Prague Medical Report / Vol. 106 (2005) No. 1, p. 5–26

Quality Measurement of Lossy Compression in Medical Imaging

Smutek D.
Third Department of Medicine of the First Faculty of Medicine,
Charles University in Prague, Czech Republic

Received December 23, 2004, Accepted February 11, 2005

Abstract: At the time when most of image data in hospitals are stored in digital form using picture archiving and communication systems (PACS), telemedicine goes through its boom, and demand for data storage and bandwidth requirements increases, lossy compression techniques become necessity. This review article summarizes different methods for quality measurement of image compression in radiology. After brief compression techniques description, technical and medical measurements (including Receiver Operating Characteristic Curves) of image compression are described. Employing of these methods in practice and their results are shown on sample studies. The article concludes with basic recommendations for experimental protocols when performing quality measurement of medical images.

Key words: Images quality – Data compression – Receiver operating characteristic (ROC) curve – Picture archiving and communication system (PACS).

Paper was supported by Grant agency of Academy of Science of the Czech Republic by project IET 101050403.

Mailing Address: Ing. Daniel Smutek, MD., PhD., Third department of Medicine of the First Faculty of Medicine, Charles University, U Nemocnice 1, 128 08 Prague 2, Czech Republic, Phone: +420 224 962 958, Fax: +420 224 919 780, e-mail: smutek@cesnet.cz

Introduction

The objective of radiological image compression is to reduce the volume of data and to achieve a low bit rate in the digital representation of radiological images without perceived loss of image quality. However, the demand for transmission bandwidth and storage space in the digital radiology environment, especially picture archiving and communication systems (PACS) and teleradiology, and the proliferating use of various imaging modalities, such as magnetic resonance imaging, computed tomography, ultrasonography, nuclear medicine, computed radiography, digital subtraction angiography, positron emission tomography, single photon emission computerised tomography, and digital fluorography continue to outstrip the capabilities of existing technologies. The availability of lossy coding techniques for clinical diagnoses further implicates many complex legal and regulatory issues.

The overall goal of compression 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 [1].

Compression techniques

Technically, all image data compression schemes can be broadly categorised into two types. One is **reversible compression**, also referred to as "lossless". A reversible scheme achieves only modest compression ratios, but allows exact recovery of the original image from the compressed version. Lossless coding techniques (a predictive model, a multi-resolution model or both) are well understood, readily available, e.g., [2–4], and typically yield compression ratios of 2:1 to 3:1 on still-frame medical images.

An **irreversible scheme**, or a "lossy" scheme (2D discrete cosine transform, full-frame discrete cosine transform, lapped orthogonal transform, subband coding, vector quantization, quadtrees, and adaptive predictive coding schemes) [5–6], does not allow exact recovery after compression, but can achieve much higher compression ratios, ranging from ten to fifty or more. Generally speaking, more compression is obtained at the expense of more image degradation, i.e., the image quality declines as the compression ratio increases. Lossy coding is unavoidable if the original image is analogue, as is ordinary X-ray film. Digitisation of an analog signal causes a loss of information and hence a possible deterioration of the signal. Analog information is converted into a relatively small number of bits. This operation is nonlinear and noninvertible. The conversion can operate on individual pixels (scalar quantization) or groups of pixels (vector quantization). Quantization is fundamentally lossy [7] and can include throwing away some of the components of the signal decomposition

step [8]. An advantage of a well-designed lossy compression system is that it minimises information loss or image distortion for a given allotted storage space or communication rate. When bits are scarce, good compression schemes devote the available bits to the information of greatest importance. As a result, lossy compression schemes are capable of enhancing specific structures of importance to the viewer.

Another type of compression which is used in medical imaging is **clinical image compression**, which stores a few medically relevant images, as determined by the physicians, out of a series of real-time images and thus reduces the total image size. The stored images may be further compressed [7].

Image degradation may be visually apparent. The term "visually lossless" has been used to characterize lossy schemes that result in no visible loss under normal radiological viewing conditions [6]. A related term used by the American College of Radiology and National Electrical Manufacturing Association (ACR/NEMA) is "information preserving." The ACR/NEMA standard report defines a compression scheme to be information preserving if the resulting image retains all of the significant information of the original image. Both "visually lossless" and "information preserving" are subjective definitions and extreme caution must be taken in their interpretations [7].

Performance Measure of Image Compression

There are different ways for technical measuring image compression. The bit rate of a compression system is the average number of bits produced by the encoder for each image pixel. Compression ratios must be interpreted with care as they depend crucially on the image type (most medical images are dominated by high frequency regions against a flat, low frequency background), original bit rate, sampling density, how much background is in the image, and how much coding of the background figures into the calculation [1]. Compression ratio is defined as the word-length of the image data divided by the bit-rate in b/pixel (bpp) of the compressed data.

The instrumentation engineers characterise the digital image by three physical parameters: density resolution, spatial resolution, and signal-to-noise ratio (SNR). The density resolution is the total number of discrete grey level values in the digital image, the spatial resolution measures the number of pixels used to represent the objects. For two images of fixed density and spatial resolutions, a high signal-to-noise ratio means that the image is very pleasing to the eyes. Typically, a power spectrum is used to study the noise of reconstructed images. The relationships between compression ratio and the three parameters are discussed in [9].

The manufacturers use the frequency representation of an image to measure the quality of its sharpness. This leads to the notions of point spread function, line spread function, edge spread function, and modulation transfer function [9], which

measure the sharpness of points, lines, and edges, as well as the system response at a spatial frequency.

Medical measurements (and Receiver Operating Characteristic Curves)
The images of interest to the clinicians are complex anatomical objects. Simulated measures resulting from the engineer's and the manufacturer's experiments do not always correlate well with human subjective testing and perception. To remedy this situation, numerical statistical analyses based on the receiver operating characteristic (ROC) curves [10, 11] are used to examine medical image quality for individual applications.

For medical images, subjective quality is often identified with diagnostic accuracy. The limitations of diagnostic "accuracy" as a measure of decision performance require introduction of the concepts of the "sensitivity" and "specificity" of a diagnostic test. These measures and the related indices "true positive fraction" and "false positive fraction", are more meaningful than "accuracy", yet do not provide a unique description of diagnostic performance because they depend on the arbitrary selection of a decision threshold. The ROC curve is shown to be a simple complete empirical description of this decision threshold effect, indicating all possible combinations of the relative frequencies of the various kinds of correct and incorrect decisions. ROC curve is a plot of the true positive rate against the false positive rate for the different possible cutpoints of a diagnostic test.
ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making and can be employed to optimise various diagnostic strategies [12].

In practical implementation a filtered version of a signal (target) with superimposed Gaussian noise is sampled and thresholded. If the sample value exceeds the threshold, the target is assumed to be present, otherwise not present. A conditional probability, which is a function of the threshold, can be associated with the decision, once stated whether the signal was present or not. An ROC curve plots the true positive fraction (TPF, also called sensitivity, the complement of the probability of Type I error) versus the false positive fraction (FPF, the complement of specificity). TPF and FPF respectively represent the probability that the target is detected when present, and the probability that it is detected when it is absent. Depending on the threshold, TPF and FPF may assume different values. In particular, if the threshold is very low, both of them equal 1, since every occurrence of the target is correctly detected, but the target is always claimed also when absent. On the contrary, when the threshold is very high, both FPF and TPF are 0; in fact the target is never discovered when absent, but it is nev er detected even when present. These considerations suggest an ROC curve representing a trade-off between true positive and false positive decisions. Obviously, detection is as much accurate, as much TPF is high, and FPF is low. The area underlying the ROC curve (dashed line in Fig. 1) ranges between 0.5 and 1 and represents a

quality index of the efficiency of the detection process. In particular, the value of 0.5 for the area is obtained when TPF and FPF are equal irrespective of the threshold value (heads-and-tails decision). In this case the ROC curve is just a straight diagonal line (chance line) partitioning the ROC plane [13]. Swets [11] has experimentally demonstrated the relation between the false positive rate (FPR) and the true positive rate (TPR).

The extension of ROC analysis to the medical field is not quite immediate. In fact, decision thresholds are not easily definable as in signal detection theory, where the origins of ROC are, but are implicit in the judgement of physicians. In a ROC study, experts of the field, or typical users are asked to review the reconstructed images after compression, which either did or did not possess an abnormality and to provide a binary decision, i.e., abnormality present or not, along with a quantitative value for their degree of certainty, usually a number from 1 to 5 (subjective trials). A subjective confidence rating of diagnoses is then used as if it were a threshold to adjust for detection accuracy [11]. The diagnostic accuracy of reconstructed images is compared with that of the original plain films. Since the viewers rate diagnostic usefulness rather than general appearance or simply line or edge patterns, these studies relate diagnostic accuracy to compression level. It should be emphasised that the diagnostic quality is evaluated, not the efficiency of observers. Radiologists can be trained to use the rating scale and the results can be combined with assumptions on the nature of the data to produce summary statistics reflecting the diagnostic accuracy [14–16].

The observers' ratings can be brought back to thresholds adjustable to higher detection accuracy, thereby originating points in the ROC plane [12]. A continuous curve can then be produced by interpolating these points, in order to easily define and obtain the requested quality measurement.

Certain statistical model, such as the bivariate binormal distribution, would be used to test differences between ROC curves based on correlated data sets. ROC analyses are used to quantify the compression levels in a specific medical

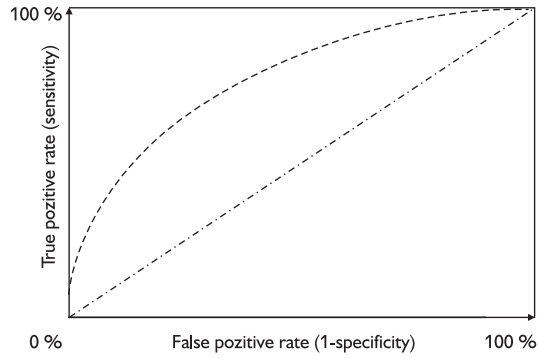

Figure 1 – Sample receiver operating characteristic curve.

application that can be used without a statistically significant change in diagnostic accuracy. ROC analyses are, however, expensive and time-consuming to perform. Caution must be expressed in the ROC evaluation, as its degree of significance depends on the number of test images, and on the number of radiologists taking part to the experiments [13]. A typical ROC study would require over 300 images to obtain a reasonable statistical confidence level, five or more radiologists to view these images, and a full-time statistician to coordinate and analyse the data [7].

Evaluating/Measure of Medical Image Quality According to Compression
Methods of evaluating image quality can be divided into two groups: 1) subjective ratings such as statistical analyses of viewers' scores on quality (e.g., analysis of variance (ANOVA)), paired comparisons, receiver operating characteristics (ROC) curves, sensitivity and positive predictive value and 2) objective rating such as numerical Signal-to-Noise Ratios (SNRs), average distortion, average difference, structural content, normalised cross-correlation, correlation quality, maximum difference, image fidelity, weighted distance, Laplacian mean square error, peak mean square error, (normalised) absolute error, (normalised) mean square error, L_p-norm [17] and graphical Visual Differences Predictor (VDP) [18], histograms of the compression error [19], Hosaka plots [20], and Eskicioglu charts [19].

Quantitative measures for image quality can be classified according to two criteria: 1) the number of images used in the measurement; 2) the nature or type of measurement. According to the first criterion, the measures are divided into two classes: **univariate** and **bivariate**. A univariate measure uses a single image, whereas a bivariate measure is a comparison between two images. According to the second criterion, there are again two classes: **numerical** and **graphical**. A numerical measure takes one (if the measure is univariate) or two (if the measure is bivariate) images as input, and processes the pixel values by an integration rule. The output of this processing is a single integer or real number.

Table 1 – Classification of image quality criteria [21]

Subjective	Objective	
	Numerical	Graphical
Absolute	Mean Square Error	Visual Differences Predictor
Comparative	Lp-norm	Histograms
	Power spectrum	Hosaka plots
	Other	Eskicioglu charts

A typical approach is to obtain useful statistics (activity in blocks, errors on edges, block distortions, etc.) about the impairments in a compressed image. These statistics are then combined in a weighted sum to represent the error characteristics. Reduction of the output to a single value, however, is a major drawback because much of the useful information is lost. Graphical measures do not reduce their output into a scalar value. They are multi-dimensional measures in the form of images, histograms, plots, or charts [21].

Quality measurements are usually made using the pixel elements of digitized images. For more accurate assessment, a continuous image field can be generated by two-dimensional interpolation of the pixel matrix [21].

Mean Square Error (MSE) measures punctual variations of the image intensity by averaging the squared differences between couples of corresponding pixels. Signal to Noise Ratio (SNR) and Peak Signal to Noise Ratio (PSNR), can be directly derived from the MSE using equations, that assume that distortion introduced by the coding-decoding operation can be modelled as a kind of noise [22]. Laplacian mean square error captures information relating to edge features. Edge information is known to be an image property to which the human visual system is highly sensitive [23].

Subjective ratings are possible to divide to absolute and comparative evaluation techniques. Absolute evaluation is a process whereby the observer assigns to an image a category in a given rating scale, whereas comparative evaluation is the ranking of a set of images from best to worst. Another technique belonging to comparative group is bubble sort – observer takes two images A and B from a group, and compares them. If his order is AB, he picks a third image to establish the order ABC or ACB. If the order is ACB, then a comparison is needed between A and C. The procedure ends with the best image at the top if no ties are allowed [21].

Subjective methods are usually preferred in quality measurement in medical applications. According to Cosman [8] lossy compressed images should be judged by their use in making accurate diagnoses, i.e., a more natural and fundamental aspect of relative image quality. Classical ROC analyses are the most credible and acceptable way to measure the image quality by the radiologists, because they include subjective appraisals of the value of an image for a particular application [7]. In a medical application it does not suffice for an image to simply “look good” or to have a high 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 [1]. Subjective rating results may not be reproducible as they can be affected by a number of factors including such as type, size and range of images, observers’ background and motivation or experimental conditions (lighting, display quality, etc.) [21].

Table 2 – Objective measurements of image quality [17, 22, 24]

Image Quality Measure	Formula
MSE	$\sum_{i=1}^{m}\sum_{j=1}^{n}[f(i,j)-f'(i,j)]^2/m\cdot n$
SNR	$\frac{\sigma_x^2}{MSE}$; $\sigma_x^2=\frac{1}{m\cdot n}\sum_{i=1}^{m}\sum_{j=1}^{n}(f(i,j)-\bar{f})$ where $\bar{f}=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}f(i,j)}{m\cdot n}$
PSNR	$\frac{(2^b)^2}{MSE}$
Average Distance/Difference	$\sum_{i=1}^{m}\sum_{j=1}^{n}[f(i,j)-f'(i,j)]^2/m\cdot n$
Structural Content	$\sum_{i=1}^{m}\sum_{j=1}^{n}[f(i,j)^2/\sum_{i=1}^{m}\sum_{j=1}^{n}f'(i,j)^2$
Normalised Cross Correlation	$\sum_{i=1}^{m}\sum_{j=1}^{n}f(i,j)f'(i,j)/\sum_{i=1}^{m}\sum_{j=1}^{n}f(i,j)^2$
Correlation Quality	$\sum_{i=1}^{m}\sum_{j=1}^{n}f(i,j)f'(i,j)/\sum_{i=1}^{m}\sum_{j=1}^{n}f(i,j)$
Maximum Difference	$\max\{\lvert f(i,j)-f'(i,j)\rvert\}$
Image Fidelity	$1-\left[\sum_{i=1}^{m}\sum_{j=1}^{n}[f(i,j)-f'(i,j)]^2/\sum_{i=1}^{m}\sum_{j=1}^{n}f(i,j)^2\right]$
Weighted Distance	Every element of the difference matrix is normalised in some way and L_1-norm is applied
Laplacian MSE*	$\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}[O\{f(i,j)\}-O\{f'(i,j)\}]^2/\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}[O\{f(i,j)\}]^2$
Peak MSE	$\frac{1}{m\cdot n}\sum_{i=1}^{m}\sum_{j=1}^{n}[f(i,j)-f'(i,j)]^2/[\max(f'(i,j))]$
Normalised Absolute Error*	$\sum_{i=1}^{m}\sum_{j=1}^{n}\lvert O\{f(i,j)\}-O\{f'(i,j)\}\rvert/\sum_{i=1}^{m}\sum_{j=1}^{n}\lvert O\{f(i,j)\}\rvert$
Normalised MSE*	$\sum_{i=1}^{m}\sum_{j=1}^{n}[O\{f(i,j)\}-O\{f'(i,j)\}]^2/\sum_{i=1}^{m}\sum_{j=1}^{n}[O\{f(i,j)\}]^2$
L_p-norm	$\left\{\frac{1}{m\cdot n}\sum_{i=1}^{m}\sum_{j=1}^{n}\lvert f(i,j)-f'(i,j)\rvert^p\right\}^{\frac{1}{p}}$, $p=1, 2, 3$

*Note that for LMSE $O\{f(i,j)\}=f(i+1,j)+f(i-1,j)+f(i,j+1)+f(i,j-1)-4f(i,j)$. For NAE, NMSE and L_2-norm $O\{f(i,j)\}$ is defined in 3 ways: 1) $O\{f(i,j)\}=f(i,j)$; 2) $O\{f(i,j)\}=f(i,j)^{1/3}$; 3) $O\{f(i,j)\}=H\{(u^2+v^2)^{1/2}\}f(i,j)$ where u and v are co-ordinates in the DCT transform domain, $r=(u^2+v^2)^{1/2}$ and $H(r)=\left\{0.05e^{r^{0.554}} \text{ for } r<7;\ e^{-9[\lvert\log_{10}r-\log_{10}\rvert]^{2.3}} \text{ for } r\geq 7\right\}$.

Sample Studies

Studies using either subjective, either objective methods

Some studies involve either objective either subjective methods, they use SNRs and statistical analyses on viewers' scores. Examples of such approaches may be found in [15, 24–27]. When viewers rate diagnostic usefulness rather than simply general appearance, these studies relate compression level to diagnostic accuracy. In other studies, radiologists are asked to view an image which either does or does not possess an abnormality and to provide a binary decision (abnormality present or not) along with a quantitative value for their degree of certainty. Subsequent statistical analyses, usually ROC-based, attempted to quantify the levels of compression in a specific application that can be used without a statistically significant change in diagnostic accuracy. There are numerous examples of such approaches, e.g. [14, 16,17, 26–28]. In these studies, the basic experiments were subjective and did not simulate the ordinary tasks of radiologists. The observers were asked to rate numerically their confidence or their opinion of image quality or usefulness rather than to make diagnoses as they would under ordinary clinical conditions. This rating resulted in data useful for ROC analysis, but it constitutes an artificial diagnostic task. Furthermore, radiologists often face images, which may contain one or more abnormalities, and the diagnostic task is to find any and all that are present. In this case the task is not binary, and is not amenable to traditional ROC analysis techniques. Lastly, some studies used paired comparisons, where an original and a compressed image were displayed simultaneously and a radiologist was asked to rate the difference. This procedure differs markedly from ordinary clinical practice [8].

Cosman [8] applies a lossy compression algorithm to medical images and quantifies the quality of the images by the diagnostic performance of radiologists, as well as by traditional signal-to-noise ratios ("The traditional manner for comparing the performance of different lossy compression systems is to plot distortion rate or SNR versus bit rate curves" [8]). Her study is unlike previous studies of the effects of lossy compression. She considers non-binary detection tasks, simulates actual diagnostic practice instead of using paired tests or confidence rankings, uses statistical methods that are more appropriate for non-binary clinical data than are the popular ROC curves, and uses low-complexity predictive tree-structured vector quantization for compression rather than DCT-based transform codes combined with entropy coding.

In [29] Cosman describes three dominating approaches to the measurement of medical image quality: 1) computable objective distortion measures such as mean squared error or signal-to-noise ratio (SNR) and peak signal-to-noise ratio (PSNR), 2) subjective quality as measured by psychophysical tests or questionnaires with numerical ratings, and 3) clinical simulation and statistical analysis of a specific application of the medical images, e.g., diagnostic accuracy.

In **SNR approach** Cosman emphasises that a distortion measure (in Cosman's work segmental and conventional SNR are used) should have three desirable properties: easing computation, reflecting perceptual quality, and tractability in analysis. Ideal subjectively meaningful distortion measure could be incorporated into the system design. There are techniques for incorporating subjective criteria into compression system design, but these tend to be somewhat indirect. For example, it is possible to transform the image and to assign bits to transform coefficients according to their perceptual importance or use postfiltering to emphasise important subbands before compression [30]. These simple computable measures have a role in the design algorithms and in the evaluation of quality because they are quickly and cheaply obtainable, and tractable in analysis. On the other hand one of the drawbacks of computable measures is that they do not take in account the medical nature of the images.

In **Subjective Ratings** approach formalised subjective testing methods such as Mean Opinion Score (MOS) and Diagnostic Acceptability Measure (DAM) are mentioned with the reservation that there is no standardisation for rating still images. For medical images, more important than subjective quality may be a computable objective measure for predicting diagnostic accuracy. The common method is plotting mean values of SNR against mean values of the corresponding subjective quality and then fitting a curve to the resulting points. Many curves have been considered, including polynomial splines, quadratics, and exponentials. The residual sum of errors then provides an indication of the goodness of the fit. Another popular method is to measure the correlation coefficient between the fitted and actual data points [31–33]. A drawback of assessment of medical image quality by perceptual measures is that it requires the detailed, time-consuming, and expensive efforts of human observers, typically highly trained radiologists.

The most common mean of measuring **diagnostic accuracy** for computer-processed medical images is based on receiver operating characteristic (ROC) analysis. A variety of summary statistics such as the area under the ROC curve can be computed and used for evaluation.

Cosman names several shortcomings of ROC: the necessity for the radiologists to assign specific values to their confidence departs from ordinary clinical practice, methods relying on Gaussian assumptions when image data are non-Gaussian (this problem can be overcome by using computer-intensive statistical sample reuse techniques which can help get around the failures of Gaussian assumptions). Many clinical detection tasks are non-binary, in which case specificity does not make sense because it has no natural or sensible denominator, as it is not possible to say how many abnormalities are absent. Another two drawbacks are mentioned in [1]: traditional ROC analysis does not come equipped to distinguish among the various possible notions "ground truth" or "gold standard" in clinical experiments (different kinds of standards are discussed in one of next paragraphs) and 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). ROC studies are also too specific to cover the wide range of medical imaging modalities and applications [19].

Genuinely non-binary detection tasks (locating any and all abnormalities that are present) are not amenable to ordinary ROC analysis techniques. Extensions to ROC to permit consideration of multiple abnormalities have been developed [34], but they require the use of confidence ratings as well as Gaussian or Poisson assumptions on the data. Finally, ROC analysis has no natural extension to the evaluation of measurement accuracy in compressed medical images.

Further Cosman points out that for measuring diagnostic accuracy, a "gold standard" is needed first. She distinguishes four different gold standards: 1) a consensus gold standard which is determined by the consensus of the three judges on the original; 2) a personal gold standard which uses each judge's readings on an original (uncompressed) image as the gold standard for the readings of that same judge on the compressed versions of that same image; 3) an independent gold standard which is formed by the agreement of the members of an independent panel of particularly expert radiologists and 4) a separate gold standard that is produced by the results of autopsy, surgical biopsy, reading of images from a different imaging modality, or subsequent clinical or imaging studies. She concludes that the personal and consensus gold standards are most useful for comparing the various compressed levels among themselves.

Because acquiring perceptual measures is very demanding (highly trained specialists), it is desirable to find computable measures such as SNR that strongly correlate with or predict the perceptual measures. Cosman's work suggests that cross-validated fits to the data using generalized linear models can be used to examine the usefulness of SNR (or other computable measure) as a predictor for subjective quality (or other perceptual measure).

Interesting Cosman's conclusion is that radiologists are trained to interpret only certain kinds of images, and when they are asked to look at another type of image (e.g., compressed or highlighted) they may not do as well just because they were not trained on those. But with image enhancement techniques or slightly compressed images, perhaps a radiologist trained on those would do better when reading those than someone trained on originals would do reading originals. This corresponds to findings in [1] that the observers in particular expressed dissatisfaction with the fact that the background in the digitally produced films was not as dark as that of the photographic films, even though this ideally had nothing to do with their diagnostic and management decisions.

Perlmutter [1] suggests that the protocol for subjective evaluation should simulate ordinary clinical practice as closely as possible. Participating 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 should include examples of images containing the full range of possible findings. Perlmutter describes a general protocol for performing clinical experiment simulating ordinary practice and suitable statistical analysis for image equality (quantifiable manner that a specific image mode is at least equal to another). In comparison with others Perlmutter also takes in account that learning and fatigue are both processes that might change the score of an image depending upon when it was seen. In his work they looked for whether learning effects were present in the management outcomes using "runs" test. They also segmented the sample image into a region of interest (ROI) and a background. The background was coded using the same algorithm, but at only lower bit rate. They report SNRs and bit rates for both the full image and for the ROI. One of Perlmutter's very important conclusions is that all the differences due to digitisation and lossy compression were small with respect to the differences among individual radiologists (inter-observer variability). It suggests that great care must be taken with any statistical analysis, which attempts to draw conclusions based on the pooling of radiologists.

Studies using only subjective methods

Wu [35] uses both subjective and objective measurement for evaluating compression algorithm in his research. Two radiologists verify that images are acceptable for practical application (subjective quality of decoded image – they use ROC protocol suggested in [36]) and peak signal-to noise ratio (PSNR), expressed in decibels is used as objective measure. Wu states that "PSNR has been accepted as a widely used quality measurement in the fields of image compression".

Van Schelven [37] use ROC in comparing between Advanced Multiple Beam Equalization Radiography (AMBER) and conventional screen-film radiography. His study comprises patients with interstitial processes of lungs such as cryptogenic fibrosing alveolitis and sarcoidosis without hiliar lymphadenopathy. ROC is performed by nine readers (six senior and three resident radiologists) which use 5 point rating scale. One of interesting findings is that residents performed better than radiologists ($p = 0.054$) in conventional radiography.

Gooley [38] uses ROC for comparison of statistical methods of image reconstruction. He considers ROC to be an objective (not subjective as it is according to others) as method for comparisons. Based on the responses of ten observers, an ROC curve was constructed for each observer and each method, and a value for the performance index known as the area under the curve (AUC) was obtained from each of these curves. The AUC was estimated by use of the non-parametric Mann-Whitney test as suggested Hanley [39]. The value for the AUC for each method was then taken as the average value across all observers for each method. Gooley uses "properly" trained students as observers and they use six-point scale for image description. Interesting finding is that student observers

perform as well as physicians (on simple detection tasks). Use of training before observing is in discrepancy with Betts [40] and Perlmutter [1], who suggest that such kind of studies should be performed without any special training and that they should be as close to common clinical practice as possible.

Baudin [36] uses ROC for validating a compression scheme applied on digitized wrist radiographs. He states that physical measurements as the error image or SNR do not inform about critical artefacts, which are unacceptable from a medical point of view. Thus he uses ROC, which takes into account the diagnostic quality of the reconstructed images. ROC in his study involves five graduate radiologists. They investigate the diagnostic performance of both original digitised films and associated compressed/decompressed images using ROC methodology to analyse the detectability of fractures of the scaphoid bone. Baudin pays great attention to realisation of a medical image database, dedicated to the evaluation of the diagnostic quality. The images in the database must represent the actual clinical reality a medical expert is likely to encounter, whereas their capacity of diagnosis must neither be too easy nor too difficult in order to bring the expert to use the full range of his diagnostic capability. Original and reconstructed images were not mixed during individual evaluation sessions, only one type of images was evaluated per session without the reader knowing it. A minimum of one week was maintained between two sessions of the same reader in order to minimise bias due to the learning effect.

In [40] Betts uses statistic methods similar to ROC to determine whether lossy compression affected the ability of a doctor to make a correct diagnosis. Six board-certified radiologists were presented original and compressed images. They were asked to determine whether any dominant findings were present in images and if findings were present, to associate a classification. Independent two expert radiologists established a gold standard. Separate examinations of sensitivity and specificity were performed with respect to the gold standard and not independent biopsy results. The statistics used in their experiment included McNemar's test for agreement counts, the Wilcoxon signed-rank test and a permutation t-test for testing variance of sensitivity and positive predictive value across modalities. Betts achieved some surprising results such as that lossy compression tended to have a beneficial affect on sensitivity and specificity. He tried to explain it by fact that compression artefacts forced the judges to study compressed images more carefully, thus improving their sensitivity and specificity, or that wavelet coders enhanced visual cues needed for correct diagnosis.

Slone [41] evaluates the degree of irreversible image compression by subjective assesment of five observers (two imaging scientists and three board-certified radiologists). Two versions of an image were compared on a single monitor by using an interactive soft-copy feature on an image-comparison workstation. This method is supposed to be very sensitive. It exploites the observer's temporal sensitivity to differences in the image, because the human visual system is naturally

drawn to changes in structure or brightness. This technique allows detection of subtle differences and provides a mechanism for comparing image quality loss caused by different kinds of distortion (JPEG and wavelet-based trellis-coded quantization (WTCQ) algorithms were used in this study). Slone uses three different image evalutation methods: 1) two-alternative forced choice where no ties were allowed (observer was forced to choose even if he percieved no difference between images); 2) original-revealed two-alternative forced choice, in which the noncompressed image was identified to the observer; and 3) a resolution-metric method where observers decided which level of blurring most closely matched, with respect to clinical utility, the level of compression. This method alows to compare various compression techniques.

Perlmutter [42] shows another approach for investigation of the effects of lossy image compression. He checks measurement accuracy in magnetic resonance images compressed to five different levels using predictive pruned tree-structured vector quantization (predictive PTSVQ). Three radiologists measured the diameters of the four principal blood vessels on each image. Using t and Wilcoxon tests no significant differences in measurement were found up to compression 16:1. The approach such as protocol, establishing gold standard, etc. were similar as in his and his co-authors other papers [1, 29, 40].

Quality judgement across various algorithms and graphical measurements

Quality judgement across various algorithms such as JPEG (using discrete cosine transform technique, public release of the Independent JPEG Group's JPEG software), EPIC (using wavelet transform technique, Vision Science Group, The Media Laboratory, MIT), RLPQ (wavelet transform/fractals, Department of Computer Sciences, University of North Texas), and SLPQ (wavelet transform/fractals, Department of Computer Sciences, University of North Texas) is much more complex since a wide range of image impairments is involved.

The choice of the compression techniques for an investigation of the performance of quality measures (especially those that are graphical) is important since it is desirable to include techniques, which produce different types of impairments in the reconstructed images. The major types of degradation in the images are blockiness with JPEG, blurriness with EPIC, both fuzziness and blockiness with RLPQ, and fuzziness with SLPQ (The term fuzziness is used in the sense of equal amount of blurriness over the entire image) [19].

In Eskicioglu's [19] opinion a major problem in evaluating lossy techniques is the extreme difficulty in describing the type and amount of degradation in reconstructed images. He tries to find a quantitative measure, either in numerical or graphical form, not only for judging the quality of images obtained by a particular algorithm, but also for quality judgement across various algorithms where a wide range of image impairments is involved. He compares the outcomes of scalar objective quality measures with graphical measures called Hosaka plots

and histograms. He states that the first mentioned quality measures are all bivariate, exploiting the differences between corresponding pixels in the original and degraded images. He shows that although some numerical measures correlate well with the observers' response (at first the images were subjectively evaluated by ten observers) for a given compression technique and that they can reliably be used to specify the magnitude of degradation in reconstructed images for a given compression technique, an evaluation across different techniques is not possible. This is because a single scalar value cannot be used to describe a variety of impairments. He concludes that a graphical measure called Hosaka plots, can be used to appropriately specify not only the amount, but also the type of degradation in reconstructed images (particularly useful when the impairment is blockiness – JPEG and RLPQ compressions) and that the Hosaka plots provide a good indication of how images lose their fidelity. Histograms did not prove as useful.

To construct a **Hosaka plot**, or an h-plot, original image is segmented by quadtree decomposition into certain activity regions. Five classes of blocks are formed with this decomposition. Two features are computed for each class – the average standard deviation of the blocks and the average mean of the blocks less the average mean of the classes. With the same segmentation, these features are also computed for the reconstructed image. Then features of reconstructed and original image are compared [20, 43]. The difference between the two feature vectors generates a vector error measure, which, unlike scalar quantities, allows a description not only of the amount, but also of the type of degradation. Such information is extremely helpful considering the sensitivity of the human observer to the location of the image error.

In spite of these advantages, Hosaka plots could not properly describe the type of loss, i.e., the nature and distribution of error. They provide limited information concerning the activity levels in different areas of the reconstructed images. A major drawback of Hosaka plots is its absolute measurement of the error in the two features. It is not possible to know whether there is an increase or a decrease in the standard deviations or the means of the blocks when the degradation level in an image is changed [19].

Despite Eskicioglu does not recommend using simple scalar quality measures (mainly because they do not represent different degradation of image caused by different compression schemas, but also because their performance is poor with higher compression ratio), he divides them into three groups: 1) Average Difference and Structural Content; 2) N.Cross-Correlation, Correlation Quality, Laplacian Mean Square Error and Maximum Difference; 3) Weighted Distance, Peak Mean Square Error, Image Fidelity, N. Absolute Error, N. Mean Square Error and L_p-norm. The measures in Group 1 cannot be reliably used with all techniques as the sign of the correlation coefficient does not remain the same. Group 2 measures are consistent, but nevertheless have poor correlation with the observers' response for some of the techniques. Among the useful measures in

Group 3, NMSE is the best one for all the test images. Except for a single case, the incorporation of the aspects of Human Visual System into NMSE makes the correlation slightly stronger. Similar conclusion, that Human Visual System (HVS) does not always improve the correlation, and when it does, the gain is small, can be found in [44] and [45]. Nowadays there is enough evidence to show that simple HVS models incorporated into numerical quality measures result in higher correlation with subjective assessment. But some of the numerical measures exploiting the HVS have limited use and scope as they are able to detect only a particular type of distortion (mostly blockiness) [21].

Another (apart from Hosaka plots) measures, which take in account human perception parameters, are information content and perceptual distortion measure. **Information content** (IC) is based on the evaluation of the perceptual distortion. It consists of five stages: 1) the original image is re-mapped by a non-linear transformation; 2) a linear transformation on the DCT domain is applied to 8×8 image blocks; 3) a matrix of coefficients is calculated at fixed resolution; 4) the DCT coefficients are multiplied by the weights; 5) IC is computed by summing the coefficient magnitudes. The **perceptual distortion measure** is based on an empirical model of the human perception of spatial patterns. The model consists of four stages: 1) front-end linear filtering; 2) squaring; 3) normalization; 4) detection [22].

Hermiston [46] presents a method of graphical and scalar image quality bivariate measurement utilizing integer wavelet transformations. The measure can perform a similar function to the Hosaka plot whilst not requiring segmentation and threshold parameters. The measure can represent separately the components of image distortion such as noise and blur through relative energy in the wavelet transform subbands. However, according to Hermiston, the application of graphical measures remains limited. He says it is more popular to use scalar image quality measures, which represent image distortion in a more concise manner. Through weighted summation, the graphical measure can be degenerated into a single number. The scalar measure was then compared with other image quality measures and found to present consistently high correlation with subjective image quality assessment (active image analysts from military intelligence centres within UK participated) using National Imagery Interpretability Rating Scale (NIIRS). Not only proposed measurement, but also MSE, Image Fidelity, PMSE and normalised MSE showed correlation with NIIRS assessment. The unexpectedly good performance of MSE appears to negate much of the criticism of its low correlation with subjective assessment. This finding is in concordance with [17]. When subjective quality assessment was performed, it was generally recognised that interpreting compressed imagery would become easier with experience of the artefact characteristics that a particular compression algorithm can produce.

Another paper by Eskicioglu [19] evaluates the performance of Hosaka plots and Eskicioglu charts. Mimicking the human visual system, they compute local features,

and produce a graphical output. They are quantitative and general, facile, inexpensive, and quick to apply, they are not affected by dc-shifts, and more informative than scalar measures. They have the potential of determining the perceived image quality, which eliminates the need for elaborate computations to simulate the human visual system.

Eskicioglu charts are described in [17] and [19]. Eskicioglu starts with a similar decomposition as for Hosaka plots, but he uses only four classes. Three features are computed for each class and then normalised by the number of pixels divided by the number of pixels in the entire image, the number of distinct pixel values divided by the number of possible pixel values and by the average standard deviation of the blocks divided by a preset maximum standard deviation. The fourth feature for each class is the average of the end of block disturbances normalised by a preset maximum disturbance. The features are displayed in bar charts. The segmentation for the reconstructed image is obtained separately, and the resulting bar chart is compared with that of the original. An alternative way of displaying the characteristics of distorted images is to use the quadtree decomposition of the original image. This modification, known as improved Eskicioglu charts, eliminates the first dimension, allowing a direct comparison of the respective features.

Eskicioglu concludes that the features, used in charts, represent the essential characteristics of the image and that charts are suitable for classifying the major types of impairment and expressing the nuances between the artefacts exhibited in images. They can be used by researchers and manufacturers with confidence in applying the lossy compression technology to medical imaging systems.

Not all authors use graphical quality measures for comparing between different algorithms. For example Aiazzi [13] uses for comparing of his encoder (based on an enhanced Laplacian pyramid) and JPEG compression two objective measures, SNR and percentage peak error (PE), and subjective measure, ROC analysis, but no graphical measure.

Giusto [22] suggests four innovative methods for blockness distortion measurement, two based on DCT analysis, and two on differential Sobel operator. Block distortion, or tiling effect is typical of any kind of block-based coding systems. It consists of a visual mosaic effect produced by the imperfect matching of neighbouring approximated blocks. His methods evaluate the amount of this particular but very usual image degradation. The proposed methods are evaluated on standard, not medical images.

Cen [47] presents experimental and statistical framework for comparing progressive progressive image compression algorithm (JPEG, EZW, and SPIHT) which represent an image in such a way that the decoder can reconstruct the image with increasing quality as more bits arrive. He says that traditional measurements such as SNR or subjective quality judgments may be inappropriate for evaluating progressive compression methods (In order to achieve good

performance as measured by PSNR, progressive algorithms have often focused on sending information on the largest DCT or wavelet coefficients first, in order to minimise MSE distortion at a given bit rate.). Cen's comparisons use response time studies in which human observers view a series of progressive transmissions, and respond to questions about the images as they become recognizable. His study involves no subjective opinions; it directly assesses image recognition by having observers respond to questions whose answer can only be known by recognising the image content. The images he used for his experiment were not medical images and observers were untrained persons.

Various Results Related to the Compression Level

Cosman [29] concludes that compression level with bit rate 0.56 bpp is unacceptable for diagnostic use. Since the blocking and prediction artefacts became quite noticeable at this level, the judges tended not to attempt to mark any abnormality unless they were quite sure it was there. This explains the initially surprising result level 0.56 bpp did well for positive predictive value, but very poorly for sensitivity. Since no differences were found among 1.8 bpp, 2.2 bpp, 2.64 bpp, and original images at 12 bpp these three compressed levels are clearly acceptable for diagnostic use in applications. The decision concerning levels 1.18 bpp and 1.34 bpp is less clear, and requires further tests involving a larger number of detection tasks, more judges, and use of an independent gold standard that in principle should remove at least one of the biases against compression that are present in their study.

Wu [35] demonstrates the effectiveness of the employment of the adaptive sampling algorithm to the DCT spectral domain. He concludes employing adaptive sampling to the spatial domain can achieve a bit rate of 0.33 bpp with a PSNR value of 37.85 dB. The bit rate achieved from spectral domain is 0.18 bpp with a PSNR value of 42.82 dB (with processing size 16×16). Because his method achieves a higher PSNR value than JPEG does, he deduces that the performance of his method is better than JPEG under the same compression ratio. He also suggests that different compression ratios are suitable for different imaging modalities: for the decoded X-ray image compression ratio below 20 is acceptable for practical applications, for an angiogram image the compression ratio of 45, for the CT bone image the compression ratio 35, and for sonogram image the compression ratio below 15.

Perlmutter [1] showed that digital mammograms and lossy compressed digital mammograms using an embedded wavelet code at 0.15 bpp yielded image quality with no statistically significant differences from the analog original (measured by an appropriate clinical experiment and statistical analyses).

In one of Baudin's [36] experiment, an area under the ROC curve constructed for compressed images was higher than the one computed on an evaluation of the diagnostic quality of the original, non-compressed images. This result can be

explained by the fact that the compression method eliminates the visual effect of high frequency noise, which is present in the original images. Other of Baudin's conclusion is that the average difference found between the area under the two ROC curves for digitised original images and compressed (by their method) images with 0.2 bpp is less than the average intra-expert variation (3–5%).

Slone [41] found the visually lossless threshold between 8:1 and 16:1 for JPEG and wavelet-based trellis-coded quantization algorithm. JPEG baseline algorithm resulted in performance as good as that with WTCQ compression at these ratios. At x2 magnification, images compressed with either JPEG or WTCQ algorithms were indistinguishable from unaltered original images for most observers at compression ratios between 8:1 and 16:1, indicating that 10:1 compression is acceptable for primary image interpretation.

Experimental Protocols

Choosing the proper protocol is necessary. There are several basic principles for protocol design [40]: The protocol should simulate ordinary clinical practice as closely as possible. In particular, participating 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 should include examples of images containing the full range of possible findings, all but extremely rare conditions. 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 subjects should be sufficient to ensure satisfactory size and power for the principal statistical tests of interest.

Also big attention should be paid to inter-observer and intra-observer variability. The former essentially summarises the idea that different doctors might give different diagnoses for a patient. Intra-observer variability captures the notion that a doctor might give a different diagnosis for a patient upon a second or third reading. No additional information is made available to the doctor but the diagnosis changes. A good study will take both of these sources of variability into account.

Prevalence of studied cases should be taken in account too – radiologists might behave differently if they knew that the prevalence in an experiment were different from that ordinarily encountered in a clinic. This effect could be analysed in a quantifiable manner by varying the prevalence at different sites in a controlled manner not known to the judges or assistants.

Conclusion

Crucial task is to define an objective measure of image quality for the lossy compression technology and put it into clinical environment. Ideal measure would evaluate the quality of radiological images, covering not only the global parameters, such as noise and bit-rate measurements, but also the local parameters, such as texture and sharpness.

The traditional noise and bit-rate measurements are insufficient, because they do not provide any information regarding the type of loss. On the other hand, besides costly and time consuming to perform, ROC studies are too specific to cover the wide range of medical imaging modalities and applications.

References

1. PERLMUTTER S. M., COSMAN P. C., GRAY R. M., OLSHEN R. A., IKEDA D., ADAMS C. N., BETTS B. J., WILLIAMS M. B., PERLMUTTER K. O., LI J., AIYER A., FAJARDO L., BIRDWELL R.: Image quality in lossy compressed digital mammograms. *Signal Processing* 59: 189–210, 1997.
2. DAVISSON L. D., MCELIECE R. J., PURSLEY M. B., WALLACE M. S.: Efficient Universal Noiseless Source Codes. *Ieee Transactions on Information Theory* 27: 269–279, 1981.
3. LYNCH T. J.: Data Compression: Techniques and Applications. In Lifetime Learning. Belmont C., Ed. Wadsworth, 1985.
4. STORER J.: Data Compression. Rockville, Computer Sci. Press, 1988.
5. NETRAVALI A., HASKELL B.: Digital Pictures: Representation and Compression. New York, Plenum, 1988.
6. RABBANI M. J. P.: Image compression techniques for medical diagnostic imaging systems. *J. Digit. Imaging. 4*: 65–78, 1991.
7. WONG S., ZAREMBA L., GOODEN D., HUANG H. K.: Radiologic Image Compression – A Review. *Proceedings of the Ieee* 83: 194–219, 1995.
8. COSMAN P. C., TSENG C., GRAY R. M., OLSHEN R. A., MOSES L. E., DAVIDSON H. C., BERGIN C. J., RISKIN E. A.: Tree-Structured Vector Quantization of Ct Chest Scans – Image Quality and Diagnostic-Accuracy. *Ieee Transactions on Medical Imaging* 12: 727–739, 1993.
9. HUANG H. K.: Elements of Digital Radiology. In A Professional Handbook and Guide. Englewood Cliffs N. J., Prentice-Hall, 1987.
10. METZ C. E.: Some Practical Issues of Experimental-Design and Data-Analysis in Radiological Roc Studies. *Investigative Radiology* 24: 234–245, 1989.
11. SWETS J.: ROC analysis applied to the evaluation of medical imaging techniques. *Investigative Radiology* 14: 109–121, 1979.
12. METZ C. E.: Basic principles of ROC analysis. *Semin. Nucl. Med.* 3: 282–298, 1988
13. AIAZZI B., ALPARONE L., BARONTI S., CHIRO G., LOTTI F., MORONI M.: Pyramid-based error-bounded encoder: An evaluation on X-ray chest images. *Signal Processing* 59: 173–187, 1997.
14. ISHIGAKI T., SAKUMA S., IKEDA M., ITOH Y., SUZUKI M., IWAI S.: Clinical-Evaluation of Irreversible Image Compression – Analysis of Chest Imaging with Computed Radiography. *Radiology* 175: 739–743, 1990.
15. MACMAHON H., DOI K., SANADA S., MONTNER S. M., GIGER M. L., METZ C. E., NAKAMORI N., YIN F. F., XU X. W., YONEKAWA H., TAKEUCHI H.: Data-Compression – Effect on Diagnostic-Accuracy in Digital Chest Radiography. *Radiology* 178: 175–179, 1991.
16. SAYRE J. W., ABERLE D. R., BOECHAT M. I., HALL T. R., HUANG H. K., BRUCE K., KASHFIAN P.,

RAHBAR G.: Effect of data compression on diagnostic accuracy in digital hand and chest radiography. 1653: 232–240, 1992. Newport Beach, CA, USA. Proceedings of SPIE – The International Society for Optical Engineering; Medical Imaging VI: Image Capture, Formatting, and Display. 1992.

17. ESKICIOGLU A. M., FISHER P. S.: Image quality measures and their performance. *Ieee Transactions on Communications* 43: 2959–2965, 1995.
18. DALY S.: The Visible Differences Predictor: An Algorithm for the Assessment of Image Fidelity. 1616: 2–15,1992. Proceedings of SPIE.
19. ESKICIOGLU A. M.: Multi-dimensional measure for image quality. 469. 1995. Snowbird, UT, USA. Proceedings of the Data Compression Conference. 28–10, 1995.
20. HOSAKA K.: A new picture quality evaluation method. 17–18. 1986. Tokyo, Japan. *Proc. Int. Picture Coding Symp*. 1986.
21. ESKICIOGLU A. M.: Quality measurement for monochrome compressed images in the past 25 years. 4: 1907–1910, 2000. IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings. 2000.
22. GIUSTO D. D., PERRA M.: Estimating blockness distortion for performance evaluation of picture coding algorithms. 1: 318–321, 1997. Victoria, Can. IEEE Pacific RIM Conference on Communications, Computers, and Signal Processing – Proceedings. 1997.
23. BRADLEY A. P.: A wavelet visible difference predictor. *Ieee Transactions on Image Processing* 8: 717–730, 1999.
24. BRAMBLE J. M., COOK L. T., MURPHEY M. D., MARTIN N. L., ANDERSON W. H., HENSLEY K. S.: Image Data-Compression in Magnification Hand Radiographs. *Radiology* 170: 133–136, 1989.
25. GOLDBERG M. A., PIVOVAROV M., MAYOSMITH W. W., BHALLA M. P., BLICKMAN J. G., BRAMSON R. T., BOLAND G. W. L., LLEWELLYN H .J., HALPERN E.: Application of Wavelet Compression to Digitized Radiographs. *American Journal of Roentgenology* 163: 463–468, 1994.
26. LEE H., ROWBERG A. H., FRANK M. S., CHOI H. S., KIM Y.: Subjective evaluation of compressed image quality. 1653: 241–251, 1992. Newport Beach, CA, USA. Proceedings of SPIE – The International Society for Optical Engineering; Medical Imaging VI: Image Capture, Formatting, and Display. 1992.
27. WILHELM P., HAYNOR D. R., KIM Y. M., RISKIN E. A.: Lossy Image Compression for Digital Medical Imaging-Systems. *Optical Engineering* 30: 1479–1485, 1991.
28. CHEN J., FLYNN M. J., GROSS B., SPIZARNY D.: Observer detection of image degradation caused by irreversible data compression processes. 1444: 256–264, 1991. San Jose, CA, USA. Proceedings of SPIE – The International Society for Optical Engineering; Medical Imaging V: Image Capture, Formatting, and Display. 1991.
29. COSMAN P. C., GRAY R. M., OLSHEN R. A.: Evaluating Quality of Compressed Medical Images – Snr, Subjective Rating, and Diagnostic-Accuracy. *Proceedings of the Ieee* 82: 919–932, 1994.
30. SAFRANEK R. J., JOHNSTON J. D., ROSENHOLTZ R. E.: Perceptually tuned sub-band image coder. 1249: 284–293, 1990. Santa Clara, CA, USA. Proceedings of SPIE – The International Society for Optical Engineering; Human Vision and Electronic Imaging: Models, Methods, and Applications. 1990.
31. MARMOLIN H.: Subjective Mse Measures. *Ieee Transactions on Systems Man and Cybernetics* 16: 486–489, 1986.
32. QUACKENBUSH S., BARNWELL T., CLEMENTS M.: Objective Measures of Speech Quality. 1988. Englewood Cliffs, N.J., Prentice-Hall. Prentice-Hall Signal Processing Series. 1988.
33. WANG S. H., SEKEY A., GERSHO A.: An Objective-Measure for Predicting Subjective Quality of Speech Coders. *Ieee Journal on Selected Areas in Communications* 10: 819–829, 1992.
34. CHAKRABORTY D. P., WINTER L. H. L.: Free-Response Methodology – Alternate Analysis and A New Observer-Performance Experiment. *Radiology* 174: 873–881, 1990.

35. WU Y. G.: Medical image compression by sampling DCT coefficients. *Ieee Transactions on Information Technology in Biomedicine* 6: 86–94, 2002.
36. BAUDIN O., BASKURT A., MOLL T., PROST R., REVEL D., OTTES F., KHAMADJA M., AMIEL M.: ROC assessment of compressed wrist radiographs. *European Journal of Radiology* 22: 228–231, 1996.
37. VANSCHELVEN I. H., WINTER L. H. L., CHAKRABORTY D. P., KOOL L. J. S.: A ROC Study of AMBER and conventional chest imaging in the detection of simulated interstitial lung disease. *European Journal of Radiology* 21: 67–71, 1995.
38. GOOLEY T. A., BARRETT H. H.: Evaluation of Statistical-Methods of Image-Reconstruction Through Roc Analysis. *Ieee Transactions on Medical Imaging* 11: 276–283, 1992.
39. HANLEY J. A. M. B.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143: 29–36, 1982.