LOSSY COMPRESSED MEDICAL IMAGES AND MEASUREMENT ACCURACY

BY

SHARON M. PERLMUTTER  PAMELA C. COSMAN
CHIEN-WEN TSENG  RICHARD A. OLSHEN  ROBERT M. GRAY
KING C. P. LI  COLLEEN J. BERGIN

TECHNICAL REPORT NO. 185
OCTOBER 1996

PREPARED UNDER THE AUSPICES
OF
PUBLIC HEALTH SERVICE GRANT
5 R01 CA59039-20

DIVISION OF BIOSTATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Lossy Compressed Medical Images and Measurement Accuracy

By

Sharon M. Perlmutter, Pamela C. Cosman, Chien-Wen Tseng
Richard A. Olshen, Robert M. Gray, King C. P. Li, Colleen J. Bergin

Technical Report No. 185
October 1996

Prepared Under the Auspices

Of

Public Health Service Grant
5 R01 CA59039-20

Division of Biostatistics
Stanford University
Stanford, California
Lossy Compressed Medical Images and Measurement Accuracy

Sharon M. Perlmutter Pamela C. Cosman Chien-Wen Tseng
Richard A. Olshen Robert M. Gray King C. P. Li Colleen J. Bergin *

October 18, 1996

Abstract

Lossy image compression can significantly improve the efficiency of transmission and storage of medical images, and it holds promise for speeding image processing to assist in screening and diagnosis. Because lossy compressed images do not perfectly reproduce the original image, however, it is necessary to quantifiably demonstrate their quality and utility in clinical applications. The fundamental ideas of lossy signal compression are presented and exemplified by a scheme in which the complexity of implementation and performance are balanced: predictive pruned tree-structured vector quantization of image pixel blocks, a simple progressive compression algorithm based on ideas from statistical clustering and classification trees. Methods are developed for quantifying the effects of lossy compression of magnetic resonance images on the accuracy of measurement of blood vessels in the chest. They are applied to data from a clinical experiment in which the proposed algorithm was used. Thirty scans were compressed to five different levels and three radiologists measured the diameters of the four principal blood vessels on each image. Data are displayed graphically with plots of percent measurement error as a function of bit rate, and they are analyzed quantitatively. Both error relative to an independent standard and personal performance on uncompressed images were considered. We conclude that for the purpose of measuring blood vessels in the chest, there is no significant difference in measurement accuracy when images are compressed up to 16:1 with our algorithms.

* S. M. Perlmutter is with Johnson-Grace Co., Newport Beach, Calif. P. C. Cosman is with the Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92093-0407. C. Tseng is attending University of California, Los Angeles Medical School. She was formerly with the Department of Electrical Engineering at Stanford University. R. A. Olshen is with the Division of Biostatistics of the Stanford University School of Medicine and, by courtesy, the Department of Statistics and the Department of Electrical Engineering of Stanford University, Stanford, CA 94305. R. M. Gray is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305. K. C. P. Li is with the Department of Diagnostic Radiology and Nuclear Medicine, Stanford University School of Medicine, Stanford, CA 94305, and C. J. Bergin is with the Department of Radiology, University of California, San Diego, La Jolla, CA 9203. This work was supported by the National Institutes for Health under Grants CA49697-02, 5 RO1 CA55325, and 5 RO1 CA59039-20, and by the National Science Foundation under Grant DMS-9101528 and a postdoctoral fellowship.
Key Words: lossy image compression, measurement accuracy, image quality, evaluation

1 Introduction

Recent years have seen an increasing use of imaging technologies such as magnetic resonance imaging (MRI) and computerized tomography (CT) that render images in digital format. Although such digital images are usually reduced to hard copy analog form for display, their digital properties hold promise for a significant expansion of their uses in diagnosis. These images are amenable to signal processing, including image enhancement and classification, the combining of images obtained from different modalities, 3-D modeling, and motion video. One fundamental difficulty in working with digital medical images, however, is the size of individual files involved (CT: 0.5 Mb, MR: 0.13 Mb, digitized X-rays: 8 Mb); this difficulty is compounded by the immense number of images produced. Each year Stanford University Medical Center alone generates some 423,200 Mb of CT and MR images that require digital storage on magnetic tape. Owing to space constraints, the tapes are erased and reused after two years; and images remain available only in analog hardcopy format if at all. If the digital properties of medical images are to be exploited, it is clear that compression will be needed to allow longterm and cost-efficient storage, as well as rapid access and transmission.

When a standard lossless (reversible) compression scheme such as a Lempel-Ziv algorithm is applied to MR and CT scans, typically compression ratios of about 2:1 are achieved. Recent studies have shown that with more complex lossless compression, compression ratios of 3:1 or 4:1 are possible [1, 2, 3, 4]. Lossless compression alone is in general insufficient to attain ratios better than 4:1. It is then natural to turn to schemes for lossy (irreversible) compression that have provided excellent results on non-medical images. One goal of this paper is to survey the basic theory and algorithmic ideas underlying lossy compression, especially the tradeoffs between common engineering measures of image quality, the bit rate required for transmission and storage, and the complexity of implementing the compression algorithms. The basic problem formulation and many of the techniques used to design compression systems directly parallel ideas in statistical classification and regression, and these parallels have proved useful in designing simple and effective codes.

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:

- **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, thereby 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. The transformed data may provide a useful data structure.

- **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 image compression systems have been proposed in the literature and incorporated into standards and software and hardware products. These differ primarily by the different choices made for three basic components: signal decomposition, quantization, and lossless coding. A variety of systems and algorithms for compression are described to provide context, but the method chosen for the current study is a compromise of a variety of considerations. The algorithm used was predictive pruned tree-structured vector quantization [14, 15, 16, 17], involving fast encoding and decoding, and providing additional advantages such as simple progressive transmission and potential incorporation of other signal processing techniques such as classification [18, 19]. The algorithm does not perform a signal decomposition such as a DCT or wavelet and that produces directly a variable
length code without separate entropy coding. Our reasons for this are threefold. The first is simplicity, by which is meant to use a compression algorithm that operates directly on the individual pixels and produces a variable rate bitstream without the need to compute transforms and inverse transforms or to do separate entropy coding. In particular, this leads to a simple decompression algorithm that depends mostly on table lookups with few arithmetic operations. The second reason is that our emphasis in this work is on judging the quality and utility of lossy compressed medical images, and the protocol for evaluating quality does not depend on the compression algorithm used. The third is that the tree-structured algorithms used inherently provide a natural progressive structure to the code, which incorporates the ability for progressive reconstruction of an improved image as bits arrive.

The idea is to code an information source, such as a sequence of pixel blocks making up an image, into a sequence of binary integers or bits, which can then be decoded or decompressed to reproduce the original source with the best possible fidelity. The goal is to have the best possible fidelity between the reproduction and original subject to a constraint on the average number of bits transmitted or stored.

With medical images, however, the common engineering measures of quality such as signal-to-noise ratio are insufficient; in medical applications the primary concern is that the diagnostic accuracy of the lossy compressed images remain not less than that of the original images. Signal-to-noise ratios or mean squared error (MSE) may or may not be a good indication of diagnostic accuracy, but the accuracy itself must be demonstrated. Beyond diagnostic accuracy, however, the images must appear nearly identical to the originals or the radiologists will not use them no matter their other features. A wide variety of diagnostic tasks must be studied, including measurement of structures, detection of lesions, and interpretation of texture. We have provided and implemented protocols for experimentation by which the diagnostic accuracy of radiologists who make use of images, compressed or not, can be quantified.

Most previous studies have focused on the effects of lossy compression on detection tasks [5, 6, 7, 8, 9, 10]. We are not aware of any other studies on the effects of lossy compression on the accuracy of measurement. Measurement tasks on structures such as blood vessels, other organs, and tumors take a central role in the evaluation of aneurysms, especially prior to surgery. Abdominal aortic aneurysms are evaluated routinely with ultrasound; and thoracic aortic aneurysms are evaluated by CT or MRI. The aortic diameter is usually measured manually with calipers. A measured value of diameter in excess of 4 cm entails a diagnosis of aneurysm. A larger aneurysm carries a greater risk of rupture. About 10% of those between 5 and 10 cm in diameter and about 50% of those with values greater than 10 cm do eventually rupture [11]. Because rupture is invariably fatal, measured values more than 5 or 6 cm indicate operative repair [12, 13]. Of course the clinical decision depends not only on the size of the aneurysm but also on the clinical status of the patient (especially as concern pain and hemodynamic instability). Dilation less than 5 cm in diameter may be followed conservatively by serial MR imaging studies at 6-month intervals. Observing an increase in the aortic diameter of 0.5 cm over the course of a 6 month interval would be
indication for surgical repair. Comparison films are imperative for appropriate management of these patients.

The goal of the study reported here was to quantify the effects of lossy compression on measurement accuracy through experiments that follow closely the clinical tasks of radiologists evaluating aortic aneurysms. We wished to examine whether compression maintains the information required for accurate measurements, or whether it leads to inaccuracies by blurring edges or distorting structures. The task to be studied is the measurement of four primary blood vessels in the mediastinum: the ascending aorta, descending aorta, right pulmonary artery (RPA), and superior vena cava (SVC). Clearly, if compression at a certain bit rate caused a 0.5 cm error in the aortic measurement, it would be unacceptable. Although we focused on the medical problem of thoracic aortic aneurysms as seen on MR scans, the methodology developed in this research is broadly applicable to any medical task requiring the measurement of structures. Our project is divided into four general areas as follows:

- Establishing a protocol for obtaining measurements and subjective scores in a clinical setting.
- Establishing a gold standard for the "correct vessel sizes," selecting parameters for quantifying measurement error, and choosing a set of statistical methods to determine to what bit rate these images can be compressed without loss in measurement accuracy.
- Determining whether subjective scores and measurement error vary similarly with decreasing bit rates.
- Comparing the effects of lossy compression on measurement accuracy to other diagnostic tasks such as tumor detection.

A set of 9-bit original MR chest images containing aneurysms and normal vessels was compressed to five bit rates between 0.36 and 1.7 bits per pixel (bpp). Figure 2(a) shows an original 9.0 bpp MR chest scan. Figure 2(b) shows the same image compressed to 1.14 bpp and Figure 2(c) shows the image compressed to 0.36 bpp. Radiologists measured the four vessels on each image. As a separate task, the radiologists also rated the subjective quality of each image by assigning a score of 1 (worst) - 5 (best) to each image.

In our statistical analyses, we set two gold standards, a "personal" one [6, 7], and an "independent" one. These are two methods of establishing the correct size of each blood vessel, that is, the underlying "truth" of each image. The personal gold standard is derived for individual radiologists based on their own measurements of the same image at the uncompressed level. If there is random error in the measurement process, since the personal gold standard defines the measurements on the originals to be correct (for that image and that judge), the compressed images cannot be as good as the originals according to this standard. For this reason, we also defined an independent gold standard. This is based on the consensus measurements taken by two radiologists on the original images.
Figure 2: (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
These two radiologists are different from the three radiologists whose judgments are used to determine diagnostic accuracy. This does introduce interobserver variability into that portion of the analysis, but it also allows the original images to be compared fairly with the compressed ones.

For each of these gold standards, we quantify the accuracy of the measurements at each compression level by taking the percent measurement error for each image, defined to be the difference between a radiologist’s measurement and the gold standard, scaled by the gold standard measurement. This error is reported as a function of bit rate. Other parameters such as subjective scores and signal-to-noise ratios (SNRs) are also analyzed as functions of bit rate. Variabilities of the measurements by (judge, image) pairs are quantified by two-way analyses of variance in which the effects are level of compression and structure, and there was also a (one degree of freedom) term for non-additivity. These ANOVAs are thought of as descriptive statistics, and they are summarized by various box plots. Evidently structures are somewhat different, while judges are very different.

2 Signal Compression

We review the basic notions of sources, codes, fidelity, and optimal performance and describe both general and specific compression systems, including the particular algorithm — predictive tree-structured vector quantization — emphasized in the image quality experiments considered here.

Source Coding

The Shannon model for a compression system is a source code with a fidelity criterion [20, 21]. The source to be coded, \( \{X(n); n \in \mathcal{Z}\} \), is considered a random process, where \( \mathcal{Z} \) is the integers. The \( X(n) \) are assumed to take values in \( k \)-dimensional Euclidean space as described by a stationary probability distribution \( P_X \). This assumption is never exactly right for our applications and it is made primarily for convenience. Our algorithms do not depend in any explicit way upon stationarity, however, and the underlying mathematics all extends to more general models such as asymptotically mean stationary processes [22]. This distribution might be parametric, but in practice we usually work with empirical distributions estimated from a training or learning set \( \mathcal{L} = \{x_i; l = 1, \ldots, |\mathcal{L}|\} \) by

\[
P_{\mathcal{L}}(G) = \frac{1}{|\mathcal{L}|} \sum_{x \in \mathcal{L}} 1(x \in G)
\]

for any event \( G \), where \( 1(x \in G) \) is an indicator function that equals 1 if \( x \in G \) and equals 0 otherwise. If the process is assumed to have suitable stationarity properties (e.g., to be asymptotically mean stationary), then these empirical distributions will stabilize as \( |\mathcal{L}| \) grows.

The dimension \( k \) is a parameter of the particular application. Shannon information theory [20, 21] indicates that improved performance can be achieved using larger vector
dimensions at the expense of added complexity in terms of memory and computation. As our example of primary interest, the vectors are rectangular blocks of pixel intensities within a sampled image and the distribution will be an empirical distribution based on a learning set.

A source code or compression code for the source \( \{X(n)\} \) consists of a pair \((\alpha, \beta)\) of encoder and decoder. An encoder \( \alpha : A \to \{0,1\}^* \) is a mapping from the input alphabet \( A \) (typically a subset of \( \mathcal{R}^k \)) into the set of all binary sequences of finite length. The random vector \( W \equiv \alpha(X) \) is discrete with a probability mass function (pmf) \( p_W(w) = P_X(\{x : \alpha(x) = w\}) \). Of particular importance is the range space \( \mathcal{W} \equiv \alpha(A) \), which we refer to as the channel codebook. This is the set of binary sequences that are stored in a digital storage medium or are transmitted from the transmitter to the receiver via a digital communication link. In order to ensure that a sequence of symbols (variable length binary vectors) drawn from the channel codebook can be decoded uniquely if the starting point is known, we require that the \( \mathcal{W} \) be prefix-free or satisfy the prefix condition: no word in the codebook is a prefix of any other word in the codebook. It is a standard exercise in information theory to demonstrate that any uniquely decodable channel codebook can be made into a channel codebook with the same codeword lengths that also satisfies the prefix condition, and hence no essential generality is lost by the assumption.

The decoder \( \beta : \{0,1\}^* \to C \) is a mapping from the space of finite-length binary sequences onto a set \( C \equiv \{\beta(w); w \in \mathcal{W}\} \) called the reproduction codebook, with members called reproduction codewords or templates. The members of \( C \) are chosen from a reproduction alphabet \( \hat{A} \) which typically, but not always, is simply the input alphabet, \( A \). For a given encoder we care about the definition of \( \beta(w) \) only for \( w \in \mathcal{W} \). It can be defined arbitrarily outside this set.

This model of a compression system is general in the sense that it includes any code operating on disjoint blocks of data functionally independent of past or future coding operations. In other words, it models codes that have no memory of previous vectors or anticipation of future vectors. These codes are sometimes referred to as block source codes to distinguish them from codes that can vary the dimension of input blocks or that can operate on overlapping input blocks in a "sliding-block" fashion.

Tree-structured Codebooks

Any binary prefix-free channel codebook \( \mathcal{W} \) can be depicted by a binary tree as shown in Figure 3. The binary tree \( \mathcal{T} \) consists of nodes connected by branches. Each branch is labeled by a 1 or 0. We arbitrarily place the root node at the left (or top) and grow the tree to the right (or down). Equally arbitrarily, we label the upper branch emanating from a node by 1 and the lower branch by 0. Given a tree, every node is uniquely determined by a finite length binary sequence giving the pathmap from the root node to the given node. If the node is \( n \in \mathcal{T} \), then the binary pathmap from the root to \( n \) is denoted \( b(n) \). The length of \( b(n) \) is \( l(b(n)) \), which is also the depth of the node within the tree. The pathmap of the root node is the empty string, which has length 0.

Given a prefix-free channel codebook \( \mathcal{W} \), the corresponding tree is constructed by tracing
Figure 3: A Binary Code Tree
out all the corresponding paths in the tree and terminating the tree at the nodes where the
codewords of \( \mathcal{W} \) end. These nodes are terminal nodes or leaves of the tree. Thus the binary
pathmaps to the terminal nodes constitute the channel codebook, and the index set \( \mathcal{W} \) can be
used to index the terminal nodes.

The leaves of a tree \( \mathcal{T} \) are denoted by \( \mathcal{T} \). Any node of a tree that is not a leaf is called
an internal node of the tree. Thus each codeword determines the leaf and, conversely,
knowing the leaf determines the codeword by tracing the path from the root node to the
leaf. Equivalently, \( \mathcal{W} = \{ b(n); \ n \in \mathcal{T} \} \).

Any binary tree of this type with a finite (or countable) number of leaves corresponds
to a binary prefix-free channel codebook since no leaf can be an internal node of the tree.
Thus every prefix-free binary channel codebook is equivalent to a binary tree-structured
channel codebook.

Every input vector \( x \) will yield through \( \alpha \) a channel codeword that can be viewed as a
binary pathmap
\[
\alpha(x) = (\alpha(x)_0, \alpha(x)_1, \ldots, \alpha(x)_{l(\alpha(x)) - 1})
\]
through the tree, where \( l(\alpha(x)) \) is the length of the binary vector and hence the depth in the
tree of the corresponding leaf, and where \( \alpha(x)_k \) is the binary label of the \( (k + 1) \)st branch
encountered in the pathmap, \( k = 0, \ldots, l(\alpha(x)) - 1 \).

A tree \( \mathcal{T}_1 \) is a subtree of another \( \mathcal{T}_2 \) and we write \( \mathcal{T}_1 \subset \mathcal{T}_2 \) if the two trees share a root
node and all nodes of the smaller tree. Leaves of the smaller tree can be internal nodes or
leaves of the tree that contains it.

Two nodes in the same level emanating from a common node in the previous level will
be called siblings. The previous node will be called a parent of the siblings and the siblings
are children of the parent. A node reachable as a sequence of children from another is called
a descendant of the ancestor.

Associated with every node \( n \in \mathcal{T} \) is a set of all input vectors \( x \in A \) with pathmaps to
that node:
\[
S_n = \{ x : (\alpha(x)_0, \ldots, \alpha(x)_{l_n}) = b(n) \}.
\]  
(1)

Each internal node has associated with it a binary partition \( B_n = \{ B_n(0), B_n(1) \} \) dividing
the vectors \( S_n \) that can reach that node into those that follow the 0 branch to one child
node or the 1 branch to the other. This provides the binary node test for that node. If the
current path map is \( b(n) \), then the pathmap to the next node will be \( (b(n), 1(x \in B_n(1))) \).

In general this binary test can be constructed by working backwards from the description of
a source code: each input vector is mapped into a binary pathmap by the source encoder.
The binary test at the root node sends any vector having a 0 as the first symbol in its
binary pathmap to the child node reached via the branch labeled 0. Similarly, at any node
the binary test is determined by the relevant coordinate of the overall path map of the
input vector. So far this is simply a change of perspective. It does not necessarily provide a
better method of implementing an encoder. That will happen if the binary node tests can
be made simple.
Quality vs. Cost

To measure the fidelity or lack thereof between an input vector and its reproduction, we assume that we have a distortion measure \( d(x, y) \geq 0 \) defined for every possible \( x, y \). \( d(x, \hat{x}) \) measures the distortion or loss resulting if an original input \( x \) is reproduced as \( \hat{x} \). The overall goal of a compression system is to keep distortion and bit rate small. The distortion measure need not be a metric, but ideally it should possess the following properties.

- It should be easy to compute so that the distortion can be monitored easily.
- It should be tractable for theoretical analysis.
- It should be meaningful in the desired application, e.g., large or small average distortion should correspond to an image that looks bad or good, respectively, in an entertainment application, or to an image that lends itself poorly or well to further analysis, e.g., recognizing tumor tissue in medical images.

No single distortion measure accomplishes all of these goals, although the ubiquitous squared error distortion defined by

\[
d(x, y) = ||x - y||^2 = \sum_{l=0}^{k-1} |x_l - y_l|^2,
\]

where \( x = (x_0, x_1, \ldots, x_{k-1}) \), accomplishes the first two goals and occasionally correlates with the third. Unfortunately, distortion measures that have been found to be good measures of perceived quality in speech and images have proved to be quite complicated and have usually violated the first two properties. As a result, the squared error distortion has dominated the literature with relatively rare extensions to other measures.

The distortion resulting from applying a source code \((\alpha, \beta)\) to a specific input vector \( x \) is \( d(x, \beta(\alpha(x))) \). A code \((\alpha, \beta)\) will be said to be lossless if \( \beta(\alpha(x)) = x \), for all inputs \( x \in A \). For a lossless code, \( \beta \) is the inverse of \( \alpha \); in fact

\[
\beta = \alpha^{-1} \quad \text{iff code is lossless.} \tag{2}
\]

Given that \( d(x, y) > 0 \) if \( x \neq y \), the code is lossless iff \( d(x, \beta(\alpha(x))) = 0 \), \( x \in A \). Lossless codes are also called invertible codes, noiseless codes, or entropy codes.

A code that is not lossless is lossy and it is usually called a quantizer. Since the alphabet is in general a \( k \)-dimensional vector, the more common notion is a vector quantizer or VQ. The overall mapping \( Q: A \rightarrow C \) defined as the composite \( Q(x) = \beta(\alpha(x)) \) is often referred to as a VQ, and the term is also used somewhat more generally to denote any mapping from a continuous or large discrete space into a smaller space. The smaller space is usually required to be finite, but sometimes it is useful to allow it to be countably infinite.

Lossless codes are an important special case of compression codes. When compressing a computer program or bank statement, for example, it is critically important to make no errors. A single wrong digit could have catastrophic implications. We argue that in some
examples, such as medical and scientific images, lossy compression may be quite useful even though the utility of the images is strongly related to the quality of the reproductions.

The “cost” of encoding an input vector \( x \) in terms of the memory occupied by the stored channel codeword or the communications channel capacity required for its transmission is given by the length of the encoded input vector \( \alpha(x) \) in binary symbols, which we denote by \( l(\alpha(x)) \). This quantity is also referred to as the instantaneous rate \( r(x) = l(\alpha(x)) \) in bits per input vector. It is convenient to normalize both distortion and rate into units per input symbol by dividing by the dimension \( k \) of the input vectors and to report \( r \) in terms of bits per input symbol. Obviously the distortion resulting from encoding an input vector depends on the encoder and decoder, while the instantaneous rate depends only on the encoder.

The performance of a compression system is measured by the expected values of the distortion and rate. The average distortion corresponding to a specific source code on a specific source is denoted

\[
D(\alpha, \beta) = E[d(X, \beta(\alpha(X)))]
\]

It is often reported in logarithmic form as a signal to noise ratio \( \text{SNR} = 10 \log_{10}(D_0/D) \) dB, where \( D_0 \) is a reference value, often the average distortion resulting from the optimum zero rate code, which is the variance of the input vector in the case of a squared error distortion measure. Other normalizations are used, including the noncentral second moment (energy) and the square of the maximum possible input value (yielding what is called a “peak SNR” (PSNR)). The average rate is defined as

\[
R(\alpha) = E(r(X)) = E(l(\alpha(X)))
\]

in bits per vector. This is usually normalized by dividing by the dimension \( k \), which provides average bits per symbol (or pixel in the case of sampled images). The distribution with respect to which expectations are computed is always determined by context, though in practice it can only be a suitable empirical distribution.

Every source code operating on a source will be characterized by a point in the two-dimensional rate-distortion plane: \( (R(\alpha), D(\alpha, \beta)) \). Both \( D(\alpha, \beta) \) and \( R(\alpha) \) can be considered as cost functions for a source code: all else being equal, one code is better than another if it has smaller \( D \) or smaller \( R \) than another. Thus if we were to consider a plot of all achievable distortion-rate pairs, the only points of practical interest would be those not dominated by any other points in the sense of having a smaller \( D \) (or \( R \)) given the same or smaller \( R \) (or \( D \)). This leads to the definitions of optimal source codes, just as it does to admissibility in statistics.

**Optimal Source Codes**

The goal of source coding is to characterize the optimal tradeoff between average bit rate and average distortion and to design codes that compare well to the optimal performance. The optimization problem can be formulated in several ways. The distortion-rate approach constrains the average rate and minimizes the average distortion; the rate-distortion approach
constrains the average distortion and minimizes the average rate; and the Lagrangian formulation minimizes a weighted sum of distortion and rate. All formulations are analogous to those that bear upon the construction of statistical tests, in Neyman-Pearson style for the first two and as Bayesian tests for the third. In the first approach we define the optimal code (if it exists) for a given rate $R$ as the $(\alpha, \beta)$ minimizing $D(\alpha, \beta)$ subject to $R(\alpha) \leq R$. Define the operational distortion-rate function

$$\hat{D}(R) = \inf_{\alpha, \beta: R(\alpha) \leq R} D(\alpha, \beta).$$

$\hat{D}(R)$ is obviously monotonically nonincreasing in $R$. The rate-distortion function is defined by reversing the roles of rate and distortion. The Lagrangian approach incorporates the rate constraint into a modified distortion measure $\rho$ defined between the input vector and the channel codeword by

$$\rho(x, i) = d(x, \beta(i)) + \lambda(i),$$

for a Lagrange multiplier $\lambda$, so that the average modified distortion measure is given by

$$E[\rho(X, \alpha(X))] = D(\alpha, \beta) + \lambda R(\alpha).$$

The constrained minimization of $D$ is equivalent to an unconstrained minimization of $E\rho$. We focus on the Lagrangian formulation as the more natural for the problem at hand.

The case of 0 rate code is of course artificial, but it is useful as a step towards describing optimality properties for the general case. In order to achieve the minimum possible average distortion with a 0 rate codebook having a single word $y$, a $y$ yielding the minimum average distortion (if such exists) should be chosen as the output of $\beta$ operating on the empty string $\emptyset$. Thus $\beta(\emptyset) = \arg\min_{y \in A} E[d(X, y)]$. If the minimum indeed exists, we have

$$\hat{D}(0) = \min_y E[d(X, y)].$$

The vector achieving this minimum will be called the centroid of the alphabet $A$ with respect to the distribution $P_X$. If the average distortion is given by the squared error $E[||X - y||^2]$, then this is the expected value $EX$ since for any $y$

$$E[||X - y||^2] \geq E[||X - EX||^2].$$

If the distribution is an empirical distribution described by the training sequence, then this is simply the sample mean or Euclidean centroid

$$\frac{1}{L} \sum_{n=1}^{L} x_n.$$

As the selection of the best 0 rate codeword is simply the generalization of the well known fact that the Euclidean centroid minimizes the moment of inertia about a point, we will call the optimal 0 rate codeword the centroid when it exists.
The zero rate result extends easily to describing the optimal decoder in general for a given encoder. Suppose in particular that the tree-structured code representation is used. $S_n$ denotes the set of input vectors that code into a node $n$. If $n$ is a leaf, the best choice for the decoder output given the channel codeword that terminates at node $n$ is the best estimate of the original vector given $X \in S_n$. This is simply the centroid solving the 0 rate problem with the conditional node distribution replacing the full distribution: given a node $n$ in the tree-structured representation of a source code $\alpha$, the best reproduction value $y$ to represent the node support set $S_n$ in the sense of minimizing the average conditional distortion $E[d(X,y)|X \in S_n]$ is $\arg\min_{\delta \in \mathcal{A}} E[d(X,y)|X \in S_n]$. For the squared error distortion, this is $E[X|X \in S_n]$. If $Pr(X \in S_n) = 0$, then the decoder can be defined in an arbitrary fashion.

If $n$ is an internal node, then this optimal decoder output for $X \in S_n$ provides an interim value of the output before all of the bits of the channel codeword have arrived. A code with the ability to provide early reproductions before all of the bits are available is said to be progressive or a successive approximation if, on average, the early reproductions improve with additional bits.

This provides a general optimality condition describing the best decoder for a given encoder, a condition originally formulated by Lloyd for scalar quantization [23]: Given an encoder $\alpha$, the optimal decoder $\beta$ is given by

$$\beta(i) = \arg\min_{y \in \mathcal{A}} E[d(X,y)|\alpha(X) = i]$$

for each $i \in \mathcal{W}$. The optimal decoder for any encoder is also defined for any internal nodes in the tree structured representation, permitting a progressive reconstruction as the bits arrive.

In a similar fashion, one can define an optimal encoder $\alpha$ for a fixed decoder $\beta$ with respect to the Lagrangian distortion measure. Given $\beta$, any encoder $\alpha$ must satisfy the inequality

$$E[\rho(X,\beta(\alpha(X)))] = \int dP_X(x)[d(x,\beta(\alpha(x))) + \lambda(\alpha(X))]$$

$$\geq \int dP_X(x) \min_{i \in \mathcal{W}} [d(x,\beta(i)) + \lambda(\alpha(i))].$$

This lower bound is achievable by the minimum modified distortion encoder

$$\alpha(x) = \arg\min_{i \in \mathcal{W}} [d(x,\beta(i)) + \lambda(\alpha(i))].$$

Thus given the reproduction codebook $\beta$, the optimal encoder is the minimum distortion encoder with respect to the modified distortion measure.

These optimality properties suggest an iterative design algorithm for compression given an initial encoder/decoder pair. Any code can be improved (at least made no worse) by successively applying the above two properties. First optimize the reproduction codebook
(or decoder) for the given encoder, then optimize the encoder for the given reproduction codebook and channel codebook. This iteration is a variation of an algorithm of Lloyd (1957) [23] for the design of optimum pulse coded modulation (PCM) systems, scalar quantizers with fixed rate codebooks, and of the similar methods in statistical clustering such as $k$-means clustering.

Since the distortion is nonnegative and nondecreasing, the algorithm is a descent algorithm. In general the algorithm converges only to a stationary point, and there is no guarantee that the resulting code will be globally optimal. (It is guaranteed to be globally optimal for all codebook sizes if $X$ is univariate and the distribution is absolutely continuous with log concave density [24, 35]. This is equivalent to the distribution being strongly unimodal, that is, for its convolution with every univariate unimodal distribution to be unimodal (see [33]; see also [34]). In addition, neither step changes the basic codebook tree shape, only the labels and the selection rule. It might improve the tree to split one leaf in the tree, allowing for a possible decrease in distortion at the expense of an increase in bit rate, and to prune off a branch, to compensate by a decrease in bit rate but an increase in distortion. This might improve a tree by allowing more bits where they are needed, while not wasting bits where they do not help much. We shall consider a design algorithm that combines all these considerations.

In practice it is often of interest to optimize over a constrained subset of possible codes rather than over all of them. Unconstrained codes may prove difficult or impossible to implement, and added structure may provide gains in practical simplicity that more than compensate for loss of optimality. A useful simplification is to consider a suboptimal encoder that provides a greedy search of the code tree instead of a full search of all leaves to find the minimum modified distortion. In this case each node is considered to be labeled by its optimal reproduction, and the binary decision function becomes a simple comparison of the distortion resulting from the two children nodes available. In other words, the binary node decision is a simple pairwise nearest neighbor selection. Since one bit is added regardless of which node is selected, bits are not explicitly taken into account during encoding. They are taken into account when the tree itself is designed. The decision is a minimum distortion selection between two available reproductions for a given input vector. A code with this structure is called a tree-structured VQ or TSVQ. A TSVQ uses a tree-structured representation of the channel codebook and uses a minimum distortion test as a binary node test. The minimum distortion binary decision is equivalent to a hyperplane test or, in engineering applications, a correlation or matched filter detector. The channel codeword is thus selected by a sequence of simple binary decisions. Vector reproductions are stored at each node in the tree. The search begins at the root node. The encoder compares the input vector to two possible candidate reproductions, chooses the one with the minimum distortion, and advances to the selected node. If the node is not a terminal leaf, the encoder continues and chooses the best available node of the new pair presented. The encoder produces binary symbols to represent its sequence of left/right decisions. The stored index is then a path map through the tree to the terminal node, which is associated with the final codeword.
As a simple example of a TSVQ, consider the labeled tree of Figure 4. Each node is labeled by the reproduction vector used to represent any data coded to that node. Suppose that this tree is used to encode the image of Figure 5, with intensities from an alphabet of size 8 or 3 bits per pixel (bpp).

\[
\begin{array}{c|c|c}
0 & 1 & 2.25 \\
& & 4.0 \\
& & 2.375 \\
& & 5.875 \\
1.75 & 5.75 & \\
2.25 & 5.75 & \\
& & \\
& & 0.5 \\
& & 5.5 \\
& & 1.5 \\
& & 6.5 \\
3.0 & 6.0 & \\
3.0 & 5.0 & \\
& & \\
\end{array}
\]

Figure 4: Simple TSVQ Example

\[
\begin{array}{cccccccc}
0 & 1 & 2 & 3 & 5 & 3 & 1 & 2 & 3 \\
1 & 7 & 2 & 6 & 5 & 5 & 1 & 7 & \\
1 & 5 & 3 & 3 & 3 & 2 & 3 & 7 & \\
2 & 6 & 2 & 6 & 2 & 6 & 4 & 4 & \\
\end{array}
\]

Figure 5: Example Image

If only the root node is used for a zero rate code, then the average squared error resulting will be $D_0 = 7.53$. If the nearest neighbor from the two codewords labeling the first level nodes is chosen for each input vector, then the rate is one bit per input vector (1/4 bpp); and the resulting average squared error is $D = 4.19$, yielding SNR = $10 \log_{10} D_0/D = 2.55$ dB. If the complete tree is used so that the left node in the first level is split and each image vector reaching this node is further encoded by choosing the nearest neighbor label from the children nodes, then the distortion drops to 2.81 and the SNR rises to 4.3 dB, but the cost is an increase in the average bit rate to 1.5 bpp.

A TSVQ has the same form as a classification or regression tree, such as those designed by the CART$^\text{TM}$ algorithm [25]. The squared error of TSVQ becomes Bayes risk in a classification tree, and the cost might be the number of leaves or the total number of nodes in the channel codebook in a classification problem. The Euclidean minimum distortion binary decision rule in a TSVQ typically becomes a one dimensional threshold rule on a single vector coordinate in CART, although one can make the decisions multidimensional by first transforming the input vector, or by allowing more general hyperplane splits.
TSVQ Design

TSVQs, like classification trees, can be designed by means of a gardening metaphor. The trees are first grown from a root node by successively splitting nodes and running a Lloyd algorithm on the new leaf set, which now becomes a 2-means design. As growing may overfit the data, the resulting tree can then be pruned so as to optimally trade off average distortion and rate. This trading of average distortion and average bit rate is simply a variation on the design strategy for classification and regression trees [25] embodied in the CART algorithm. The simplest growing technique is to split every leaf simultaneously to form a new tree. In particular, the root node is split and the Lloyd algorithm run to produce a one-bit tree. Then both nodes are split and the Lloyd algorithm run to produce a two-bit tree with four leaves. The Lloyd algorithm is modified in that it is really two clustering algorithms with two codewords each rather than a single clustering algorithm with four words. Each Lloyd algorithm is run for the conditional distributions of each parent node. Continuing in this way will yield a balanced tree and fixed-rate code.

Alternatively, nodes can be split individually and selectively. For example, every leaf in a code tree has some conditional average distortion, say $D(n)$, which is the average distortion resulting when the input vectors coded to this node are encoded into the centroid of the node. If this node is split and the Lloyd algorithm run on the conditional node distribution, then the two children nodes, $n_0$ and $n_1$, will have conditional distortions $D(n_0)$ and $D(n_1)$ with probabilities $p_0$ and $p_1 = 1 - p_0$, respectively. This will result in a new conditional distortion for input vectors reaching node $n$ of

$$p_0 D(n_0) + p_1 D(n_1) \leq D(n);\quad (8)$$

that is, the average distortion drops due to the node split. On the other hand, all vectors reaching node $n$ will now have an additional bit added to their path map so that the average rate will increase. Thus one strategy for splitting is to split the node that results in the greatest drop in average conditional distortion per average additional bit. This is the most common growing strategy, but it is by no means the only one. For example, one could split the node with the largest contribution to the overall average distortion.

Once grown, the tree can be pruned by removing all descendents of any internal node, thereby making it a leaf. This will increase average distortion, but will also decrease the rate. Once again, one can select for pruning the node that offers the best tradeoff in terms of the least increase in distortion per decrease in bits. It can be shown that, for quite general measures of distortion, pruning can be done in optimal fashion and the optimal subtrees of decreasing rate are nested.

Predictive vector quantization (PVQ)

One method for incorporating memory or context into coding is to predict the current vector based on its neighbors and then quantize the prediction residual [15]. PVQ is a straightforward vector extension of traditional scalar predictive quantization (DPCM). The basic encoder and decoder structures are shown in Figure 6. The encoder makes a prediction
of the incoming vector based on previously encoded vectors. The difference between the actual input vector and its prediction is called the residual vector. This residual is vector quantized. Because the encoder only uses the previously decoded outputs in making its prediction, the decoder is able to make the same prediction. After dequantizing the residual vector, the decoder adds the prediction to it to form the reproduction vector. The prediction is often a simple linear predictor that takes a weighted average of nearby previously encoded coefficients.

![Diagram](image)

**Figure 6:** Encoder and decoder for a predictive vector quantizer

In predictive TSVQ, the residual quantizer is a TSVQ. For each residual vector, 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 it to its prediction of the block. The selection of vector dimension or block size is important in predictive TSVQ. 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 tradeoff 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 the block size with an emphasis on achieving low complexity.

The coefficients for the predictor have in the past been calculated by Wiener-Hopf techniques. They are simple and have worked reasonably 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 prediction coefficients are used with encoded values of adjacent previously encoded blocks rather than with original pixel values.
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.

An advantage of the predictive TSVQ system is that explicit entropy coding is not needed as the code is designed directly to minimize average bit rate. Additional compression could be achieved by not using the natural tree-structured code representation and instead designing an optimal entropy code for the final code indexes. If this is to be done, then better performance could be achieved by designing the original TSVQ to minimize average entropy instead of average length.

An additional advantage is the natural progressive character of the code: on the average distortion diminishes with additional bits of the path map.

Recent work has shown that the Wiener-Hopf technique can be improved upon in some applications by a variation of ridge regression [26, 27]. However, one should not lose track of the fact that the goal is good ultimate codes rather than good prediction for its own sake. One can imagine prediction that is dreadful in an MSE sense, but that makes for trivial encoding of residuals. In statistical terms, bias is not the issue here. Instead, it is the simplicity of the range of the predictor.

3 Study Design

We turn now to a particular clinical experiment we conducted and that was described earlier.

To develop a tree-structured residual encoder and decoder, 20 MR chest scans were picked to be the training set; they included a wide range of normal and aneurysmal vessels. 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 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. To simulate normal clinical practice, test images were selected from 30 sequential thoracic MR examinations of diagnostic quality that were obtained after February 1, 1991. The patients studied included 16 females and 14 males, with ages ranging from 1 to 93 years and an average age of 48.0 ± 24.7 years (mean ± s.d.). Clinical indications for the thoracic scans included suspected aortic aneurysm (11), thoracic tumors (11), evaluation of patients before or after lung transplant (5), constrictive pericarditis (1) and subclavian artery rupture (1). From each examination, one image which best demonstrated all four major vessels of interest was selected. The training images were selected similarly from different examinations. All analyses are based solely on measurements made on the test images.
In our study, the 256 × 256 pixel MR scans were broken into 2 × 4 pixel blocks for encoding and decoding. The coefficients for the Wiener-Hopf linear predictor were calculated from the training set images and used to produce a residual (prediction error) training set. Using predictive TSVQ, a residual tree was grown to 2.25 bpp and pruned back to a set of optimally pruned subtrees representing bit rates from 0 bpp to 2.25 bpp. Five subtrees were chosen to produce average bit rates of 0.36, 0.55, 0.82, 1.14, and 1.70 bpp on the 30 test images and to have no overlap between the bit rates of the image set encoded with two different subtrees. These subtrees and their average bit rates are termed compression levels 1-5. The original scans at 9.0 bpp are termed level 6.

The 30 test scans compressed to 5 bit rates plus the originals give rise to a total of 180 images. These images were arranged in a randomized sequence and presented on separate hardcopy films to three radiologists. The viewing protocol consisted of 3 sessions held at least 2 weeks apart. Each session included 10 films viewed in a pre-determined order with six scans on each film. The radiologists began viewing films at different starting points in the randomized sequence. To minimize the probability of remembering measurements from past images, a radiologist saw only 2 of the 6 levels of each image in each session, with the second level of each image spaced at least 4 films after the first.

Following standard clinical methods for detecting aneurysms, the radiologists used calipers and a mm scale available on each image to measure the four blood vessels appearing on each scan. Although the use of digital calipers might have allowed more accurate measurements, this would have violated one of our principal goals, namely to follow as closely as possible actual clinical practice. It is the standard practice of almost all radiologists to measure with manual calipers. This is especially true for radiologists in private practice, who represent more than 90% of the radiologist population in the U.S. Even in a tertiary referral setting, manual calipers are used routinely. We asked radiologists to make all measurements between the outer walls of the vessels along the axis of maximum diameter. It is this maximum diameter measurement that is used to make clinical decisions. If measurements were made only in the straight anterior-posterior direction or the orthogonal transverse direction, it would not be possible to determine whether the compression impacts clinical decisions. Both the measurements and axes were marked on the film with a grease pencil. A subjective score of 1 (worst) to 5 (best) was also assigned to each image based on the radiologist’s opinion of the quality of that image for the measurement task. The subjective scores were used purely as a measure of subjective quality and not as a measure of diagnostic accuracy. Relationships among subjective score, SNR, and diagnostic accuracy are further elaborated in [28, 29].

4 Univariate Analyses

Measurement Standards and Error Parameters

In order to quantify the accuracy of measurements at each level of compression, we set two “gold standards” to represent the “correct measurement” for each vessel. One gold
standard was set by having two expert radiologists, not the judges, come to an agreement on vessel sizes on the uncompressed scans. This provides an "independent standard." The two radiologists first independently measured the vessels on each scan. For those vessels on which they differed, they remeasured until an agreement was reached. A "personal standard" was also derived for each judge by taking their own measurements on the uncompressed image to be the gold standard for their measurements on the compressed scans. Comparison with the personal gold standard quantifies individual consistency, or lack of it, over bit rates, rather than performance relative to "absolute truth."

Once the gold standard measurement for each vessel in each image was assigned, the analysis of a radiologist's performance was made by comparing the errors made on compressed and on uncompressed images. The measurement error can be quantified in a variety of ways. If \( z \) is the radiologist's measurement and \( g \) represents the gold standard measurement, then some potential error parameters are \( (z - g) \), \( \log(z/g) \), \( (z - g)/g \), and \( |(z - g)/g| \). These parameters have invariance properties that bear upon understanding the data. For example, \( z - g \) is invariant to the same additive constant (that is, to a change in origin), \( \log(z/g) \) is invariant to the same multiplicative constant (that is, to a change in scale), and \( (z - g)/g \) is invariant to the same multiplicative constant and to the same sign changes. For simplicity and appropriateness in the statistical tests carried out, the error parameters chosen for this study are percent measurement error \( (z - g)/g \times 100\% \) and absolute percent measurement error \( |(z - g)/g| \times 100\% \), both of which scale the error by the gold standard measurement to give a concept of error relative to the size of the vessel being measured.

**Parameters and Tests**

The differences in error achieved at each bit rate for our paired data could be quantified in many ways. We use both the t and Wilcoxon tests. We also accounted for the multiplicity of comparisons. If \( x_1 \) is the measurement of a vessel at bit rate 1, \( x_2 \) its measurement at bit rate 2, and \( g \) the vessel's gold standard measurement, then the percent measurement errors at bit rates 1 and 2 are \( (x_1 - g)/g \times 100\% \) and \( (x_2 - g)/g \times 100\% \), and their difference is \( (x_1 - x_2)/g \times 100\% \). In such a two-level comparison, percent measurement error more accurately preserves the difference between two errors than does absolute percent measurement error. A vessel that is over-measured by \( \alpha\% \) (positive) on bit rate 1 and under-measured by \( \alpha\% \) (negative) on bit rate 2 will have an error distance of \( 2\alpha\% \) if percent measurement error is used but a distance of zero if absolute percent measurement is used. Therefore both the t-test and the Wilcoxon signed rank test were computed using only percent measurement error. Absolute percent measurement error is used to present a more accurate picture of average error across the 30 test images plotted against bit rate.

The size of our data set (4 vessels \( \times 30 \) images \( \times 6 \) levels \( \times 3 \) judges = 2160 data points) makes a formal test for normality nearly irrelevant. Q-Q plots of percent measurement error differences that were made for comparisons of other levels exhibit varying degrees of linearity. In general, the Q-Q plots indicate a moderate fit to a Gaussian model.
5 Results

Distortion-Rate Performance

Figure 7 shows SNR versus bit rate for the 30 test images compressed to the 5 bit rates. A quadratic spline with a single knot at 1.0 bpp was fit through the data points to show the general trend. Generally, images with lower visual distortion have higher SNR.

Figure 7: SNR as a function of bit rate. The x’s indicate data points for all images, pooled across judges and compression levels. The solid curve is a quadratic spline fit to the data with a single knot at 1.0 bpp.

Results using the independent gold standard

Figures 8-11 are plots of trends in measurement error as a function of bit rate. In all cases, the general trend of the data is indicated by fitting the data points with a quadratic spline having one knot at 1.0 bpp. Figure 8 plots the average percent measurement error against the mean bit rate for all radiologists pooled (i.e., the data for all radiologists, images, levels, and structures, with each radiologist’s measurements compared to the independent gold standard) and for each of the three radiologists separately. In Figure 9, the percent measurement error versus actual achieved bit rate is plotted for all data points. The relatively flat curve begins to increase slightly at the lowest bit rates, levels 1 and 2 (0.36, 0.55 bpp). It is apparent that except for measurement at the lowest bit rates, accuracy does not vary greatly with lossy compression. Possibly significant increases in error appear only at the lowest bit rates, whereas at the remaining bit rates measurement accuracy is similar to that
obtained with the originals. The average performance on images compressed to level 5 (1.7 bpp) is actually better than performance on originals.

![Graph showing mean percent measurement error vs. mean bit rate for judges 1, 2, 3, and pooled judges.](image)

Figure 8: Mean percent measurement error vs. mean bit rate using the independent gold standard. The dotted, dashed, and dash-dot curves are quadratic splines fit to the data points for Judges 1, 2, and 3, respectively. The solid curve is a quadratic spline fit to the data points for all judges pooled. The splines have a single knot at 1.0 bpp.

While the trends in percent measurement error vs. bit rate are useful, over-measurement (positive error) can cancel under-measurement (negative error) when these errors are being averaged or fitted with a spline. For this reason, we turn to absolute percent measurement error which measures the error made by a radiologist regardless of sign. Figure 10 is a plot of average absolute percent measurement error versus average bit rate for each radiologist and for all radiologists pooled. Figure 11 shows actual absolute percent measurement error versus actual bit rate achieved. These plots show trends similar to those already seen. The original level has about the same absolute percent measurement error as compression levels 3, 4, and 5 (0.82, 1.14, 1.7 bpp). Levels 1 and 2 (0.36, 0.55 bpp) show slightly higher measurement error.

The t-test was used to test the null-hypothesis that the "true" percent measurement error between two bit rates is zero. None of the levels down to the lowest bit rate of 0.36 bpp was found to have a significantly higher percent measurement error when compared to the error of measurements made on the originals. Among the compressed levels however, level 1 (0.36 bpp) was found to be significantly different from level 5 (1.7 bpp). As was mentioned, the performance on level 5 was better than that on all levels, including the uncompressed level.
Figure 9: Percent measurement error vs. actual bit rate using the independent gold standard. The x's indicate data points for all images, pooled across judges and compression levels. The solid curve is a quadratic spline fit to the data with a single knot at 1.0 bpp.

When using the Wilcoxon signed rank test to compare compressed images against the originals, only level 1 (0.36 bpp) differed significantly in the distribution of percent measurement error. Within the levels representing the compressed images, levels 1, 3, and 4 (0.36, 0.82, 1.14 bpp) had significantly different percent measurement errors than those at level 5 (1.7 bpp). Since measurement accuracy is determined from the differences with respect to the originals only, a conservative view of the results of the analyses using the independent gold standard is that measurement accuracy is retained down to 0.55 bpp (level 2).

Results using the personal gold standard

As previously described, the personal gold standard was set by taking a radiologist's recorded vessel size on the uncompressed image to be the correct measurement for judging performance on the compressed images. Using a personal gold standard in general accounts for a measurement bias attributed to an individual radiologist, thereby providing a more consistent result among the measurements of each judge at the different compression levels. The personal gold standard thus eliminates the interobserver variability present with the independent gold standard. However, it does not allow us to compare performance at compressed bit rates to performance at the original bit rates since the standard is determined from the original bit rates. As before, we first consider visual trends.

Figure 12 shows average percent measurement error vs. mean bit rate for the 5 com-
Figure 10: Mean absolute percent measurement error vs. mean bit rate using the independent gold standard. The dotted, dashed, and dash-dot curves are quadratic splines fit to the data points for Judges 1, 2, and 3, respectively. The solid curve is a quadratic spline fit to the data points for all judges pooled. The splines have a single knot at 1.0 bpp.
Figure 11: Absolute percent measurement error vs. actual bit rate using the independent gold standard. The x's indicate data points for all images, pooled across judges and compression levels. The solid curve is a quadratic spline fit to the data with a single knot at 1.0 bpp.
Figure 12: Mean percent measurement error vs. mean bit rate using the personal gold standard. The dotted, dashed, and dash-dot curves are quadratic splines fit to the data points for Judges 1, 2, and 3, respectively. The solid curve is a quadratic spline fit to the data points for all judges pooled. The splines have a single knot at 1.0 bpp.
pressed levels for each judge separately and for the judges pooled, whereas Figure 13 is a display of the actual percent measurement error vs. actual achieved bit rate for all the

![Graph](image)

Figure 13: Percent measurement error vs. actual bit rate using the personal gold standard. The x’s indicate data points for all images, pooled across judges and compression levels. The solid curve is a quadratic spline fit to the data with a single knot at 1.0 bpp.

data points. The data for the judges pooled are the measurements from all judges, images, levels, and vessels, with each judge’s measurements compared to her or his personal gold standard. In each case, quadratic splines with a single knot at 1.0 bpp were fit to the data. Figures 14 and 15 are the corresponding figures for the absolute percent measurement error. As expected, with the personal gold standard, the errors are less than those obtained with the independent gold standard. The graphs indicate that whereas both Judges 2 and 3 overmeasured at all bit rates with respect to the independent gold standard, only Judge 3 consistently overmeasured with respect to the personal gold standard.

The t-test results indicate that levels 1 (0.36 bpp) and 4 (1.14 bpp) have significantly different percent measurement error associated with them. The results of the Wilcoxon signed rank test on percent measurement error using the personal gold standard are similar to those obtained with the independent gold standard. In particular, only level 1 at 0.36 bpp differed significantly from the originals. Furthermore, levels 1, 3, and 4 were significantly different from level 5.

With Bonferroni considerations, the percent measurement error at level 4 (1.14 bpp) is not significantly different from the uncompressed level. The simultaneous test indicates that only level 1 (0.36 bpp) has significantly different percent measurement error from the uncompressed level. This agrees with the corresponding result using the independent gold
Figure 14: Mean absolute percent measurement error vs. mean bit rate using the personal gold standard. The dotted, dashed, and dash-dot curves are quadratic splines fit to the data points for Judges 1, 2, and 3, respectively. The solid curve is a quadratic spline fit to the data points for all judges pooled. The splines have a single knot at 1.0 bpp.
Figure 15: Absolute percent measurement error vs. actual bit rate using the personal gold standard. The x's indicate data points for all images, pooled across judges and compression levels. The solid curve is a quadratic spline fit to the data with a single knot at 1.0 bpp.

standard. Thus, percent measurement error at compression levels down to 0.55 bpp does not seem to differ significantly from the error at the 9.0 bpp original.

Subjective score

In the previous sections, we looked at how measurement performance changes with bit rate in terms of the distribution of percent measurement error. In addition to characterizing such objective aspects of error, we would like to examine the effect of compression on subjective opinions. In particular, does a radiologist's perception of image quality change with bit rate, and does it change in a manner similar to the way percent measurement error changes? At the time of measurement, radiologists were asked to assign subjective scores of 1 (worst) - 5 (best) 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.

We do not know 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, 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.

Figure 16 shows the general trend of mean subjective score versus mean bit rate. A

![Figure 16: Mean subjective score vs. mean bit rate. The dotted, dashed, and dash-dot curves are splines fit to the data points for Judges 1, 2, and 3, respectively. The solid curve is a spline fit to the data points for all judges pooled.](image)

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 17 shows a spline fit of subjective score plotted against actual bit rate for the compressed levels only. The general conclusion from the plots is that the subjective scores at level 5 (1.7 bpp) and level 6 (9 bpp) were quite close (with level 6 slightly higher) but at lower levels there was a steep drop-off of scores with decreasing bit rate.

The Wilcoxon signed rank test shows that the subjective scores at all of the five 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.

Although the subjective impressions of the radiologists were that the accuracy of measurement tasks degraded rapidly with decreasing bit rate, their actual measurement performance on the images as shown by both the t-test and Wilcoxon signed rank test (or the Bonferroni simultaneous test) remained consistently high down to 0.55 bpp. Thus, their subjective opinion of an image's usefulness for diagnosis was not a good predictor of
Figure 17: Subjective score vs. bit rate. The x’s indicate data points for all images, pooled across judges and compression levels. The solid curve is a quadratic spline fit to the data with a single knot at 1.0 bpp.

measurement accuracy.

6 Discussion, Further Analysis, Conclusions

There are a number of issues to consider in order to determine which gold standard is preferable. One disadvantage of an independent gold standard is that since it is determined by the measurements of radiologists who do not judge the compressed images, significant differences between a compressed level and the originals may be due to differences between judges. For example, a judge who tends to overmeasure at all bit rates may have high percent measurement errors which will not be entirely reflective of the effects of compression. In our study, we determined that two judges consistently overmeasured relative to the independent gold standard. This is not an issue with the personal gold standard. A personal gold standard also has the advantage of reducing percent and absolute percent measurement error at the compressed levels, one result being a clarification of trends in a judge’s performance across different compression levels. Differences are based solely on compression level and not on differences between judges. Another argument in favor of a personal gold 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. This is of interest because the systematic biases of a radiologist are well recognized and corrected for by the referring physicians who regularly send cases to
that radiologist.

One disadvantage with the personal gold standard, however, is that by defining the measurements on the original images to be "correct," we are not accounting for the inherent variability of a judge's measurement on an uncompressed image. For example, if a judge makes an inaccurate measurement on the original and accurate measurements on the compressed images, these correct measurements will be interpreted as incorrect. Thus the method is biased against compression. An independent gold standard reduces the possibility of this situation occurring since we need an agreement by two independent radiologists on the "correct" measurement.

Most analyses presented thus far were based on data for which judges, vessels, and images were pooled. Other analyses in which the performances of judges on particular vessels and images are separated demonstrate additional variability. Judges seem to have performed significantly differently from each other. Judges 2 and 3 consistently overmeasured. As a result, the Wilcoxon signed rank test using the independent gold standard indicates significant differences between the gold standard and the measurements of Judges 2 and 3 at all compression levels, including the original. Judge 1, however, does not have any significant performance differences between the gold standard and any compression levels. In addition, certain vessels and images had greater variance in percent measurement error than others.

We turn now to matters of describing variabilities we cite. Whether compression degrades clinical performance is of fundamental importance to policy. We believe that at least within broad ranges it does not. However, this is not to obscure the finding that radiologists themselves are different. They are different in the impact compression has on their performance. They are different in how they measure vessels, on compressed or original images. And they differ in how compression may degrade their performance for different vessels, to the extent that it does.

The outcome here is percent measurement error relative to an independent gold standard. One approach to this analysis might have been a (necessarily complex) random or mixed effects analysis of variance. Inferences therefrom tend to be heavily dependent upon Gaussian assumptions, and even when they hold, distributions of some relevant statistics have been computed only approximately. In fact, the "signals" in our data are simple to describe and can be summarized well with a far simpler approach. Thus, we fix (judge, image) pairs, of which there are 90 in all, and analyze them by fixed effects two-way analyses of variance with one observation per cell in which there are two fixed effects: levels and structures. Though we are not ordinarily entitled from such data to make inferences on interactions, we compute Tukey's one degree of freedom for non-additivity [31, 32]. If \( y_{ij} \) is the percent measurement error, \( i \) refers to levels and \( j \) to structures, then we model \( y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij} \) where \( \varepsilon_{ij} \) are taken to be iid mean 0 and constant variance, though not constant across judges or images; \( i = 1, \ldots, 6; j = 1, \ldots, 4 \). We assume the usual constraints, that is \( \sum_i \alpha_i \equiv \sum_j \beta_j \equiv \sum_i \gamma_{ij} \equiv \sum_j \gamma_{ij} \equiv 0 \), and further that \( \gamma_{ij} = G\alpha_i\beta_j \) for some constant \( G \). Gaussian assumptions on the \( \varepsilon_{ij} \) are not in force here, for we view the F-statistics and residual mean squares as descriptive statistics that are summarized by box
plots of Figures 18–21. Details of the computations are given in Section 4.8 of the book by Scheffé [32].

![Box Plots of F for Levels](image)

**Figure 18: Box Plots of F for Levels**

The box plots are self-explanatory, and dramatic! Clearly Judge 2 was affected by compression more than was Judge 3, who was affected more than was Judge 1, whose F-statistics surround the null value 1. Variability obviously increased by judge as the impact of compression increased. Structures differed less for Judge 1 than for the others, and variability was less, too. The influence of structure upon variability in level was less for Judge 1 than for the others, yet residual variability was less for Judge 2 than for Judge 1, and was highest for Judge 3. Though we are reluctant to infer much by way of performance from our study, it does appear that Judge 1 fared better than did the others of our capable judges.

This study focused primarily on characterizing the measurement accuracy of images subjected to lossy compression, whereas others of our studies have been concerned with the detection of lesions. Since a CT scan of the chest that was taken primarily for detecting the presence of a tumor might also be used to detect the size(s) of anything found, one would like to know which of these related but different tasks might suffer more the effects of compression. Our study of detection in a CT context [6] involved the location of single or multiple abnormal mediastinal lymph nodes or lung nodules in chest scans. A predictive TSVQ encoder was used to compress a set of 12-bit original scans to 0.56, 1.18, 1.34, 1.80, 2.20, and 2.64 bpp. Both compressed and original images were presented to radiologists who marked the locations of all found lesions. The gold standards were taken to be the consensus of the radiologists on the uncompressed images, and (separately) each
Figure 19: Box Plots of F for Structures

Figure 20: Box Plots of F for Interaction
radiologist’s personal markings on the uncompressed images. The accuracy of detection for each image was quantified by “sensitivity,” the fraction of lesions found, and by “predictive value positive,” the fraction of true lesions among those reported [30]. A Behrens-Fisher (Welch) t-statistic was used to compare accuracy at different levels. Thus the CT and MR diagnostic accuracy studies use similar versions of the predictive TSVQ compression algorithm as well as the same method for generating and viewing a randomized sequence of images and of determining a personal gold standard. The studies differ in the radiologists involved; the specific parameters that quantify accuracy and corresponding analyses; the determination of a (non-personal) gold standard; and, of course, the imaging modality. This latter entails differences in the bit rate of the original, the resolution, and the distributions of pixel intensities. Given these differences, a useful comparison of diagnostic accuracy is at best problematical with these data. We would like to be able to address in a definitive manner whether CT or MR chest scans are more easily compressed, and whether accuracy in detection is maintained to the same bit rates as accuracy in measurement. In terms of diagnostic accuracy, the Behrens-Fisher t-test applied to sensitivity and predictive value positive showed that 12-bit original CT images could be compressed to 1 to 2 bpp without a loss in diagnostic accuracy. Our results from the MR study show that MR chest scans can be compressed from 9 bpp down to 0.55 bpp. It cannot be said whether this difference is due to differences in inherent compressibility for the two modalities, or because there is more inherent fluctuation for one problem than for the other. Larger fluctuations in measurement error would overshadow the variations due to bit rate and would allow for greater compression of MR images before a degradation in measurement accuracy is seen as
statistically significant. Some of these interesting questions might be addressed by a study in which one image type is used for multiple diagnostic interpretation tasks and by a study in which two different imaging modalities are used for very similar aspects of interpretation.

Acknowledgements

We gratefully acknowledge the assistance of Charles H. McDonnell III, M.D. and D.V. Paranjpe, M.D. in judging the images, Jorg Debatin, M.D. and Herbert Wegmuller, M.D. for establishing the independent gold standard, and the assistance of H. Christian Davidson, M.D. in filming the images. We also acknowledge useful discussions with Adam Olshen, Jonathan Buckheit, and Keren Perlmutter.

References


