Title

Image quality in lossy compressed digital mammograms

Journal

Signal Processing, 59(2)

ISSN

01651684

Authors

Perlmutter, S. M.
Cosman, P. C.
Gray, R. M.
et al.

Publication Date

1997-06-01

DOI

10.1016/S0165-1684(97)00046-7

Peer reviewed

ELSEVIER

Signal Processing 59 (1997) 189–210

SIGNAL PROCESSING

Image quality in lossy compressed digital mammograms*

S.M. Perlmutter[a], P.C. Cosman[b], R.M. Gray[c,*], R.A. Olshen[d], D. Ikeda[e], C.N. Adams[f], B.J. Betts[c], M.B. Williams[g], K.O. Perlmutter[a], J. Li[c], A. Aiyer[c], L. Fajardo[g], R. Birdwell[e], B.L. Daniel[e]

[a] *Johnson-Grace Company, Inc., 2 Corporate Plaza, Suite 150, Newport Beach, CA, USA*
[b] *Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92093-0407, USA*
[c] *Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA*
[d] *Division of Biostatistics of the Stanford University School of Medicine and the Department of Statistics of Stanford University, Stanford, CA 94305, USA*
[e] *Department of Radiology, Stanford University School of Medicine, Stanford, CA 94305, USA*
[f] *Apple Computer, 3515 Monroe St., MS: 70-IS, Santa Clara, CA 95015, USA*
[g] *Department of Radiology, University of Virginia, Charlottesville, VA, USA*

Received 4 July 1996; revised 8 January 1997

Abstract

The substitution of digital representations for analog images provides access to methods for digital storage and transmission and enables the use of a variety of digital image processing techniques, including enhancement and computer assisted screening and diagnosis. Lossy compression can further improve the efficiency of transmission and storage and can facilitate subsequent image processing. Both digitization (or digital acquisition) and lossy compression alter an image from its traditional form, and hence it becomes important that any such alteration be shown to improve or at least not damage the utility of the image in a screening or diagnostic application. One approach to demonstrating in a quantifiable manner that a specific image mode is at least equal to another is by clinical experiment simulating ordinary practice and suitable statistical analysis. In this paper we describe a general protocol for performing such a verification and present preliminary results of a specific experiment designed to show that 12 bpp digital mammograms compressed in a lossy fashion to 0.015 bpp using an embedded wavelet coding scheme result in no significant differences from the analog or digital originals. © 1997 Elsevier Science B.V.

Zusammenfassung

Die Ersetzung analoger Bilder durch digitale Darstellungen erlaubt eine digitale Speicherung und Übertragung sowie den Einsatz einer Vielzahl von Methoden der digitalen Bildverarbeitung, z.B. zur Verbesserung der Bildqualität und zum computerunterstützten Screening bzw. zur computerunterstützten Diagnose. Eine verlustbehaftete Kompression kann die Effizienz der Übertragung oder Speicherung weiter steigern und eine nachfolgende Bildverarbeitung erleichtern. Sowohl die Digitalisierung (oder digitale Aufnahme) als auch die verlustbehaftete Kompression ändern ein Bild bezüglich seiner ursprünglichen Form. Deswegen ist es wichtig, zu zeigen daß eine solche Veränderung die Nützlichkeit des Bildes bei Screening- oder diagnostischen Anwendungen steigert oder wenigstens nicht beeinträchtigt. Eine Möglichkeit,

* Corresponding author. Tel.: 1 41 5723 4001; fax: 1 41 5723 8473; e-mail: rmgray@stanford.edu.

PII S0165-1684(97)00046-7

auf quantifizierbare Weise zu zeigen, daß eine bestimmte Bilddarstellung einer anderen zumindest äquivalent ist, ist ein die gewöhnliche Praxis simulierendes klinisches Experiment und eine geeignete statistische Analyse. In diesem Artikel beschreiben wir ein allgemeines Protokoll für die Durchführung einer solchen Verifikation. Wir präsentieren weiters vorläufige Resultate eines spezifischen Experiments, welches zeigt, daß die verlustbehaftete Kompression digitaler Mammogramme von 12 bpp auf 0.15 bpp mittels einer eingebetteten Wavelet-Codierung zu keinen signifikanten Unterschieden von den analogen oder digitalen Originalen führt. © 1997 Elsevier Science B.V.

Résumé

La substitution d'images analogiques par des représentations numériques donne accès à des méthodes de stockage et de transmission numériques, et permet l'utilisation d'une grande variété de techniques de traitement d'images, incluant le rehaussement, les tests de dépistage assisté ordinateur et le diagnostic. La compression avec pertes peut encore améliorer l'efficacité de la transmission et du stockage, et peut faciliter le traitement ultérieur des images. La numérisation et la compression avec pertes altérant toutes deux une image par rapport à sa forme traditionnelle, il devient important de montrer qu'une telle altération améliore, ou du moins ne réduit pas, l'utilité de l'image dans un screening ou une application de diagnostic. Une approche pour démontrer d'une manière quantifiable qu'un mode d'image spécifique est au moins égal à un autre est l'expérimentation clinique simulant la pratique ordinaire jointe à une analyse statistique adaptée. Dans cet article, nous décrivons un protocole général pour effectuer une telle vérification et présentons les résultats préliminaires d'une expérience faite pour montrer que des mamogrammes numérisés à 12 bpp et comprimés avec pertes à 0.15 bpp à l'aide d'une technique de codage par ondelettes incluses ne présentent pas de différences significatives par rapport aux versions originales analogique ou numérique. © 1997 Elsevier Science B.V.

Keywords: Lossy compression; Image quality; Digital mammography

1. Introduction

X-ray mammography is the most sensitive technique for detecting breast cancer [2], with a reported sensitivity of 85–95% for detecting small lesions. Most non-invasive ductal carcinomas, or DCIS, are characterized by tiny non-palpable calcifications detected at screening mammography [16, 25, 46]. Traditional mammography is essentially analog photography using X-rays in place of light and analog film for display. For a variety of reasons, digital technologies are likely to change and eventually replace most of the existing analog methods. The digital format is required for access to modern digital storage, transmission, and digital computer processing. Hardcopy films use valuable hospital space and are prone to loss and damage, which undermine the ability of radiologists to carry out comparisons with subsequent studies. Images in analog format are not easily distributed to multiple sites, either in-hospital or off-site. Currently only 30% of women get regular mammograms, and the storage problems will be compounded if this number increases with better education or wider insurance coverage. Digital image processing provides the possibilities for easy image retrieval, efficient storage, rapid image transmission for off-site diagnoses, and the maintenance of large banks for purposes of teaching and research. It allows filtering, enhancement, classification, and combining images obtained from different modalities, all of which can assist screening, diagnosis, research, and treatment. Retrospective studies of interval cancers (carcinomas detected in the time intervals between mammographic screenings which were interpreted as normal) show that observer error can comprise up to 10% of such cancers. That is to say, carcinomas present on the screening mammograms were missed by the radiologist because of fatigue, misinterpretation, distraction, obscuration by a dense breast, or other reasons [18, 24, 32]. To this end, schemes for computer-aided diagnosis (CAD) may assist the radiologist in the detection of clustered micro-calcifications and masses [10, 27, 28, 37, 56]. Virtually all existing CAD schemes require images in digital format.

To take advantage of digital technologies, either analog signals such as X-rays must be converted into a digital format, or the signals must be directly acquired in digital form. Digitization of an analog

signal causes a loss of information and hence a possible deterioration of the signal. In addition, with the increasing accuracy and resolution of analog-to-digital converters, the quantities of digital information produced can overwhelm available resources. A typical digitized mammogram with 4500×3200 picture elements (pixels) with 50 μm spot size and 12 bit per pixel depth requires approximately 38 Mbytes of data. Complete studies can easily require unacceptably long transmission times through crowded digital networks and can cause serious data management problems in local disk storage. Advances in technologies for transmission and storage do not solve the problem. In recent years these improvements on the Internet have been swamped by the growing volume of data. Even with an ISDN line, a single X-ray can take several minutes for transmission. Compression is desirable and often essential for efficiency of storage and communication. The overall goal is to represent an image with the smallest possible number of bits, or to achieve the best possible fidelity for an available communication or storage bit rate capacity.

A digital compression system typically consists of a signal decomposition such as Fourier or wavelet, a quantization operation on the coefficients, and finally lossless or entropy coding such as Huffman or arithmetic coding. Decompression reverses the above process; although if quantization is used, the system will be lossy because quantization is only approximately reversible. Theory and experience argue that good compression can be designed by focusing separately on each individual operation, though simpler implementations may be obtained by combining some operations. Lossless coding is well understood, readily available [47], and typically yields compression ratios of 2:1 to 3:1 on still frame greyscale medical images. This modest compression is often inadequate. Lossy coding does not permit perfect reconstruction of the original image but can provide excellent quality at a fraction of the bit rate [9, 26, 29, 31, 40]. The *bit rate* of a compression system is the average number of bits produced by the encoder for each image pixel. If the original image has 12 bits per pixel (bpp) and the compression algorithm has rate R bpp, then the *compression ratio* is $12:R$. Compression ratios must be interpreted with care as they depend crucially on the image type, original bit rate, sampling density, how much background is in the image, and how much coding of the background figures into the calculation.

Early studies of lossy compressed medical images performed compression using variations on the standard discrete cosine transform (DCT) coding algorithm combined with scalar quantization and loseless (typically Huffmann and run-length) coding. These are variations of the international standard Joint Photographic Experts Group (JPEG) compression algorithm [36, 51]. The standard permits a user-specified quantization table that describes the uniform quantizers used to quantize the transform coefficients. Although the standard suggests specific values, performance can be improved by customizing these tables for a specific application. The American College of Radiology–National Electrical Manufacturers Association (ACR–NEMA) standard [6] has not yet firmly recommended a specific compression scheme, but transform coding methods are suggested. These algorithms are well understood and have been tuned to provide good performance in many applications.

More recent studies of efficient lossy image compression algorithms have used subband or wavelet decompositions combined with scalar or vector quantization [3, 30, 38, 39, 41, 42, 49, 55]. These signal decompositions provide several improvements, including better concentration of energy, better decorrelation for a wider class of signals, better basis functions for images than the smoothly oscillating sinusoids of Fourier analysis because of diminished Gibbs and edge effects and better localization in both time and frequency. Because of their sliding-block operation using 2-dimensional linear filters, they do not produce blocking artifacts (although other artifacts arise at low rates).

Since lossy coding can degrade the quality of an image, making precise the notion of 'excellent quality' of a compressed or processed image is a serious issue. Analog mammography remains the gold standard against which all other imaging modalities can be judged. In a medical application it does not suffice for an image to simply 'look good' or to have a high signal-to-noise ratio (SNR), nor should one necessarily require that original and processed

images be visually indistinguishable. Rather it must be convincingly demonstrated that essential information has not been lost and that the processed image is at least of equal utility for diagnosis or screening as the original. Image quality is typically quantified objectively by average distortion or SNR, and subjectively by statistical analyses of viewers' scores on quality (e.g., analysis of variance (ANOVA) and receiver operating characteristic (ROC) curves). Examples of such approaches may be found in [4, 7, 20, 29, 31, 40, 53].

ROC analysis is the dominant technique for evaluating the suitability of radiologic techniques for real applications [23, 33, 34, 48]. Its origins are in the theory of signal detection: a filtered version of signal plus Gaussian noise is sampled and compared to a threshold. If the threshold is exceeded, then the signal is said to be there. As the threshold varies, the probability of erroneously declaring a signal absent and the probability of erroneously declaring a signal there when it is not vary too, and in opposite directions. The plotted curve is a summary of the tradeoff in these two quantities; more precisely, it is plot of *true positive rate* or *sensitivity* against *false positive rate*, the complement of *specificity*. Summary statistics, such as the area under the curve, can be used to summarize overall quality. In typical implementations, radiologists or other users are asked to assign integer confidence ratings to their diagnoses, and thresholds in these ratings are used in computing the curves.

We have argued in our previously cited references (summarized in Section 2) that traditional ROC analysis violates several reasonable guidelines for designing experiments to measure quality and utility in medical images because of the use of artificial confidence ratings as thresholds in a binary detection problem and because of the statistical assumptions of Gaussian or Poisson behavior. In addition, traditional ROC analysis is not well suited to the study of the accuracy of detection and location when a variety of abnormalities are possible. Although extensions of ROC designed to handle location and multiple lesions have been proposed [8, 45], they inherit many of the more fundamental problems of the approach and are not widely used. Traditional ROC analysis also does not come equipped to distinguish among the various possible notions of 'ground truth' or 'gold standard' in clinical experiments.

During the past decade our group at Stanford University has worked to develop an alternative approach to evaluating the diagnostic accuracy of lossy compressed medical images (or any digitally processed medical images) that mimics ordinary clinical practice as closely as is reasonably possible, does not require special training or artificial subjective evaluations, applies naturally to the detection of multiple abnormalities and to measurement tasks, and requires no assumptions of Gaussian behavior of crucial data. While some departures from ordinary practice are necessary and some additional information may be gathered because it is of potential interest, the essential goal remains the imitation of ordinary practice and the drawing of diagnostic conclusions based only on diagnostic simulations. The methods are developed in detail for CT and MR images [12–15, 35]. Extensions to digital mammography were described in [21, 22], and preliminary results for a pilot study are described in [1] (a reprint of which can be found at the World Wide Web site [44]). This paper expands on the description, discussion, and data analysis of the results of [1]. In particular, we here emphasize the lossy compression performance using both traditional engineering methods of image quality and the diagnostic accuracy measurement approach.

2. Methods

2.1. Study design

The general methods used are extensions to digital mammography and elaborations of techniques developed for CT and MR images by our group and reported in [12–15, 35], where all details regarding the data, compression code design, clinical simulation protocols, and statistical analyses may be found. We here describe extensions [1, 21, 22] of these methods to digital mammography. Further results are available in the Final Project Report (available at [44]) and other papers in progress. The design of the proposed mammogram evaluation study incorporates elements from both the CT and MR studies, as well as many new aspects.

The following general principles for protocol design have evolved from our earlier work. Although they may appear self-evident in hindsight, they provide a useful context for evaluating protocols for judging image quality in medical imaging applications and they represent an accumulation of over eight years of discussion and experience among electrical engineers, statisticians, radiologists, and medical physicists. The protocol should *simulate ordinary clinical practice as closely as possible.* In particular, participating radiologists (judges, observers) should perform in a manner that *mimics their ordinary practice* as closely as reasonably possible given the constraints of good experimental design. The studies should require *little or no special training* of their clinical participants. The clinical studies *include examples of images containing the full range of possible findings*, all but extremely rare conditions. The findings should be *reportable using a subset of the American College of Radiology (ACR) Standardized Lexicon.* Any standardized nomenclature would do. *Statistical analyses of the trial outcomes should be based on assumptions as to the outcomes and sources of error that are faithful to the clinical scenario and tasks.* '*Gold standards*' for evaluation of equivalence or superiority of algorithms *must be clearly defined and consistent with experimental hypotheses.* Careful *experimental design should eliminate or minimize any sources of bias in the data* that are due to differences between the experimental situation and ordinary clinical practice, e.g., learning effects that might accrue if a similar image is seen using separate imaging modalities. The number of patients should be sufficient to ensure *satisfactory size and power* for the principal statistical tests of interest.

The ROC assumptions and approach generally differ from clinical practice. Digitization of an analog image and lossy compression are not equivalent to the addition of signal-independent noise. Radiologists are not threshold detectors. Using ROC curves to compare computer aided diagnosis (CAD) schemes is appropriate because such schemes almost always depend on a threshold, albeit in a possibly complicated way. No hard evidence exists, however, to support the contention that human radiologists behave in this way and, even if they did, that the ROC method of asking them for confidence ratings to interpret as thresholds in fact measures whatever internal threshold they might have. We believe this to be a fundamental flaw in using ROC curves to draw conclusions about quality comparisons among radiologists or among images read by radiologists. Because of the need for confidence ratings, the traditional ROC approach requires special training to familiarize a radiologist with the rating system. On the statistical side, image data are not well modeled as known signals in Gaussian noise, and hence methods that rely on Gaussian assumptions are suspect. This is particularly true when Gaussian approximations are invoked to compute statistical size and power on a data set clearly too small to justify such approximations. Modern computer-intensive statistical sample reuse techniques can help get around the failures of Gaussian assumptions, but this does not address the more fundamental issues.

Traditional ROC methods are not location specific, and if an actual lesion is missed, a diagnosis can be considered correct if an incorrect lesion is spotted elsewhere. Extensions of ROC have been extended to address this [45], but the method is cumbersome and inherits the remaining faults of ROC. For clinical studies that involve other than binary tasks, specificity does not make sense because it has no natural or sensible denominator as it is not possible to say how many abnormalities are absent. This can be done for a truly binary diagnostic task for if the image is normal then exactly one abnormality is absent. Previous studies were able to use ROC analysis by focusing on detection tasks which were either truly binary or could be rendered binary. Extensions of ROC such as FROC to permit consideration of multiple abnormalities have been developed [8], but these still require the use of confidence ratings as well as Gaussian or Poisson assumptions on the data In our view they attempt to fit the method (ROC analysis) to clinical practice in an artificial way, rather than trying to develop more natural methods for measuring how well radiologists perform ordinary clinical functions on competing image modalities.

Traditional ROC analysis has no natural extension to problems of estimation or regression instead of detection. For example, measurement

plays an important role in some diagnostic applications and there is no ROC analysis for measurement error.

Lastly, traditional ROC applications have often been lax in clarifying the 'gold standard' used to determine when decisions are 'correct', when in fact a variety of gold standards are possible, each with its own uses and shortcomings. We focus on three definitions of diagnostic truth as a basis of comparison for the diagnoses on all lossy reproductions of that image. These are:

Personal: Each judge's readings on an original analog image are used as the gold standard for the readings of that same judge on the digitized version of that same image,

Independent: formed by the agreement of the members of an independent expert panel, and

Separate: produced by the results of further imaging studies (including ultrasound, spot and magnification mammogram studies), surgical biopsy, and autopsy.

The first two gold standards are usually established using the analog original films. As a result, they are extremely biased in favor of the established modality, i.e., the original analog film. Thus statistical analysis arguing that a new modality is equal to or better than the established modality will be conservative since the original modality is used to established 'ground truth'. The personal gold standard is in fact hopelessly biased in favor of the analog films. It is impossible for the personal gold standard to be used to show that digital images are *better* than analog ones. If there is any component of noise in the diagnostic decision, the digital images cannot even be found equal to analog. The personal gold standard is often useful, however, for giving some indication of the diagnostic consistency of an individual judge. The independent gold standard is also biased in favor of the analog images, but not hopelessly so, as it is at least possible for the readings of an individual judge on either the digital or analog images to differ from the analog gold standard provided by the independent panel. If the independent panel cannot agree on a film, the film could be removed from the study; but this would forfeit potentially valuable information regarding difficult images. By suitable gathering of data, one can instead define several possible independent gold standards and report the statistics with respect to each. In particular, a cautious gold standard declares a finding if any of the panel do so. An alternative is that the panel designates a chair to make a final decision when there is disagreement.

Table 1
Data test set: 57 studies, 4 views per study

6	Benign mass
6	Benign calcifications
5	Malignant mass
6	Malignant calcifications
3	Malignant combination of mass and calcifications
3	Benign combination of mass and calcifications
4	Breast edema
4	Malignant architectural distortion
2	Malignant focal asymmetry
3	Benign asymmetric density
15	Normals

Whenever a believable separate gold standard is available, it provides a more fair gold standard against which both old (analog) and new (digital, compressed digital) images can be compared. In future work we plan to use histologic data and long-term followup to establish a separate gold standard.

Our image database was generated in the Department of Radiology of the University of Virginia School of Medicine and is summarized in Table 1. The studies were digitized using a Lumisys Lumiscan 150 at 12 bpp with a spot size of 50 μm. Good quality directly acquired digital mammograms were not yet available when the experiment was begun, so digitized mammograms were used. The films were printed using a Kodak 2180 X-ray film printer, a 79 μm 12 bit greyscale printer which writes with a laser diode of 680 nm bandwidth. The 57 studies included a variety of normal images and images containing benign and malignant objects. We have corroborative biopsy information on at least 31 of the test subjects, which will later be used for a separate gold standard.

2.2. Experimental protocol

Images were viewed on hardcopy film on an alternator by judges in a manner that simulates

ordinary screening and diagnostic practice as closely as possible, although patient histories and other image modalities were not provided. Two views were provided of each breast (CC and MLO), so four views were seen simultaneously for each patient. Each of the judges viewed all the images in an appropriately randomized order over the course of nine sessions. Two sessions were held every other week, with a week off in between. A clear overly was provided for the judge to mark on the image without leaving a visible trace. For each image, the judge either indicated that the image was normal, or, if something was detected, had an assistant fill out the Observer Form (see Appendix A) using the American College of Radiology (ACR) Standardized Lexicon by circling the appropriate answers or filling in blanks as directed. The instructions for assistants and radiologists along with suggestions for prompting and a CGI web data entry form may be found at the project Web site [44]. The judges used a grease pencil to circle the detected item. The instructions to the judges specified that ellipses drawn around clusters should include all microcalcifications seen, as if making a recommendation for surgery, and outlines drawn around masses should include the main tumor as if grading for clinical staging, without including the spicules (if any) that extend outward from the mass. This corresponds to what is done in clinical practice except for the requirement that the markings be made on copies. The judges were allowed to use a magnifying glass to examine the films.

Although the judging form is not standard (there is no standard form for evaluating mammograms), the ACR Lexicon is used to report findings, and hence the judging requires no special training. The reported findings permit subsequent analysis of the quality of an image in the context of its true use, finding and describing anomalies and using them to assess and manage patients.

To confirm that each radiologists identifies and judges a specific finding, the location of each lesion is confirmed both on the clear overlay and the judging form. Many of these lesions were judged as 'A' (assessment incomplete), since it is often the practice of radiologists to obtain additional views in two distinct scenarios: (1) to confirm or exclude the presence of a finding, that is, a finding that may or may not represent a true lesion, or (2) to further characterize a true lesion, that is, to say a lesion clearly exists but is incompletely evaluated.

The judging form allows for two meanings of the 'A' code. If the judge believes that the findings is a possible lesion, this is indicated by answering 'yes' to the question 'are you uncertain if the finding exists?' Otherwise, if the lesion is definite, the judges should give their best management decision based on the standard two-view mammogram.

The initial question requesting a subjective rating of diagnostic utility on a scale of 1–5 is intended for a separate evaluation of the general subjective opinion of the radiologists of the images. The degree of suspicion registered in the Management portion also provides a subjective rating, but this one is geared towards the strength of the opinion of the reader regarding the cause of the management decision. It is desirable that obviously malignant lesions in a gold standard should also be obviously malignant in the alternative method.

2.3. *Statistical analysis*

Although long term analysis focuses on lesion-by-lesion accuracy of detection, the preliminary results reported here focus on patient management, the decisions that are made based on the radiologists' reading of the image. Management is a key issue in digital mammography. There is concern that artifacts could be introduced, leading to an increase in false positives and hence in unnecessary biopsies. The management categories we emphasize are the following four, given in order of increasing seriousness:

RTS incidental, negative, or benign with return to screening,

F/U probably benign but requiring six month follow-up,

C/B call back for more information, additional assessment needed,

BX Immediate biopsy.

These categories are formed by combining categories from the basic form of Appendix A: RTS is any study that had assessment = 1 or 2, F/U is assessment = 3, C/B is assessment = indeterminate/incomplete with best guess either unsure it

Table 2
Agreement 2 × 2 table

II/I	R	W
R	$N(1, 1)$	$N(1, 2)$
W	$N(2, 1))$	$N(2, 2)$

exists, 2 or 3, and BX is assessment = indeterminate/incomplete with best guess either 4*L*, 4*M*, 4*H* or 5, or assessment = 4*L*, 4*M*, 4*H* or 5.

We also consider the binarization of these four categories into two groups: normal and not normal. But there is controversy as to where the F/U category belongs, so we make its placement optional with either group. The point is to see if lossy compression makes any difference to the fundamental decision made in screening: does the patient return to ordinary screening as normal, or is there suspicion of a problem and hence the demand for further work?

Truth is determined by agreement with a gold standard. The raw results are plotted as a collection of 2 × 2 tables, one of each category or group of categories of interest and for each radiologist. A typical table is shown in Table 2.

The columns correspond to image modality or method I and the rows to II; I could be original analog and II original digitized, or I could be original digitized and II compressed digitized, '*R*' and '*W*' correspond to 'right' (agreement with gold standard) and 'wrong' (disagreement with gold standard). The particular statistics could be, for example, the decision of 'normal', i.e., return to ordinary screening. Regardless of statistic, the goal is to quantify the degree, if any, to which differences exist.

One way to quantify the existence of statistically significant differences is by an exact McNemar test, which is based on the following argument. If there are $N(1, 2)$ entries in the (1, 2) place and $N(2, 1)$ in the (2, 1) place, and the technologies are equal, then the conditional distribution of $N(1, 2)$ given $N(1, 2) + N(2, 1)$ is binomial with parameters $N(1, 2) + N(2, 1)$ and 0.5; that is,

$$P(N(1, 2) = k | N(1, 2) + N(2, 1) = n) = \binom{n}{k} 2^{-n},$$

$$k = 0, 1, \ldots, n.$$

This is the conditional distribution under the null hypothesis that the two modalities are equivalent. The extent to which $N(1, 2)$ differs from $(N(1, 2) + N(2, 1))/2$ is the extent to which the technologies were found to be different in the quality of performance with their use. Let $B(n, 1/2)$ denote a binomial random variable with these parameters. Then a statistically significant difference at level 0.05, say, will be detected if the observed k is so unlikely under the binomial distribution that a hypothesis test with size 0.05 would reject the null hypothesis if k were viewed. Thus if $\Pr(|B(n, 1/2) - n/2| \geqslant |N(1, 2) - n/2|) \leqslant 0.05$, then we declare a statistically significant difference has occurred.

Whether and how to agglomerate the multiple tables is an issue. Generally speaking, we stratify the data so that any test statistics we apply can be assumed to have sampling distributions that we could defend in practice. It is always interesting to simply pool the data within a radiologist across all gold standard values, though it is really an analysis of the off-diagonal entries of such a table that is of primary interest. If we look at such a 4 × 4 table in advance of deciding upon which entry to focus, then we must contend with problems of multiple testing, which would lower the power of our various tests. Pooling the data within gold standard values but across radiologists is problematical because our radiologists are patently different in their clinical performances. This is consistent with what we found in an earlier study of MR and the measurement of the sizes of vessels in the chest [13, 35]. Thus, even if one does agglomerate, there is the issue of how.

The counts can also be used to estimate a variety of interesting statistics, including sensitivity, predictive value positive (PVP), and specificity with respect to the personal and independent gold standards. An ROC-style curve can be produced by plotting the (sensitivity, specificity) pairs for the management decision for the levels of suspicion. Sample reuse methods (rather than common Gaussian assumptions) could be applied to provide confidence regions around the sample points [19].

A Wilcoxon signed rank test [43] can be employed to assess whether the subjective scores given to the analog originals, the uncompressed digitals,

and the compressed images differ significantly from each other. With the Wilcoxon signed rank test, the significance of the difference between the bit rates is obtained by comparing a standardized value of the Wilcoxon statistic to two-tailed standard Gaussian probabilities. (The distribution of this standardization Wilcoxon is nearly Gaussian if the null hypothesis is true for samples as small as 20.) Our previous criticism of Gaussian assumptions are not relevant when they are applied to statistics for which the Central Limit Theorem is applicable.

Several other approaches are planned, including estimating sensitivity, PVP, and, when appropriate, specificity of detection and management statistics, estimated by counts with bootstrapped confidence regions for each modality [5, 11]. Simple means and variances for the management statistics are presented in Section 3.

2.4. Learning effects

The radiologists saw each study at least 5 times during the course of the entire experiment. These 5 versions were the analog originals, the digitized versions, and the 3 wavelet compressed versions. Some images would be seen more than 5 times, as there were JPEG compressed images, and there were also some repeated images, included in order to be able to directly measure intra-observer variability. We therefore needed to ascertain whether learning effects were significant. Learning and fatigue are both processes that might change the score of an image depending upon when it was seen.

In this work, we looked for whether learning effects were present in the management outcomes using what is known in statistics as a 'runs' test [17]. We illustrate the method with an example. Suppose a study was seen exactly five times. The management outcomes take on four possible values (RTS, F/U, C/B, BX). Suppose that for a particular study and radiologist, the observed outcomes were BX three times and C/B two times. If there were no learning, then all possible "words" of length five with three BX's and two C/B's should be equally likely. There are 10 possible words that have three BX's and two C/B's. These words have the outcomes ordered by increasing session number; that is, in the chronological order in which they were produced. For these 10 words, we can count the number of times that a management outcome made on one version of a study differs from that made on the immediately previous version of the study. The number ranges from one (e.g., BX BX BX C/B C/B) to four (BX C/B BX C/B BX). The expected number of changes in management decision is 2.4, and the variance is 0.84. If the radiologists had learned from previous films, one would expect that there would be fewer changes of management prescription than would be seen by chance. This is a conditional runs test, which is to say that we are studying the conditional permutation distribution of the runs.

We assume that these 'sequence data' are independent across studies for the fixed radiologist, since examining films for one patient probably does not help in evaluating a different patient. So we can pool the studies by summing over studies the observed values of the number of changes, subtracting the summed (conditional) expected value, and dividing this by the square root of the sum of the (conditional variances). The attained significance level (p-value) of the resultant Z value is the probability that a standard Gaussian is $\leqslant Z$.

Those studies for which the management advice never changes have an observed number of changes 0. Such studies are not informative with regard to learning, since it is impossible to say whether unwavering management advice is the result of perfect learning that occurs with the very first version seen, or whether it is the result of the obvious alternative, that the study in question was clearly and independently the same each time, and the radiologist simply interpreted it the same way each time. Such studies, then, do not contribute in any way to the computation of the statistic. The JPEG versions and the repeated images, which are ignored in this analysis, are believed to make this analysis and p-values actually conservative. If no learning had occurred, then the additional versions make no difference. However, if learning did occur, then the additional versions (and additional learning) should mean that there would be even fewer management changes among the 5 versions that figure in this analysis.

2.5. Compression algorithms

We use a compression algorithm of the subband/pyramid/wavelet coding class. These codes typically decompose the image using an octave subband, critically sample pyramid, or complete wavelet transformation, and then code the resulting transform coefficients in an efficient way. The decomposition is typically produced by an analysis filter bank followed by downsampling. Any or all of the resulting subbands can be further input to an analysis filter bank and downsampling operation, for as many stages as desired.

The most efficient wavelet coding techniques exploit both the spatial and frequency localization of wavelets. The idea is to group coefficients of comparable significance across scales by spatial location in bands oriented in the same direction. The early approach of Lewis and Knowles [30] was extended by Shapiro in his landmark paper on embedded zerotree wavelet coding [42] and the best performing schemes are descendents or variations on this theme. The approach provides codes with excellent rate-distortion tradeoffs, modest implementation complexity, and an embedded bit stream, which makes the codes useful for applications where scalability or progressive coding are important. Scalability implies there is a 'successive approximation' property in the bit stream. As the decoder gets more bits from the encoder, the decoder can decode a progressively better reconstruction of the image. This feature is particularly attractive for a number of applications, especially those where one wishes to view an image as soon as bits begin to arrive, and where the image improves as further bits accumulate. With scalable coding, a single encoder can provide a variety of rates to customers with different channels or display capabilities. Since images can be reconstructed to increasing quality as additional bits arrive, it provides a natural means of adjusting to changing channel capacities and a more effective means of using a relatively slow channel.

After experimenting with a variety of algorithms, we chose Said and Pearlman's variation [39] of Shapiro's EZW algorithm because of its good performance and the availability of working software for 12 bpp originals. We use the default filters (the 9-7 biorthogonal filters) in the software compression package of Said and Pearlman [39]. These filters are considered, for example, in Antonini [3] and Villasenor et al. [50]. A description and discussion of the algorithm along with access to the software may be found at the World Wide Web site [38]. The algorithm applies a succession of thresholds to each coefficient, each half the size of the preceding. Coefficients with magnitude smaller than the threshold are deemed insignificant and are effectively quantized to zero. Bits are sent only to indicate the location of pixels that fall above the thresholds, and they are sent in an order determined by a subset partitioning algorithm that takes advantage of the correlation across scales of significance according to spatial location and orientation. Once a pixel is deemed significant, further bits sent regarding that pixel are devoted to refining the accuracy of the actual location by bit plane transmission. The bits are sent so as to first describe the largest coefficients, which contribute the most to the reconstruction accuracy. In this way the bit stream can be stopped at any point with a good reproduction for the given number of bits. The system incorporates the adaptive arithmetic coding algorithm considered in Witten et al. [54].

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

The image test set was compressed in this manner to three bit rates: 1.75, 0.4 and 0.15 bpp, where the bit rates refer to rates in the ROI. The average bit rates for the full image thus depended on the size of the ROI. An example of the Said–Pearlman algorithm with a 12 bpp original and 0.15 bpp reproduction is given in Fig. 1. For comparison purposes we also compressed a few images using a perceptually optimized JPEG [52], however those results are not included in this paper.

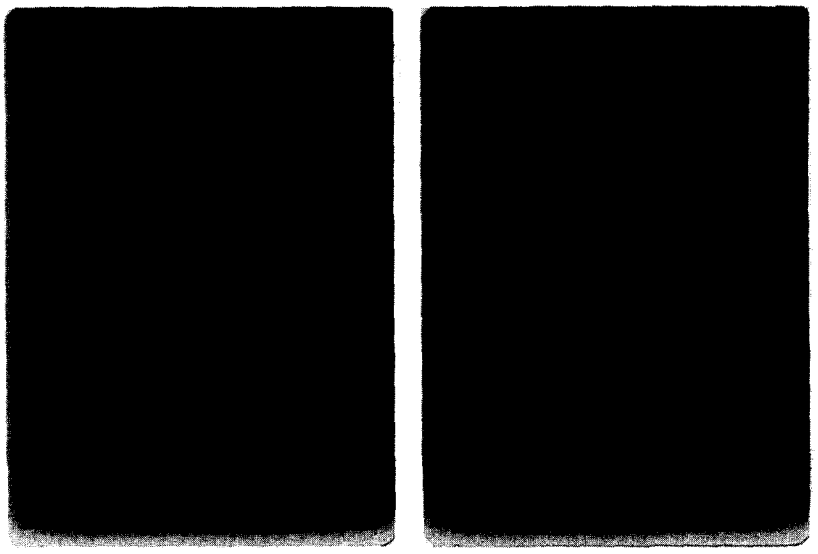

Fig. 1. Original image (left) and compressed image at 0.15 bpp in the ROI (right).

3. Results and discussion

The clinical experiment took place at Stanford University Hospital during spring 1996. The gold standard was established by E. Sickles, M.D., Professor of Radiology, University of California at San Francisco, and Chief of Radiology, Mt. Zion Hospital, and D. Ikeda, Assistant Professor and Chief, Breast Imaging Section, Department of Radiology, Stanford University, an independent panel of expert radiogists, who evaluated the test cases and then collaborated to reach agreement. The majority of the detected items were seen by both radiologists. Any findings seen by only one radiologist were included. The other type of discrepancy resolved was the class of the detected lesions. Since the same abnormality may be classified differently, the two radiologists were asked to agree on a class.

Table 3
Average SNR: ROI, wavelet coding

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

3.1. SNR versus bit rate

The SNRs are summarized in Tables 3 and 4. The SNR definition is $10\log_{10} E/\mathrm{MSE}$, where MSE denotes the average squared error and E denotes the energy of the digital original pixels. The overall averages are reported as well as the averages for the specific image types or views (left and right breast, CC and MLO view). This demonstrates the variability among various image types as well as the overall performance. Two sets of SNRs and bit rates are reported: ROI only and full image. For the ROI SNR the rates are identical and correspond to the nominal rate of the code used in the ROI. For the full images the rates vary since the ROI code is used in one portion of the image and

Table 4
Average SNR: full image, wavelet coding

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

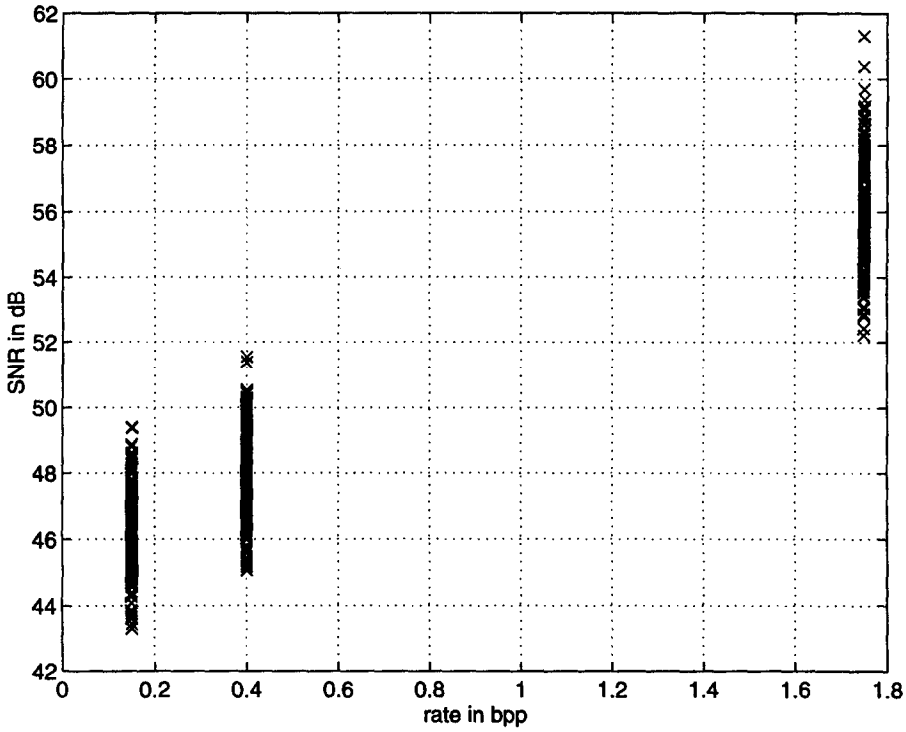

Fig. 2. Scatter plot of ROI SNR: wavelet coding.

much lower rate code is used in the remaining background and the average depends on the size of the ROI, which varies among the images. A scatter plot of the ROI SNRs is presented in Fig. 2.

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

3.2. Management differences

The focus of the statistical analysis of this paper is the screening and management of patients and how it is affected by analog versus digital and lossy compressed digital. We also consider the less important, but still informative, issue of

Table 5
Agreement 2 × 2 tables for radiologist A

II/I	R	W	II/I	R	W	II/I	R	W	II/I	R	W
R	7	2	R	0	0	R	6	4	R	14	2
W	1	2	W	0	1	W	3	5	W	2	8
	RTS			F/U			C/B			BX	

(A) Analog versus digital original

II/I	R	W	II/I	R	W	II/I	R	W	II/I	R	W
R	6	3	R	0	0	R	8	2	R	14	2
W	1	2	W	0	1	W	2	6	W	1	9
	RTS			F/U			C/B			BX	

(B) Analog versus digital lossy compressed: 1.75 bpp

II/I	R	W	II/I	R	W	II/I	R	W	II/I	R	W
R	6	3	R	0	0	R	6	4	R	12	3
W	0	3	W	0	1	W	2	6	W	4	6
	RTS			F/U			C/B			BX	

(C) Analog versus digital lossy compressed: 0.4 bpp

II/I	R	W	II/I	R	W	II/I	R	W	II/I	R	W
R	4	4	R	0	0	R	3	7	R	11	4
W	0	3	W	0	1	W	4	4	W	4	6
	RTS			F/U			C/B			BX	

(D) Analog versus digital lossy compressed: 0.15 bpp

subjective perceived quality as a function of bit rate.

In all, there were 57 studies that figure in what we report. According to the gold standard, the respective numbers of studies of each of the four types management types RTS, F/U, C/B and BX were 13, 1, 18 and 25, respectively. For each of the four possible outcomes, the analog original is compared to each of four technologies: digitized from analog original, and wavelet compressed to three different levels of compression (1.75, 0.4 and 0.15 bpp). So the McNemar 2 × 2 statistics based on the generic table of Table 2 for assessing differences between technologies were computed 48 times, 16 per radiologist, for each competing image modality (original digital and the three lossy compressed bit rates). For example, the 2 × 2 tables for a single radiologist (A) comparing analog to each of the other four modalities are shown in Table 5. For none of these tables for any radiologist was the exact binomial attained significant level (p-value) 0.05 or less. For our study and for this analysis, there is nothing to choose in terms of being 'better' among the analog original, its digitized version, and three levels of compression, one rather extreme. We admit freely that this limited study had insufficient power to permit us to detect small differences in management. The larger the putative difference, the better our power to have detected it.

Table 6 summarizes the performance of each radiologist on the analog versus uncompressed digital and lossy compressed digital. In all cases, columns are 'digital' and rows 'analog'. Table 6(A)

Table 6
Radiologist agreement tables

	RTS	F/U	C/B	BX		RTS	F/U	C/B	BX		RTS	F/U	C/B	BX
RTS	11	0	5	1	RTS	4	0	0	0	RTS	8	0	6	1
F/U	0	0	0	0	F/U	0	0	0	1	F/U	0	0	0	0
C/B	3	0	11	7	C/B	3	0	3	3	C/B	1	0	10	1
BX	2	0	2	15	BX	1	0	7	35	BX	0	0	7	23

A: Analog versus digital

	RTS	F/U	C/B	BX		RTS	F/U	C/B	BX		RTS	F/U	C/B	BX
RTS	11	0	6	0	RTS	2	1	0	1	RTS	11	0	4	0
F/U	0	0	0	0	F/U	0	1	0	0	F/U	0	0	0	0
C/B	2	0	15	4	C/B	3	1	3	2	C/B	1	1	8	2
BX	1	0	2	16	BX	1	0	4	37	BX	1	0	5	24

B: Analog versus lossy compressed digital: 1.75 bpp

	RTS	F/U	C/B	BX		RTS	F/U	C/B	BX		RTS	F/U	C/B	BX
RTS	9	0	6	2	RTS	1	0	2	1	RTS	7	0	7	1
F/U	0	0	0	0	F/U	0	0	0	1	F/U	0	0	0	0
C/B	1	0	10	10	C/B	2	0	2	5	C/B	2	0	8	2
BX	1	0	2	15	BX	2	0	5	36	BX	1	0	4	25

C: Analog versus lossy compressed digital: 0.4 bpp

	RTS	F/U	C/B	BX		RTS	F/U	C/B	BX		RTS	F/U	C/B	BX
RTS	8	0	7	1	RTS	3	1	0	0	RTS	7	0	7	0
F/U	0	0	0	0	F/U	0	0	0	1	F/U	0	0	0	0
C/B	3	1	9	8	C/B	3	0	3	2	C/B	0	0	9	3
BX	1	0	6	11	BX	1	1	5	35	BX	0	0	9	20

D: Analog versus lossy compressed digital: 0.15 bpp

Radiologist A | Radiologist B | Radiologist C

treats analog versus original digital and Tables 6(B)–(D) treat analog versus lossy compressed digital at bit rates of 1.75, 0.4 and 0.15 bpp, respectively. Statements which follow are with respect to the independent gold standard regarding which some information is implicit in Table 5. Consider as an example the analog versus digital comparison of radiologist A. Radiologist A made 20 'mistakes' of 57 studies from analog, and 24 from original digital studies. The most frequent mistake, eight for analog and seven for digital, was classifying a gold standard 'biopsy' as 'additional assessment'. Radiologist B made 28 'mistakes' from analog studies, and 24 from digital. In both cases, the most frequent mistake was to 'biopsy' what should, by the gold standard, have been 'additional assessment'. There were 15 such mistakes with analog and 13 with digital. Radiologist C made 20 'mistakes' from analog studies and 17 from digital. With the former, the most frequent mistake occurred eight times when 'biopsy' was judged when 'additional assessment' was correct. With digital, the most frequent mistake occurred six times when 'additional assessment' was judged when 'biopsy' was correct. On this basis, we cannot say that analog and digital are different beyond chance.

Both Tables 5 and 6 suggest that radiologists differ substantially from each other. However, comparing radiologists is not a goal of this study; we are interested in what happens when a particular radiologist views the same image under different modalities. The difference among radiologists merely make it more difficult to evaluate the difference among analog, digital, and lossy compressed images, since extreme care must be taken when doing any pooling or averaging of results across radiologists.

The runs test for learning did not find any learning effect at the 5% significance level for these management outcomes. For each of the 3 judges, approximately half of the studies were not included in the computation of the statistic, since the management decision was unchanging. For the 3 judges, the numbers of studies retained in the computation were 28, 28 and 27. The Z values obtained were -0.12, -0.86 and -0.22, with corresponding p-value of 0.452, 0.195 and 0.413. Further testing for learning will include an analysis of the detected findings.

3.3. *Management sensitivity and specificity*

The means and variances of the sensitivity and specificity and the mean of the PVP of the management decisions with respect to the independent gold standard are summarized in Table 7.

Level 1 refers to the analog images, level 2 to the uncompressed digital, and levels 3, 4 and 5 refer to those images where the breast section was compressed to 0.15, 0.4 and 1.75 bpp, respectively (and where the label was compressed to 0.07 bpp). In this table, sensitivity, specificity and PVP are defined relative to the independent gold standard. The table does not show any obvious trends for these parameters as a function of bit rate. Sensitivity is the ratio of the number of cases a judge calls 'positive' to the number of cases actually 'positive' according to the independent gold standard. Here 'positive' is defined as the union of categories F/U, C/B and BX. A 'negative' study is RTS. Sensitivity and specificity can be thought of as binomial issues, and so if the sensitivity is p, then the variance associated with that sensitivity is $p(1-p)$. The standard deviation calculation for PVP is somewhat more complicated and is not included here; because PVP is the ratio of two random quantities (even given the gold standard), the variance calculation requires approximate statistical methods as in analyses by 'propagation of errors'.

Table 7
Sensitivity, specificity and PVP

		Sensitivity		Specificity		
Level	Judge	mean	stdev	mean	stdev	PVP mean
1	A	0.826	0.379	0.692	0.462	0.905
1	B	1.000	0.000	0.308	0.462	0.836
1	C	0.913	0.282	0.846	0.361	0.955
2	A	0.886	0.317	0.769	0.421	0.929
2	B	0.955	0.208	0.385	0.487	0.840
2	C	0.932	0.252	0.462	0.499	0.854
3	A	0.814	0.389	0.333	0.471	0.814
3	B	0.953	0.211	0.417	0.493	0.854
3	C	0.977	0.151	0.500	0.500	0.875
4	A	0.860	0.347	0.615	0.487	0.881
4	B	0.955	0.208	0.154	0.361	0.792
4	C	0.977	0.149	0.615	0.487	0.896
5	A	0.841	0.366	0.538	0.499	0.860
5	B	0.953	0.211	0.231	0.421	0.804
5	C	0.932	0.252	0.769	0.421	0.932

Table 8
Subjective scores

Level	Judge	Mean	Stdev
1	A	3.90	0.97
1	B	4.52	0.75
1	C	4.59	0.79
2	A	3.91	0.41
2	B	3.85	0.53
2	C	3.67	0.65
3	A	3.82	0.39
3	B	4.27	0.93
3	C	3.49	0.64
4	A	3.91	0.39
4	B	3.93	0.55
4	C	3.82	0.50
5	A	3.92	0.42
5	B	3.66	0.57
5	C	3.82	0.55
	Judges pooled		
1	pooled	4.33	0.89
2	pooled	3.81	0.55
3	pooled	3.86	0.76
4	pooled	3.88	0.49
5	pooled	3.80	0.57

3.4. *Subjective ratings versus bit rate*

In the previous sections, objective measure of the quality of the compressed images were analysed via the SNR values and patient management decisions on the digitally compressed images. It is also informative to examine the effects of compression on subjective opinions. Table 8 provides the means and standard deviations for the subjective scores for each radiologist separately and for the radiologists pooled. The distribution of these subjective scores are displayed in Figs. 3–5.

Fig. 3 displays the frequency for each of the subjective scores obtained with the analog images. Fig. 4 displays the frequency for each of the subjective scores obtained with the uncompressed digital images (judges pooled), and Fig. 5 displays the frequency for each of the subjective scores obtained with the digital images at Level 3.

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

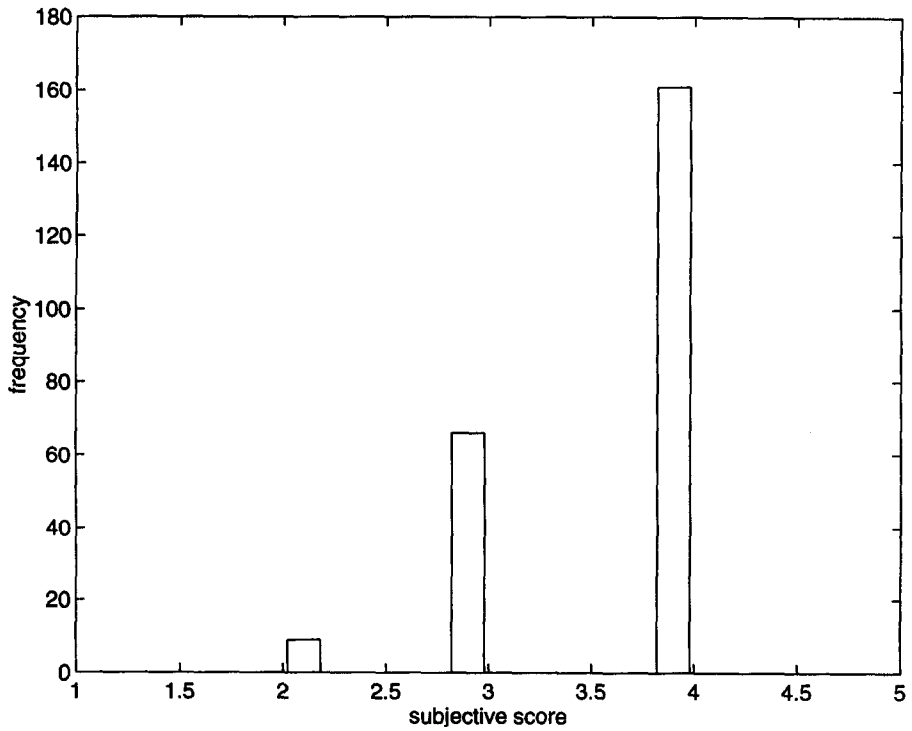

Fig. 3. Subjective scores: analog images.

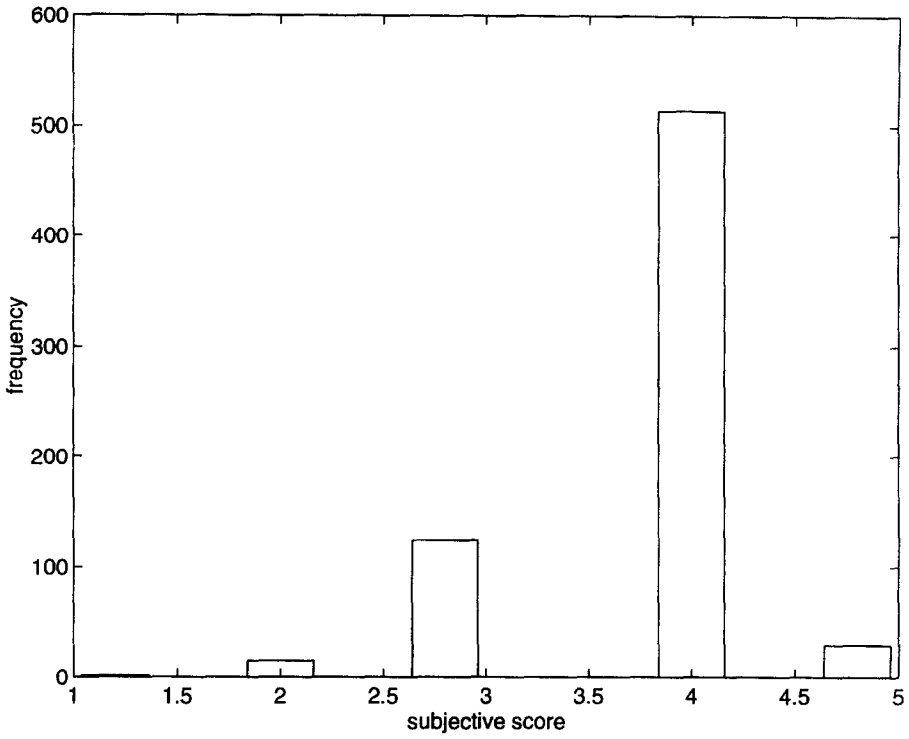

Fig. 4. Subjective scores: original digital images.

Judge A: All levels were significantly different from each other except the digital to 0.04 bpp, digital to 1.75 bpp, and 0.4 to 1.75 bpp.

Judge B: The only differences that were significant were 0.15 bpp to 0.4 bpp and 0.15 bpp to digital.

Judge C: All differences significant.

All judges pooled: All differences were significant except digital to 0.15 bpp, digital to 1.75 bpp, 0.15 to 0.4 bpp, and 0.15 to 1.75 bpp.

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

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

4. Comments

The goal of this project was to demonstrate a protocol for evaluating quality in various image modalities. The particular example was to show that digital mammograms and lossy compressed digital mammograms using an embedded wavelet code at 0.15 bpp yields image quality with no statistically significant differences from the analog original, as

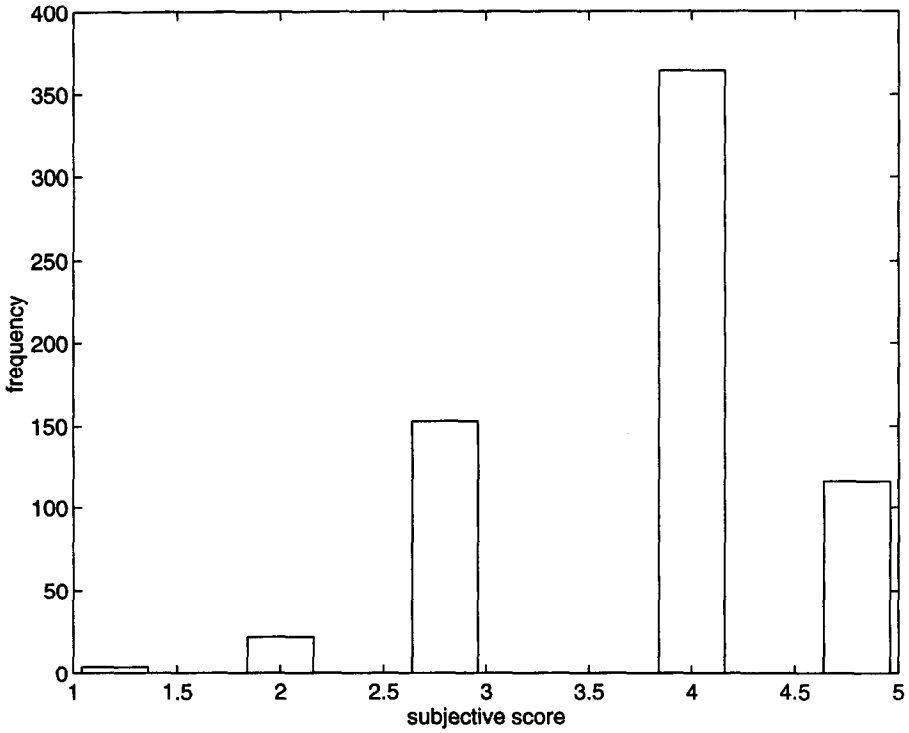

Fig. 5. Subjective scores: lossy compressed digital images at 0.15 bpp.

measured by an appropriate clinical experiment and statistical analyses that are germane to the question. We have argued that perceived subjective quality does differ significantly, but our suspicion is that much of this difference is based on portions of the image that are not important to screening or diagnosis and that this problem can be corrected with better background coding and film printing. All of the differences due to digitization and lossy compression were small with respect to the differences among individual radiologists, which suggests that great care must be taken with any statistical analysis which attempts to draw conclusions based on the pooling of radiologists.

To our knowledge, this is the largest data-gathering experiment of this kind conducted, and the only experiment of this kind to be analyzed by exact methods without Gaussian assumptions or artificial confidence ratings. The study can only be considered a pilot study, however, as the number of patients is too small to provide good statistical site and power for the tests considered We have considered elsewhere the issue of the number of patients required for a definitive demonstration of the essential equivalence of lossy compressed digital mammograms and analog originals in screening applications [1, 21], but even this issue must await the gathering of larger data sets to be resolved. Our current estimates are that a test set of 520 patient studies and 12 radiologists, each radiologist reading half the studies, would suffice.

Acknowledgements

The authors gratefully acknowledge the many hours of help contributed by Stan Rossiter, M.D., Edward Sickles, M.D., and by Sarah Horine and Dalia Gomez. This work was supported by the U.S. Army Medical Research and Materiel Command Breast Cancer Research Initiative under Grant DAMD 17-94-J-4354, and by Kodak, Inc.

Appendix A

ID number _____ Session number _____ Case number _____

Reader initials _____
Mammograms were of (**Left** **Right** **Both**) breast(s).

.....

Subjective rating for diagnostic quality:

(bad) 1–5 (good):

Left CC	Left MLO	Right CC	Right MLO

If any rating is <4 the problem is:
(1) sharpness (2) contrast (3) position (4) breast compression
(5) noise (6) artifact (7) penetration

Recommend repeat? **Yes** **No**

Breast Density: Left **1** **2** **3** **4** Right **1** **2** **3** **4**

(1) almost entirely fat (2) scattered fibroglandular densities
(3) heterogeneously dense (4) extremely dense

Findings: **Yes** **No**
Note: If there are NO findings, the assessment is: **(1) (N) negative–return to screening**

Findings(detection): **Dominant** **Incidential, focal** **Incidential, diffuse**

Individual finding side: **Left** **Right** **Both/Bilateral** Finding #_____ of _____

Finding type: **(possible, definite)**

(1) mass
(2) clustered calcifications
(3) mass containing calcifications
(4) mass with surrounding calcs
(5) spiculated mass
(6) ill defined mass
(7) architectural distortion
(8) solitary dilated duct
(9) asymmetric breast tissue
(10) focal asymmetric density
(11) breast edema

(12) multiple scattered and occasionally clustered benign appearing calcs
(13) occasional scattered benign appearing calcs
(14) multiple benign appearing masses
(15) skin lesion
(16) milk of calcium
(17) plasa cell mastitis/secretory calcs
(18) oil cysts
(19) lymph node
(20) fibroadenoma
(21) calcified fibroadenoma
(22) bascular calcs
(23) dermal/in calcs
(24) post biopsy scar
(25) reduction mammoplasty
(26) implants
(27) benign mass

(28) other

Location:

(1) UOQ	**(5) 12:00**	**(9) outer/lateral**	**(13) whole breast**	**(17) both breasts/bilateral**
(2) UIQ	**(6) 3:00**	**(10) inner/medial**	**(14) central**	
(3) LOQ	**(7) 6:00**	**(11) upper/cranial**	**(15) axillary tail**	
(4) LIQ	**(8) 9:00**	**(12) lower/inferior**	**(16) retroareolar**	

View(s) in which finding is seen: **CC** **MLO** **CC and MLO**

Associated findings include: (p = possible, d = definite)

(1) breast edema	**(p, d)**	**(8) architectural distortion**	**(p, d)**
(2) skin retraction	**(p, d)**	**(9) calcs associated with mass**	**(p, d)**
(3) nipple retraction	**(p, d)**	**(10) multiple similar masses**	**(p, d)**
(4) skin thickening	**(p, d)**	**(11) dilated veins**	**(p, d)**
(5) lymphadenopathy	**(p, d)**	**(12) asymmetric density**	**(p, d)**
(6) trabecular thickening	**(p, d)**	**(13) none**	**(p, d)**
(7) scar	**(p, d)**		

Assessment: **The finding is**

(A) indeterminate/incomplete, additional assessment needed

What? (1) spot mag (2) extra views (3) U/S (4) old films (5) mag

What is your *best guess* as to the finding's 1–5 assessment? _____ or are you uncertain if the finding exists? *Y*

(1) (N) negative – return to screening
(2) (B) benign (also negative but with benign findings) – return to screening
(3) (P) probably benign finding requiring 6-month followup
(4L) (S) suspicion of maliganancy (low), biopsy
(4M) (S) suspicion of maliganancy (moderate), biopsy
(4H) (S) suspicion of malignancy (high), biopsy
(5) radiographic maliganancy, biopsy

Comments: _____

_____

Measurements:

CC View Size: _____ cm long axis by _____ cm short axis
Distance from center of finding to: nipple _____ cm left edge
_____ cm, top edge _____ cm

MLO View Size: _____ cm long axis by _____ cm short axis
Distance from center of finding to: nipple _____ cm
left edge _____ cm, top edge _____ cm

References

[1] C.N. Adams, A. Aiyer, B.J. Betts, J. Li, P.C. Cosman, S.M. Perlmutter, K.O. Perlmutter, D. Ikeda, L. Fajardo, R. Birdwell, B.L. Daniel, S. Rossiter, R.A. Olshen, R.M. Gray, Evaluating quality and utility of digital mammograph and lossy compressed digital mammograms, in: K. Doe, M.L. Griger, R.M. Nishikawa (Eds.), Digital Mammography '96. Proc. 3rd Internat. Workshop on Digital Mammograhy, Chicago, USA, 9–12 June 1996, Elsevier, Amsterdam, pp. 169–176. (Reprint available at http://www-isl.stanford.edu/~gray/army.html.)

[2] I. Andersson, Mammography in clinical practice, Med Radiography Photography 62 (2) (1986) 2.

[3] M. Antonini, M. Barlaud, P. Mathieu, I. Daubechies, Image coding using wavelet transform, IEEE Trans. Image Proces. 1 (1992) 205–220.

[4] H.H. Barrett, T. Gooley, K. Girodias, J. Rolland, T. White, J. Yao, Linear discriminants and image quality, in: Proc. 1991 Internat. Conf. Inform. Process. in Med. Imaging (IPMI '91), Wye, UK, Springer, Berlin, July 1991, pp. 458–473.

[5] Y. Bishop, S. Feinberg, P. Holland, Discrete Multivariate Analysis, MIT Press, Cambridge, MA, 1975.

[6] H. Blume, ACR-NEMA Digital Imaging and Communications Standard Committeee, Working Group #4, MEDPACS Section, Data Compression Standard #PS2, 1989.

[7] J. Bramble, L. Cook, M. Murphey, N. Martin, W. Anderson, K. Hensley, Image data compression in magnification hand radiographs, Radiology 170 (1989) 133–136.

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

[9] K. Chan, S. Lou, H. Huang, Full-frame transform compression of CT and MR images, Radiology 171 (3) (1989) 847–851.

[10] L. Clarke, G. Blaine, K. Doi, M. Yaffe, F. Shtern, G. Brown, Digital mammography, cancer screening: Factors important for image compression, in: Proc. Space and Earth Science Data Compression Workshop, Snowbird, Utah, NASA, April 1993.

[11] J. Cohen, Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit, Psychol. Bull. 20 (1968) 213–220.

[12] P.C. Cosman, H.C. Davidson, C.J. Bergin, C. Tseng, L.E. Moses, R.A. Olshen, R.M. Gray, The effect of lossy compression on diagnostic accuracy of thoracic CT images, Radiology 190 (2) (1994) 517–524.

[13] P. Cosman, R. Gray, R. Olshen, Evaluating quality of compressed medical images: SNR, subjective rating, and diagnostic accuracy, Proc. IEEE 82 (1994) 919–932.

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

[15] H.C. Davidson, C.J. Bergin, C. Tseng, P.C. Cosman, L.E. Moses, R.A. Olshen, 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, December 1991.

[16] D. Dershaw, A. Abramson, D. Kinne, Ductal carcinorma in situ: mammographic findings and clinical implications, Radiology 170 (1989) 411–415.

[17] W. Feller, Introduction to Probability Theory, 3rd ed., Wiley, New York, 1968.

[18] J. Frisell, G. Eklund, L. Hellstrom, A. Somell, Analysis of interval brest carcinomas in a randomized screening trial in Stockholm, Breast Cancer Res Trea 9 (1987) 219–225.

[19] A. Garber, R. Olshen, H. Zhang, E. Venkatraman, Predicting high-rish cholesterol levels, Internat. Statist. Rev. 62 (2) (1994) 203–228.

[20] M. Goldberg, M. Pivovarov, W. Mayo-Smith, M. Bhalla, J. Blickman, R. Bramson, G. Boland, H. Llewellyn, E. Halpem, Application of wavelet compression to digitized radiographs, Amer. J. Radiology 163 (1994) 463–468.

[21] R.M. Gray, R.A. Olshen, D. Ikeda, P. Cosman, S. Perlmutter, C. Nash, K. Perlmutter, Evaluating quality and utility in digital mammography, in: Proc. 1995 IEEE Internat. Conf. on Image Process., IEEE, October 1995. Vol. II, Washington, DC, October 1995, pp. 5–8.

[22] R.M. Gray, R.A. Olshen, D. Ikeda, P.C. Cosman, S.M. Perlmutter, C. Nash, K.O. Perlmutter, Measuring quality in computer processed radiological images, in: Proc. 29th Asilomar Conf. on Signals Systems Comput., October 1995.

[23] J.A. Hanley, B.J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC)curve, Diagnostic Radiology 143 (1982) 29–36.

[24] D. Ikeda, I. Andersson, Atypical mammographic presentations of ductal carcinoma in situ, Radiology 172 (1989) 661–666.

[25] D. Ikeda, I. Andersson, L. Janzon, F. Linell, Radiographic appearance and prognostic consideration of interval carcinorma in the Malmo mammographic screening trial, Amer. J. Roentgenology 159 (1992) 287–294.

[26] T. Ishigaki, S. Sakuma, M. Ikeda, Y. Itoh, M. Suzuki, S. Iwai, Clinical evaluation of irreversible image compression: Analysis of chest imaging with computed radiography, Radiology 175 (1990) pp. 739–743.

[27] W. Kegelmeyer, Software for image analysis aids in breast cancer detection, OE Reports, February 1993, p. 7.

[28] W.P. Kegelmeyer, Jr., Evaluation of stellate lesion detection in a standard mammogram data set, in: Proc. IS&T/SPIE Annual Symp. on Electronic Imaging Sci. Technol., San Jose, CA, January–February 1993.

[29] H. Lee, A.H. Rowberg, M.S. Frank, H.S. Choi, Y. Kim, Subjective evaluation of compressed image quality, in: Proc. Med. Imaging VI: Image Capture, Formatting, and Display, Vol. 1653, SPIE, February 1992, pp. 241–251.

[30] A.S. Lewis, G. Knowles, Image compression using the 2-D wavelet transform, IEEE Trans. Image Process. 1 (1992) 244–250.

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

[32] J. Martin, M. Moskowitz, J. Milbrath, Breast cancer missed by mammography, Amer. J. Roentgenology 132 (1979) 737–739.

[33] B.J. McNeil, J.A. Hanley, Statistical approaches to the analysis of receiver operating characteristic (ROC) curve, Med. Dec. Making 4 (1984) 137–150.

[34] C.E. Metz, Basic principles of ROC analysis, Seminars Nucl. Med. VIII (October 1978) 282–298.

[35] S. Perlmutter, C. Tseng, P. Cosman, K. Li, R. Olshen, R. Gray, Measurement accuracy as a measure of image quality in compressed mr chest scans, in: Proc. IEEE 1994 Internat. Symp. on Image Process., Vol. 1, San Antonio, TX, October 1994, pp. 861–865.

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

[37] W.B. Richardson, Jr., Nonlinear filtering and multiscale texture discrimination for mammograms, in: Proc. SPIE, Vol. 1768 of Mathematical Methods in Medical Imaging, San Diego, California, SPIE, July 1992, pp. 293–305.

[38] A. Said, W.A. Pearlman, Set partitioning in hierarchical trees, http://ipl.rpi.edu/SPIHT/.

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

[40] J. Sayre, D.R. Aberle, M.I. Boechat, T.R. Hall, H.K. Huang, B.K. Ho, P. Kashifian, G. Rahbar, Effect of data compression on diagnostic accuracy in digital hand and chest radiography, in: Proc. Med. Imaging VI: Image Capture, Formatting and Display, Vol. 1653, SPIE, February 1992, pp. 232–240.

[41] T. Senoo, B. Girod, Vector quantization for entropy coding of image subbands, IEEE Trans. Image Process. (1992) 526–532.

[42] J. Shapiro, Embedded image coding using zerotrees of wavelet coefficients, IEEE Trans. Signal Process. 41 (December 1993) 3445–3462.

[43] G.W. Snedecor, W.G. Cochran, Statistical Methods, Iowa State University Press, Ames, IA, 1989.

[44] Stanford University Compression and Classification Group, USARMC Digital Mammography Image Quality Project, http://www-isl.stanford.edu/~gray/army.html, 1996.

[45] S.J. Starr, C.E. Metz, L.B. Lusted, D.J. Goodenough, Visual detection and localization of radiographic images, Radiology 116 (1975) 533–538.

[46] P. Stomper, J. Connolly, J. Meyer, J. Harris, Clinically occult ductal carcinoma in situ detected with mammography: analysis of 100 cases with radiographic-pathologic correlation, Radiology 172 (1989) 235–241.

[47] J. Storer, Data Compression, Computer Science Press, Rockville, MD, 1988.

[48] J.A. Swets, ROC analysis applied to the evaluation of medical imaging techniques, Invest. Radiology 14 (1979) 109–121.

[49] M. Vetterli, J. Kovacevic, Wavelets and Subband Coding, Prentice-Hall, Englewood Cliffs, NJ, 1995.

[50] J. Villasenor, B. Belzer, J. Liao, Wavelet filter evaluation for image compression, IEEE Trans. Image Process. 4 (8) (1995) 1053–60.

[51] G. Wallace, The JPEG still picture compression standard, Comm. ACM 34 (1991) 30–44.

[52] A.B. Watson, Visually optimal DCT quantization matrices for individual mages, in: Proc. 1993 IEEE Data Compression Conf. (DCC), 1993, pp. 178–187.

[53] P. Wilhelm, D.R. Haynor, Y. Kim, E.A. Riskin, Lossy image compression for digital medical imaging system, Opt. Eng. 30 (1991) 1479–1485.

[54] I.H. Witten, R.M. Neal, J.G. Cleary, Arithmetic coding for data compression, Comm. ACM 30 (1987) 520–540.

[55] J.W. Woods (Ed.), Subband Image Coding, Kluwer Academic Publishers, Boston, 1991.

[56] K. Woods, J. Solka, C. Priebe, C. Doss, K. Bowyer, L. Clarke, Comparative evaluation of pattern recognition techniques for detection of microcalcifications, in: Proc. IS&T/SPIE Annual Symp. on Electronic Imaging Sci. Technol., San Jose, CA, January–February 1993.