Tree-Structured Vector Quantization of CT Chest Scans: Image Quality and Diagnostic Accuracy

P. C. Cosman, C. Tseng, R. M. Gray, R. A. Olshen, L. E. Moses, H. C. Davidson, C. J. Bergin, and E. A. Riskin

Abstract— **The quality of lossy compressed images is often characterized by signal-to-noise ratios, informal tests of subjective quality, or receiver operating characteristic (ROC) curves that include subjective appraisals of the value of an image for a particular application. We believe that for medical applications, lossy compressed images should be judged by a more natural and fundamental aspect of relative image quality: their use in making accurate diagnoses. We apply a lossy compression algorithm to medical images, and quantify the quality of the images by the diagnostic performance of radiologists, as well as by traditional signal-to-noise ratios and subjective ratings. Our study is unlike previous studies of the effects of lossy compression in that we consider non-binary detection tasks, simulate actual diagnostic practice instead of using paired tests or confidence rankings, use statistical methods that are more appropriate for non-binary clinical data than are the popular ROC curves, and use low-complexity predictive tree-structured vector quantization for compression rather than DCT-based transform codes combined with entropy coding.**

Our diagnostic tasks are the identification of nodules (tumors) in the lungs and lymphadenopathy in the mediastinum from computerized tomography (CT) chest scans. Radiologists read both uncompressed and lossy compressed versions of images. For the image modality, compression algorithm, and diagnostic tasks we consider, the original 12 bit per pixel (bpp) CT image can be compressed to between 1 bpp and 2 bpp with no significant changes in diagnostic accuracy. The techniques presented in this paper for evaluating image quality do not depend on the specific compression algorithm and are useful new methods for evaluating the benefits of any lossy image processing technique.

I. Introduction

DIGITAL image processing in recent years has shown tremendous potential for application to digital medical images. At the foreground are the possibilities for easy image retrieval, efficient storage, rapid image transmission for off-site diagnoses, and the maintenance of large image banks for teaching and research. In addition, images in digital format are accessible for digital signal processing such as filtering, enhancement, classification, 3-D modeling, and motion video. Although these techniques may in the future become an integral part of medical research, diagnosis, and treatment, they are hampered by the difficulty of handling multi-megabyte image files (X-ray: 2 Mb, CT: 0.5 Mb, MRI: 0.13 Mb). The typical compression ratios of 2:1 or 3:1 achieved by lossless schemes may be insufficient, thereby making lossy schemes attractive for consideration in any digital medical image processing system. With medical images, however, the determination of image quality is far more sensitive than the issue of image quality for entertainment or personal communications.

Medical image quality has been examined in many recent studies. Typically, quality is quantified objectively by using signal-to-noise ratios (SNRs), and subjectively by performing statistical analyses on viewers' scores. When viewers rated diagnostic usefulness rather than simply general appearance, these studies related diagnostic accuracy to compression level. In other studies, radiologists were asked to view an image which either did or did not possess an abnormality and to provide a binary decision (abnormality present or not) along with a quantitative value for their degree of certainty, typically a number from 1 to 5. Subsequent statistical analyses, usually ROC-based, attempted to quantify the levels of compression in a specific application that can be used without a statistically significant change in diagnostic accuracy. There are numerous examples of such approaches [1]–[10].

In most of these studies, the basic experiments were subjective and did not simulate the ordinary tasks of radiologists. The observers were asked to rate numerically their confidence or their opinion of image quality or usefulness rather than to make diagnoses as they would under ordinary clinical conditions. This rating resulted in data useful for ROC analysis, but it constitutes an artificial diagnostic task. Furthermore, radiologists often face images which may contain one or more abnormalities, and the diagnostic task is to find any and all that are present. In this case the task is not binary, and is not amenable to traditional ROC analysis techniques. Lastly, some studies used paired comparisons, where an original and a compressed image were displayed simultaneously and a radiologist was asked to rate the difference. This procedure differs markedly from ordinary clinical practice.

Manuscript received August 17, 1992; revised February 23, 1993. 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. The associate editor responsible for coordinating the review of this paper and recommending its publication was Dr. M. A. Viergever.

P. C. Cosman and R. M. Gray are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305.

C. Tseng was with the Department of Electrical Engineering, Stanford University. She is currently a medical student at the UCLA School of Medicine, Los Angeles, CA 90025.

R. A. Olshen and L. E. Moses are with the Division of Biostatistics of the Stanford University School of Medicine and the Department of Statistics of Stanford University, Stanford, CA 94305.

H. C. Davidson is with the Department of Diagnostic Radiology and Nuclear Medicine, Stanford University School of Medicine, Stanford, CA 94305.

C. J. Bergin was with the Division of Diagnostic Radiology, Stanford University School of Medicine. She is now with the Department of Radiology, University of California, San Diego, La Jolla, CA 92093.

E. A. Riskin is with the Department of Electrical Engineering, University of Washington, Seattle, WA 98195.

IEEE Log Number 9213396.

Our goal was to use the performance of radiologists evaluating compressed CT chest scans in order to quantify the impact of a specific lossy compression algorithm on diagnosis. Images could contain multiple abnormalities, and the radiologists were asked to make diagnoses in their usual fashion by locating and marking visible abnormalities. Images were not seen in a paired fashion, and the judges were not asked to associate confidence rankings with their diagnoses. The compression algorithm used was predictive pruned tree-structured vector quantization (PTSVQ). This technique combines statistical clustering methods for reproduction quality with classification and regression tree design techniques for efficient search and progressive transmission properties [11]–[14]. Image quality was quantified in three ways: SNRs (or average normalized distortion as measured by mean squared error), average scores of quality ratings by three radiologists, and accuracy of diagnosis of mediastinal adenopathy and lung nodules. The images were presented at six different compressed levels ranging from 2.6 bpp to 0.56 bpp plus the uncompressed original at 12 bpp. Evaluation of compression performance in terms of diagnostic accuracy was based on sensitivity and predictive value positive, as will be described. Preliminary results of our study have been reported in [15] and [16].

Our public domain vector quantization algorithms use low complexity software implementations. This is in contrast to the full-frame or large block DCT used in all the cited references except [3]. For a specific bit rate, however, vector quantization approaches generally do not yield as high a quality as the transform codes based on much larger block sizes, especially when the transform code also incorporates noiseless coding (entropy coding, reversible coding) such as Huffman coding. Our compression algorithms operate on square blocks of $2 \times 2 = 4$ pixels and directly perform variable rate encoding without transforms for compression or decompression. The noiseless coding that can be cascaded after the DCT quantization step for an extra reduction in bit rate of typically 1.5:1 or 2:1 can also be used with vector quantization algorithms, at the cost of added complexity.

Verification of diagnostic accuracy by clinical testing is detailed and expensive. It is highly desirable to find easily computable quantitative features of images that strongly correlate with or predict the results of more extensive diagnostic studies. Whether by computable features or clinical testing, such studies are necessary to develop reasonable policies for the use of lossy processing on medical images. The medical community is concerned about the legal consequences of an incorrect diagnosis based on a lossy processed image. However, since diagnoses based on original images can also be incorrect, clinical experiments are necessary to establish the limits of image processing for which the diagnostic accuracy is at least equal to that from an unprocessed image.

II. Tree-Structured Vector Quantization

A general vector quantizer (VQ) for image compression consists of an encoder and a decoder. An input image is parsed into a sequence of groups of pixels, often 2×2 squares, but larger squares and rectangles are also used. The encoder views an input vector X_n at time n and produces a binary vector or channel codeword i_n. In fixed rate systems, the binary vectors all have dimension R. In variable rate systems, the binary vectors have an average dimension R. The decoder is a table lookup, and upon receiving a channel codeword i_n, it puts out a stored codeword or template $\hat{Y}_{i_n}$, that is, a word in memory indexed by the channel codeword. Given a codebook containing all of the possible codewords, the decoder is completely described. The encoder operates according to a nearest neighbor or minimum distortion rule, possibly with constraints on how the search is done. To minimize overall average distortion, a good encoder will try to minimize distortion at each step. Suppose that $d(X, \hat{X}) \geq 0$ measures the distortion or cost of reproducing an input vector X as a reproduction $\hat{X}$. With no structural constraints, the optimal encoder for a given codebook selects the vector Y_i if

$$d(X_n, Y_i) \leq d(X_n, Y_j), \text{ for all } j.$$

The distortion measure permits us to quantify the performance of a VQ in a manner that can be computed, used in analysis, and used in design optimization. The theory focuses on average distortion in the sense of a probabilistic average or expectation. Practice emphasizes the time average or sample average distortion

$$D = \frac{1}{L} \sum_{n=1}^{L} d(X_n, \hat{X}_n)$$

for large L. With well-defined stationary random process models instead of real signals, the sample average and expectation are effectively equal. We here focus on the simple squared error distortion which is the square of the Euclidean distance between the input and reproduction vectors. Given two k-dimensional vectors $X = (X(0), \ldots, X(k-1))$ and $\hat{X} = (\hat{X}(0), \ldots, \hat{X}(k-1))$, then

$$d(X, \hat{X}) = ||X - \hat{X}||^2 = \sum_{l=0}^{k-1} |X(l) - \hat{X}(l)|^2.$$

There are a variety of approaches to VQ design. We concentrate on a class of constrained structure quantizers that are relatively easy to design and implement and which provide good performance. Codebooks are designed with a binary tree structure, and the encoder selects the codeword by a sequence of fast binary minimum distortion comparisons. The input vector is first compared to two templates representing the first two nodes in the tree. Since the squared error distortion measure is used, this is equivalent to a hyperplane or correlation test. The minimum distortion template is chosen and the algorithm advances to the corresponding node. At each step in the search, either one arrives at a terminal node (in which case the channel codeword is the binary sequence indicating the path from the root to that node), or the algorithm again faces a choice of two templates.

The code is designed by first growing a tree in a greedy manner [17]. One begins with a root note which has associated with it a label or template equalling the Euclidean centroid (or vector average) of the input distribution (usually an empirical distribution, that is, a learning or training set). The root node is split into two child nodes, one labeled by the original centroid and the other labeled by a close, but distinct, vector. The Lloyd (or k-means) clustering algorithm is then run on the two word code to produce a good 1-bit-per-vector codebook. At this point, or any point in the tree growing, one has a collection of labeled terminal nodes or leaves. Any of these leaves could be made into an internal node by splitting it into two new leaves and running the clustering algorithm again. Splitting a leaf results in a decrease of average distortion and an increase in the average length of the binary codeword. The greedy growing algorithm finds the leaf that when split yields the largest decrease in distortion per increase in average length, i.e., best trade-off rate and distortion in a greedy sense. The tree is grown until some target rate is achieved or until the leaves have so few vectors from the training set that they become untrustworthy.

Once grown, a tree may then be pruned to find better codes at lower bit rates than appeared in the growing stage. Here the operation is not greedy, but is a true optimum. The algorithm considers subtrees of the full-grown tree with a common root node. Pruning the full tree to a subtree yields a new code with a reduced average length, but an increased average distortion. Here the goal is to minimize the increase in average distortion per decrease in average length. Unlike growing, pruning can remove large numbers of nodes at a time. The combination of growing and pruning is the extension to vector quantization of a classification and regression tree design technique of Breiman, Friedman, Olshen, and Stone [18].

The design algorithm outlined here yields a nested (embedded) family of codes with several useful properties. Pruned TSVQ usually yields lower distortion than fixed-rate full-search VQ for a given *average* rate, has a natural successive approximation (progressive) property, and is well matched to variable-rate environments such as storage or packet communications. A TSVQ can have its tree tailored by using input-dependent weighted distortion measures that permit the incorporation of enhancement and classification into the tree [19], [20]. Further details which would allow a reader to implement this algorithm may be found in [21], [17], [14], [11].

A. Predictive VQ

If the codebook of reproductions is fixed for all input vectors, then the VQ operates in a memoryless fashion on each input vector; that is, each vector is encoded independently of previous inputs and outputs of the encoder. In general an encoder can have memory by varying the set of possible reproduction vectors according to past actions. Predictive and finite-state vector quantizers are examples of VQ with memory.

Predictive PTSVQ works through a combination of a linear predictor and a residual quantizer [11]–[14]. The encoder predicts the current pixel block using past blocks, forms the residual from the prediction error, and quantizes it with a TSVQ. The encoding path through the tree is sent to the decoder. Given the same tree, the decoder decodes the quantized residual and reconstructs the pixel block by adding its prediction of the block to the quantized residual. The coefficients for the predictor are calculated from Wiener-Hopf equations, a simple method that has worked well experimentally. From the training set, the correlation matrix between the current block and its neighbors is estimated and inverted to obtain the prediction coefficients. These coefficients are thus based upon correlations between original pixel values and neighboring original pixel values. During compression, however, the encoder is constrained to predict on the encoded versions of surrounding blocks rather than unquantized versions. This matches the encoder block prediction with the decoder block prediction.

Once the prediction coefficients are fixed, a training sequence of residuals is generated from the training sequence of original pixel values by calculating the differences between actual values and predicted values. The tree-structured encoder is developed using these residual vectors as a training set. By encoding the lower energy residual signal, fewer bits can be used to encode to a desired distortion level than would be needed for encoding the original higher energy signal.

Both the predictor and residual quantizer were designed for 2×2 pixel blocks. For the predictor, a larger block size results in a more tenuous prediction, since pixels being predicted are farther apart from pixels used in the prediction. For the residual quantizer, on the other hand, larger pixel blocks better exploit Shannon's theory on the ability of vector quantizers to asymptotically outperform scalar quantizers. The block size choice is a trade-off among prediction accuracy, algorithmic complexity, storage memory, and quantization performance. While performance theoretically improves with block size, large block sizes can introduce block artifacts into an image that can outweigh any improvement in quantitative performance. Here we have chosen block size with an emphasis on achieving low complexity.

III. Data Set and Study Design

The diagnostic tasks chosen were the detection of mediastinal adenopathy and lung nodules, both of crucial importance in chest imaging. Abnormally enlarged lymph nodes in the mediastinum (the central portion of the chest that contains the heart and major blood vessels) are frequently caused by lymphoma or metastatic disease. Typically radiologists can easily locate lymph nodes in a chest scan, and the detection task is therefore to determine which of them are enlarged. Lung nodules, commonly caused by primary or metastatic cancer, range in size from undetectably small to large enough to fill an entire segment of the lung. There can be multiple nodules in one or both lungs. In contrast to the mediastinal task, both the presence and size of lung nodules are detection issues.

In our study, the prediction coefficients and residual quantizer for compressing CT images are designed on a training

set of representative CT images, thereby freeing the algorithm from pre-formed models of the source, either before or during compression. Twenty CT images of the mediastinum were used in the training set for detecting enlarged lymph nodes and twenty CT lung images were used in the training set for detecting lung nodules. All images were 512 × 512 pixels, and were obtained using a GE 9800 scanner (120 kV, 140 mA, scan time 2 seconds per slice, bore size 38 cm, field of view 32–34 cm). Although no formal research was undertaken to accurately determine 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. Images with very large nodules were excluded as it was felt that these would be too easily detected. 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), a codebook tree for the prediction residuals was grown to a depth of 2.8 bpp and pruned back to 6 different levels, ranging from 0.56 bpp to 2.64 bpp. Then 30 test images were encoded with the 6 subtrees. Patient studies represented in the training set were not used as subsequent test images, and the SNR, subjective quality, and diagnostic results are based only on test images.

A. Protocol for Judging Images

The compressed and original images were viewed by 3 radiologists. For each of the 30 images in a study, each radiologist viewed the original and 5 of the 6 compressed levels, and thus 360 images were seen by each judge. Images were seen on hardcopy film on a lightbox. A usual adjustment to the dynamic range of the image, called "windows and levels," was applied to each image before filming. A radiologist who was not involved in the judging applied standard settings for "windows and levels" for the mediastinal images, and different standard settings for the lung nodule images. The compressed and original images were filmed in standard 12-on-1 format on 14″ × 17″ film using the scanner that produced the original images.

The viewings were divided into 3 sessions during which the judges independently viewed 10 pages, each with 6 lung nodule images and 6 mediastinal images. The judges marked abnormalities directly on the films with a grease pencil. No constraints were placed on the viewing time, the viewing distance, or the lighting conditions. Each judge was encouraged to simulate the conditions he or she would use in everyday work. The judges were, however, constrained to view the 10 pages in the predetermined order, and could not go back to review earlier pages. At each session, each judge saw each image at exactly 2 of the 7 levels of compression (7 levels includes the original). The two levels never appeared on the same film, and the ordering of the pages ensured that they never appeared with fewer than 3 pages separating them. This was intended to reduce learning effects. A given image at a given level was never seen more than once by any one judge, and so intra-observer variability was not explicitly measured. Of the 6 images in one study on any one page, only one image was shown as the original, and exactly 5 of the 6 compressed levels were represented.

B. Determination of a Gold Standard

To measure the preservation of diagnostic accuracy, it is necessary to determine a "gold standard" for each image that can serve as the standard for comparison with readings of the compressed versions. For our study, the gold standard was determined by consensus of the 3 judges on the original image. For more than half of the images, examination of the readings by the 3 judges revealed complete agreement in the number and location of abnormalities, and the gold standard was established. In the cases of disagreement, each judge was separately informed that a disagreement had occurred, and was asked to review his or her reading of that original. In a small number of cases where this did not produce agreement, the judges were brought together to discuss the image. As a result of this process, the 30 images in each study were winnowed to 24, with elimination of the ones that generated the most irreconcilable controversy. If judges do not agree on what structures are abnormal in an uncompressed image, there are other ways to determine what constitutes a correct or incorrect judgment on a compressed image [33]. The 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.

This process for determining the gold standard allows the study to be most useful for comparing the various compressed levels among themselves. The gold standard was established by consensus on the original images, and this consensus was achieved on only 24 images out of 30, the remaining 6 being eliminated from the study. 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 diagnostic accuracy relative to that gold standard, the original images have an advantage when compared with the others; hence the comparative statements we make later are conservative.

IV. Distortion-Rate Performance

The traditional manner for comparing the performance of lossy compression algorithms is to plot the average distortion of the codes as a function of bit rate. Often the distortion per pixel is normalized by either the energy or the variance of the input distribution, or by the maximum pixel intensity squared. The ratio of normalization constant to average distortion is called a signal-to-noise ratio (SNR) or signal-to-quantization-noise ratio if the input variance or energy is the normalization:

$$SNR = 10 \log_{10} \frac{E(X^2)}{D}$$

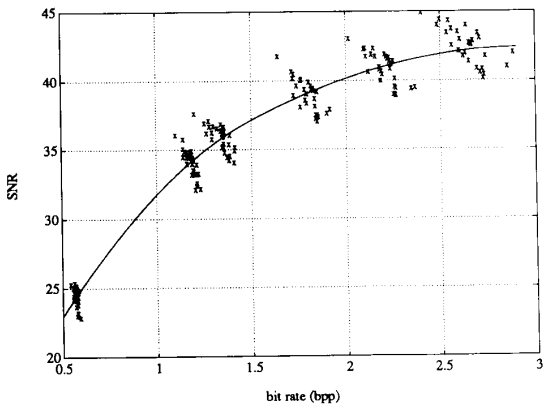


Fig. 1. SNR for the 24 lung test images at the 6 compressed levels.

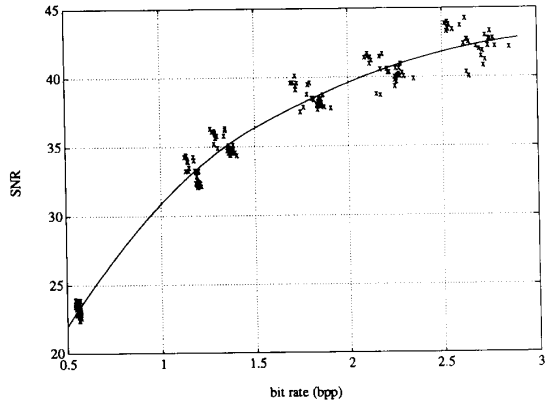


Fig. 2. SNR for the 24 mediastinal test images at the 6 compressed levels.

where D is the average distortion, $E(X^2)$ is the input energy, and SNR is measured in dB. Figures 1 and 2 depict a scatter plot of the SNR and average bit rate for each of the 24 images in the mediastinum and lung studies that had a gold standard, together with a quadratic spline fit with a single knot at 1.5 bpp [22] that provides a good indication of the overall distortion-rate performance of the code family on the test data. The SNR plots take into account the differences in distortion due to differences in input signal intensity; log plots are commonly used for such comparisons.

Long-term experience has led to the conclusion that SNR is a reasonable indicator of relative visual distortion for most types of images, but it is not based on visual models. The lack of context and the quadratic nature of the error response are not thought to be consistent with models of human observers considered as measuring instruments. Some compression algorithms can introduce artifacts into the decompressed image that perceptually degrade the image even though the SNR seems high. For example, any compression algorithm operating on blocks of pixels (including transform codes and VQ) can introduce block artifacts such as staircases and texturing if the block size is too large. Due to its simplicity, however, SNR is a useful method for comparing different algorithms or approximately evaluating visual distortion at different bit rates. SNR has yet to qualify, however, as a successful measure of the preservation of diagnostic accuracy in lossy compression of medical images.

V. Subjective Quality

Distortion-rate plots are the most commonly quoted indicator of performance in data compression systems, but they are often accompanied by the caveat that they do not necessarily represent subjective quality. In this section we consider the issue of subjective quality in two ways. First we give examples of images of both lung and mediastinum at all bit rates so that readers can form their own opinions of the usefulness of the compressed images. The images are collected in Appendix B. Little degradation is evident in any but the lowest rate of 0.56 bpp. The lung image contains 3 nodules with complete initial agreement on the gold standard. One judge had perfect readings at all levels, a second judge missed the smallest tumor at the most compressed level, and the third judge missed the two smallest tumors at the most compressed level. The mediastinal test image contains one abnormal lymph node. Its gold standard had to be determined by asking the judges to independently review their decisions on the uncompressed image, as one judge differed from the other two judges on the original. The judges scored perfectly in diagnosis at all other bit rates except for a second judge who had a wrong diagnosis on the least compressed version at 2.75 bits.

The assessment of subjective quality is based on a questionnaire provided the radiologists following the experiments. Each radiologist was asked to assign a score of 1 to 5 for each of the last 12 images in each session. The score was based on evaluating the diagnostic quality of the image and responding to the question "How good is the quality of this image for diagnostic purposes?" The allowed responses were as follows:

1. Image is of excellent quality.
2. Standard image quality. Image usable and very adequate.
3. Image quality passable. Image usable but below standard.
4. Low image quality. Image usable but difficult to use.
5. Very low quality image. Image not usable.

These questions resembled the subjective ranking of Barrett *et al.* [5] who asked how certain the observer was that a lesion was present and who then based their statistical analyses on these responses. In our study, however, the subjective questionnaire was aimed only at getting an overall appraisal from the radiologist for comparison with the diagnostic accuracy results. Quality score/bit rate pairs for all of the test images are plotted in Figs. 3 and 4. Scores for individual images are displayed as x's. The o's mark the average of the x's for each bit rate. For both mediastinum and lung images, uncompressed images at 12 bpp received scores of 1 and 2. Images compressed to lower bit rates received worse quality scores as expected.

The qualitative scores are not as smooth as the SNR, predictive value positive, or sensitivity data. To guarantee that fits to these non-negative data are themselves non-negative (indeed positive), we take logarithms of the data, fit a model

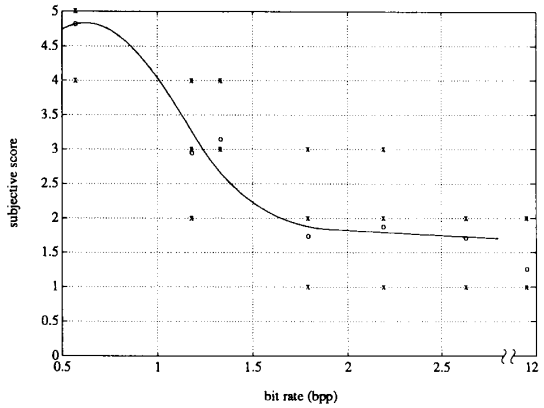


Fig. 3. Subjective quality vs. bit rate for lungs: 1 = excellent; 5 = poor.

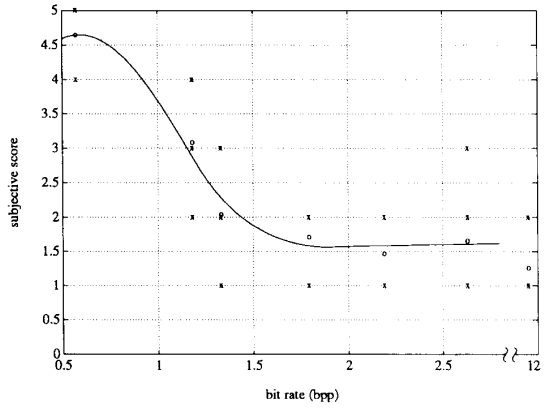


Fig. 4. Subjective quality vs. bit rate for mediastina: 1 = excellent; 5 = poor.

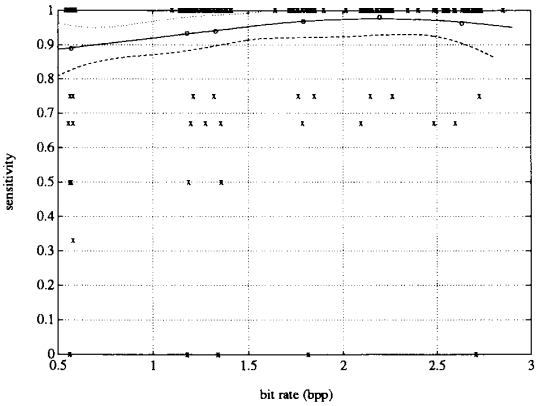


Fig. 5. Lung Sensitivity: RMS = .177.

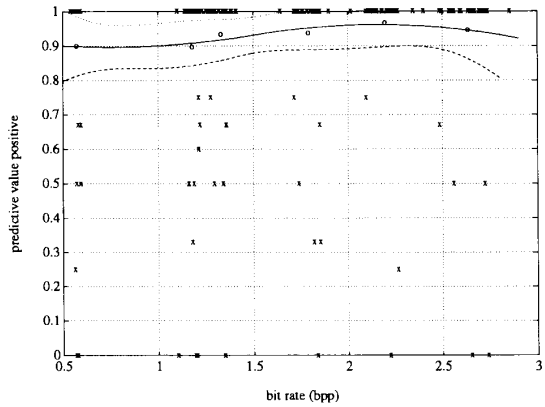


Fig. 6. Lung Predictive Value Positive (PVP): RMS = .215.

(for example, a spline model), and exponentiate the answer. It follows from Jensen's inequality that if the usual assumptions for linear models are in force when we fit the logged data, then there is a systematic bias in the exponentiated answers we report. However, corrections for that bias are negligible, and have not been incorporated into our algorithms. This approach has been studied in detail and applied in some settings by C. J. Stone and C. Kooperberg [23], [24]. We found their "log-spline" approach simple but very powerful when adapted to our regression problem. Figs. 3 and 4 include quadratic splines with knots at 1.2 and 1.9 bpp fit to the logged qualitative scores, and then exponentiated.

VI. Statistical Methods and Results

A. Sensitivity and Predictive Value Positive

Many analyses of studies in clinical radiology involve ROC or "receiver operating characteristic" curves [25], [26]. They summarize a trade-off between *true positive* and *false positive rates*, typically as a threshold for detection varies. The *true positive rate* is also called *sensitivity*, the probability something is detected given that it is present. The complement of the *false positive rate* is termed *specificity*, the probability something is not detected given that it is actually absent. For our realistic clinical studies, specificity does not make sense because there is no sample space in which to do the computation; that is, specificity has no natural denominator, as it is not possible to say how many abnormalities are absent. On the other hand, once our protocol for determining the gold standard is concluded, sensitivity not only makes sense, but also is a crucial statistic that quantifies results. However, a judge who finds nodules or abnormal nodes everywhere in an image could have perfect sensitivity. The additional descriptive statistic we use, therefore, should penalize false positive reporting. What we can and do measure is termed *predictive value positive* (PVP), the chance something marked as an abnormality is indeed an abnormality [27]. So 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 sensitivity and PVP will be 1 if the decision is perfect. In Fig. 5 we plot sensitivity as a function of bit rate for all lung images, judges, compression levels (not including the original), and sessions. There are 360 x's: 360 = 24 images $\times$ 3 judges $\times$ 5 compressed levels seen for each image. Figure 6 is analogous

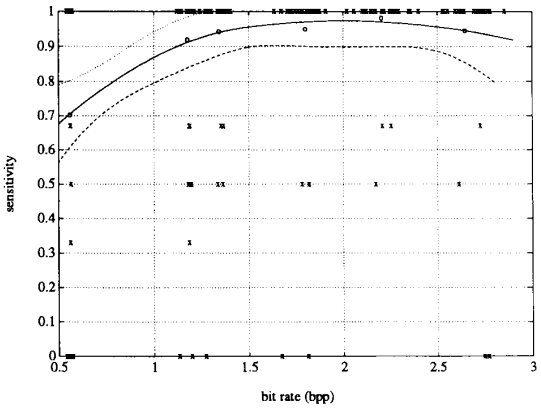


Fig. 7. Mediastinum Sensitivity: RMS = .243.

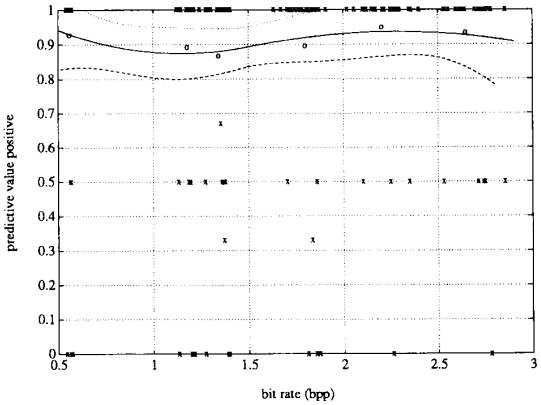


Fig. 8. Mediastinum Predictive Value Positive (PVP): RMS = .245.

for lung PVP. Figures 7 and 8 are the corresponding figures for the mediastinum. The values of the sensitivity and PVP are simple fractions such as 1/2 and 2/3 because there are only 0–4 abnormalities found in each image. The o's mark the average of the x's for each bit rate. The six achieved bit rates (averaged across 24 test images for a given codebook) for compressed images of both lung and mediastinum were nearly identical: for the lung .57, 1.18, 1.33, 1.79, 2.19, 2.63, and for the mediastinum 0.56, 1.18, 1.34, 1.80, 2.20, 2.64 bpp.

The curves are least squares quadratic spline fits to the data with a single knot at 1.5 bpp [22], together with the two-sided 95% confidence regions. With regression splines, observations at particular bit rates tend to have most influence nearby, and observations at remote bit rates have little influence. This is what we prefer since the functional form of PVP (or sensitivity) as it varies with bit rate is utterly unknown beyond our knowing that the relationship is a smooth one. In view of the highly nonGaussian nature of the data, the Scheffé simultaneous confidence regions were obtained by a bootstrapping procedure. The algorithm for this is adapted from [28], [29], and is described in Appendix A. 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. Though the text that follows gives various detailed analyses of our data, its basic messages are transparent from Figures 5 through 8. Namely, sensitivity for the lung seems to be nearly as good at low rates of compression as at high rates, while sensitivity for the mediastinum falls off noticeably for the lowest bit rate. PVP for both lung and mediastinum is roughly constant across the bit rates.

B. Behrens-Fisher t Statistic

The comparison of sensitivity and PVP at different bit rates was carried out using a variation of the two-sample t-test, sometimes called the Behrens-Fisher test [30], in which inequality of variances is accounted for in the (approximate) test; the test is quite robust when data are not Gaussian, as our data clearly are not. 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 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 $\overline{\Delta_i}$ be the average difference: $\overline{\Delta_i} = \frac{1}{N_i}\sum_j \Delta_{ij}$. We define

$$S_i^2 = \frac{1}{N_i - 1}\sum_j (\Delta_{ij} - \overline{\Delta_i})^2$$

and then the Behrens-Fisher t statistic is given by

$$t_{BF} = \frac{\sum_i \overline{\Delta_i}}{\sqrt{\sum_i \frac{S_i^2}{N_i}}}.$$

Our Δ_{ij} are fractions with denominators not more than 4, so are utterly nonGaussian. 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. As discussed later, 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. For example, in comparing levels B and G for mediastinal sensitivity, Judge 1 saw 19 images at both levels, and had identical readings on 16 of them. Three times the reading on level G led to a higher value of sensitivity than the reading on the same image seen at level B. However, a 3:0 split cannot be considered significant at the 5% level. For Judges 2 and 3, there were only 2 differences apiece out of 21 and 20 image pairs seen, which again is not a significant result. In pooling the judges, however, there were 7 cases where level G outperformed level B, a result which achieves significance. When the data were examined this way, again level A was inferior to almost all of 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 vs. 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).

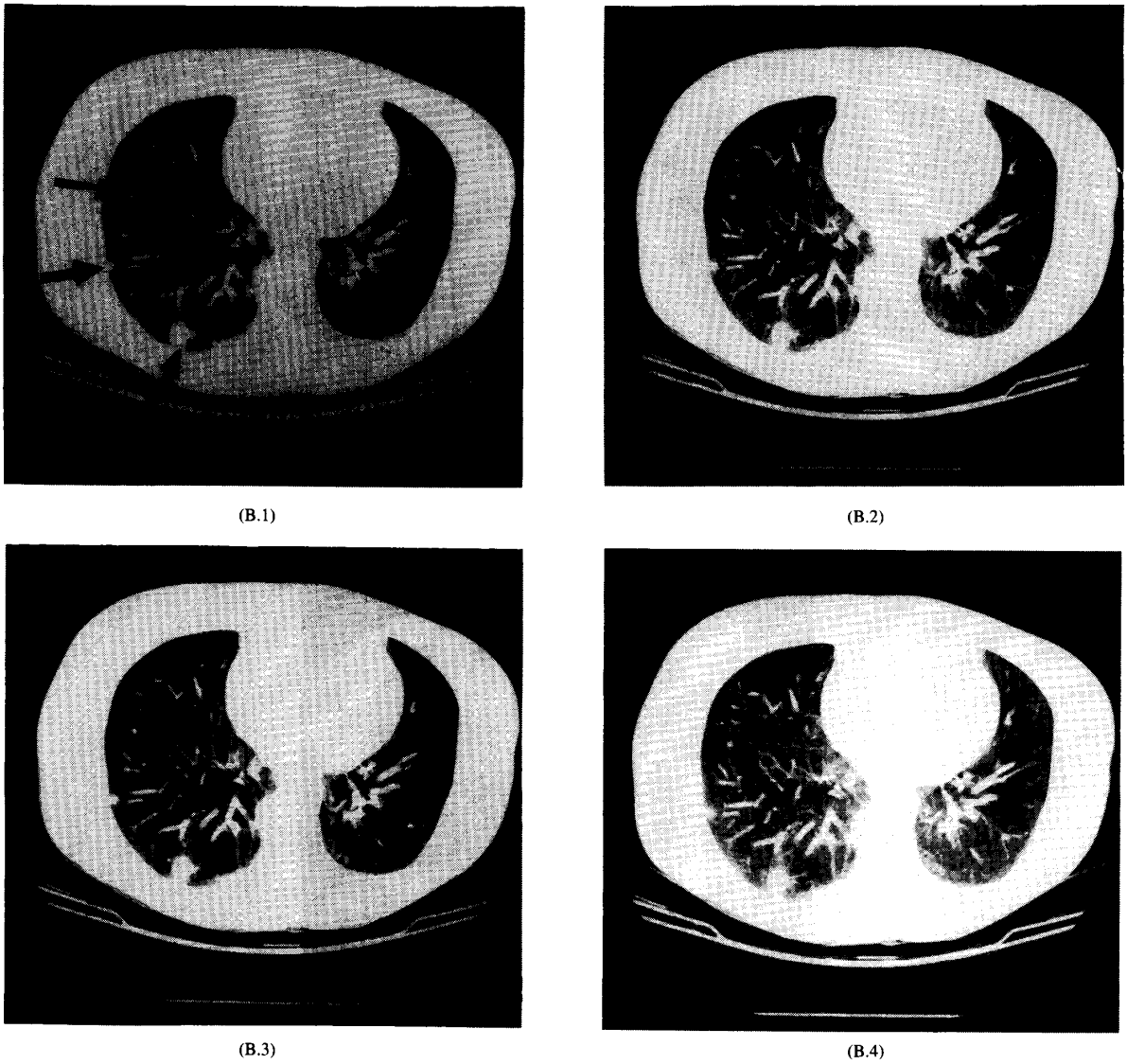
The results indicate that level A (0.56 bpp) is unacceptable for diagnostic use. Since the blocking and prediction artifacts became quite noticeable at level A, the judges tended not to attempt to mark any abnormality unless they were quite sure it was there. This explains the initially surprising result that level A was not different from most other levels for PVP, whereas it is different from all other levels 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 3 compressed levels are clearly acceptable for diagnostic use. 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 a reformulation of the gold standard protocol that would remove at least one of the biases against compression that are present in this study. This latter goal could be accomplished, for example, by having the gold standard determined by an independent panel of radiologists, rather than by the consensus of the judging radiologists on level G.

As we study schemes for compressing data, we would like to conclude that their implementation would not degrade clinical practice, but to make our point, we must argue as our own devil's advocates. This criterion is met by the fact that the statistical approach described here contains 4 identifiable biases, none of which favors compression. The first is the bias in the gold standard described earlier, which specifically confers an advantage upon level G relative to the compressed levels. Secondly there is the problem of multiple comparisons [29]. Since we perform comparisons for all possible pairs out of the 7 levels, for both sensitivity and PVP, for both lung and mediastinal images, and for both 3 judges separately and for judges pooled, we are reporting on $21 \times 2 \times 2 \times 4 = 336$ tests. One would expect that, even if there were no effect of compression upon diagnosis, 5% of these comparisons would show significant differences at the 5% level. A third element which argues against compression is the use of a 1-sided test instead of a 2-sided test. In most contexts, for example when a new and old treatment are being compared and subjects on the new treatment do better than those on the old, we do a two-sided test of significance. Such two-sided tests implicitly account for both possibilities: that new interventions may make for better or worse outcomes than standard ones. For us, a two-sided test would implicitly recognize the possibility that compression improves, not degrades, clinical practice. In fact, we believe this can happen, but to incorporate such beliefs in our formulation of a test would make us less our own devil's advocates than would our use of a one-sided test. Our task is to find when compression might be used with clinical impunity, not when it might enhance images. The fourth bias stems from the fact that the summands in the numerator of t_{BF} may well be positively correlated (in the statistical sense), though we have no way to estimate this positive dependence from our data. If we did, the denominator of t_{BF} would typically be smaller, and we are nearly certain that such incorporation would make finding "significant" differences between compression levels more difficult. For all of these reasons, we believe that the stated conclusions are conservative.

Other tests that were carried out were the pairwise comparison of judges using the permutation distribution of Hotelling's paired T^2 statistic [31], and a comparison of images. The judges were found to be different in judging the mediastinum but not the lung, and the lung images were found to be very different for all judges, but mediastinal images were different for only 2 of the judges. These results are reported in [32]. Consistency of individual judges is examined in [33].

C. Learning: Analyses by ANOVA and McNemar Statistics

Since the radiologists would see an image at different compression levels six times during the course of the study, we needed to ascertain whether learning effects were significant. Learning and fatigue are both processes that might change the score of an image depending on when it was seen. In each session, each image was seen at exactly 2 levels, and the ordering of the pages ensured that they never appeared with fewer than 3 pages separating them. At least two weeks separated the sessions. To examine the possibility of intrasession learning and fatigue, we examined the paired data in which the first occurrence of a given image in a session was paired with the second occurrence of that same image (at a different compression level) in the same session. Each image in the pair was either "perfect" (sensitivity = 1, PVP = 1) or not. There were thus four types of pairs, those with both members perfect, those with the first occurrence perfect and the second not, those with the second occurrence perfect and the first not, and those with neither one perfect. In the McNemar analysis [34], we concern ourselves with 2 of the 4 types: those pairs in which the members differ. If it did not matter whether an image was seen first or second, then conditional on the numbers of the other two types, each would have a binomial distribution with parameter $1/2$. Examining the data according to this formulation showed no differences at the 5% significance level between images seen first and those seen second, whether the



(B.1) (B.2)

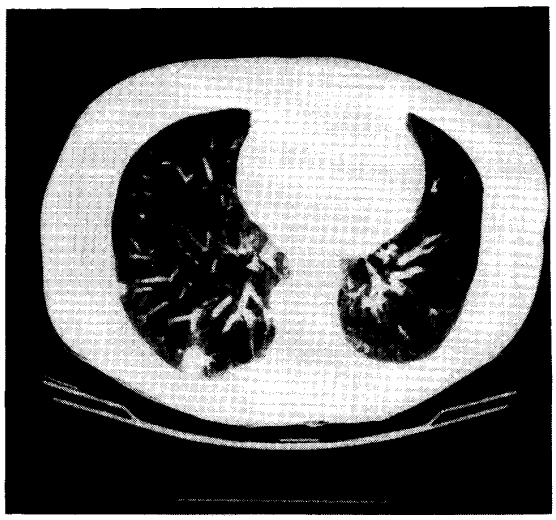
(B.3) (B.4)

Lung nodule images. (B.1) Original lung nodule image. The black arrows indicate the three tumors of the gold standard. (B.2) Lung nodule image compressed to 2.66 bpp. (B.3) Lung nodule image compressed to 2.23 bpp. (B.4) Lung nodule image compressed to 1.83 bpp.

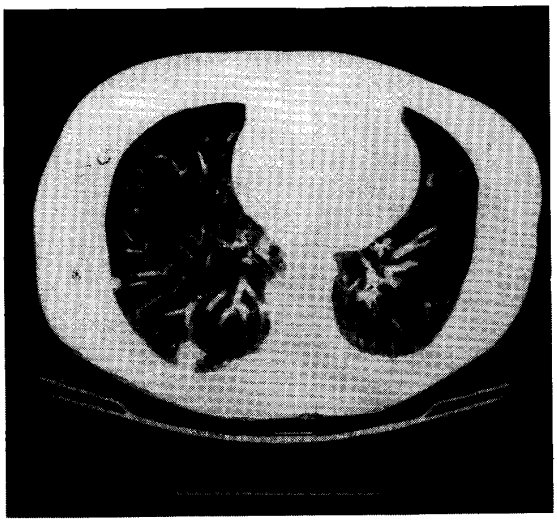
judges were considered separately or pooled together. As an example of the calculation, Judge 1 in evaluating lung nodules over the course of 3 sessions saw 71 pairs of images, in which an image seen at one compression level in a given session is paired with the same image seen at a different level in the same session. Of the 71 pairs, 53 times both images in the pair were judged perfectly, and 5 times both images were judged incorrectly. We concern ourselves with the other 13 pairs: 9 times the image seen first was incorrect while the second one was correct, and 4 times the image seen second was incorrect when the first one was correct. The probability that a fair coin flipped 13 times will produce a heads/tails split at least as great as 9 to 4 is 0.267; thus this result is not significant. Considering both image types with judges pooled or separate, the probabilities ranged from 0.06 to 1.0. In no case was a significant difference found. An analysis of variance (ANOVA) using the actual sensitivity and PVP observations similarly also indicated that page order and session order had no significant effect on the diagnostic result.

VII. Conclusion

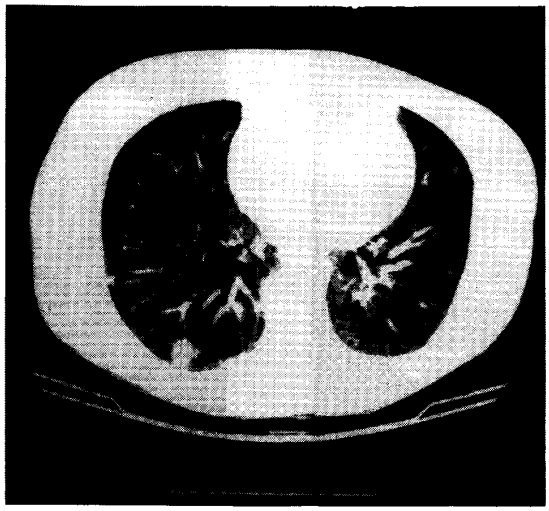
This paper examines diagnostic accuracy testing on compressed images for non-binary diagnostic tasks, and presents statistical tools appropriate for analyzing such data. Some diagnostic detection tasks are genuinely binary, e.g., in examining a chest radiograph for a pneumothorax, on a given half there will always be either 0 or 1 occurrences. ROC



(B.5)



(B.6)



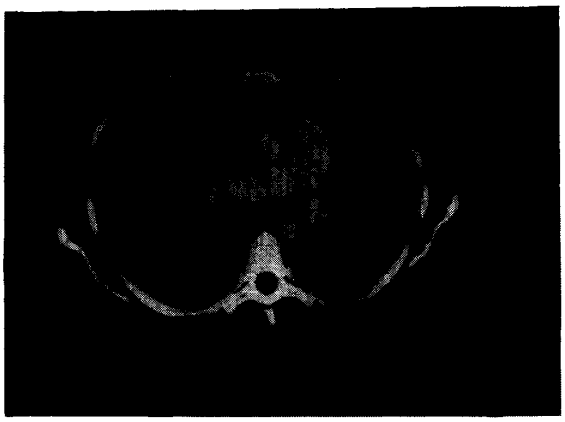
(B.7)

Lung nodule images (continued). (B.5) Lung nodule image compressed to 1.38 bpp. (B.6) Lung nodule image compressed to 1.21 bpp. (B.7) Lung nodule image compressed to 0.58 bpp.

analysis for such binary diagnoses is well-developed, and an extension [37] has been developed for non-binary tasks such as the search for lung nodules and lymphadenopathy presented here. We believe that our approach, including the definition of sensitivity and predictive value positive relative to the gold standard, the fitting of quadratic splines with associated bootstrapped confidence regions, and the pairwise comparison of levels via the Behrens-Fisher t statistic, represents a useful and appropriate new way of analyzing this type of detection task that does not suffer the problems of ROC analysis for this application. In fact, the methods presented here can be used for analyzing data from binary detection tasks as well.

The primary conclusion on diagnostic accuracy is that for the image modality, compression algorithm, and diagnostic tasks considered, the original 12 bpp CT image can be compressed to between 1 and 2 bpp with no statistically significant changes in diagnostic accuracy as measured by sensitivity and PVP. For a variety of reasons outlined above, we believe that our statistical methodology is biased against compression, and so the stated conclusions regarding the implementation of our lossy codes in clinical practice are conservative. A visual comparison of these sensitivity and PVP curves with the ones representing SNR and subjective quality seems to indicate that with increasing compression ratios, distortion increases and subjective quality degrades sooner than diagnostic accuracy falls off. Although not yet a conclusive result, this is certainly a promising sign, as researchers in lossy compression would like to reassure physicians and medical policy makers that

(B.8)



(B.9)

(B.10)

(B.11)

Mediastinal images. (B.8) Original mediastinal adenopathy image. The white arrow indicates the one abnormal lymph node of the gold standard. (B.9) Mediastinal image compressed to 2.75 bpp. (B.10) Mediastinal image compressed to 2.27 bpp. (B.11) Mediastinal image compressed to 1.86 bpp.

diagnoses will not suffer even when the general appearance of the compressed image is perceptibly degraded.

Many studies in the published literature do not find significant differences until rates are well below 1 bpp. This difference from our study is due both to the difference in compression algorithms (which is difficult to verify when proprietary algorithms are used) and to the statistical methods used to verify diagnostic accuracy. Although other compression algorithms may achieve comparable quality at lower bit rates than PTSVQ, the latter does have several potential advantages not explicitly exploited in this study: the ability to incorporate other types of signal processing tasks such as enhancement or classification [35], [20], [36], the progressive transmission capability, and the speed of encoding and decoding without the need for any specialized hardware. However, in any comparison of the diagnostic accuracy results of one study with another, it is essential to include such factors as dynamic range and image size that relate to the chosen modality, as well as differences between the diagnostic tasks undertaken.

We are currently pursuing several areas of improvement in both the algorithmic and methodologic components of this study. The former includes using larger VQ block sizes and improving prediction schemes for larger blocks. Improvements in the methodology should rectify the bias involved in determining a gold standard. Additional information can be gained by examining the vectors of spline coefficients to make various comparisons though we have not reported on such applications here. Different spline techniques, such as smoothing splines, could also be applied to our problems. These various improvements are currently being applied to a study of the effects of compression on a size measurement task.

ACKNOWLEDGMENT

The authors gratefully acknowledge the assistance of D. Brown, M.D., and D. Lentz, M.D., in judging the images. We also thank the anonymous referees for their many helpful comments.

(B.12)

(B.13)

(B.14)

Mediastinal images (continued). (B.12) Mediastinal image compressed to 1.38 bpp. (B.13) Mediastinal image compressed to 1.20 bpp. (B.14) Mediastinal image compressed to 0.57 bpp.

APPENDIX A

A. Bootstrapping Confidence Regions for Spline Fits (PVP or Sensitivity)

1. A quadratic spline equation can be written as $y = a_0 + a_1 x + a_2 x^2 + b_2(\max(0, x - x_0))^2$, where x_0 is the "knot" (in our study, x = bit rate and x_0 = 1.5 bpp). This gives rise to the linear model $\mathbf{Y} = \mathbf{D}\beta + \mathbf{e}$, with one entry of $\mathbf{Y}$ (and corresponding row of $\mathbf{D}$) per observation. $\mathbf{D}$ is the "design matrix." It has four columns, the first having the multiple of a_0 (always 1), the second the multiple of a_1 (that is the bit rate), and so on.
2. Write least squares estimate of β as $\hat{\beta}$ $(= (\mathbf{D}'\mathbf{D})^{-1}\,\mathbf{D}'\mathbf{Y})$.
3. For a given bit rate b, write the row vector dictated by the spline as $\mathbf{d}' = \mathbf{d}'(b)$.
4. The confidence region will be of the form $\mathbf{d}\hat{\beta} - S\sqrt{F}\sqrt{\mathbf{d}'(\mathbf{D}'\mathbf{D})^{-1}\mathbf{d}} \leq y \leq \mathbf{d}\hat{\beta} + S\sqrt{F}\sqrt{\mathbf{d}'(\mathbf{D}'\mathbf{D})^{-1}\mathbf{d}}$,
where S is the square root of the residual mean square from an analysis of variance of the data. So, if $\mathbf{Y}$ is $n \times 1$ and β is $k \times 1$, then
$$S = \sqrt{\frac{1}{n-k}\|\mathbf{Y} - \mathbf{D}\hat{\beta}\|^2}.$$
The region will be truncated, if necessary, so that always $0 \leq y \leq 1$.
5. Construct the empirical distribution $\hat{F}_n$ of the residuals (the coordinates of $\mathbf{Y} - \mathbf{D}\hat{\beta}$).
6. Sample, successively, n times *with replacement* from $\hat{F}_n$, obtaining $\varepsilon_1^*, \cdots, \varepsilon_n^*$. The motivation for this bootstrap sampling is simple: the bootstrap sample (in this case of residuals) bears the same relationship to the original sample (the "true" residuals) that the original sample bears to nature. We do not know the real relationship between the true residuals and nature; if we did, we would use it in judging coverage probabilities in Steps 9 and 10. However, we do know the true residuals themselves, and so we can imitate the relationship between the true residuals and nature by examining the observed relationship between a sample from the true residuals and themselves [28].

7. Construct the fictitious bootstrap data $\mathbf{Y}^*$, whose i^{th} coordinate is $\mathbf{d}_i'\hat{\beta}+\varepsilon_i^*$, where $\mathbf{d}_i'$ is the ith row of $\mathbf{D}$. The bootstrap process will be carried out $n_b = 1000$ times.
8. For the jth bootstrap sample compute a new $\hat{\beta}$ and S; denote them, respectively, by $\hat{\beta}_{B,j}$ and $S_{B,j}$.
9. Compute, for each $\sqrt{F}$,

$$\begin{aligned}\hat{G}_B(\sqrt{F}) &= (n_b)^{-1}\{\#j : \mathbf{d}'\hat{\beta} - S_{B,j}\sqrt{F}\sqrt{\mathbf{d}'(\mathbf{D}'\mathbf{D})^{-1}\mathbf{d}} \\ &\qquad \le \mathbf{d}'\hat{\beta}_{B,j} \le \mathbf{d}'\hat{\beta} + S_{B,j}\sqrt{F}\sqrt{\mathbf{d}'(\mathbf{D}'\mathbf{D})^{-1}\mathbf{d}}\ \forall \mathbf{d}\} \\ &= (n_b)^{-1}\{\#j : (\hat{\beta}_{B,j} - \hat{\beta})'(\mathbf{D}'\mathbf{D})(\hat{\beta}_{B,j} - \hat{\beta}) \le FS_{B,j}^2\}\end{aligned}$$

   Note that the latter expression is what is used in the computation. This is the standard Scheffé method, as described in [29].
10. For a 100p% confidence region compute $(\sqrt{F})_p = \min\{\sqrt{F} : \hat{G}_B(\sqrt{F}) \ge p\}$ and use that value in the equation in step 4.

APPENDIX B

The images [(B.1)–(B.14)] in this appendix are examples of both lung and mediastinum at all bit rates.

REFERENCES

[1] 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 Med. Imaging*, vol. 767, pp. 298–305, 1987.

[2] R. 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*, vol. 4, pp. 945–953, 1987.

[3] 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.

[4] 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.

[5] 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), pp. 458–473, Springer-Verlag, July 1991.

[6] 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.

[7] 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.

[8] 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, Feb. 1992.

[9] 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.

[10] J. Chen, M. Flynn, B. Gross, and D. Spizarny, "Observer detection of image degradation caused by irreversible data compression processes," in *Proc. Medical Imaging V: Image Capture, Formatting, and Display*, vol. 1444, pp. 256–264, SPIE, 1991.

[11] T. Lookabaugh, E. A. Riskin, P. A. Chou, and R. M. Gray, "Variable rate vector quantization for speech, image, and video compression," *IEEE Trans. Commmunications*, vol. 41, pp. 186–199, Jan. 1993.

[12] E. A. Riskin, T. Lookabaugh, P. A. Chou, and R. M. Gray, "Variable rate vector quantization for medical image compression," *IEEE Trans. Medical Imaging*, vol. 9, pp. 290–298, Sept. 1990.

[13] E. A. Riskin, "Variable Rate Vector Quantization of Images," Ph.D. dissertation, Stanford University, 1990.

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

[15] R. A. Olshen, P. C. Cosman, C. Tseng, C. Davidson, L. Moses, R. M. Gray, and C. Bergin, "Evaluating compressed medical images," in *Proc. Third Int. Conf. on Advances in Communication and Control Systems (COMCON III)*, (Victoria, B.C., Canada), pp. 830–840, Oct. 1991.

[16] H. C. Davidson, C. J. Bergin, C. Tseng, P. C. Cosman, L. E. Moses, R. A. Olshen, and R. M. Gray, "The effect of lossy compression on diagnostic accuracy of thoracic CT images," presented at the 77th Scientific Assembly of the Radiological Society of North America, Chicago, IL, Dec. 1991.

[17] E. A. Riskin and R. M. Gray, "A greedy tree growing algorithm for the design of variable rate vector quantizers," *IEEE Trans. Signal Process.*, vol. 39, pp. 2500–2507, Nov. 1991.

[18] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.

[19] P. C. Cosman, K. L. Oehler, A. A. Heaton, and R. M. Gray, "Tree-structured vector quantization with input-weighted distortion measures," in *Proc. Visual Communications and Image Processing '91*, vol. 1605, (Boston, MA), pp. 162–171, SPIE, Nov. 1991.

[20] K. Oehler, P. C. Cosman, R. M. Gray, and J. May, "Classification using vector quantization," in *Proc. Twenty-Fifth Annual Asilomar Conf. on Signals, Systems, and Computers*, (Pacific Grove, CA), pp. 439–445, Nov. 1991.

[21] P. A. Chou, T. Lookabaugh, and R. M. Gray, "Optimal pruning with applications to tree-structured source coding and modeling," *IEEE Trans. Information Theory*, vol. 35, pp. 299–315, Mar. 1989.

[22] M. Powell, *Approximation theory and methods*. Cambridge, England: Cambridge Univ. Press, 1981.

[23] C. Stone, "Large-sample inference for log-spline models," *Ann. Statist.*, vol. 18, pp. 717–741, 1990.

[24] C. Kooperberg and C. Stone, "A study of logspline density estimation," *Comp. Statist. Data Anal.*, vol. 12, pp. 327–347, 1991.

[25] C. E. Metz, "Basic principles of ROC analysis," *Seminars in Nuclear Medicine*, vol. {VIII}, pp. 282–298, Oct. 1978.

[26] J. A. Swets, "ROC analysis applied to the evaluation of medical imaging techniques," *Investigative Radiology*, vol. 14, pp. 109–121, Mar.–Apr. 1979.

[27] M. Weinstein and H. Fineberg, *Clinical Decision Analysis*. Philadelphia: W. B. Saunders, 1980.

[28] B. Efron, *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics, 1982.

[29] R. Miller, Jr., *Simultaneous Statistical Inference*, 2nd ed. New York: Springer-Verlag, 1981.

[30] P. Armitage, *Statistical Methods in Medical Research*. Oxford: Blackwell Scientific, 1971.

[31] E. Lehmann, *Testing Statistical Hypotheses*, 2nd ed. New York: Wiley, 1986.

[32] P. C. Cosman, "Perceptual aspects of vector quantization," Ph.D. dissertation, Stanford University, 1993.

[33] P. C. Cosman, H. C. Davidson, C. J. Bergin, C. Tseng, R. A. Olshen, L. E. Moses, E. A. Riskin, and R. M. Gray, "The effect of lossy compression on diagnostic accuracy of thoracic CT images," *Radiology*, to appear, 1994.

[34] I. McNemar, "Note on the sampling errors of the differences between correlated proportions of percentages," *Psychometrika*, vol. 12, pp. 153–157, 1947.

[35] P. C. Cosman, E. A. Riskin, and R. M. Gray, "Combining vector quantization and histogram equalization," *Information Processing and Management*, vol. 28, pp. 681–686, Nov.–Dec. 1992.

[36] K. Oehler and R. M. Gray, "Combining image classification and image compression using vector quantization," in *Proc 1993 IEEE Data Compression Conference (DCC)*, Snowbird, UT, Mar. 1993, pp. 2–11.

[37] D. P. Chakraborty and L. H. L. Winter, "Free response methodology: Alternate analysis and a new observer-performance experiment," *Radiology*, vol. 174, pp. 873–881, 1990.