical Engineering, Stanford University, Stanford, CA 94305-4055, USA.

R. A. Olshen is with the Division of Biostatistics, Stanford University School of Medicine, and the Department of Statistics, Stanford University, Stanford, CA 94305, USA.

IEEE Log Number 9400793.

bits produced per image pixel. If the original image has 12 bits per pixel (bpp) and the compression algorithm has rate R bpp, then the *compression ratio* is $12 : R$. Compression ratios must always be taken with a grain of salt, however, as they are strongly dependent on the image type, the original bit rate, and the sampling density.

Lossless coding techniques are well understood, readily available [4]–[6], and typically yield compression ratios of 2:1 to 3:1 on still-frame medical images. Although scientists and physicians prefer to work with uncorrupted data, the modest compression offered by lossless coding is often inadequate for storage or communication facilities. *Lossy* coding, though it does not permit perfect reconstruction of the original image, can provide excellent quality at a fraction of the original bit rate [7], [8]. Lossy coding is unavoidable if the original image is analog, as is ordinary X-ray film. An advantage of a well-designed lossy compression system is that it works to minimize information loss or image distortion for a given allotted storage space or communication rate. When bits are scarce, good compression schemes devote the available bits to the information of greatest importance. As a result, lossy compression schemes are capable of enhancing specific structures of importance to the viewer.

Utility depends critically on the quality of the processed images, but quality is itself an attribute with many possible definitions and interpretations, depending on the use to which the images will be put. A "good" processed image might be one that is perceptually pleasing or useful in a specific application. No single approach to quality measurement has gained universal acceptance, but three general approaches have come to dominate:

- Computable objective distortion measures such as mean squared error or signal-to-noise ratio (SNR),
- subjective quality as measured by psychophysical tests or questionnaires with numerical ratings, and
- simulation and statistical analysis of a specific application of the images, e.g., diagnostic accuracy in medical images measured by clinical simulation and statistical analysis.

The intent of this paper is to compare and contrast these three types of distortion measures in a particular application, lossy compression of medical images. The application is a controversial one; the concern is often expressed that radiologists will never accept lossy compressed images because of liability issues or because of fear of impaired quality. Many in the medical image compression community argue that lossy compression is both necessary and helpful in the long run [9]–[16]. It is necessary in order to preserve rapid access for follow-up and recall studies because the overwhelming quantity of medical image data requires the remote storage of hardcopy films, which frequently results in loss or damage and always requires significant time to locate and transfer. It is helpful because it permits efficient storage and rapid communication of images among hospitals, clinics, and radiologists' offices and homes. Image communication can be within a local-area network or to remote locations using satellite, ISDN, or ordinary phone links. The issue is not so much whether lossy compression should be used, but rather how many bits are needed to ensure sufficient accuracy for a particular use. To answer this question one needs acceptable quantitative measures of image quality and protocols for careful studies of the tradeoffs of quality and bit rate in specific applications.

II. Image Distortion and Quality

A. Average Distortion and SNR

Suppose that one has a system in which an input pixel block or vector $X = (X_0, X_1, \cdots, X_{k-1})$ is reproduced as $\hat{X} = (\hat{X}_0, \hat{X}_1, \cdots, \hat{X}_{k-1})$ and that one has a measure $d(X, \hat{X})$ of distortion or cost resulting when X is reproduced as $\hat{X}$. A natural measure of the quality or fidelity (actually the lack of quality or fidelity) is then the average distortion $D = E[d(X, \hat{X})]$, where the average might be with respect to a probability model for the images or, more commonly, a sample or time-average distortion. It is common to normalize the distortion in some fashion to produce a dimensionless quantity D/D_0, to form the inverse D_0/D as a measure of quality rather than distortion, and to describe the result in decibels. One way to do this is to normalize by the average distortion resulting by the best possible 0 bit rate reproduction $D_0 = \min_y E[D(X, y)]$. With the usual squared error measure of distortion

$$d(X, Y) = ||X - Y||^2 = \sum_{i=0}^{k-1} |X_i - Y_i|^2 \tag{1}$$

$$D_0 = E[||X - E(X)||^2] = \sigma_X^2 \tag{2}$$

the variance of the input vector, and we have the *signal-to-noise ratio* (SNR)

$$\text{SNR} = 10 \log_{10} \frac{D_0}{D} = 10 \log_{10} \frac{\sigma_X^2}{E[||X - \hat{X}||^2]}. \tag{3}$$

A common alternative normalization when the input is itself an r-bit discrete variable is to replace the variance or energy by the maximum input symbol energy $(2^r - 1)^2$. This is called the *peak signal-to-noise ratio* (PSNR). For the common case of 8-bpp input images this yields

$$\text{PSNR} = 10 \log_{10} 255^2 / D. \tag{4}$$

A key attribute of useful distortion measures is ease of computation, but other properties are also important. We would like a distortion measure that reflects perceptual quality or usefulness in a particular application. No easily computable distortion measure such as squared error is generally agreed to have this property. Common faults of squared error are that a slight spatial shift of an image causes a large numerical distortion but no visual distortion and, conversely, a small average distortion can result in a damaging visual artifact if all the error is concentrated in a small important region. It is because of such shortcomings

that many other quality measures have been studied. The pioneering work of Budrikus [17], Stockham [18], and Mannos and Sakrison [19] was aimed at developing computable distortion measures that emphasized perceptually important attributes of an image by incorporating knowledge of human vision. Their work and subsequent work has provided a bewildering variety of candidate measures of image quality or distortion [20]–[44], [56]–[62]. Examples are general l_p norms such as the absolute error (l_1), the cube root of the sum of the cubed errors (l_3), and maximum error (l_∞), as well as variations on such error measures which incorporate linear weighting. A popular form is the linearly weighted quadratic distortion measures that attempt to incorporate properties of the human visual system such as sensitivity to edges, insensitivity to textures, and other masking effects. The image and the original can be transformed prior to forming the distortion, providing a wide family of spectral distortions, which can also incorporate linear weighting in the transform domain to reflect perceptual importance. Alternatively, one can capture the perceptual aspects by linearly filtering the original and reproduction images prior to forming a distortion, which is equivalent to weighting the distortion in the transform domain. A simple variation of SNR that has proved popular in the speech and audio field is the segmental SNR which is an average of local SNR's on a log scale [45], [46], effectively replacing the arithmetic average of distortion by a geometric average.

In addition to easing computation and reflecting perceptual quality, a third desirable property of a distortion measure is tractability in analysis. The popularity of squared error is partly owed to the wealth of theory and numerical methods available for the analysis and synthesis of systems which are optimal in the sense of minimizing mean squared error. One might design a system to minimize mean squared error because it is a straightforward optimization but then use a different, more complicated, distortion measure to evaluate quality because it does better at predicting subjective quality. Ideally, one would like to have a subjectively meaningful distortion measure that could be incorporated into the system design. There are techniques for incorporating subjective criteria into compression system design, but these tend to be somewhat indirect. For example, one can transform the image and assign bits to transform coefficients according to their perceptual importance or use postfiltering to emphasize important subbands before compression [18], [47], [48].

B. Subjective Ratings

Subjective quality of a reconstructed image can be judged in many ways. A suitably randomized set of images can be presented to experts or typical users who rate them, often on a scale of 1 to 5. Subsequent statistical analysis can then highlight averages, variability, and other trends in the data. Such formalized subjective testing is common in speech and audio compression systems as in the Mean Opinion Score (MOS) and the descriptive rating called the diagnostic acceptability measure (DAM) [46], [49], [50]. These rating systems are common in speech research, and some effort has been devoted to developing a rating system for entertainment video [51], [52], but there has been no standardization for rating still images. Considerable variation exists in the allowed range of numerical responses, in the decision of whether or not to provide a corresponding descriptive phrase for each number, in the choice of those descriptive phrases, and in the attempt to have the subjective rating reflect image utility for a specific application. Some ratings were for paired comparisons in which a viewer assigns a number to the degree of similarity or dissimilarity between two images.

A useful attribute of an objective quality measure such as SNR would be the ability to predict subjective quality. This is particularly important in applications such as digital audio and entertainment video where subjective quality is of paramount importance. For medical images, it may be more important that a computable objective measure be able to predict diagnostic accuracy rather than subjective quality. Many methods have been considered for quantifying the degree of correlation between two such quality measures or the ability of one to predict another for audio, images, and video [26], [46], [53]. Among common methods are to plot mean values of SNR against mean values of the corresponding subjective quality and then fit a curve to the resulting points. Many curves have been considered, including polynomial splines, quadratics, and exponentials. The residual sum of errors then provides an indication of the goodness of the fit. Another popular method is to measure the correlation coefficient between the fitted and actual data points.

A potential pitfall in relating objective distortion measures to subjective quality is the choice of image distortions used in the tests. Some of the literature on the subject has considered signal-independent distortions such as additive noise and blurring, yet it has been implied that the results were relevant for strongly signal-dependent distortions such as quantization error. Experiments should imitate closely the actual distortions to be encountered.

C. Diagnostic Accuracy

The most common means of measuring diagnostic accuracy for computer-processed medical images is based on receiver operating characteristic (ROC) analysis, which has its origins in signal detection theory. A filtered version of signal plus Gaussian noise is sampled and compared to a threshold. If the sample is greater than the threshold, the signal is declared to be there; otherwise, it is declared absent. As the threshold is varied in one direction, the probability of erroneously declaring a signal absent when it is there goes down, but the probability of erroneously declaring a signal there when it is not (a false alarm) goes up. An ROC curve plots the variation of these two quantities or, equivalently, of the tradeoffs between *true positive rate* (sensitivity, the complement of the probability of Type I error) and *false positive rate* (the false alarm rate, the complement of *specificity*). A variety of summary statistics such as the area under the ROC curve can be

computed and interpreted to compare the quality of different detection techniques.

Applications of ROC analysis to other fields require the creation of some form of threshold whose variation allows a similar tradeoff. For radiological applications, this involves asking radiologists to provide a subjective confidence rating of their diagnoses (typically on a scale of 1–5) which is then used as if it were a threshold to adjust for detection accuracy [54], [55]. Radiologists can be trained to use the rating scale and the results can be combined with assumptions on the nature of the data to produce summary statistics reflecting the diagnostic accuracy [9]–[14].

Although by far the dominant technique for quantifying diagnostic accuracy in radiology, ROC analysis possesses several shortcomings for this application. By and large, the necessity for the radiologists to assign specific values to their confidence departs from ordinary clinical practice. Further, as image data are non-Gaussian, methods that rely on Gaussian assumptions are suspect. Modern computer-intensive statistical sample reuse techniques can help get around the failures of Gaussian assumptions. Many clinical detection tasks are nonbinary, in which case specificity does not make sense because it has no natural or sensible denominator, as it is not possible to say how many abnormalities are absent. This can be done for a truly binary diagnostic task such as detection of a pneumothorax, for if the image is normal then exactly one abnormality is absent. Previous studies were able to use ROC analysis by focusing on detection tasks which were either truly binary or could be rendered binary. In these studies, the specificity is defined for the entire image set as the conditional average of these binary-valued specificities that apply for individual images. Genuinely nonbinary detection tasks (locating any and all abnormalities that are present) are not amenable to ordinary ROC analysis techniques. Extensions to ROC to permit consideration of multiple abnormalities have been developed [63], but as presented thus far, these require the use of confidence ratings as well as Gaussian or Poisson assumptions on the data. Finally, ROC analysis has no natural extension to the evaluation of measurement accuracy in compressed medical images. By means of specific examples we describe an approach that closely simulates ordinary clinical practice, applies to nonbinary and non-Gaussian data, and extends naturally to measurement data.

III. Compression Algorithms

Almost all of the previously referenced studies of lossy compression of medical images performed the compression using variations on the standard discrete cosine transform (DCT) coding algorithm combined with scalar quantization and lossless coding. (For treatments of DCT coding, see the tutorial book by Rabbani and Jones [8] or [7], [45], [64]–[68].) These are variations of the international standard ISO/CCITT Joint Photographic Experts Group (JPEG) compression algorithm [69]. Some groups used full frame transforms and differing algorithms for bit allocation and noiseless coding. These algorithms are well understood and have been tuned to provide good performance in many applications. Alternative algorithms incorporating vector quantization (VQ) can provide advantages in some applications in terms of simplicity, speed, performance, natural progressive reconstruction, and ease of combining with additional signal processing such as enhancement and classification. Furthermore, VQ can be used in combination with traditional techniques by replacing the scalar quantizers often used in transform and subband techniques by vector quantizers. We emphasize that the clinical simulation protocols and statistical methods considered here are applicable to any compression algorithm.

Vector quantization is the conversion of vectors (typically a block of pixel intensity values in the original image such as a 2×2 square) into binary vectors that tell the decompressor which reproduction template (or codeword) from a limited set called a *codebook* should be used to best approximate the original vector. The compression or encoding is accomplished by a nearest neighbor or minimum distortion matching of the input vector with the available codewords, where the distortion or cost function can be a simple squared error or a more complicated measure which weights perceptual importance. Basic VQ decompression is simply table lookup, yielding extremely fast image reconstruction. To improve efficiency and performance, the encoder is usually constrained to perform its selection in a computationally efficient way and the decoder is allowed to do some simple linear computation. Surveys of the general approach and many of its variations may be found in [68] and in several tutorial articles [70]–[73] and in the IEEE Reprint Collection [74].

Stated in this generality, VQ includes the common transform coding techniques as a special case. More commonly, however, the name is associated with a particular family of design techniques based on information-theoretic and statistical techniques such as clustering and classification trees. The codebooks or collections of templates are often designed using statistical clustering techniques which attempt to find a small number of representatives for a large data set that do a good job of representing the entire set in the sense of minimizing the average distortion between the original and the representative. A common example is the generalized Lloyd algorithm (also called the Forgey or k-means algorithm), which has a variety of forms. (See, e.g., [68], [74].)

In order to make the codebook search low-complexity, techniques from the design of statistical classification trees can be extended to design codebooks with a tree structure, thus codebooks that can be searched by a sequence of simple comparisons (hyperplane or correlation tests) instead of a large number of distortion computations. The complexity of tree-structured codes grows linearly in bit rate instead of exponentially, as is the case with unstructured codes. This approach combines ideas from the classification and regression tree (CART^{TM}) design technique of Breiman, Friedman, Olshen, and Stone [75] with those of clustering to provide a class of compression codes called pruned

tree-structured VQ (PTSVQ) with several nice properties. PTSVQ generally yields lower distortion than fixed-rate full search VQ for a given *average* rate, has a simple encoder and a simple design algorithm, has a natural successive approximation (progressive) property, is well matched to variable-rate environments such as storage or packet communications, and can have its tree tailored by an input-dependent importance weighting by using weighted distortion measures. This latter attribute permits the optional incorporation of enhancement or highlighting into compression by using distortion measures that assign increased importance to specified features, where the features can be automatically classified or marked by a human expert in a learning data set. Such enhancement can force the compression algorithm to devote its bits and quality to features of primary importance and can highlight regions of an image that the computer finds suspicious because they use codewords that had been usually used for pathologies during the codebook design stage. PTSVQ can be improved by incorporating prediction. This technique first performs a simple linear prediction of the current pixel block given previously coded blocks, subtracts the prediction to form a residual error vector, and then applies a tree-structured VQ to the residual.

VQ alone provides a compression technique for sampled images, but it can also be used as the quantization step in a general compression system by combining it with signal decomposition techniques such as transforms or subband or wavelet filter banks and with lossless coding. Such cascades can provide better performance at the cost of added complexity.

IV. Study Design

We now turn to compressed CT and MR chest scans as specific examples for demonstrating the three quality measures. Two CT image types were considered: mediastinal and lung images for diagnoses of lymphadenopathy and lung nodules, respectively. Abnormally enlarged lymph nodes (adenopathy) in the mediastinum (the central portion of the chest containing the heart and major blood vessels), can be caused by primary or metastatic cancer, tuberculosis, and noninfectious inflammatory diseases. Radiologists can usually locate lymph nodes in a CT chest scan. The detection task is therefore to determine which of the located nodes are enlarged. Lung nodules can be caused by fungal and bacterial infections, and by malignancy, primary or metastatic. The latter can cause multiple nodules in one or both lungs. Nodules can be undetectably small or large enough to fill an entire segment of the lung. In contrast to the mediastinal task, both the presence and size of lung nodules must be ascertained, although this study has not yet considered size. This was the focus of the MR study, in which blood vessel sizes were measured. Such measurements are useful for detecting aneurysms and assessing cardiovascular physiology.

Each study (CT lymph nodes, CT lung nodules, MR blood vessels) used predictive PTSVQ with the prediction

Fig. 1. Original MR chest scan at 9 bpp.

Fig. 2. MR chest scan compressed to 0.56 bpp.

coefficients and residual quantizers designed on training sets of representative images. The three training sets consisted, respectively, of 20 CT mediastinum images, 20 CT lung images, and 20 MR chest scans. All sets included images containing pathology and normal images. For each study, 30 test images were chosen. Patient studies represented in the training sets were not used as test images. The SNR, subjective quality, and diagnostic results all are based on test and not training images. Figure 1 shows an original 9-bpp MR chest image. Figure 2 shows that image compressed to 0.56 bpp.

The CT studies used 2 × 2 pixel blocks. The 12-bpp original test images were encoded at 6 compression levels: 0.57, 1.18, 1.33, 1.79, 2.19, and 2.63 bpp. The compressed and original images were viewed by three radiologists. For each of the 30 images in a study, each radiologist independently viewed the original and 5 of the 6 compressed levels, so a total of 360 images were seen by each judge. Images were seen on hardcopy film on a lightbox, with a standard "windows and levels" adjustment to the dynamic range applied to each image before filming. The viewings were divided into three sessions, which were at least two weeks apart. The judges marked abnormalities directly on the hardcopy films with a grease pencil. No constraints were placed on the viewing time, the viewing distance, or the lighting conditions. As the task did not differ from those encountered in daily work, the judges were given no special training for this experiment. They

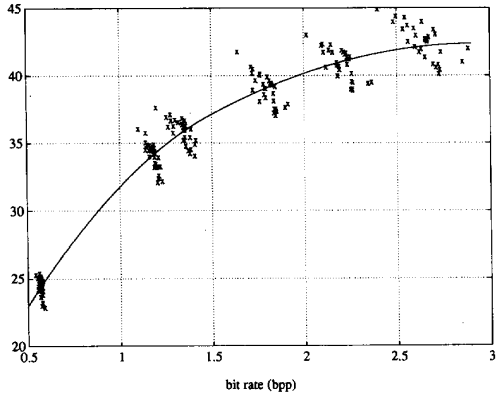


Fig. 3. SNR for CT lung images.

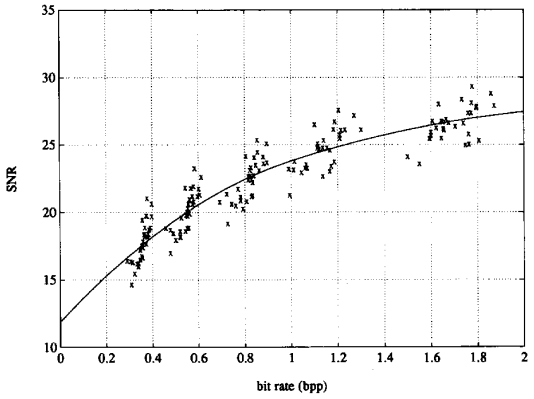


Fig. 4. SNR for MR images.

were, however, constrained to view the 10 pages in the predetermined order. At each session, each judge saw each image at exactly 2 of the 7 levels of compression (7 levels includes the original) with at least 3 pages separating them. This was intended to reduce learning effects.

The MR study used 2 × 4 pixel blocks and 5 compression levels: 0.36, 0.55, 0.82, 1.13, and 1.7 bpp. The test images were taken at a level in the chest where the four major blood vessels (ascending aorta, descending aorta, right pulmonary artery (RPA), and superior vena cava (SVC)) appeared. In a protocol similar to that of the CT studies, the six levels of each test image were presented separately to three radiologists. For each image, each radiologist marked the axis of measurement and measured each blood vessel at its widest diameter to the nearest millimeter using calipers. Aortic diameters ranged from 10 to 60 mm and the SVC and RPA from 15 to 40 mm.

V. Image Quality Results

The traditional manner for comparing the performance of different lossy compression systems is to plot distortion rate or SNR versus bit rate curves. Figure 3 shows a scatter plot of the rate–SNR pairs for the 24 images in the lung CT study that had a consensus gold standard (as discussed below). It includes a quadratic spline fit with a single knot at 1.5 bpp [76]. Quadratic spline fits provide good indications of the overall distortion-rate performance of the code family on the test data. The SNR results for the CT mediastinal images (not shown) were very similar to those for the lung task. Figure 4 shows a scatter plot of the rate–SNR pairs for the 30 images in the MR study. The quadratic spline has a single knot at 1.0 bpp.

The assessment of subjective quality attempted to relate subjective image quality to diagnostic utility. For the MR study, each radiologist was asked at the time of measuring the vessels to "assign a score of 1 (worst) to 5 (best) to each image based on its usefulness for the measurement task." The CT subjective assessment was performed separately from the diagnostic task by three different radiologists. The subjective rating/bit rate pairs for the CT studies and MR study are plotted in Figs. 5–7. Images compressed to lower

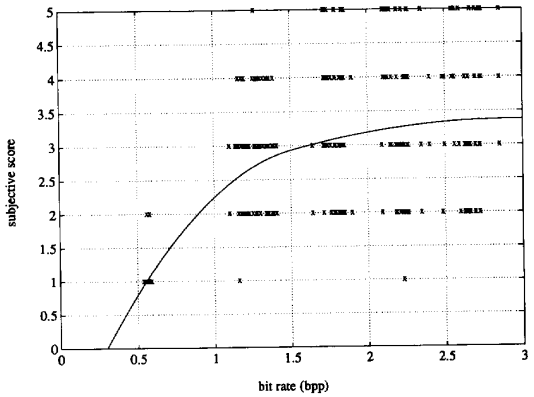


Fig. 5. Subjective ratings versus bit rate for CT lung study.

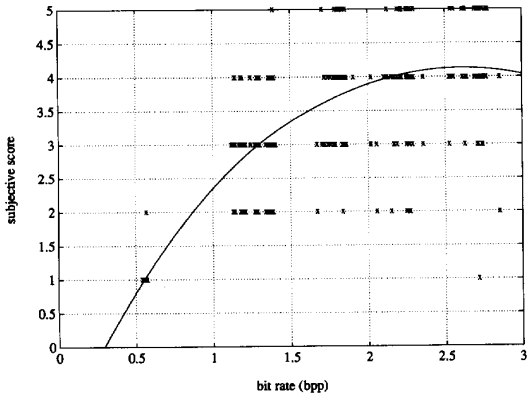


Fig. 6. Subjective ratings versus bit rate for CT mediastinum study.

bit rates received worse quality scores as was expected. The data are fit with a quadratic spline with a single knot.

To measure diagnostic accuracy, we first needed to determine a "gold standard" that would represent the diagnostic truth of each original image, and could serve as a basis of comparison for the diagnoses on all versions of that image. There are many possible choices for the gold standard:

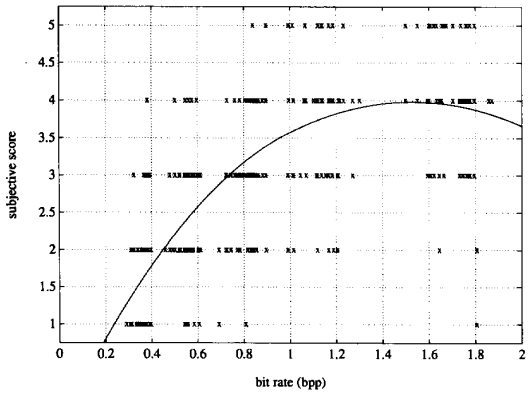


Fig. 7. Subjective ratings versus bit rate for MR chest study.

- A *consensus* gold standard is determined by the consensus of the three judges on the original.
- A *personal* gold standard uses each judge's readings on an original (uncompressed) image as the gold standard for the readings of that same judge on the compressed versions of that same image.
- An *independent* gold standard is formed by the agreement of the members of an independent panel of particularly expert radiologists.
- A *separate* gold standard is produced by the results of autopsy, surgical biopsy, reading of images from a different imaging modality, or subsequent clinical or imaging studies.

A separate standard is generally not available, as there may be no further procedures or studies of that patient made or available; for these studies, we did not have records of further procedures. In the CT studies we used both the consensus and personal gold standards; in the MR study we used both the independent and personal standards. In the cases where the initial CT readings disagreed in the number or location of abnormalities, the judges were asked separately to review their readings of that original. If this did not produce agreement, the judges discussed the image together. Six images in each CT study could not be assigned a consensus gold standard due to irreconcilable disagreement. This was a fundamental drawback of the consensus gold standard and subsequent studies did not use this method. Since the consensus was clearly more likely to be attained for those original images where the judges were in perfect agreement initially and thus where the original images would have perfect accuracy relative to that agreed gold standard, the original images have an advantage when compared to the others. The consensus gold standard for the lung determined that there were, respectively, 4 images with 0 nodules, 9 with 1, 4 with 2, 5 with 3, and 2 with 4 among those images retained. For the mediastinum, there were 3 images with 0 abnormal nodes, 17 with 1, 2 with 2, and 2 with 3. The personal gold standard is even more strongly biased against compression. It defines a judge's reading on an original image to be perfect, and uses that reading as the basis of comparison for the compressed versions of that image. In the presence of any random noise in the process of judging, with compressed images performance is less good than with the originals. Thus the personal and consensus gold standards are most useful for comparing the various compressed levels among themselves. Comparisons of the original images with the compressed ones are conservative. One argument for the personal standard is that in some clinical settings a fundamental question is how the reports of a radiologist whose information is gathered from compressed images compare to what they would have been on the originals, the assumption being that systematic biases of a radiologist are well recognized and corrected for by the referring physicians who regularly send cases to that radiologist. The personal gold standard thus concentrates on consistency of individual judges. The independent gold standard has the advantage that it allows the diagnoses of the judges on the originals to be compared against those on the compressed images in an unbiased way, since in both cases errors will be determined by reference to the standard created by the independent panel. In the MR study, the panel was composed of two senior radiologists who first measured the vessels separately and then discussed and remeasured in those cases where there was initial disagreement.

A. Detection Accuracy (CT Study)

Once a gold standard is established, a value can be assigned to the sensitivity, the probability that something is detected given that it is present in the gold standard. Sensitivity makes sense for nonbinary detection tasks, and is a crucial statistic that quantifies results. However, a judge who labels abnormalities everywhere in an image could have perfect sensitivity. *Predictive value positive* (PVP), the chance an abnormality is actually present given that it is marked [77], fills the role of specificity in penalizing false positive reporting. A judge who is too aggressive in finding abnormality could have high sensitivity at the expense of low PVP while a judge who is too stringent about what defines abnormality could have a high PVP at the expense of low sensitivity. As is the case with the ROC parameters of true positives and false positives, both true positive and PVP will be 1 if the decision is perfect. The plots below show sensitivity and PVP relative to the consensus gold standard.

Figures 8 and 9 display all data for lung sensitivity and lung PVP for all 24 images, judges, and compressed levels. There are 360×'s: 360 = 3 judges × 24 images × 5 compressed levels seen for each image. Figures 10 and 11 are the corresponding figures for the mediastinum. The o's mark the average of the ×'s for each bit rate. The values of the sensitivity and PVP are simple fractions such as 1/2 and 2/3 because there are at most a few abnormalities in each image. The curves are least squares quadratic spline fits to the data with a single knot at 1.5 bpp, together with the two-sided 95% confidence regions. In view of the highly non-Gaussian nature of the data, the confidence regions were obtained by a bootstrapping procedure [78],

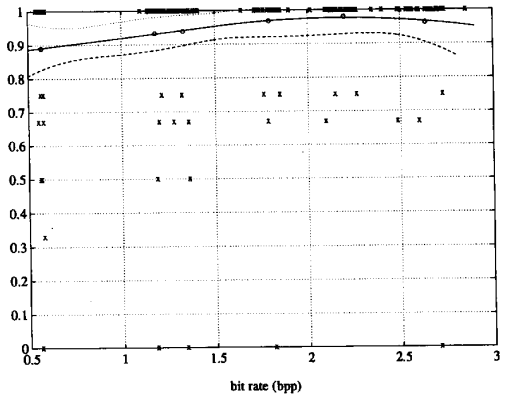


Fig. 8. Lung sensitivity: rms = 0.177.

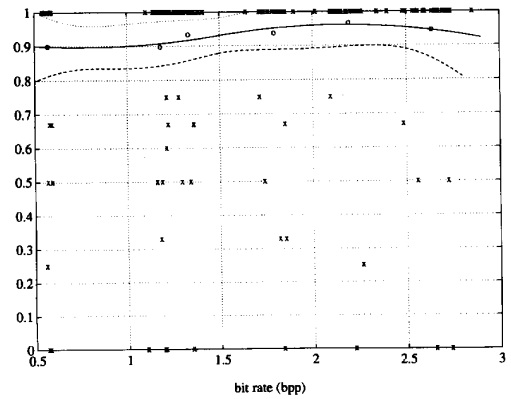


Fig. 9. Lung PVP: rms = 0.215.

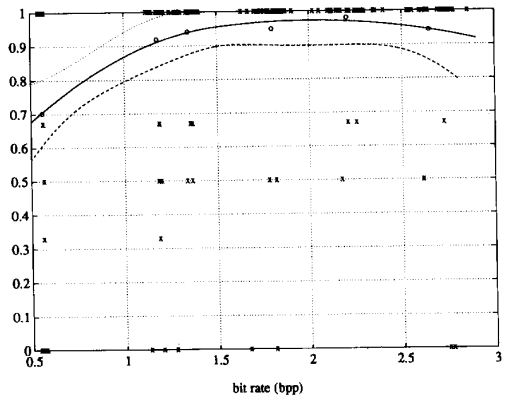


Fig. 10. Mediastinum sensitivity: rms = 0.243.

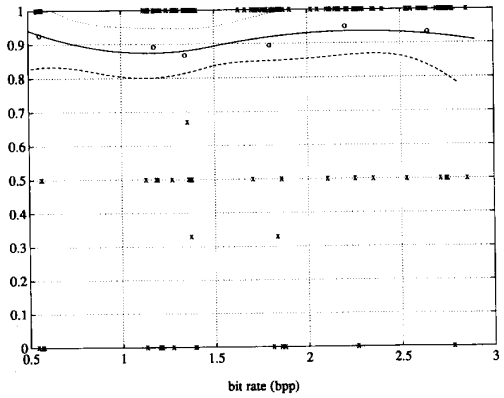


Fig. 11. Mediastinum PVP: rms = 0.245.

[79]. Since the sensitivity and PVP cannot exceed 1, the upper confidence curve was thresholded at 1. The residual root mean square (rms) is the square root of the residual mean square from an analysis of variance of the spline fits. Sensitivity for the lung seems to be nearly as good at low rates of compression as at high rates, but sensitivity for the mediastinum drops off at the lower bit rates, driven primarily by the results for one judge. PVP for the lung is roughly constant across the bit rates, and the same for the mediastinum.

Behrens–Fisher t-Statistic: The comparison of sensitivity and PVP at different bit rates was carried out using a permutation distribution of a two-sample t-test that is sometimes called the Behrens–Fisher test [80], [81]. The statistic takes account of the fact that the within group variances are different. The test is exact and does not rely on Gaussian assumptions that would be patently false for this data set. The use of this statistic is illustrated by the following example. Suppose Judge 1 has judged N lung images at both levels A and B. These images can be divided into 5 groups, according to whether the consensus gold standard for the image contained 0, 1, 2, 3, or 4 abnormalities. Let N_i be the number of images in the ith group. Let Δ_{ij} represent the difference in sensitivities (or PVP) for the jth image in the ith group seen at level A and at level B. Let $\bar{\Delta}_i$ be the average difference:

$$\bar{\Delta}_i = \frac{1}{N_i} \sum_j \Delta_{ij}.$$

We define

$$S_i^2 = \frac{1}{N_i - 1} \sum_j (\Delta_{ij} - \bar{\Delta}_i)^2$$

and then the Behrens–Fisher t statistic is given by

$$t_{\mathrm{BF}} = \sum_i \bar{\Delta}_i / \sqrt{\sum_i S_i^2 / N_i}.$$

Our Δ_{ij} are fractions with denominators not more than 4, so are utterly non-Gaussian. Therefore, computations of attained significance (p values) are based on the restricted permutation distribution of t_{BF}. For each of the N images, we can permute the results from the two levels [$A \rightarrow B$ & $B \rightarrow A$] or not. There are 2^N points possible in the full permutation distribution, and we calculate t_{BF} for each one. The motivation for the permutation distribution is that if there were no difference between the bit rates, then in computing the differences Δ_{ij}, it should not matter whether

we compute level A – level B or *vice versa*, and we would not expect the "real" t_{BF} to be an extreme value among the 2^N values. If k is the number of permuted t_{BF} values that exceed the "real" one, then $(k+1)/2^N$ is the attained one-sided significance level for the test of the null hypothesis that the lower bit rate performs at least as well as the higher one. The one-sided test of significance is chosen to be conservative and to argue most strongly against compression.

When the judges were evaluated separately, level A (the lowest bit rate) was found to be significantly different at the 5% level against most of the other levels for two of the judges, for both lung and mediastinum sensitivity. No differences were found among levels B through G. There were no significant differences found between any pair of levels for PVP. When judges were pooled, more significant differences were found. Level A was generally inferior to the other levels for both lung and mediastinal sensitivity. Also levels B and C differed from level G for lung sensitivity (p = 0.016 for both) and levels B and C differed from level G for mediastinal sensitivity (p = 0.008 and 0.016, respectively). For PVP, no differences were found against level A with the exception of A versus E and F for the lungs (p = 0.039 and 0.012, respectively), but B was somewhat different from C for the lungs (p = 0.031), and C was different from E, F, and G for the mediastinum (p = 0.016, 0.048, and 0.027, respectively).

The results indicate that level A (0.56 bpp) is unacceptable for diagnostic use. Since the blocking and prediction artifacts became quite noticeable at level A, the judges tended not to attempt to mark any abnormality unless they were quite sure it was there. This explains the initially surprising result that level A did well for PVP, but very poorly for sensitivity. Since no differences were found among levels D (1.8 bpp), E (2.2 bpp), F (2.64 bpp), and G (original images at 12 bpp), despite the biases against compression contained in our analysis methods, these three compressed levels are clearly acceptable for diagnostic use in our applications. The decision concerning levels B (1.18 bpp) and C (1.34 bpp) is less clear, and would require further tests involving a larger number of detection tasks, more judges, and use of an independent gold standard that in principle should remove at least one of the biases against compression that are present in this study.

We also compared compression levels using sensitivity and PVP defined relative to the personal gold standard, we used a Hotelling T^2 statistic to compare the judges and the images, and we used a McNemar statistic applied to paired data (in which the first occurence of a given image in a session was paired with the second occurence of that same image, at a different compression level, in the same session) to ascertain that learning effects were not significant at the 5% level [15], [16].

B. Measurement Accuracy (MR Images)

The measurement error for the MR images can be quantified in several ways, the first being *percent measurement error*, which is the difference between the gold standard measurement and that of the judge, scaled by the gold standard. Though analyses of variance suggest that for this task judges are not the same—indeed, differences among them may be the largest source of variability in the measurements—we could argue that differences in bit rates are not confounded with differences in judges, and that for present purposes nothing essential in our comparison of bit rates is lost if we pool the judges. Figure 12 is a graph of percent error versus bit rate for the independent gold standard when all judges and images are pooled. The curve, a quadratic spline with a single knot at 1.5 bpp, is nearly flat. This could owe to some sort of cancellation, which possibility is addressed by Fig. 13, where the outcome is absolute percent measurement error. The displayed spline fit is nearly flat. The corresponding figures for the personal gold standards are similar. From these figures we tentatively conclude that for this measurement task, error does not seem to vary with bit rate.

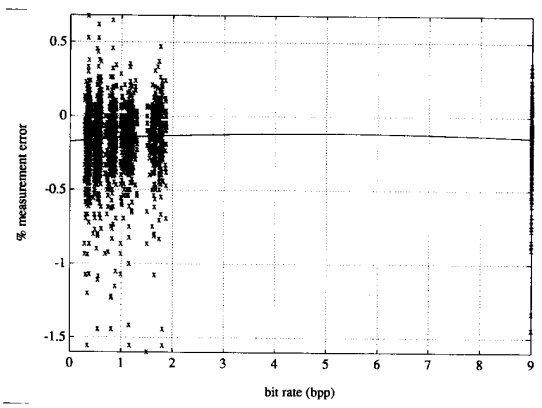


Fig. 12. Spline fit to percent measurement error versus bit rate.

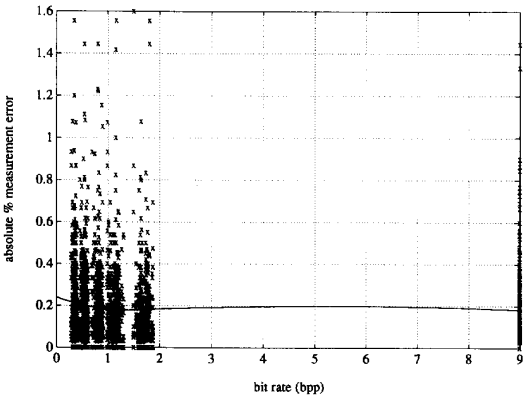


Fig. 13. Spline fit to absolute percent measurement error versus bit rate.

We used two statistical tests of significance on the error differences: a paired t test and a Wilcoxon signed rank test. As the data are, the conclusions are identical up to noise; the Gaussian assumption is close to but not exactly confirmed (by a plot of data quantiles against corresponding Gaussian ones, an empirical Shapiro–Wilk test). With

the Wilcoxon, measurement seemed significantly better in terms of percentage error at 1.7 bpp than at 0.82 or 1.13 bpp for the independent gold standard ($p < 0.05$), though in no case was measurement at any of our levels of compression different from that on the original images. With the personal gold standard, images compressed all the way to 0.36 bpp definitely seem not as good as the originals, and, further, bit rates 0.36, 0.55, and 1.13 bpp seem different from 1.7 bpp. Results are qualitatively similar for comparisons with the t test. From these arguments and others, we assert confidently that compression with our algorithms down to 0.55 bpp is possible without significantly affecting the accuracy of this vessel measurement task [82].

VI. Relationships Between Quality Measures

As image quality can be quantified by diagnostic accuracy, subjective ratings, or computable measures such as SNR, one key question concerns the degree to which these different measures agree. Verifications of medical image quality by perceptual measures require the detailed, time-consuming, and expensive efforts of human observers, typically highly trained radiologists. Therefore, it is desirable to find computable measures that strongly correlate with or predict the perceptual measures. Our work suggests that cross–validated fits to the data using generalized linear models can be used to examine the usefulness of SNR (or other computable measure) as a predictor for subjective quality (or other perceptual measure).

In the classical linear regression model, the "predictor" x is related to the outcome y by $y = \beta^t x + \epsilon$, where β is a vector of unknown coefficients, and the error ϵ at least has mean zero and constant variance, or may even be normally distributed. In the regression problem of using SNR to predict subjective quality scores, the response variable y takes on integer values between 1 and 5, and so the assumption of constant variance is inappropriate because the variance of y depends on its mean. Furthermore, y takes on values only in a limited range, and the linear model does not follow that constraint without additional untenable assumptions. We turn to a generalized linear model that is designed for modeling binary and, more generally, multinomial data [83].

A generalized linear model requires two functions: a link function that specifies how the mean depends on the linear predictors, and a variance function that describes how the variance of the response variable depends on its mean. If $X_1, X_2, \cdots, X_n$ are independent Poisson variables, then conditional upon their sum, their joint distribution is multinomial. Thus the regression can be carried out with the Poisson link and variance functions: $\beta^t x = \ln \mu$ and $\text{var}(y) = \mu$ in which case the mean of the response variable is $\mu = e^{\beta^t x}$. The results of this approach are shown in Fig. 14. The predictors are a quadratic spline in SNR. In Fig. 14, the x symbols denote the raw data pairs (subjective score, SNR) for the judges pooled, and the curve is the regression fit. The model parameters were estimated using the statistical software S, which uses iteratively reweighted

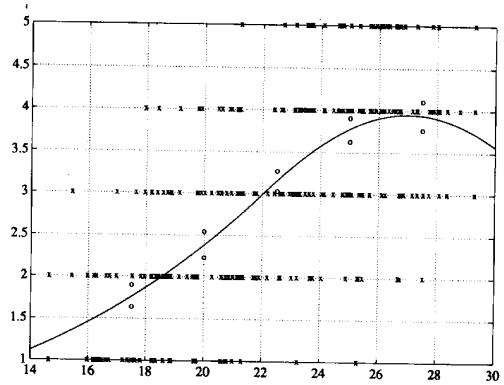


Fig. 14. Expected subjective score versus SNR (Permission for reprint, courtesy Society for Information Display).

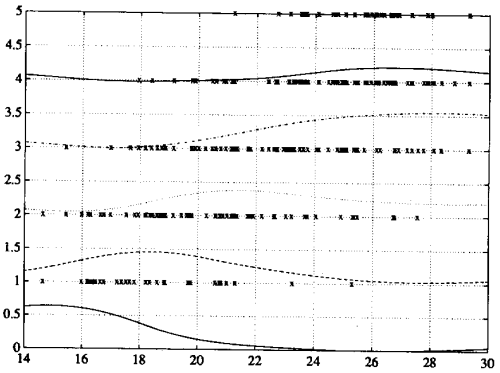


Fig. 15. Response probabilities versus SNR (Permission for reprint, courtesy Society for Information Display).

least squares to generate a maximum-likelihood estimate. The o symbols denote the 95% confidence intervals [84] obtained from the bootstrapped BC_a method [79].

Instead of fitting directly to the expectation of the response, a second way to approach this problem looks for the probability p_i of obtaining the response i, for each of the five possible responses ($i = 1, \cdots, 5$). The expectation can then be calculated from the probabilities. We can transform the responses y into binary outcomes:

$$y_i = \begin{cases} 1, & \text{if } y = i \\ 0, & \text{otherwise.} \end{cases}$$

The binary response variables y_i can then each be fitted using the *logit* link, for which the mean of the response variable is

$$\mu = \frac{e^{\beta^t x}}{1 + e^{\beta^t x}}$$

which guarantees that μ is in the interval [0, 1]. The logit link together with the binomial variance function $\mu(1 - \mu)$ defines the *logistic regression* model. For each y_i the

Table 1 Comparison of Computable Quality Measures

Block Size	M
256 × 256	42.96
128 × 128	42.29
64 × 64	34.49
32 × 32	46.48
16 × 16	47.72
8 × 8	48.10
4 × 4	46.62
2 × 2	47.21
l_1	38.60
l_3	35.08

predictor x was a quadratic spline in SNR, with the knots located in each case at the mean value of the SNR's which produced that response (18.2, 20.12, 22.57, 24.61, 25.56). The probabilities p_i are shown in Fig. 15 with vertical offsets so they are not superimposed. As the five probabilities have been determined from separate regressions, they are scaled so that they add to one before calculating $E(y)$ from them. The expectation is almost indistinguishable from the curve of Fig. 14, thereby validating the Poisson model.

Having established the appropriateness of the Poisson model, we use it to compare SNR against segmental SNR in their ability to predict subjective quality ratings. Segmental SNR, often used in speech quality evaluation, compensates for the under-emphasis of weak-signal performance in conventional SNR. An image is divided into blocks, the SNR is calculated for each block on a log scale, thresholded below at 0 and above at 45, and the values are averaged. By converting component SNR values to decibel values prior to averaging, very high SNR values corresponding to well-coded large-signal segments do not camouflage coder performance with the weak segments, as in conventional SNR. We examined block sizes of all powers of 2 between 2 × 2 and 256 × 256. Since the images are of size 256 × 256, the segmental SNR for that block size equals the conventional SNR. The usefulness of the computable metric in predicting subjective quality was examined as follows: For $n = 20$ times, the 30 MR images were put in a different random order. Each time, a ten-fold cross validation was performed in which 3 images at a time were left out, and the other 27 images were used to fit the model. All judges and levels corresponding to those 27 images were used. The 3 images not involved in determining the parameters of the fit comprise 45 data points (3 images × 3 judges × 5 compression levels). For these data we compute the mean outcome and the sum of squared deviations from this overall mean. This value is called S_1. Then we calculate the fitted values for these data, and take the sum of squared deviations of observed and fitted, called S_2. If the model is good and the test set of 3 images is not unlike the set of 27 images used to fit the model, we expect S_2 to be smaller than S_1. The percent reduction in mean-squared error that owes to fitting the model (beyond fitting an overall constant) is a statistic that summarizes the model's predictive power: $M = 100(1-(S_2/S_1))\%$. This statistic is a cross-validated analog of the multiple correlation. The results are presented in Table 1.

It appears that segmental SNR at several different block sizes outperforms conventional SNR. The best of these (on 8 × 8 blocks) produced a 48% reduction compared to the 43% reduction for SNR. In ongoing research we are examining the statistical significance of these differences by sampling from the permutation distribution, and comparing SNR against perceptually based computable quality measures.

VII. Discussion and Conclusion

There are many different perspectives from which these different measures of image quality can be viewed. They vary in the extent to which they explicitly consider the application for which the images are used. At one extreme are the computable measures such as SNR, which in no way take account of the medical nature of the images. Subjective ratings in which a radiologist is asked to rate the medical usefulness of an image begin to address the issue. ROC analysis, which includes both a (generally) binary diagnostic decision and a subjective confidence ranking associated with that diagnosis, are serious attempts to capture the medical interest of the images through their diagnostic value. Studies such as the CT detection task and MR measurement task presented in this paper attempt to reproduce very closely some actual clinical diagnostic tasks of radiologists, and to ask the fundamental question of whether a diagnosis made on a compressed image is as good as one made on an original. By this measure, an image has high quality if the number and locations of lesions one finds there precisely match the number and locations one finds on the original (or what the independent panel finds on the original). But is that really the fundamental question? A diagnosis is made on a patient's scan in order to make a decision about medical care for that patient, so perhaps image quality could be defined in terms of medical care. That is, an image has high quality if the decision on medical care is unchanged from that determined upon the original. So if the original image has 6 nodules and the compressed one has 9, that may still be an extremely high quality image according to this particular measure, because the medical care decision may be unaltered in the case of many tumors with a few more or less. One can step back further to look at patient outcome rather than medical care decision. Suppose hypothetically that one designs a classification scheme to highlight suspected tumors in an image. And perhaps, unbeknownst to the designers, precancerous cells which have an overlapping intensity distribution with that of cancerous cells also tend to get highlighted, causing the surgeon to make a wider resection and have lower recurrence rates. Then the processed image might rate as poorer quality than an original based on the previous measures (because both diagnosis and medical care decision would be different from those based on the original image), yet the processed image would rate as top quality according to the measure

of improved patient outcome. No one would seriously propose these as measures of image quality. The decision on medical care and the patient outcome both depend on far too many factors other than just image quality. And yet, if one considers the true measure of medical image quality to be simply whether a diagnosis on the processed image is unchanged from the diagnosis on the original, one denies the possibility that the processing may in fact *enhance* the image. This is not a worrisome consideration with image compression, although there is some indication that in fact slightly vector-quantized images are superior to originals, because noise is suppressed by a clustering algorithm. However, this may soon be a difficult issue in evaluating the quality of digitally processed medical images where the processing is, for example, a highlighting based on pixel classification, or a pseudo-colored superposition of images obtained from different modalities. There is a need to develop image evaluation protocols for medical images that explicitly recognize the possibility the processed image can be *better*.

Such issues arise in the area of speech quality measurement as well. Articulation tests score the percentage of correct identification of the sounds, words, or sentences in transmitted speech, and intelligibility tests look for correct identification of the meaning conveyed by transmitted speech [85]. Although requiring the efforts of human subjects, these tests are both objective and quantitative. Even the communicability tests, which examine the effort and level of attention required for understanding, can be conducted in a way that does not involve the subjective opinions of the listeners, for example by measuring the length of time required for subjects to perform a specified task that requires communicating over the system under test. Analogous to the objective diagnostic accuracy tests for image coding systems, these speech quality measures address the fundamental issues of what makes a speech communication system acceptable to the user—the ability to extract correct meaning from the sounds, and the ease of communication. However, the intelligibility and articulation tests are only appropriate for systems that produce moderate to severely degraded speech, since they are useless for distinguishing between speech signals that are highly intelligible. Communicability tests, while not completely useless for good-quality systems, do have low resolving power, and would require either a very large number of subjects, or unpleasantly difficult tasks. They also suffer from lack of reproducibility. All these tests also lack the ability to provide insight to the researcher into the speech perception process or the causes of unintelligibility. For all these reasons, the speech-quality research community has tended to prefer subjective measures such as the Diagnostic Acceptability Measure and the Mean Opinion Score to these intelligibility tests. However, for the reasons outlined above, we believe that diagnostic accuracy tests, the "intelligibility" tests for medical images, are more useful for evaluating medical image quality than are subjective tests.

In addition to the advantages which the evaluation protocol confers on the originals, physician training also provides a bias for existing techniques. Radiologists are trained in medical school and residency to interpret certain kinds of images, and when asked to look at another type of image (e.g., compressed or highlighted) they may not do as well just because they were not trained on those. Highly compressed images have lower informational content than do originals, so even a radiologist carefully trained on those could not do as well as a physician looking at original images. But with image enhancement techniques or slightly compressed images, perhaps a radiologist trained on those would do better when reading those than someone trained on originals would do reading originals.

We have presented several different ways of evaluating medical image quality. Simple computable measures have a role in the design algorithms and in the evaluation of quality simply because they are quickly and cheaply obtainable, and tractable in analysis. The actual diagnostic quality is determined by various statistical protocols which can evaluate diagnostic accuracy in the context of specific detection and measurement tasks. The analysis of subjective quality is of interest mostly for the fact that it shows a different trend from actual diagnostic quality, which can reassure physicians that diagnostic utility is retained even when a compressed image is peceptually distinguishable from the original. There is considerable future work to be done both in evaluation studies of image quality for different types of images and diagnostic tasks, and in searching for computable measures of image quality which can accurately predict the outcome of such studies, and perhaps be incorporated into the code design algorithm to yield better quality compression.

ACKNOWLEDGMENT

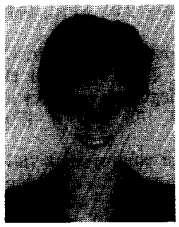
The authors acknowledge the assistance of the many colleagues with whom the research surveyed here was accomplished. In particular, they are grateful for the contributions of C. Tseng, S. Perlmutter, K. C. P. Li, L. Moses, H. C. Davidson, and C. Bergin.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.
[2] I. Csiszár and J. Körner, *Coding Theorems of Information Theory*. Budapest, Hungary: Academic Press/Hungarian Academy of Sciences, 1981.
[3] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.
[4] T. J. Lynch, *Data Compression: Techniques and Applications*. Belmont, CA: Lifetime Learning, Wadsworth, 1985.
[5] J. Storer, *Data Compression*. Rockville, MD: Computer Sci. Press, 1988.
[6] L. D. Davisson, R. J. McEliece, M. B. Pursley, and M. S. Wallace, "Efficient universal noiseless source codes," *IEEE Trans. Informat. Theory*, vol. IT-27, pp. 269–279, 1981.
[7] A. N. Netravali and B. G. Haskell, *Digital Pictures: Representation and Compression*. New York: Plenum, 1988.
[8] M. Rabbani and P. W. Jones, *Digital Image Compression Techniques*, vol. TT7 of *Tutorial Texts in Optical Engineering*. Bellingham, WA: SPIE Opt. Eng. Press, 1991.
[9] J. Sayre, D. R. Aberle, M. I. Boechat, T. R. Hall, H. K. Huang, B. K. Ho, P. Kashfian, and G. Rahbar, "Effect of data compression on diagnostic accuracy in digital hand and chest radiography," in *Proc. Medical Imaging VI: Image Capture,*

Formatting and Display, vol. 1653, pp. 232–240, SPIE, Feb. 1992.
[10] H. MacMahon, K. Doi, S. Sanada, S. Montner, M. Giger, C. Metz, N. Nakamori, F. Yin, X. Xu, H. Yonekawa, and H. Takeuchi, "Data compression: Effect on diagnostic accuracy in digital chest radiographs," *Radiology*, vol. 178, pp. 175–179, 1991.
[11] S. Lo and H. Huang, "Radiological image compression: full-frame bit-allocation technique," *Radiology*, vol. 155, no. 3, pp. 811–817, 1985.
[12] ——, "Compression of radiological images with 512, 1,024 and 2,048 matrices," *Radiology*, vol. 161, no. 2, pp. 519–525, 1986.
[13] T. Ishigaki, S. Sakuma, M. Ikeda, Y. Itoh, M. Suzuki, and S. Iwai, "Clinical evaluation of irreversible image compression: Analysis of chest imaging with computed radiography," *Radiology*, vol. 175, pp. 739–743, 1990.
[14] K. Chan, S. Lou, and H. Huang, "Full-frame transform compression of CT and MR images," *Radiology*, vol. 171, no. 3, pp. 847–851, 1989.
[15] P. Cosman, C. Tseng, R. Gray, R. Olshen, L. E. Moses, H. C. Davidson, C. Bergin, and E. Riskin, "Tree-structured vector quantization of CT chest scans: Image quality and diagnostic accuracy," *IEEE Trans. Med. Imag.*, vol. 12, pp. 727–739, Dec. 1993.
[16] P. Cosman, H. C. Davidson, C. Bergin, C. Tseng, L. E. Moses, E. Riskin, R. Olshen, and R. Gray, "Thoracic CT images: Effect of lossy image compression on diagnostic accuracy," *Radiology*, vol. 190, pp. 517–524, 1994.
[17] Z. Budrikus, "Visual fidelity criteria and modeling," *Proc. IEEE*, vol. 60, pp. 771–779, July 1972.
[18] T. G. Stockham, Jr., "Image processing in the context of a visual model," *Proc. IEEE*, vol. 60, pp. 828–842, July 1972.
[19] J. L. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion of the encoding of images," *IEEE Trans. Informat. Theory*, vol. IT-20, pp. 525–536, July 1974.
[20] A. Netravali and B. Prasada, "Adaptive quantization of picture signals using spatial masking," *Proc. IEEE*, vol. 65, pp. 536–548, 1977.
[21] D. Sakrison, "On the role of the observer and a distortion measure in image transmission," *IEEE Trans. Commun.*, vol. COM-25, pp. 1251–1267, 1977.
[22] J. Limb, "Distortion criteria of the human viewer," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, pp. 778–793, 1979.
[23] F. Lukas and Z. Budrikis, "Picture quality prediction based on a visual model," *IEEE Trans. Commun.*, vol. COM-30, pp. 1679–1692, July 1982.
[24] N. Nill, "A visual model weighted cosine transform for image compression and quality assessment," *IEEE Trans. Commun.*, vol. COM-33, pp. 551–557, 1985.
[25] K. Ngan, K. Leong, and H. Singh, "Cosine transform coding incorporating human visual system model," in *Proc.SPIE*, vol. 707, pp. 165–171, 1986.
[26] H. Marmolin, "Subjective mse measures," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-16, pp. 486–489, May/June 1986.
[27] A. Watson, "Efficiency of an image code based on human vision," *J. OSA A*, vol. 4, pp. 2401–2417, 1987.
[28] E. Shlomot, Y. Zeevi, and W. Pearlman, "The importance of spatial frequency and orientation in image decomposition and coding," in *Proc. SPIE*, vol. 845, pp. 152–158, 1987.
[29] C. Zetzsche and G. Hauske, "Multiple channel model for the prediction of subjective image quality," in *Proc. SPIE*, vol. 1077, pp. 209–216, 1989.
[30] J. Sagrhi, P. Cheatham, and A. Habibi, "Image quality measure based on a human visual system model," *Opt. Eng.*, vol. 28, pp. 813–818, July 1989.
[31] P. Barten, "Evaluation of subjective image quality with the square foot integral method," *J. OSA A*, vol. 7, pp. 2024–2031, 1990.
[32] M. Duval-Destin, "A spatio-temporal complete description of contrast," in *SID'91 Dig. Tech. Papers* (Society for Information Display), vol. 22, pp. 615–618, 1991.
[33] J. Farrell, H. Trontelj, C. Rosenberg, and J. Wiseman, "Perceptual metrics for monochrome image compression," in *SID'91 Dig. Tech. Papers* (Society for Information Display), vol. 22, pp. 631–634, 1991.
[34] V. Algazi, Y. Kato, M. Miyahara, and K. Kotani, "Comparison of image coding techniques with a picture quality scale," in *Proc. SPIE Applications of Digital Image Processing XV* (San Diego, CA, July 1992), vol. 1771, pp. 396–405 .
[35] H. Barrett, "Evaluation of image quality through linear discriminant models," in *SID'92 Dig. Tech. Papers* (Society for Information Display), vol. 23, pp. 871–873, 1992.
[36] S. Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," in *Proc. SPIE* , vol. 1666, pp. 2–14, 1992.
[37] T. Grogan and D. Keene, "Image quality evaluation with a contour-based perceptual model," in *Proc. SPIE* , vol. 1666, pp. 188–197, 1992.
[38] T. Pappas, "Perceptual coding and printing of gray-scale and color images," in *SID Dig.*, vol. 23, pp. 689–692, 1992.
[39] N. Nill and B. Bouzas, "Objective image quality measure derived from digital image power spectra," *Opt. Eng.*, vol. 31, pp. 813–825, Apr. 1992.
[40] S. Klein, A. Silverstein, and T. Carney, "Relevance of human vision to JPEG–DCT compression," in *Proc. SPIE* , vol. 1666, pp. 200–215, 1992.
[41] A. Ahumada, Jr., "Computational image-quality metrics: a review," in *SID'93 Dig. Tech. Papers* (Society for Information Display, Seattle, WA, May 1993), pp. 305–308.
[42] A. Eskioglu and P. Fisher, "A survey of quality measures for gray scale image compression," in *Computing in Aerospace 9*, (AIAA, San Diego, CA, Oct. 1993), pp. 304–313.
[43] J. Lubin, "The use of psychovisual data and models in the analysis of display system performance," in *Visual Factors in Electronic Image Communications* A. Watson, Ed. Cambridge, MA: MIT Press, 1993.
[44] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proc. IEEE*, vol. 81, pp. 1385–1422, Oct. 1993.
[45] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Englewood Cliffs, NJ: Prentice-Hall, 1984.
[46] S. Quackenbush, T. Barnwell III, and M. Clements, *Objective Measures of Speech Quality* (Prentice-Hall Signal Processing Series). Englewood Cliffs, NJ: Prentice-Hall, 1988.
[47] R. J. Safranek and J. D. Johnston, "A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression," in *Proc. ICASSP* (Glasgow, UK, 1989), pp. 1945–1948.
[48] R. J. Safranek, J. D. Johnston, and R. E. Rosenholtz, "A perceptually tuned sub-band image coder," in *Proc. SPIE—Int. Soc. Opt. Eng.* (IEEE, Santa Clara, CA, Feb. 1990), pp. 284–293.
[49] "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, pp. 227–246, Sept. 1969.
[50] W. Voiers, "Diagnostic acceptability measure for speech communication systems," in *Proc. ICASSP*, pp. 204–207, 1977.
[51] CCIR, "Recommendation 500, method for the subjective assessment of the quality of television pictures," in *XIIIth Plenary Assembly* (Geneva, Switzerland), vol. XI, 1974.
[52] R. Fish and T. Judd, "A subjective visual quality comparison of NTSC, VHS, and compressed DS1-compatible video," in *Proc. SID* (Society for Information Display), vol. 32, no. 2, pp. 157–163, 1991.
[53] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Sel. Areas Commun.*, vol. 10, pp. 819–829, June 1992.
[54] C. E. Metz, "Basic principles of ROC analysis," in *Seminars in Nuclear Medicine*, vol. VIII, pp. 282–298, Oct. 1978.
[55] J. A. Swets, "ROC analysis applied to the evaluation of medical imaging techniques," *Investigative Radiology*, vol. 14, pp. 109–121, Mar.–Apr. 1979.
[56] R. Fiete, H. Barrett, E. Cargill, K. J. Myers, and W. E. Smith, "Psychophysical validation of the Hotelling trace criterion as a metric for system performance," in *Proc. SPIE Medical Imaging*, vol. 767, pp. 298–305, 1987.
[57] R. Fiete, H. H. Barrett, W. E. Smith, and K. J. Meyers, "The Hotelling trace criterion and its correlation with human observer performance," *J. Opt. Soc. Amer. A*, vol. 4, pp. 945–953, 1987.
[58] P. Wilhelm, D. R. Haynor, Y. Kim, and E. A. Riskin, "Lossy image compression for digital medical imaging system," *Opt. Eng.*, vol. 30, pp. 1479–1485, Oct. 1991.
[59] H. Lee, A. H. Rowberg, M. S. Frank, H. S. Choi, and Y. Kim, "Subjective evaluation of compressed image quality," in *Proc. Medical Imaging VI: Image Capture, Formatting, and Display*, vol. 1653, pp. 241–251, SPIE, Feb. 1992.

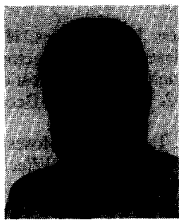
[60] H. H. Barrett, T. Gooley, K. Girodias, J. Rolland, T. White, and J. Yao, "Linear discriminants and image quality," in *Proc. 1991 Int. Conf. on Information Processing in Medical Imaging (IPMI 91)* (Wye, United Kingdom). New York: Springer-Verlag, July 1991, pp. 458–473.
[61] J. Bramble, L. Cook, M. Murphey, N. Martin, W. Anderson, and K. Hensley, "Image data compression in magnification hand radiographs," *Radiology*, vol. 170, pp. 133–136, 1989.
[62] J. Chen, M. Flynn, B. Gross, and D. Spizarny, "Observer detection of image degradation caused by irreversible data compression processes," in *Proc. of Medical Imaging V: Image Capture, Formatting, and Display*, vol. 1444, pp. 256–264, SPIE, 1991.
[63] D. Chakraborty and L. Winter, "Free-response methodology: alternate analysis and a new observer-performance experiment," *Radiology*, vol. 174, no. 3, pp. 873–881, 1990.
[64] W. K. Pratt, *Image Transmission Techniques*. New York: Academic Press, 1979.
[65] A. K. Jain, "Image data compression: A review," *Proc. IEEE*, vol. 69, pp. 349–389, Mar. 1981.
[66] B. G. Haskell and H.-M. Hang, "Comparison of discrete cosine transform and vector quantization of medical imagery," in *Proc. SPIE PACS IV* (Newport Beach, CA, Feb. 1986).
[67] B. G. Haskell, "International standards activities in image data compression," in *Proc. NASA/JPL Data Compression Workshop* (Snowbird, UT, May 1988).
[68] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Norwell, MA: Kluwer, 1992.
[69] G. Wallace, "The JPEG still picture compression standard," *Commun. ACM*, vol. 34, pp. 30–44, Apr. 1991.
[70] R. M. Gray, "Vector quantization," *IEEE ASSP Mag.*, vol. 1, pp. 4–29, Apr. 1984.
[71] A. Gersho and V. Cuperman, "Vector quantization: A pattern-matching technique for speech coding," *IEEE Commun. Mag.*, vol. 21, pp. 15–21, Dec. 1983.
[72] J. Makhoul, S. Roucos, and H. Gish, "Vector quantization in speech coding," *Proc. IEEE*, vol. 73, pp. 1551–1587, Nov. 1985.
[73] N. M. Nasrabadi and R. A. King, "Image coding using vector quantization: A review," *IEEE Trans. Commun.*, vol. 36, pp. 957–971, Aug. 1988.
[74] H. Abut, Ed., *Vector Quantization* (IEEE Reprint Collection). New York: IEEE Press, May 1990.
[75] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.
[76] M. Powell, *Approximation Theory and Methods*. Cambridge, England: Cambridge Univ. Press, 1981.
[77] M. Weinstein and H. Fineberg, *Clinical Decision Analysis*. Philadelphia, PA: Saunders, 1980.
[78] B. Efron, *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia, PA: Soc. Industrial Appl. Math., 1982.
[79] ——, "Better bootstrap confidence intervals and bootstrap approximations," *J. Amer. Stat. Assoc.*, vol. 82, pp. 171–185, 1987.
[80] P. Armitage, *Statistical Methods in Medical Research*. Oxford, England: Blackwell, 1971.
[81] E. Lehmann, *Testing Statistical Hypotheses*, 2nd ed. New York: Wiley, 1986.
[82] C. Tseng, S. Perlmutter, P. Cosman, K. Li, C. Bergin, R. Olshen, and R. Gray, "Effect of tree-structured vector quantization on the accuracy of vessel measurements in MR chest scans," submitted to *Trans. Med. Imag.*.
[83] T. Hastie and D. Pregibon, "Generalized linear models," in *Statistical Models in S*, J. Chambers and T. Hastie, Eds. (Cole Advanced Books and Software). Pacific Grove, CA: Wadsworth, 1992, ch. 6, pp. 196–246.
[84] P. Cosman, C. Tseng, R. Olshen, K. Li, and R. M. Gray, "Predicting perceptual quality from SNR in compressed medical images," in *SID'93 Dig. Tech. Papers* (Society for Information Display, Seattle, WA, May 1993), pp. 309–312.
[85] W. Voiers, "Diagnostic evaluation of speech intelligibility," in *Speech Intelligibility and Speaker Recognition*, M. Hawley, Ed. Stroudsberg, PA: Dowden Hutchinson Ross, 1977.



Pamela C. Cosman (Member, IEEE) received the B.S. degree in electrical engineering (with Honor) from the California Institute of Technology, Pasadena, in 1987 and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Staford, CA, in 1989 and 1993, respectively.

She is currently a postdoctoral fellow and lecturer at Stanford University. Her awards include a National Science Foundation Graduate Fellowship ('87–'90), a Society of Women Engineers/General Electric Scholarship ('84–'87), an AT&T Information Systems Scholarship ('85), and a National Science Foundation Postdoctoral Fellowship ('94). Her research interests are in image processing.

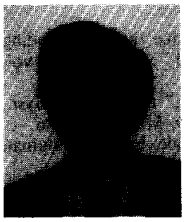
Dr. Cosman is a member of Tau Beta Pi and Sigma Xi.



Robert M. Gray (Fellow, IEEE) received the B.S. and M.S. degrees from the Massachusetts Institute of Technology, Cambridge, in 1966 and the Ph.D. degree from the University of Southern California, Los Angeles, in 1969, all in electrical engineering.

Since 1969 he has been at Stanford University, where he is currently a Professor and Acting Vice Chair of the Department of Electrical Engineering. His current research interests include image compression, enhancement, and classification. He is the author or coauthor of five books, including *Entropy and Information Theory*, Springer-Verlag (1990), and, with A. Gersho, *Vector Quantization and Signal Compression*, Kluwer Academic Press (1992). He was co-recipient or recipient of the 1976 IEEE Information Theory Group Paper Award, the 1983 IEEE ASSP Senior Award, the IEEE Centennial Medal, and the IEEE Signal Processing Society 1993 Society Award. He has held fellowships from the Japan Society for the Promotion of Science at the University of Osaka, the Guggenheim Foundation at the University of Paris XI, and NATO/Consiglio Nazionale delle Ricerche at the University of Naples. He is a former Associate Editor and Editor-in-Chief of the IEEE TRANSACTIONS ON INFORMATION THEORY and was co-chair of the 1993 IEEE International Symposium on Information Theory.

Dr. Gray is a Fellow of the IMS and a member of Sigma Xi, Eta Kappa Nu, AAAS, AMS, and the Société des Ingénieurs et Scientifiques de France.



Richard A. Olshen received the A.B. degree in 1963, from the University of California, Berkeley, and the M.S. and Ph.D. degrees in 1965 and 1966, from Yale University, New Haven, CT, all in statistics.

He is Professor of Biostatistics in the Department of Health Research and Policy and, by Courtesy, Professor of Statistics at Stanford. He has held faculty positions at the University of California, San Diego, and at the University of Michigan, Ann Arbor. His research is in statistics and mathematics and their applications to medicine and biology. Much of it has concerned tree-structured methods for classification, regression, survival analysis, and clustering. With L. Breiman, J. H. Friedman, and C. J. Stone he coauthored *Classification and Regression Trees*, which contains algorithms, examples, and theoretical background to their CART^{TM} software. His research has also concerned modeling and sample reuse methods for longitudinal data, including those that arise in studies of the development of mature walking; cholesterol; renal physiology; and, recently, molecular genetics. He has been the recipient of a Guggenheim Fellowship and of a Research Scholar in Cancer award from the American Cancer Society.

Dr. Olshen is a Fellow of the Institute of Mathematical Statistics and of the American Association for the Advancement of Science, and has been a plenary speaker at the IEEE International Symposium on Information Theory.