Quality evaluation for compressed medical images: Fundamentals

Pamela Cosman, Robert Gray, Richard Olshen

1 Introduction

As radiology becomes increasingly digital and picture archive and communication systems (PACS) move from research to development and practice, the quantity of digital information generated threatens to overwhelm available communication and storage media. While these media will improve with technology, the need for efficiency will remain for inherently narrow band links such as satellites, wireless, and existing media such as twisted pair which will remain useful for many years. The expected growth in digital data as xrays become digital will balance much of the expected gain in transmission bandwidth and local storage. Typical high resolution digital mammograms require tens of megabytes for each image. The transfer of a collection of studies for research or education across the Internet can take hours.

Image compression can provide increases in transmission speed and in the quantity of images stored on a given disk. Lossless compression, in which an original image is perfectly recoverable from the compressed format, can be used without controversy. However, its gains are limited, ranging from a typical 2:1 compression (i.e., producing computer files of half the original size) to an optimistic 4:1. Serious compression of 10:1 or more must be lossy in that the original image cannot be recovered from the compressed format; one can only recover an approximation.

How does one evaluate the approximation? Clearly the usefulness of image compression depends critically on the quality of the processed images. Quality is an attribute with many possible definitions and interpretations, depending on the use to which the images will be put. Sometimes it is felt that for a compressed image to be considered "high quality," it should be visually indistinguishable from the original. This is sometimes referred to as "transparent quality" or "perceptually lossless" since the use of compression on the image is transparent to the viewer. Upon first consideration, this concept of visually indistinguishable would seem to be a simple definition of a perceptually important quality threshold that everyone could agree upon. This is not the case. Two images that are visually indistinguishable when seen by a certain person may be distinguishable when seen by someone else. For example, a pair of medical images viewed by lay people may appear identical, but a radiologist trained in viewing those images might detect differences. Similarly, a pair of images seen by the same person under certain viewing conditions may appear identical, but when seen under different conditions of ambient lighting, viewing distance, or display characteristics might be easily seen to differ. A third issue is that a compressed image might differ from the original without *necessarily* being *worse*. To hold up transparent quality as the ultimate quality goal is to ignore the possibility that certain types of computer processing, including compression, can in some cases make images more pleasing perceptually than the originals. A fourth issue is that images have different applications, and the term "high quality" may be used to denote usefulness for a specific application rather than to indicate that an image is perceptually pleasing or visually identical to an original. For all of these reasons, the measurement of image quality is a diffiult task, and only a few researchers consider quality as a binary quantity that either meets the transparent quality standard or does not. No single approach to quality measurement has gained universal acceptance. The various approaches can be categorized into the following three groups:

- computable objective distortion measures such as squared error or signal-to-noise ratio,
- 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.

Within this latter category of evaluation methods, the methodology of receiver operating characteristic (ROC) curves has dominated historically, but a variety of other approaches have been used in which radiologists may be

called upon to perform various interpretive tasks. Radiologists detect and localize the disease, make measurements of various structures, and make recommendations for patient management. The utility of a medical image can be evaluated in terms of how well it contributes to these functions.

In this chapter, we begin with a brief introduction to image compression, and to the 3 different sets of medical images which form the basis of our studies. We discuss signal-to-noise ratios and subjective quality ratings in the context of these data sets, as well as ROC methodology. In the next chapter, we present the clinical studies including detection, measurement, and management tasks, and in the following chapter, we discuss a number of statistical issues which arise in this sort of clinical experiment.

2 Image Compression

Image compression seeks to reduce the number of bits involved in representing an image. Most compression algorithms in practice are digital, beginning with an information source that is discrete in time and amplitude. If an image is initially analog in space and amplitude, one must first render it discrete in both space and amplitude before compression. Discretization in space is generally called sampling– this consists of examining the intensity of the analog image on a regular grid of points called *picture elements* or *pixels*. Discretization in amplitude is simply scalar quantization: a mapping from a continuous range of possible values into a finite set of approximating values. The term analog-to-digital (A/D) conversion is often used to mean both sampling and quantization– that is, the conversion of a signal that is analog in both space and amplitude to a signal that is discrete in both space and amplitude. Such a conversion is by itself an example of lossy compression.

A general system for digital image compression is depicted in Figure 1. It consists of one or more of the following operations, which may be combined with each other or with additional signal processing:

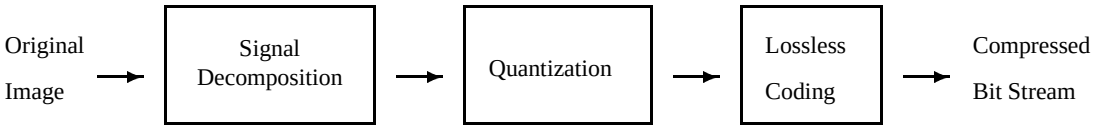

Figure 1: Image compression system

- *Signal decomposition*: The image is decomposed into several images for separate processing. The most popular signal decompositions for image processing are linear transformations of the Fourier family, especially the discrete cosine transform (DCT), and filtering with a subband or wavelet filter bank. Both methods can be viewed as transforms of the original images into coefficients with respect to some set of basis functions. There are many motivations behind such decompositions. Transforms tend to "mash up" the data so that the effects of quantization error are spread out and ultimately invisible. Good transforms concentrate the data in the lower order transform coefficients so that the higher order coefficients can be coded with few or no bits. Good transforms tend to decorrelate the data with the intention of rendering simple scalar quantization more efficient. The eye and ear are generally considered to operate in the transform domain, so that it is natural to focus on coding in that domain where psychophysical effects such as masking can be easily incorporated into frequency dependent measures of distortion. Lastly, the transformed data may provide a useful data structure, as do the multiresolution representations of wavelet analysis.

- *Quantization*: High rate digital pixel intensities are converted into relatively small numbers 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 discarding some of the components of the signal decomposition step. Our emphasis is on quantizer design.

- *Lossless compression*: Further compression is achieved by an invertible (lossless, entropy) code such as run-length, Huffman, Lempel-Ziv, or arithmetic code.

Many approaches to systems for image compression have been proposed in the literature and incorporated into standards and products, both software and hardware. We note that the methods discussed in this chapter for evaluating

the quality and utility of lossy compressed medical images, do not depend on the compression algorithm at all. The reader is referred to the literature on the subject for more information on image compression [50, 23].

3 The three data sets

In this chapter and the following two chapters, results are presented for three data sets: computerized tomography (CT), magnetic resonance (MR), and mammographic images. As will be seen later, these three studies provide examples of the detection, localization, measurement, and management aspects of a radiologist's interpretative functions.

3.1 CT study

The CT study involved two different sets of chest images. In one, the diagnostic task was the detection of abnormally enlarged lymph nodes, and in the other, the task was to detect lung nodules. Thirty patient studies were used for each task. The CT images were compressed using pruned predictive vector quantization [23] applied to 2×2 pixel blocks [15]. This method involves no transform of the data. Vector quantizers are often designed for a training set of representative images that can provide information about the statistics such as the spatial correlations that are typically found in those images. In such a situation, the compression algorithm will perform best for images that are similar to those used in the training set. For this study twenty CT images of the mediastinum were used in the training set for detecting enlarged lymph nodes, and 20 CT lung images were used in the training set for detecting lung nodules. All 512×512 pixel images were obtained using a GE 9800 scanner (120kV, 140mA, scan time 2 seconds per slice, bore size 38 cm, field-of-view 32–34 cm). Although no formal research was undertaken to determine accurately what constitutes "representative" CT images, two radiologists were consulted concerning the typical range of appearance of adenopathy and nodules that occurs in daily clinical practice. The training and test images were chosen to be approximately representative of this range, and included images of both normal and abnormal chests. The lung nodules ranged in size from 0.4 to 3.0 cm, with almost all nodules between 0.4 and 1.5 cm, and the abnormal lymph nodes were between 0.6 and 3.5 cm. The study also had a lower percentage of normal chest images than would be encountered in daily practice.

For each study (lymph nodes, lung nodules), the original thirty test images were encoded at 6 compression levels: 0.57, 1.18, 1.33, 1.79, 2.19, and 2.63 bits per pixel (bpp). The original test images are considered to be 11-bit data. Figure 2 shows an original 11 bpp CT lung image to which the "windows and levels" contrast adjustment has been applied. Although the scanner was capable of producing 12-bit data, it was found for this data set that the 12th bit was never used. Patient studies represented in the training set were not used as test images, and the results reported on SNR, subjective quality, and diagnostic accuracy are based only on test images.

3.2 MR study

In the MR study, the diagnostic task was to measure the size of blood vessels in MR chest scans, as would be done in evaluating aortic aneurysms. The study had as its goal to quantify the effects of lossy compression on the accuracy of these measurements [47, 46]. As in the CT study, the image compression scheme was predictive pruned tree-structured vector quantization, although in this case it was applied to blocks of 2×4 pixels.

The training data of 20 MR chest scans were chosen to include a wide range of both aneurysms and normal vessel structures. An additional 30 scans were chosen as test images. All images were obtained using a 1.5 T whole body imager (Signa, GE Medical Systems, Milwaukee, WI), a body coil, and an axial cardiac gated T1 weighted spin echo pulse sequence with the following parameters: Cardiac gating with repetition time (TR) of 1 R-R interval, echo time (TE) of 15-20 msec, respiratory compensation, number of repetition (NEX) of 2, 256×192 matrix, slice thickness of 7 mm with a 3 mm interslice gap.

The compression rates for this study were 0.36, 0.55, 0.82, 1.14, and 1.70 bpp on the 30 test images. These bitrates are represented by compression levels 1-5. The original scans at 9.0 bpp are represented by level 6.

Figure 3(a) shows an original 9.0 bpp MR chest scan. Figure 3(b) shows the same image compressed to 1.14 bpp and Figure 3(c) shows the image compressed to 0.36 bpp.

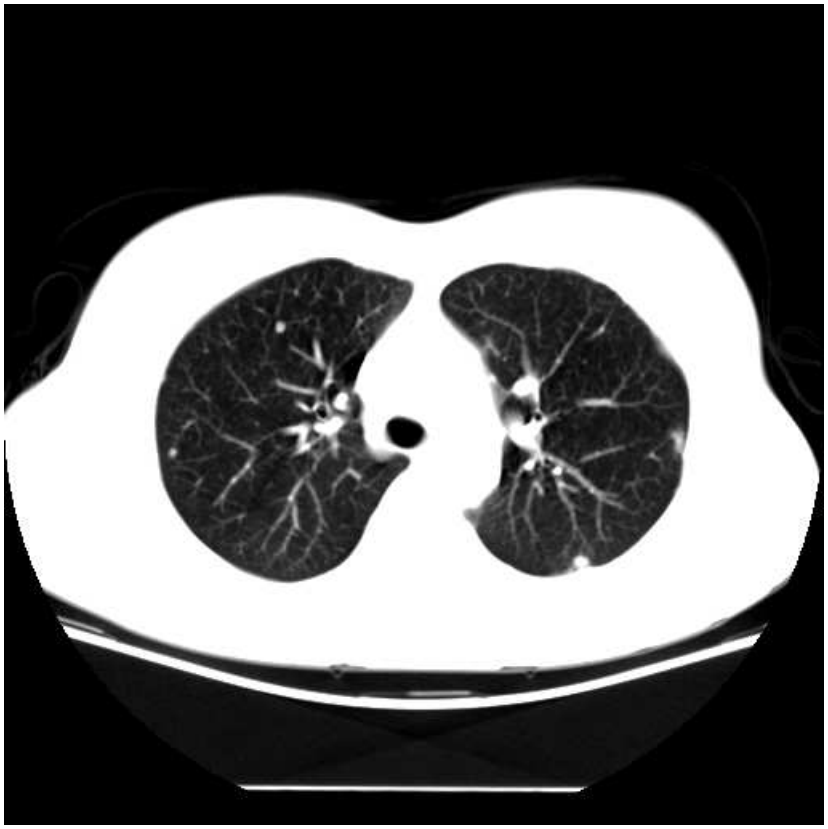

Figure 2: Original 11.0 bpp CT chest scan

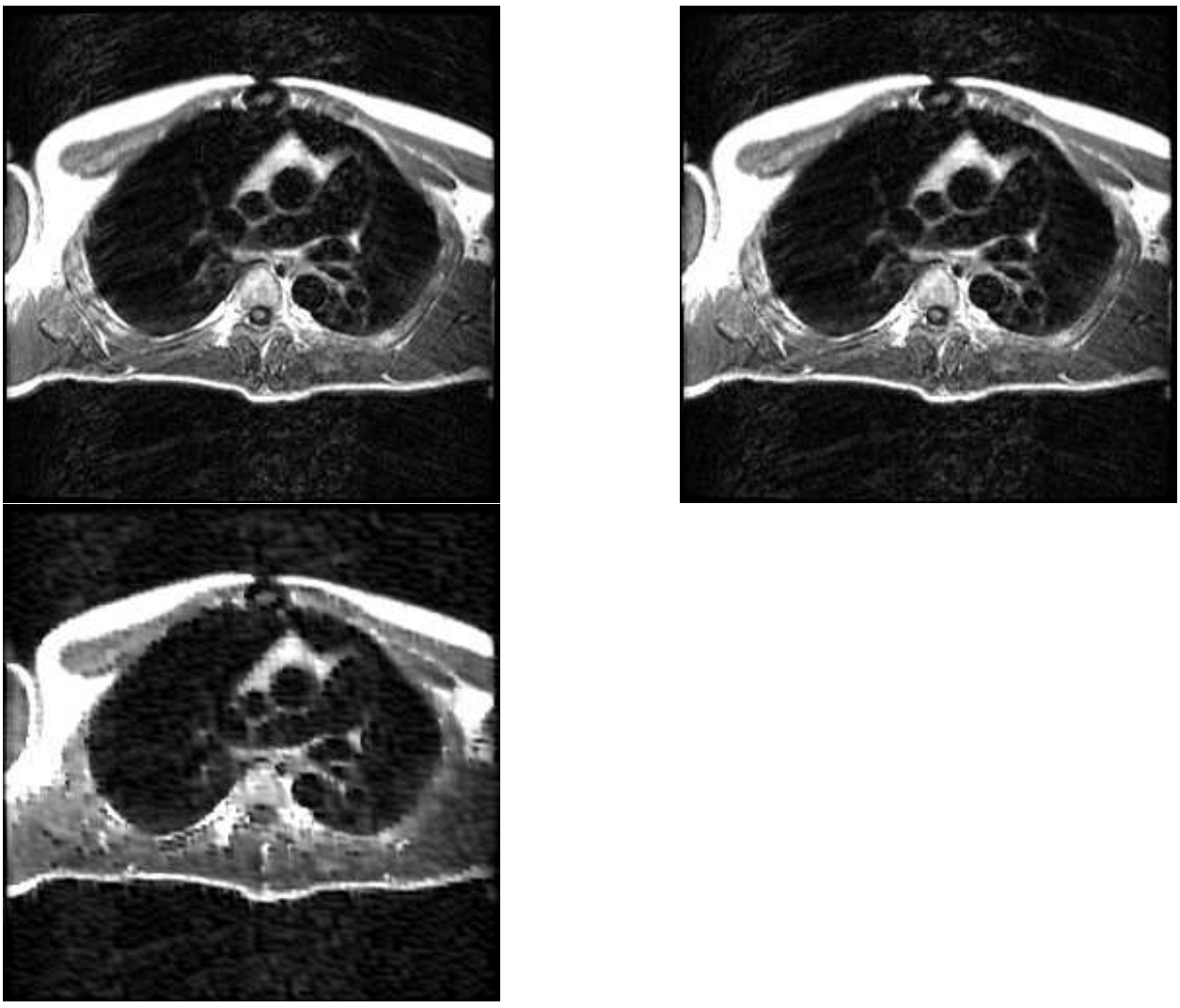

Figure 3: (a) Original 9.0 bpp MR chest scan, (b) MR chest scan compressed to 1.14 bpp, and (c) MR chest scan compressed to 0.36 bpp

3.3 Mammogram study

The mammography study involved a variety of tasks: detection, localization, measurement, and management decisions. This work has been reported upon in [45, 2, 24] as well as in the recent Stanford Ph.D. thesis of Bradley J. Betts [8] which also includes detailed analyses of a much larger trial. The image database was generated in the Department of Radiology of the University of Virginia School of Medicine and is summarized in Table 1. The 57 studies included a variety of normal images and images containing benign and malignant objects. Corroborative biopsy information was available on at least 31 of the test subjects.

6	benign mass
6	benign calcifications
5	malignant mass
6	malignant calcifications
3	malignant combination of mass & calcifications
3	benign combination of mass & calcifications
4	breast edema
4	malignant architectural distortion
2	malignant focal asymmetry
3	benign asymmetric density
15	normals

Table 1: Data Test Set: 57 studies, 4 views per study.

The images were compressed using Set Partitioning in Hierarchical Trees (SPIHT) [54], an algorithm in the subband/wavelet/pyramid coding class. These codes typically decompose the image using an octave subband, critically sampled pyramid, or complete wavelet transformation, and then code the resulting transform coefficients in an efficient way. The decomposition is typically produced by an analysis filter bank followed by downsampling.

The most efficient wavelet coding techniques exploit both the spatial and frequency localization of wavelets. The idea is to group coefficients of comparable significance across scales by spatial location in bands oriented in the same direction. The early approach of Lewis and Knowles [31] was extended by Shapiro in his landmark paper on embedded zerotree wavelet coding [57], and the best performing schemes are descendents or variations on this theme. The approach provides codes with excellent rate-distortion tradeoffs, modest complexity to implement, and an embedded bit stream, which makes the codes useful for applications where scalability or progressive coding are important. Scalability implies there is a "successive approximation" property to the bit stream. This feature is particularly attractive for a number of applications, especially those where one wishes to view an image as soon as bits begin to arrive, and where the image improves as further bits accumulate. With scalable coding, a single encoder can provide a variety of rates to customers with different capabilities. Images can be reconstructed to increasing quality as additional bits arrive.

After experimenting with a variety of algorithms, we chose Said and Pearlman's variation [54] of Shapiro's EZW algorithm because of its good performance and the availability of working software for 12 bpp originals. We used the default filters (9-7 biorthogonal filter) in the software compression package of Said and Pearlman [54]. The system incorporates the adaptive arithmetic coding algorithm considered in Witten, Neal, and Cleary [66].

For our experiment additional compression was achieved by a simple segmentation of the image using a thresholding rule. This segmented the image into a rectangular portion containing the breast — the *region of interest* or *ROI* — and a background portion containing the dark area and any alphanumeric data. The background/label portion of the image was coded using the same algorithm, but at only 0.07 bpp, resulting in higher distortion there. We report here SNRs and bit rates for both the full image and for the ROI.

The image test set was compressed in this manner to three bit rates: 1.75 bpp, 0.4 bpp, and 0.15 bpp, where the bit rates refer to rates in ROI. The average bit rates for the full image thus depended on the size of the ROI. An example of the Said-Pearlman algorithm with a 12 bpp original and 0.15 bpp reproduction is given in Figure 4.

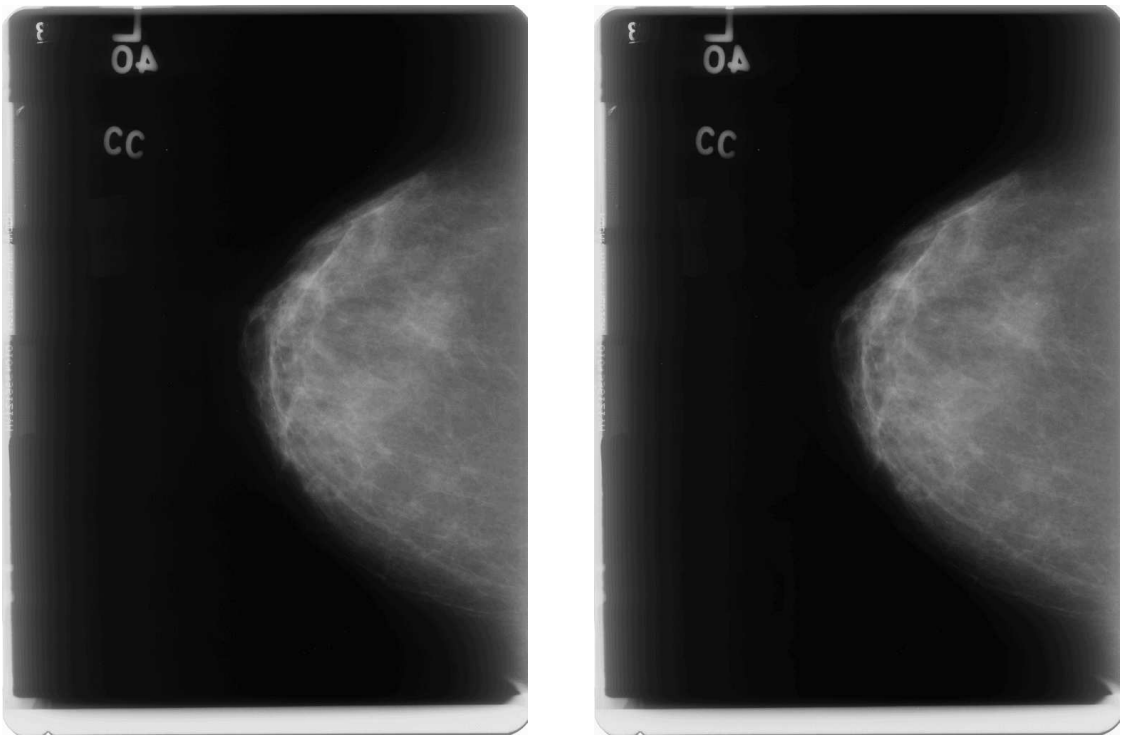

Figure 4: Original Image and Compressed Image at 0.15 bpp in the ROI

4 Average distortion and SNR

By far the most common computable objective measures of image quality are mean squared error (MSE) and signal-to-noise ratio (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) of the system is the average of the distortions for all the vectors input to that system, denoted by $D = E[d(X, \hat{X})]$. 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 dB. A common normalization is the minimum average distortion achievable if no bits are sent, $D_0 = \min_y E[D(X, y)]$. When the ubiquitous squared-error distortion given by $d(X, Y) = ||X - Y||^2 = \sum_{i=0}^{k-1}(X_i - Y_i)^2$ is used, then D_0 is simply the variance of the process, $D_0 = E[||X - E(X)||^2] = \sigma_X^2$. Using this as a normalization factor produces the *signal-to-noise ratio* (SNR)

$$\text{SNR} = 10 \log_{10} \frac{D_0}{D} = 10 \log_{10} \frac{\sigma_X^2}{E[||X - \hat{X}||^2]}. \tag{1}$$

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 $(2^r - 1)^2$, yielding the so-called *peak signal-to-noise ratio* (PSNR).

A key attribute of useful distortion measures is ease of computation, but other properties are also important. Ideally a distortion measure should reflect 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 [10], Stockham [60], and Mannos and Sakrison [36] was aimed at developing computable measures of distortion that emphasize perceptually important attributes of an image by incorporating knowledge of human vision.

Theirs and subsequent work has provided a bewildering variety of candidate measures of image quality or distortion [40, 55, 32, 34, 42, 41, 37, 63, 58, 67, 53, 7, 17, 20, 4, 5, 16, 25, 44, 43, 29, 3, 19, 33, 27]. Similar studies have been carried out for speech compression and other digital speech processing [49]. 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 that incorporate linear weighting. A popular form is weighted quadratic distortion that attempts 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 computing distortion providing a wide family of spectral distortions, which can also incorporate 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 SNRs in a log scale [28, 49], 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 that 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, 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 [51, 52, 60].

The traditional manner for comparing the performance of different lossy compression systems is to plot distortion-rate or SNR vs. bit rate curves. Figure 5(a) shows a scatter plot of the rate-SNR pairs for 24 images in the lung CT study. Only the compressed images can be shown on this plot, as the original images have by definition no noise and therefore infinite SNR. The plot includes a quadratic spline fit with a single knot at 1.5 bpp. Regression splines [48] are simple and flexible models for tracking data that can be fit by least squares. The fitting tends to be "local" in that the fitted average value at a particular bit rate is influenced primarily by observed data at nearby bit rates. The curve has 4 unknown parameters, and can be expressed as

$$y = a_0 + a_1 x + a_2 x^2 + b_2 (max(0, x - 1.5))^2. \tag{2}$$

It is quadratic "by region," and is continuous with a continuous first derivative across the knot, where the functional form of the quadratic changes. Quadratic spline fits provide good indications of the overall distortion-rate performance of the code family on the test data. In this case, the location of the knot was chosen arbitrarily to be near the center of the data set. It would have been possible to allow the data themselves guide the choice of knot location. The SNR results for the CT mediastinal images were very similar to those for the lung task. For the MR study, Figure 5(b) shows SNR versus bit rate for the 30 test images compressed to the 5 bit rates. The knot is at 1.0 bpp.

For the mammograhy study, the SNRs are summarized in Tables 2–3. The overall averages are reported as well as the averages broken out by image type or view (left and right breast, CC and MLO view). This demonstrates the variability among various image types as well as the overall performance. Two sets of SNRs and bit rates are reported: ROI only and full image. For the ROI SNR the rates are identical and correspond to the nominal rate of the code used in the ROI. For the full images the rates vary since the ROI code is used in one portion of the image and a much lower rate code is used in the remaining background and the average depends on the size of the ROI, which varies among the images. A scatter plot of the ROI SNRs is presented in Figure 6.

It should be emphasized that this is the SNR comparing the digital original with the lossy compressed versions.

5 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) [1, 49, 62]. There has been no standardization for rating still images.

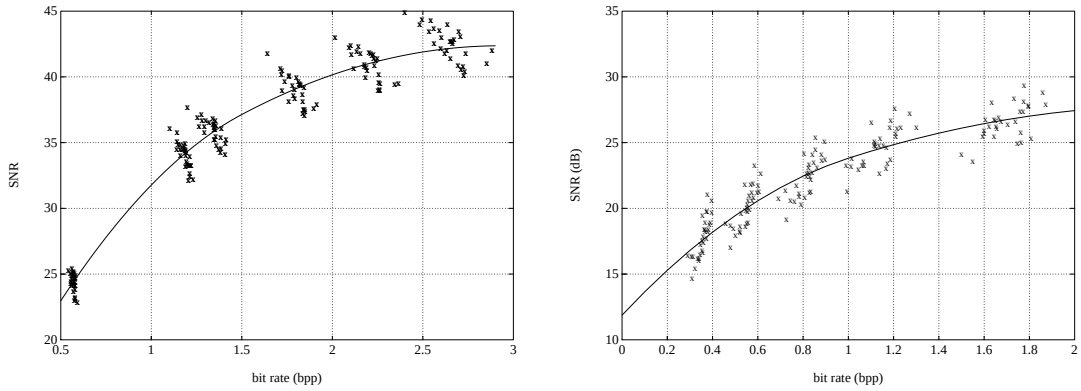

Figure 5: SNR as a function of bit rate for (a) CT lung images, and (b) MR images. The x's indicate data points for all images, judges and compression levels.

View	SNR		
	0.15 bpp ROI	0.4 bpp ROI	1.75 bpp ROI
left CC	45.93 dB	47.55 dB	55.30 dB
right CC	45.93 dB	47.47 dB	55.40 dB
left MLO	46.65 dB	48.49 dB	56.53 dB
right MLO	46.61 dB	48.35 dB	56.46 dB
left side (MLO and CC)	46.29 dB	48.02 dB	55.92 dB
right side (MLO and CC)	46.27 dB	47.91 dB	55.93 dB
Overall	46.28 dB	47.97 dB	55.92 dB

Table 2: Average SNR: ROI, Wavelet Coding

View	SNR, Bit Rate		
	0.15 bpp ROI	0.4 bpp ROI	1.75 bpp ROI
left CC	44.30 dB, 0.11 bpp	45.03 dB, 0.24 bpp	46.44 dB, 0.91 bpp
right CC	44.53 dB, 0.11 bpp	45.21 dB, 0.22 bpp	46.88 dB, 0.85 bpp
left MLO	44.91 dB, 0.11 bpp	45.73 dB, 0.25 bpp	47.28 dB, 1.00 bpp
right MLO	45.22 dB, 0.11 bpp	46.06 dB, 0.25 bpp	47.96 dB, 0.96 bpp
left side (MLO and CC)	44.60 dB, 0.11 bpp	45.38 dB, 0.24 bpp	46.89 dB ,0.96 bpp
right side (MLO and CC)	44.88 dB, 0.11 bpp	45.63 dB, 0.24 bpp	47.41 dB, 0.92 bpp
Overall	44.74 dB, 0.11 bpp	45.51 dB, 0.24 bpp	47.14 dB,0.93 bpp

Table 3: Average SNR: Full Image, Wavelet Coding

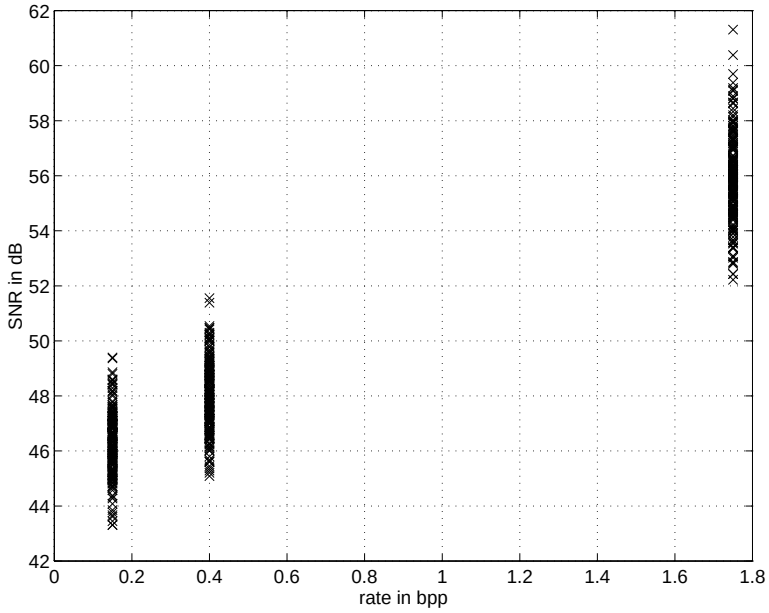

Figure 6: Scatter Plot of ROI SNR: Wavelet Coding

A useful attribute of an objective quality measure such as SNR would be the ability to predict subjective quality. For medical images, it may be more important that a computable objective measure be able to predict diagnostic accuracy rather than subjective quality. 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.

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 (least) to 5 (most) to each image based on its usefulness for the measurement task." The term "usefulness" was defined as "your opinion of whether the edges used for measurements were blurry or distorted, and your confidence concerning the measurement you took." The question was phrased in this way because our concern is whether measurement accuracy is in fact maintained even when the radiologist perceives the image quality as degraded and may have lost some confidence in the utility of the image for the task at hand. It is not clear to us whether radiologists are inculcated during their training to assess quality visually based on the entire image, or whether they rapidly focus on the medically relevant areas of the image. Indeed, one might reasonably expect that radiologists would differ on this point, and a question that addressed overall subjective quality would therefore produce a variety of interpretations from the judges. By focusing the question on the specific measurement and the radiologists' confidence in it regardless of what portion of the image contributed to that confidence level, and then by examining the relationship between actual measurement error and these subjective opinions, we hoped to obtain data relevant to the question of whether radiologists can be asked to trust their diagnoses made on processed images in which they may lack full confidence. No attempt was made to link the 5 possible scores to specific descriptive phrases, as is done with the Mean Opinion Score rating system for speech. However, the radiologists were asked to try to use the whole scale. The CT subjective assessment was performed separately from the diagnostic task by three different radiologists. The phrasing of the question was very similar.

Images compressed to lower bit rates received worse quality scores as was expected. Figure 7 shows subjective score vs. bit rate for the CT mediastinum study. The data are fit with a quadratic spline with a single knot. Figure 8 shows the general trend of mean subjective score versus mean bit rate for the MR study. A spline-like function that is quadratic from 0 to 2.0 bpp and linear from 2.0 to 9.0 bpp was fit to the data. The splines have knots at 0.6 bpp, 1.2 bpp, and 2.0 bpp. Figure 9 shows a spline fit of subjective score plotted against actual bit rate for the compressed levels only for the MR study. The general conclusion from the plots is that the subjective scores at the higher levels

were quite close to the subjective scores on the originals, but at lower levels there was a steep drop-off of scores with decreasing bit rate.

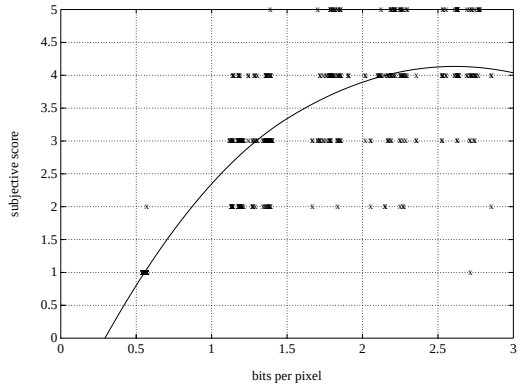

Figure 7: Subjective ratings vs. bit rate for the CT mediastinum study

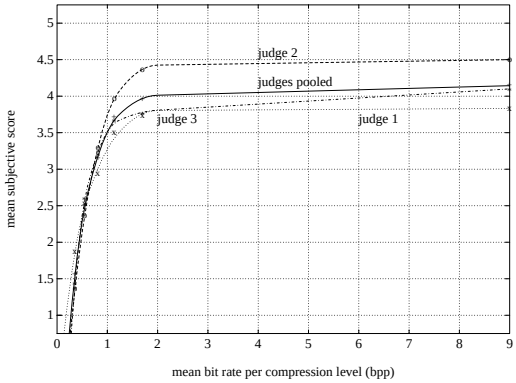

Figure 8: Mean subjective score vs. mean bit rate for the MR study. The dotted, dashed, and dash-dot curves are splines fit to the data points for Judges 1, 2, and 3, respectiviely. The solid curve is a spline fit to the data points for all judges pooled.

These scores can also be analyzed by the Wilcoxon signed rank test. The paired t-test may be slightly less applicable since the subjective scores, which are integers over a very limited range, clearly fail to fit a Gaussian model. We note that scores are assigned to the entire image rather than to any subsection of an image, such as each of the blood vessels in that image. More detailed information would be needed for a more thorough analysis since subjective score in the MR experiment is meant to reflect the quality of the image for vessel measurement, and this may differ for the different blood vessels. The Wilcoxon signed rank test showed that the subjective scores for the MR study at all of the six compression levels differ significantly from the subjective scores of the originals at $p < 0.05$ for a 2-tailed test. The subjective scores at all the compression levels also differ significantly from each other. As will be discussed later, it appears that a radiologist's subjective perception of quality changes more rapidly and drastically with decreasing bit rate than does the actual measurement error.

5.1 Mammography Subjective Ratings

For the mammography study, Table 4 provides the means and standard deviations for the subjective scores for each radiologist separately and for the radiologists pooled. The distribution of these subjective scores is displayed in

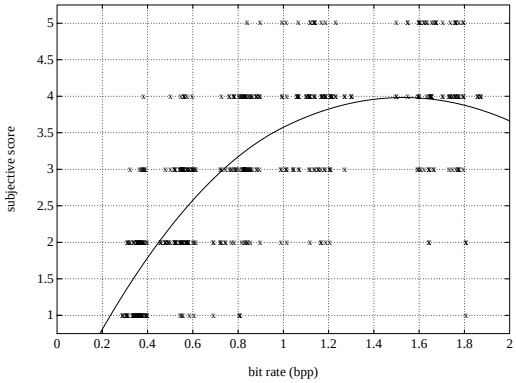

Figure 9: Subjective score vs. bit rate for the MR study. The x's indicate data points for all images, pooled across judges and compression levels.

Figures 10–12. Level 1 refers to the original analog images, Level 2 to the uncompressed digital, Level 3 to those images where the breast section was compressed to 0.15 bpp and the label to .07 bpp, Level 4 to those images where the breast section was compressed to .4 bpp and the label to .07 bpp, and Level 5 to those images where the breast section was compressed to 1.75 bpp and the label to .07 bpp.

Figure 10 displays the frequency for each of the subjective scores obtained with the analog gold standard. Figure 11 displays the frequency for each of the subjective scores obtained with the uncompressed digital images (judges pooled),and Figure 12 displays the frequency for each of the subjective scores obtained with the digital images at Level 3.

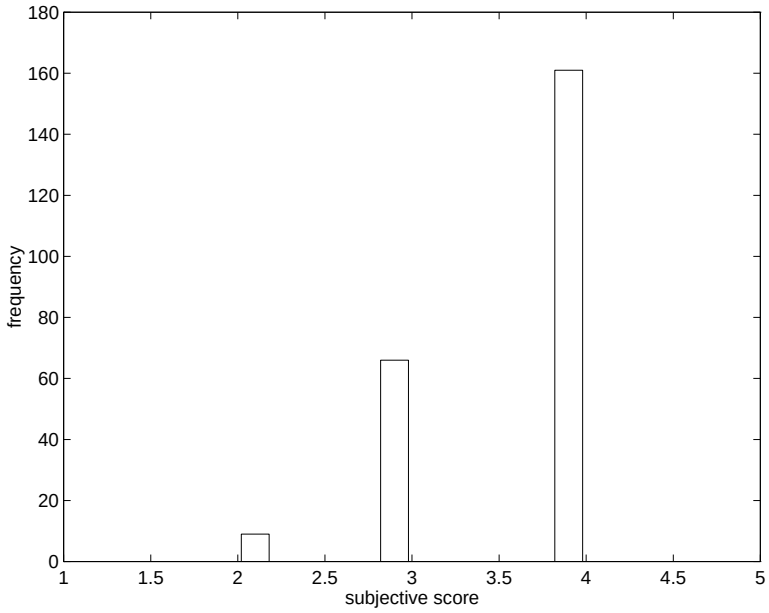

Figure 10: Subjective Scores: Analog Gold Standard

Using the Wilcoxon signed rank test, the results were as follows.

Judge A: All levels were significantly different from each other except the digital to .4 bpp, digital to 1.75 bpp, and .4 to 1.75 bpp.

Judge B: The only differences that were significant were .15 bpp to .4 bpp and .15 bpp to digital.

level	judge	mean	stdev
1	gold standard	3.6441	.5539
1	A	3.90	.97
1	B	4.52	.75
1	C	4.59	.79
2	A	3.91	.41
2	B	3.85	.53
2	C	3.67	.65
3	A	3.82	.39
3	B	4.27	.93
3	C	3.49	.64
4	A	3.91	.39
4	B	3.93	.55
4	C	3.82	.50
5	A	3.92	.42
5	B	3.66	.57
5	C	3.82	.55
judges pooled			
1	pooled	4.33	.89
2	pooled	3.81	.55
3	pooled	3.86	.76
4	pooled	3.88	.49
5	pooled	3.80	.57

Table 4: Subjective Scores

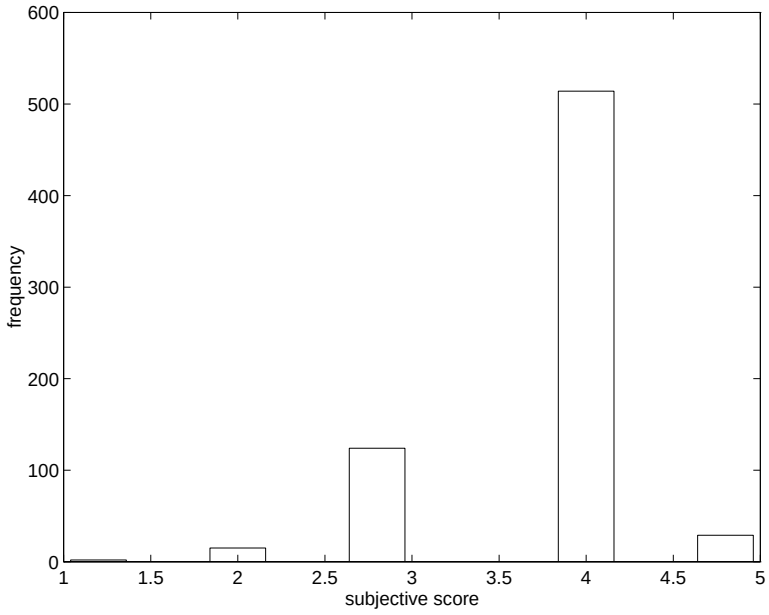

Figure 11: Subjective Scores: Original Digital

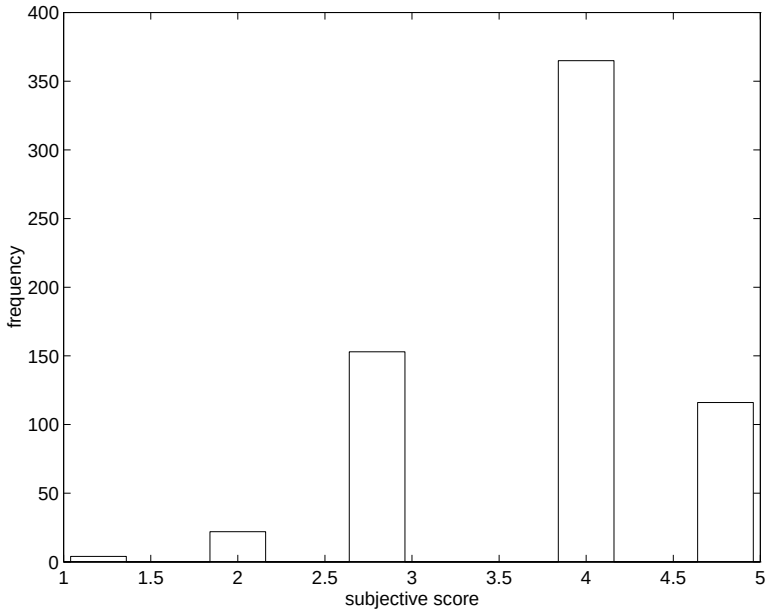

Figure 12: Subjective Scores: Lossy Compressed Digital at 0.15 bpp

Judge C: All differences were significant.

All judges pooled: All differences were significant except digital to .15 bpp, digital to 1.75 bpp, .15 to .4 bpp, and .15 to 1.75 bpp.

Comparing differences from the independent gold standard, for Judge A all were significant except digital uncompressed; for Judge B all were significant; and for Judge C all were significant except 1.75 bpp. When the judges were pooled, all differences were significant.

There were many statistically significant differences in subjective ratings between the analog and the various digital modalities, but some of these may have been a result of the different printing processes used to create the original analog films and the films printed from digital files. The films were clearly different in size and in background intensity. The judges in particular expressed dissatisfaction with the fact that the background in the digitally produced films was not as dark as that of the photographic films, even though this ideally had nothing to do with their diagnostic and management decisions.

6 Diagnostic Accuracy and ROC methodology

Diagnostic "accuracy" is often used to mean the fraction of cases on which a physician is "correct," where correctness is determined by comparing the diagnostic decision to some definition of "truth." There are many different ways that "truth" can be determined, and this issue is discussed in Section 7. Apart from this issue, this simple definition of accuracy is flawed in two ways. First, it is strongly affected by disease prevalence. For a disease that appears in less than one percent of the population, a screening test could trivially be more than 99% accurate simply by ignoring all evidence and declaring the disease to be absent. Second, the notion of "correctness" does not distinguish between the two major types of errors, calling positive a case that is actually negative, and calling negative a case that is actually positive. The relative costs of these two types of errors are generally not equal. These can be differentiated by measuring diagnostic performance using a pair of statistics reflecting the relative frequencies of the two error types.

Toward this end suppose for the moment that there exists a "gold standard" defining the "truth" of existence and locations of all lesions in a set of images. With each lesion identified in the gold standard, a radiologist either gets it correct (*true positive* or TP) or misses it (*false negative* or FN). For each lesion identified by the radiologist, either it agrees with the gold standard (TP as above) or it does not (*false positive* or FP).

The sensitivity or true positive rate (or true positive fraction (TPF)) is the probability p_{TP} that a lesion is said to be there given that it is there. This can be estimated by relative frequency

$$\text{Sensitivity} = \frac{\#\text{ TP}}{\#\text{ TP} + \#\text{ FN}} \tag{3}$$

The complement of sensitivity is the false negative rate (or fraction) $p_{FN} = 1 - p_{TP}$, the probability that a lesion is said to not be there given that it is there.

In an apparently similar vein, the false positive rate p_{FP} (or false positive fraction (FPF)) is the probability that a lesion is said to be there given that it is not there and the true negative rate p_{TN} or *specificity* is its complement. Here, however, it is not possible to define a meaningful relative frequency estimate of these probablities except when the detection problem is binary, that is, each image can have only only one lesion of a single type or no lesions at all. In this case, exactly one lesion is not there if and only if 0 lesions are present, and one can define a true negative TN as an image that is not a true positive. Hence if there are N images, the relative frequency becomes

$$\text{Specificity} = \frac{\#\text{ TN}}{\text{N}-\#\text{TP}} \tag{4}$$

As discussed later, in the nonbinary case, however, specificity cannot be defined in a meaningful fashion on an image by image basis.

In the binary case, specificity shares importance with sensitivity because perfect sensitivity alone does not preclude numerous false alarms, while specificity near 1 ensures that missing no tumors does not come at the expense of calling false ones.

An alternative statistic that is well defined in the nonbinary case and also penalizes false alarms is the *predictive value positive* (PVP), also known as positive predicted value (PPV) [64]. This is the probability that a lesion is there given that it is said to be there.

$$\text{PVP} = \frac{\#\text{ of abnormalities correctly marked}}{\text{total }\#\text{ of abnormalities marked}}. \tag{5}$$

PVP is easily estimated by relative frequencies as

$$\text{PVP} = \frac{\#\text{ TP}}{\#\text{ TP} + \#\text{ FP}}. \tag{6}$$

Sensitivity, PVP, and, when it makes sense, specificity, can be estimated from clinical trial data and provide indication of quality of detection. The next issues are

1. How does one design and conduct clinical experiments to estimate these statistics?
2. How are these statistics used in order to make judgments about diagnostic accuracy?

Together, the responses to these questions form a *protocol* for evaluating diagnostic accuracy and drawing conclusions on the relative merits of competing image processing techniques. Before describing the dominant methodology used, it is useful to formulate several attributes that a protocol might reasonably be expected to have.

- The protocol should simulate ordinary clinical practice as closely as possible. Participating radiologists should perform in a manner that mimics their ordinary practice. The trials should require little or no special training of their clinical participants.
- The clinical trials should include examples of images containing the full range of possible anomalies, all but extremely rare conditions.
- The findings should be reportable using the American College of Radiology (ACR) Standardized Lexicon.
- Statistical analyses of the trial outcomes should be based on assumptions as to the outcomes and sources of error that are faithful to the clinical scenario and tasks.

1	definitely or almost definitely negative
2	probably negative
3	possibly negative
4	probably positive
5	definitely or almost definitely positive

Table 5: Subjective confidence ratings used in ROC analysis

- The number of patients should be sufficient to ensure satisfactory size and power for the principal statistical tests of interest.
- "Gold standards" for evaluation of equivalence or superiority of algorithms must be clearly defined and consistent with experimental hypotheses.
- Careful experimental design should eliminate or minimize any sources of bias in the data that are due to differences between the experimental situation and ordinary clinical practice, e.g., learning effects that might accrue if a similar image is seen using separate imaging modalities.

Receiver operating characteristic (ROC) analysis is the dominant technique for evaluating the suitability of radiologic techniques for real applications [39, 61, 38, 26]. ROC analysis has its origins in signal detection theory. A filtered version of a 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 (a false dismissal) goes down, but the probability of erroneously declaring a signal there when it is not (a false alarm) goes up. Suppose one has a large database of waveforms, some of which actually contain a signal, and some of which do not. Suppose further that for each waveform, the "truth" is known of whether a signal is present or not. One can set a value of the threshold and examine whether the test declares a signal present or not for each waveform. Each value of the threshold will give rise to a pair (TPF,FPF), and these points can be plotted for many different values of the threshold. The ROC curve is a smooth curve fitted through these points. The ROC curve always passes through the point (1,1) because if the threshold is taken to be lower than the lowest value of any waveform, then all samples will be above the threshold, and the signal will be declared present for all waveforms. In that case, the true positive fraction is one. The false positive fraction is also equal to one, since there are no true negative decisions. Similar reasoning shows that the ROC curve must also always pass through the point (0,0), because the threshold can be set very large, and all cases will be declared negative. 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. In general, larger area under the ROC curve is better.

ROC analysis has a natural application to some problems in medical diagnosis. For example, in a blood serum assay of carbohydrate antigens (e.g., CA 125 or CA 19-9) to detect the presence of certain types of cancer, a single number results from the diagnostic test. The distribution of result values in actually positive and actually negative patients overlap. So no single threshold or decision criterion can be found which separates the populations cleanly. If the distributions did not overlap, then such a threshold would exist, and the test would be perfect. In the usual case of overlapping distributions, a threshold must be chosen; and each possible choice of threshold will yield different frequencies of the two types of errors. By varying the threshold and calculating the false alarm rate and false dismissal rate for each value of the threshold, an ROC curve is obtained.

Transferring this type of analysis to radiological applications requires the creation of some form of threshold whose variation allows a similar tradeoff. For studies of the diagnostic accuracy of processed images, this is accomplished by asking radiologists to provide a subjective confidence rating of their diagnoses (typically on a scale of 1–5) [39, 61]. An example of such ratings is shown in Table 5.

First, only those responses in the category of highest certainty of a positive case are considered positive. This yields a pair (TPF, FPF) that can be plotted in ROC space, and corresponds to a stringent threshold for detection. Next, those cases in either of the highest two categories of certainty of a positive decision are counted positive. Another (TPF, FPF) point is obtained, and so forth. The last non-trivial point is obtained by scoring any case as positive if it corresponds to any of the highest four categories of certainty for being positive. This corresponds to a very lax threshold for detection of disease. There are also two trivial (TPF, FPF) points which can be obtained, as discussed above: all cases can be declared negative (TPF=0, FPF=0) or all cases can be declared positive (TPF=1, FPF=1).

This type of analysis has been used extensively to examine the effects of computer processing on the diagnostic utility of medical images. Types of processing that have been evaluated include compression [22, 21, 65, 30, 6, 9, 35, 56, 14], and enhancement (unsharp masking, histogram equalization, and noise reduction).

Although by far the dominant technique for quantifying diagnostic accuracy in radiology, ROC analysis possesses several shortcomings for this application. In particular, it violates several of the stated goals for a clinical protocol. By and large, the necessity for the radiologists to choose 1 of 5 specific values to indicate confidence departs from ordinary clinical practice. Although radiologists are generally cognizant of differing levels of confidence in their findings, this uncertainty is often represented in a variety of qualitative ways, rather than with a numerical ranking. Further, as image data are nonGaussian, methods that rely on Gaussian assumptions are suspect. Modern computer-intensive statistical sample reuse techniques can help get around the failures of Gaussian assumptions. Classical ROC analysis is not location specific. The case in which an observer misses the lesion that is present in an image but mistakenly identifies some noise feature as a lesion in that image would be scored as a true-positive event. Most importantly, many clinical detection tasks are non-binary, in which case sensitivity can be suitably redefined, but specificity cannot. That is, sensitivity as defined in Equation 3 yields a fractional number for the whole data set. But for any one image, sensitivity takes on only the values 0 and 1. The sensitivity for the whole data set is then the average value of these binary-valued sensitivities defined for individual images. When the detection task for each image becomes non-binary, it is possible to redefine sensitivity for an individual image:

$$\text{Sensitivity} = \frac{\text{\# of true positive decisions } \textit{within 1 image}}{\text{\# of actually positive items } \textit{in that 1 image}} \tag{7}$$

or, changing the language slightly,

$$\text{Sensitivity} = \frac{\text{\# of abnormalities correctly found}}{\text{\# of abnormalities actually there}}. \tag{8}$$

In this case, the sensitivity for each individual image becomes a fractional number between 0 and 1, and the sensitivity for the entire data set is still the average of these sensitivities defined for individual images. A similar attempt to redefine the specificity leads to

$$\text{Specificity} = \frac{\text{\# of abnormalities correctly said not to be there}}{\text{\# of abnormalities actually not there}}. \tag{9}$$

This 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 definition is fine for a truly binary diagnostic task such as detection of a pneumothorax, for if the image is normal then exactly one abnormality is absent. Early studies were able to use ROC analysis by focusing on detection tasks that were either truly binary or that could be rendered binary. For example, a non-binary detection task such as "locating any and all abnormalities that are present" can be rendered binary simply by rephrasing the task as one of "declaring whether or not disease is present." Otherwise, such a non-binary task is not amenable to traditional ROC analysis techniques. Extensions to ROC to permit consideration of multiple abnormalities have been developed [18, 11, 12, 13, 59]. For example, the free-response receiver operating characteristic (FROC) observer performance experiment allows an arbitrary number of abnormalities per image, and the observer indicates their perceived locations and a confidence rating for each one. While FROC resolves the binary task limitations and location insensitivity of traditional ROC, FROC does retain the constrained 5-point integer rating system for observer confidence, and makes certain normality assumptions about the resultant 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 non-binary and nonGaussian data, and extends naturally to measurement data.

The recent Stanford Ph.D. thesis by Bradley J. Betts, mentioned earlier, includes new technologies for analyses of ROC curves. His focus is on regions of interest of the curve, that is, on the intersection of the area under the curve with rectangles determined by explicit lower bounds on sensitivity and specificity. He has developed sample

reuse techniques for making inferences concerning the areas enclosed and also for constructing rectangular confidence regions for points on the curve.

7 Determination of a Gold Standard

The typical scenario for evaluating diagnostic accuracy of computer processed medical images involves taking some database of original unprocessed images, applying some processing to them, and having the entire set of images judged in some specified fashion by radiologists. Whether the subsequent analyses will be done by ROC or some other means, it is necessary to determine a "gold standard" that can represent the diagnostic truth of each original image, and can serve as a basis of comparison for the diagnoses on all the processed versions of that image. There are many possible choices for the gold standard:

- A *consensus* gold standard is determined by the consensus of the judging radiologists 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, and
- 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.

The consensus method has the advantage of simplicity, but the judging radiologists may not agree on the exact diagnosis, even on the original image. Of course, this may happen among the members of the independent panel as well, but in that case an image can be removed from the trial or additional experts called upon to assist. Either case may entail the introduction of concerns as to generalizability of subsequent results.

In the CT study, in an effort to achieve consensus for those 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 our subsequent studies did not use this method. Although those images eliminated were clearly more controversial and difficult to diagnose than the others, it cannot be said whether the removal of diagnostically controversial images from the study biases the results in favor of compression or against it. Their failure to have a consensus gold standard defined was based only on the uncompressed versions, and it cannot be said *a priori* that the compression algorithm would have a harder time compressing such images. The consensus, when achieved, could be attained either by initial concordance among the readings of the three radiologists, or by subsequent discussion of the readings, during which one or more judge might change his or her decision. 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 diagnostic accuracy relative to that gold standard. Therefore, this gold standard has a slight bias favoring the originals, which is thought to help make the study safely conservative, and not unduly promotional of specific compression techniques.

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. If there is any component of random error in the measurement process, since the personal gold standard defines the diagnoses on the originals to be correct (for that image and that judge), the compressed images cannot possibly perform as well as the originals according to this standard. That there is a substantial component of random error in these studies is suggested by the fact that there were several images during our CT tests on which judges changed their diagnoses back and forth, marking, for example, 1 lesion on the original image as well as on compressed levels E and B, and marking 2 lesions on the in-between compressed levels F, D, and A. With a consensus gold standard, such changes tend to balance out. With a personal gold standard, the original is always right, and the changes count against compression. Because the compressed levels have this severe disadvantage, the personal gold standard is useful primarily for comparing the compressed levels among themselves. Comparisons of the original images with the compressed ones are conservative. The personal gold standard has, however, the advantage that all images can be used in the study. We no longer have to be concerned with the possible bias from the images eliminated due to

failure to achieve consensus. 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 is what many studies use, and would seem to be a good choice. However, it is not without drawbacks. First of all, there is the danger of a systematic bias appearing in the diagnoses of a judge in comparison to the gold standard. For example, a judge who consistently chooses to diagnose tiny equivocal dots as abnormalities when the members of the independent panel choose to ignore such occurrences would have a high false positive rate relative to that independent gold standard. The computer processing may have some actual effect on this judge's diagnoses, but this effect might be swamped in comparison to this baseline high false positive rate. This is an argument for using the personal gold standard as well as the independent gold standard. The other drawback of an independent gold standard is somewhat more subtle, and is discussed later. In the MR study, the independent panel was composed of two senior radiologists who first measured the blood vessels separately and then discussed and remeasured in those cases where there was initial disagreement.

A separate standard would seem to be the best choice, but it is generally not available. With phantom studies, there is of course a "diagnostic truth" that is established and known entirely separately from the diagnostic process. But with actual clinical images, there is often no autopsy or biopsy, as the patient may be alive and not operated upon. There are unlikely to be any images from other imaging modalities that can add to the information available from the modality under test, since there is typically one best way for imaging a given pathology in a given part of the body. And the image data set for the clinical study may be very difficult to gather if one wishes to restrict the image set to those patients for whom there are follow-up procedures or imaging studies which can be used to establish a gold standard. In any case, limiting the images to those patients who have subsequent studies done would introduce obvious bias into the study.

In summary, the consensus method of achieving a gold standard has a major drawback together with the minor advantage of ease of availability. The other three methods for achieving a gold standard all have both significant advantages and disadvantages, and perhaps the best solution is to analyze the data against more than one definition of diagnostic truth.

8 Concluding Remarks

We have surveyed several key components required for evaluating the quality of compressed images: the compression itself; three data sets to be considered in depth in subsequent chapters; quantitative measures of quality involving measures of pixel intensity distortion, observer judged subjective distortion, and diagnostic accuracy; and, lastly, several notions of "gold standard" with respect to which quality can be compared. In the next chapter these ideas provide a context and a collection of tools for a detailed analysis of three specific medical image modalities and tasks.

Acknowledgements

The authors gratefully acknowledge the essential assistance of many colleagues who participated in and contributed to both the performance of the research described here and the writing of the papers and reports on which these chapters are based. In particular we acknowledge and thank C.N. Adams, A. Aiyer, C. Bergin, B.J. Betts, R. Birdwell, B.L. Daniel, H.C. Davidson, L. Fajardo, D. Ikeda, J. Li, K.C.P. Li, L. Moses, K.O. Perlmutter, S.M. Perlmutter, C. Tseng, and M.B. Williams.

References

[1] IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio and Electroacoustics*, pages 227–246, Sep. 1969.

[2] C.N. Adams, A. Aiyer, B.J. Betts, J. Li, P.C. Cosman, S.M. Perlmutter, M. Williams, K.O. Perlmutter, D. Ikeda, L. Fajardo, R. Birdwell, B.L. Daniel, S. Rossiter, R.A. Olshen, and R.M. Gray. Evaluating quality and utility of digital mammograms and lossy compressed digital mammograms. In *Proceedings 3rd Intl. Workshop on Digital Mammography*, Chicago, IL, June, 1996.

[3] A.J. Ahumada, Jr. Computational image-quality metrics: a review. In *SID '93 Digest of Technical Papers*, pages 305–308, Seattle, Wa, May 1993. Society for Information Display.

[4] V.R. 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*, volume 1771, pages 396–405, San Diego, CA, July 1992.

[5] H. Barrett. Evaluation of image quality through linear discriminant models. In *SID '92 Digest of Technical Papers*, volume 23, pages 871–873. Society for Information Display, 1992.

[6] H. H. Barrett, T. Gooley, K. Girodias, J. Rolland, T. White, and J. Yao. Linear discriminants and image quality. In *Proceedings of the 1991 International Conference on Information Processing in Medical Imaging (IPMI '91)*, pages 458–473, Wye, United Kingdom, July 1991. Springer-Verlag.

[7] P. Barten. Evaluation of subjective image quality with the square foot integral method. *JOSA A*, 7:2024–2031, 1990.

[8] B.J. Betts. *A Statistical Analysis of Digital Mammography*. PhD thesis, Stanford University, Department of Electrical Engineering, 1999.

[9] J.M. Bramble, L.T. Cook, M.D. Murphey, N.L. Martin, W.H. Anderson, and K.S. Hensley. Image data compression in magnification hand radiographs. *Radiology*, 170:133–136, 1989.

[10] Z.L. Budrikus. Visual fidelity criteria and modeling. *Proc. IEEE*, 60:771–779, July 1972.

[11] P.C. Bunch, J.F. Hamilton, G.K. Sanderson, and A.H. Simmons. A free-response approach to the measurement and characterization of radiographic observer performance. *J. Appl. Photogr. Engr.*, 4:166–171, 1978.

[12] D.P. Chakraborty. Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data. *Med. Phys.*, 16:561–568, 1989.

[13] D.P. Chakraborty and L.H.L. Winter. Free-response methodology: alternate analysis and a new observer-performance experiment. *Radiology*, 174(3):873–881, 1990.

[14] J. Chen, M.J. Flynn, B. Gross, and D. Spizarny. Observer detection of image degradation caused by irreversible data compression processes. In *Proceedings of Medical Imaging V: Image Capture, Formatting, and Display*, volume 1444, pages 256–264. SPIE, 1991.

[15] P.C. Cosman, C. Tseng, R.M. Gray, R.A. Olshen, L. E. Moses, H. C. Davidson, C.J. Bergin, and E.A. Riskin. Tree-structured vector quantization of CT chest scans: image quality and diagnostic accuracy. *IEEE Trans. Medical Imaging*, 12(4):727–739, Dec. 1993.

[16] S. Daly. Visible differences predictor: an algorithm for the assessment of image fidelity. In *SPIE Proceedings*, volume 1666, pages 2–14, 1992.

[17] M. Duval-Destin. A spatio-temporal complete description of contrast. In *SID '91 Digest of Technical Papers*, volume 22, pages 615–618. Society for Information Display, 1991.

[18] J.P. Egan, G.Z. Greenberg, and A.I. Schulman. Operating characteristics, signal detectability, and the method of free response. *J. Acoust. Soc. Am.*, 33:993–1007, 1961.

[19] A.M. Eskioglu and P.S. Fisher. A survey of quality measures for gray scale image compression. In *Computing in Aerospace 9*, pages 304–313, San Diego, CA, Oct. 1993. AIAA.

[20] J. Farrell, H. Trontelj, C. Rosenberg, and J. Wiseman. Perceptual metrics for monochrome image compression. In *SID '91 Digest of Technical Papers*, volume 22, pages 631–634. Society for Information Display, 1991.

[21] R.D. Fiete, H. H. Barrett, W. E. Smith, and K. J. Meyers. The Hotelling trace criterion and its correlation with human observer performance. *J. Optical Soc. Amer. A*, 4:945–953, 1987.

[22] R.D. Fiete, H.H. Barrett, E.B. Cargill, K. J. Myers, and W. E. Smith. Psychophysical validation of the Hotelling trace criterion as a metric for system performance. In *Proceedings SPIE Medical Imaging*, volume 767, pages 298–305, 1987.

[23] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Boston, 1992.

[24] R.M. Gray, R.A. Olshen, D. Ikeda, P.C. Cosman, S. Perlmutter, C. Nash, and K. Perlmutter. Evaluating quality and utility in digital mammography. In *Proceedings ICIP-95*, volume II, pages 5–8, Washington, D.C., October 1995. IEEE, IEEE Computer Society Press.

[25] T. Grogan and D. Keene. Image quality evaluation with a contour-based perceptual model. In *SPIE Proceedings*, volume 1666, pages 188–197, 1992.

[26] J.A. Hanley. Receiver operating characteristic (ROC) methodology: The state of the art. *Critical Reviews in Diagnostic Imaging*, 29:307–335, 1989.

[27] N. Jayant, J. Johnston, and R. Safranek. Signal compression based on models of human perception. *Proc. IEEE*, 81:1385–1422, Oct. 1993.

[28] N. S. Jayant and P. Noll. *Digital Coding of Waveforms*. Prentice-Hall, Englewood Cliffs, NJ, 1984.

[29] S. Klein, A. Silverstein, and T. Carney. Relevance of human vision to JPEG–DCT compression. In *SPIE Proceedings*, volume 1666, pages 200–215, 1992.

[30] H. Lee, A. H. Rowberg, M. S. Frank, H. S. Choi, and Y. Kim. Subjective evaluation of compressed image quality. In *Proceedings of Medical Imaging VI: Image Capture, Formatting, and Display*, volume 1653, pages 241–251. SPIE, Feb. 1992.

[31] A. S. Lewis and G. Knowles. Image compression using the 2-D wavelet transform. *IEEE Trans. Image Processing*, 1(2):244–250, April 1992.

[32] J. Limb. Distortion criteria of the human viewer. *IEEE Trans. on Systems, Man, and Cybernetics*, SMC-9:778–793, 1979.

[33] J. Lubin. The use of psychovisual data and models in the analysis of display system performance. In A. Watson, editor, *Visual Factors in Electronic Image Communications*. MIT Press, Cambridge, MA, 1993.

[34] F.X.J. Lukas and Z.L. Budrikis. Picture quality prediction based on a visual model. *IEEE Trans. Comm.*, COM-30(7):1679–1692, July 1982.

[35] H. MacMahon, K. Doi, S. Sanada, S.M. Montner, M.L. Giger, C.E. Metz, N. Nakamori, F. Yin, X. Xu, H. Yonekawa, and H. Takeuchi. Data compression: effect on diagnostic accuracy in digital chest radiographs. *Radiology*, 178:175–179, 1991.

[36] J. L. Mannos and D. J. Sakrison. The effects of a visual fidelity criterion of the encoding of images. *IEEE Trans. Inform. Theory*, 20:525–536, July 1974.

[37] H. Marmolin. Subjective mse measures. *IEEE Trans. on Systems, Man, and Cybernetics*, SMC-16(3):486–489, May/June 1986.

[38] B.J. McNeil and J.A. Hanley. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Medical Decision Making*, 4:137–150, 1984.

[39] C. E. Metz. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, VIII(4):282–298, Oct. 1978.

[40] A. Netravali and B. Prasada. Adaptive quantization of picture signals using spatial masking. *Proc. IEEE*, 65:536–548, 1977.

[41] K. Ngan, K. Leong, and H. Singh. Cosine transform coding incorporating human visual system model. In *SPIE Proceedings*, volume 707, pages 165–171, 1986.

[42] N. Nill. A visual model weighted cosine transform for image compression and quality assessment. *IEEE Trans. Comm.*, COM-33:551–557, 1985.

[43] N.B. Nill and B.H. Bouzas. Objective image quality measure derived from digital image power spectra. *Optical Engineering*, 31(4):813–825, April 1992.

[44] T. Pappas. Perceptual coding and printing of gray-scale and color images. In *SID Digest*, volume 23, pages 689–692, 1992.

[45] S.M. Perlmutter, P.C. Cosman, R.M. Gray, R.A. Olshen, D. Ikeda, C.N. Adams, B.J. Betts, M. Williams, K.O. Perlmutter, J. Li, A. Aiyer, L. Fajardo, R. Birdwell, and B.L. Daniel. Image quality in lossy compressed digital mammograms. *Signal Processing*, 59:189–210, June 1997.

[46] S.M. Perlmutter, P.C. Cosman, C. Tseng, R.A. Olshen, R.M. Gray, K.C.P. Li, and C.J. Bergin. Medical image compression and vector quantization. *Statistical Science*, 13(1):30–53, Jan. 1998.

[47] S.M. Perlmutter, C. Tseng, P.C. Cosman, K.C.P. Li, R.A. Olshen, and R.M. Gray. Measurement accuracy as a measure of image quality in compressed MR chest scans. In *Proceedings ICIP-94*, volume 1, pages 861–865, Austin, TX, Nov. 1994. IEEE Computer Society Press.

[48] M.J.D. Powell. *Approximation theory and methods*. Cambridge University Press, Cambridge, England, 1981.

[49] S.R. Quackenbush, T.P. Barnwell III, and M.A. Clements. *Objective Measures of Speech Quality*. Prentice Hall Signal Processing Series. Prentice Hall, New Jersey, 1988.

[50] M. Rabbani and P. W. Jones. *Digital Image Compression Techniques*, volume TT7 of *Tutorial Texts in Optical Engineering*. SPIE Optical Engineering Press, Bellingham, WA, 1991.

[51] R. J. Safranek and J. D. Johnston. A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression. In *Proceedings ICASSP*, pages 1945–1948, Glasgow, UK, 1989.

[52] R.J. Safranek, J.D. Johnston, and R.E. Rosenholtz. A perceptually tuned sub-band image coder. In *Proceedings of the SPIE - The International Society for Optical Engineering*, pages 284–293, Santa Clara, Feb. 1990. IEEE.

[53] J.A. Sagrhi, P.S. Cheatham, and A. Habibi. Image quality measure based on a human visual system model. *Optical Engineering*, 28(7):813–818, July 1989.

[54] A. Said and W.A. Pearlman. A new, fast, and efficient image codec based on set partitioning in hierarchical trees. *IEEE Trans. on Circuits and Systems for Video Technology*, 6(3):243–250, June 1996.

[55] D.J. Sakrison. On the role of the observer and a distortion measure in image transmission. *IEEE Trans. Comm.*, 25:1251–1267, 1977.

[56] 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 *Proceedings of Medical Imaging VI: Image Capture, Formatting, and Display*, volume 1653, pages 232–240. SPIE, Feb. 1992.

[57] J. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions on Signal Processing*, 41(12):3445–3462, December 1993.

[58] E. Shlomot, Y. Zeevi, and W. Pearlman. The importance of spatial frequency and orientation in image decomposition and coding. In *SPIE Proceedings*, volume 845, pages 152–158, 1987.

[59] W.R. Steinbach and K. Richter. Multiple classification and receiver operating characteristic (roc) analysis. *Medical Decision Making*, 7:234–237, 1995.

[60] T. G. Stockham Jr. Image processing in the context of a visual model. *Proc. IEEE*, 60:828–842, July 1972.

[61] J. A. Swets. ROC analysis applied to the evaluation of medical imaging techniques. *Investigative Radiology*, 14:109–121, March–April 1979.

[62] W.D. Voiers. Diagnostic acceptability measure for speech communication systems. In *Proceedings ICASSP*, pages 204–207, 1977.

[63] A. Watson. Efficiency of an image code based on human vision. *JOSA A*, 4:2401–2417, 1987.

[64] M.C. Weinstein and H.V. Fineberg. *Clinical Decision Analysis*. W.B. Saunders Company, Philadelphia, 1980.

[65] P. Wilhelm, D. R. Haynor, Y. Kim, and E. A. Riskin. Lossy image compression for digital medical imaging system. *Optical Engineering*, 30:1479–1485, Oct. 1991.

[66] I. H. Witten, R. M. Neal, and J. G. Cleary. Arithmetic coding for data compression. *Communications of the ACM*, 30:520–540, 1987.

[67] C. Zetzsche and G. Hauske. Multiple channel model for the prediction of subjective image quality. In *SPIE Proceedings*, volume 1077, pages 209–216, 1989.