MEASUREMENT ACCURACY AS A MEASURE OF IMAGE QUALITY IN COMPRESSED MR CHEST SCANS

Sharon M. Perlmutter
Info Systems Lab
Stanford University

Chien-Wen Tseng
Medical School
Univ. of CA, LA

Pamela C. Cosman
Info Systems Lab
Stanford University

King C. P. Li
Dept. of Radiology
Stanford University

Richard A. Olshen
Dept. of Biostatistics
Stanford University

Robert M. Gray
Info Systems Lab
Stanford University

ABSTRACT

We investigated the effects of lossy image compression on measurement accuracy in magnetic resonance images. Thirty chest scans were 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. Errors were analyzed relative to both an independent standard and personal performance on uncompressed images. Data were compared with both t and Wilcoxon tests. We conclude that for the purpose of measuring blood vessels in the chest, there is no significant difference in measurement accuracy when images are compressed up to 16:1 with predictive PTSVQ.

1. INTRODUCTION

The use for diagnosis of such medical imaging technologies as magnetic resonance imaging (MRI) and computerized tomography (CT) has recently expanded. One difficulty in working with the resulting digital medical images, however, is the size of the individual image files involved. For example, a typical MR image is 0.13 Mb and a typical CT image is 0.5 Mb. This situation is further compounded by the enormous number of images produced. Compression techniques are therefore desirable and often essential for cost and time efficiency of storage and communication.

Lossless compression alone is in general insufficient to attain ratios higher than 4:1. If greater compression ratios are desired, it is necessary to use lossy compression. The primary concern with medical images is that the diagnostic accuracy of the lossy compressed images remain at least as high as the diagnostic accuracy of the original images. A wide variety of diagnostic tasks need to be considered, including the measurement of structures and the detection of lesions. A protocol is thus needed to quantify the diagnostic accuracy of radiologists who make use of an image, whether or not it is compressed. From the implementation of such a protocol, one can then determine the minimum bit rate required for a medical image to retain acceptable diagnostic accuracy relative to a putative gold standard.

Most previous studies have investigated the effects of lossy compression on detection of structures [1, 2, 3]. Measurement tasks on structures such as blood vessels and tumors also take a central role in identifying abnormalities. In the evaluation of aneurysms, for example, size is an important prognostic feature in any presurgical assessment. An aneurysm is diagnosed if the aorta exceeds 4 cm in diameter. The size of the aneurysm bears heavily upon the risk of rupture. There is a 10% risk of rupture for aneurysms between 5 and 10 cm in diameter, and about 50% for aneurysms greater than 10 cm [4]. Because rupture is invariably fatal, operative repair is usually recommended when the aorta measures more than about 5 or 6 cm in diameter [5, 6]. The clinical decision, however, depends not only on the size of the aneurysm but also on the clinical status of the patient (issues of pain and hemodynamic instability). Dilation less than 5 cm in diameter may be followed conservatively by serial MR imaging studies at 6-month intervals. An observed increase in the aortic diameter of 0.5 cm over the course of a six month interval would be an indication for surgical repair. It is therefore imperative that image compression maintain information for accurate measurements and not blur edges or distort structures.

We here describe the results of a study that quantified the effects of lossy compression on measurement

This work was supported by the National Institutes for Health under Grants CA49697-02 and 5 RO1 CA55325, and by the National Science Foundation under Grant DMS-9101528. The authors gratefully acknowledge the assistance of Charles H. McDonnell III, M.D., D.V. Paranjpe, M.D., Jorg Debatin, M.D., Herbert Wegmuller, M.D., and H. Christian Davidson, M.D.

0-8186-6950-0/94 $4.00 © 1994 IEEE

accuracy. The study included the following four tasks:

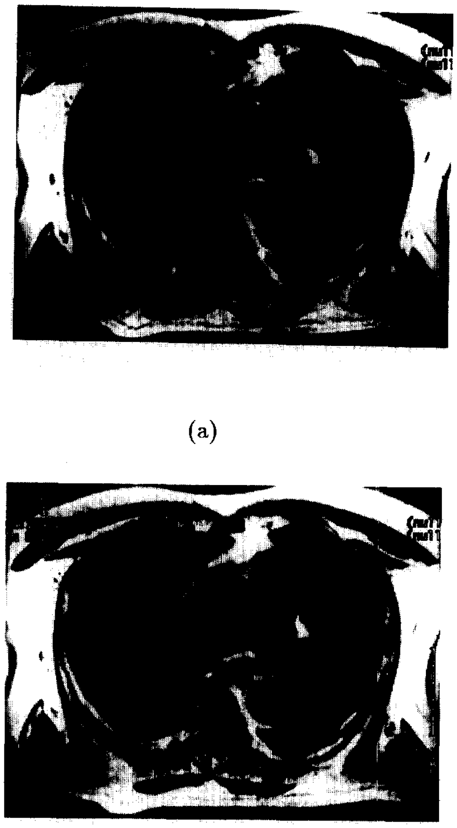
1. Establishing a protocol for obtaining measurements and subjective scores in a manner that followed closely the clinical tasks of radiologists.
2. Establishing a gold standard for the "correct vessel sizes" and choosing appropriate statistical measures for analyzing meaurement errors.
3. Determining to what bit rate the images can be compressed without a loss in measurement accuracy.
4. Determining whether subjective scores and measurement error vary similarly with decreasing bit rates.

2. STUDY DESIGN

We investigated the measurement accuracy of radiologists on four primary blood vessels in the mediastinum: the ascending aorta, descending aorta, right pulmonary artery, and superior vena cava. As previously discussed, these measurements are usually compared to a range of expected normal measurements and used to detect aneurysms, as well as to characterize overall cardiovascular physiology. Since even a 0.5 cm enlargement of the aneurysm over a 6-month period is an indication for surgery, it is important to be able to measure the size of the aneurysm accurately.

A set of 9-bit original MR chest images containing aneurysms and normal vessels was compressed to five bit rates between 0.36 and 1.7 bits per pixel (bpp). We will refer to these five rates (0.36, 0.55, 0.82, 1.14, and 1.70 bpp) as levels 1-5, respectively. We chose as our compression scheme predictive pruned tree-structured vector quantization [7]. Twenty MR chest scans were chosen as the training set based on their wide range of both aneurysms and normal vessel structures. An additional 30 scans were chosen as test images. All analyses were based solely on measurements made on the test images. The 256 × 256 pixel images were divided into 2 × 4 pixel blocks for encoding and decoding. The coefficients for the Wiener-Hopf linear predictor were calculated from the training set and used to produce a residual training set. Figure 1(a) shows an original 9.0 bpp MR chest scan and figure 1(b) shows the same image compressed to 1.14 bpp.

The test images were arranged in a randomized sequence and presented on separate hardcopy films to three radiologists. Following standard clinical methods for detecting aneurysms, the radiologists used calipers and a mm scale available on each image to measure the four blood vessels appearing on each scan. The radiologists made the measurements between the outer



(a)

(b)

Figure 1: (a) Original 9.0 bpp MR chest scan, (b) MR chest scan compressed to 1.14 bpp

walls of the vessels along the axis of maximum diameter. The viewing protocol also was designed to reduce learning effects between the different compression levels for a particular image. More details of this protocol are described in [8]. A subjective score of 1 (worst) to 5 (best) was also assigned to each image based on the radiologist's opinion of the quality of that image for the measurement task. These scores were used only as a measure of subjective quality and not as a measure of measurement accuracy.

Two gold standards were set to represent the correct size of each blood vessel. The "independent gold standard" is based on the consensus measurements made by two radiologists on the uncompressed images. These two radiologists are different from the three radiologists whose judgments on the compressed images are used to determine diagnostic accuracy. The "personal gold standard" is derived for each radiologist based on his or her own measurements on the same image at the uncompressed level. Since each gold standard has advantages and disadvantages [8], it is useful to consider both in analyzing the accuracy of measurements in compressed medical images.

The differences in error achieved at each bit rate can be quantified as statistically significant by many tests; we chose two of the most common, the t and Wilcoxon signed rank tests [9]. Both of these tests take into account the pairing of the measurements being compared. Measurements are considered paired in a comparison of two bit rates because the same vessel in the same image is measured by the radiologists at both bit rates. We also used statistical techniques that account for this multiplicity of tests.

3. RESULTS

Figure 2 plots percent measurement error vs. bit rate when the independent gold standard was used. The percent measurement error for each image is defined to be the difference between a radiologist's measurement and the gold standard, scaled by the gold standard measurement. The data are fitted with a quadratic spline with a single knot at 1.0 bpp to show the general trend of the data. A quadratic spline is a simple model for tracking data that can be fit by least squares [10]. It is quadratic "by region," and is continuous with a continuous first derivative across the *knot*, where the functional form of the quadratic changes. The graph demonstrates that except for measurement at the lowest bit rates, accuracy does not vary greatly with compression. When the t-test was used with the independent gold standard, none of the compressed images down to the lowest bit rate of 0.36 bpp was found to have a significantly higher percent measurement error when compared to the error of measurements made on the originals. When the Wilcoxon signed rank test was used with the independent gold standard, only level 1 (0.36 bpp) differed significantly in percent measurement error. Thus with the independent gold standard, measurement accuracy is preserved down to 0.55 bpp (level 2).

A plot of percent measurement error vs. bit rate when the personal gold standard was used was similar to the plot when the independent gold standard was used. The t-test results with the personal gold standard indicated that images compressed to levels 1 (0.36 bpp) and 4 (1.14 bpp) had significantly different percent measurement error from the personal gold standard. With the Wilcoxon signed rank test, only the images compressed to level 1 (0.36 bpp) had significantly different percent measurement errors from the personal gold standard. A Bonferroni simultaneous test [11] was constructed to use the significance levels of the previous two tests to obtain a significance level that is simultaneously applicable for both. With the simultaneous test, only the percent measurement error at level 1 (0.36 bpp) was determined to be significantly different from that at the uncompressed level. This agrees with the corresponding result using the independent gold standard. Thus, with both the independent and personal gold standards, we found that percent measurement error at compression levels down to 0.55 bpp did not differ significantly from the percent measurement error at the 9.0 bpp original.

It is also important to examine the effect of compression on subjective opinions. In particular, we would like to address the following two questions: Does a radiologist's perception of image quality change with bit rate? Does it change in a manner similar to the way percent measurement error changes? In order to address these issues, we asked the radiologists to assign subjective scores (at the time of measurement) of 1 (worst) - 5 (best) to each image based on "its usefulness for the measurement task." We defined "usefulness" to represent the radiologists' "opinion of whether the edges used for measurements were blurry or distorted, and your confidence concerning the measurement you took."

Figure 3 shows the relationship between mean subjective score and mean bit rate for all the radiologists pooled and for each radiologist separately. A spline-like function that is quadratic from 0 to 2.0 bpp and linear from 2.0 to 9.0 bpp was fit to the data to illustrate the general trend. The plot shows that although compression level 5 had similar scores to the original images, there was a rapid and drastic decrease in sub-

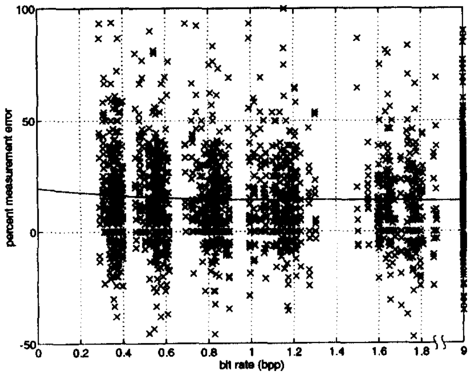


Figure 2: Percent measurement error vs. actual bit rate using the independent gold standard. The x's indicate data points for all images, pooled across judges and compression levels.

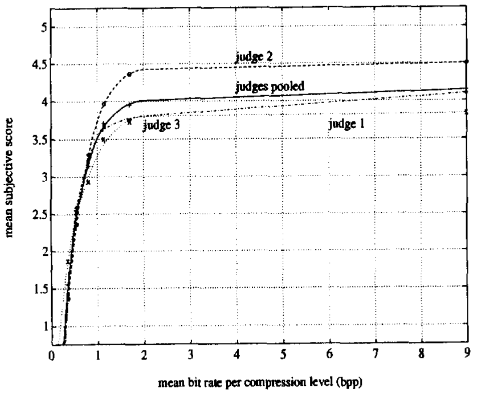


Figure 3: Mean subjective score vs. mean bit rate. The data are fitted with splines to show the general trend.

jective scores at the lower bit rates. A statistical analysis using the Wilcoxon signed rank test showed that the subjective scores at all of the compression levels differed significantly from the subjective scores of the originals at $p < 0.05$ for a 2-tailed test. Thus although actual measurement performance on the compressed images remained consistently high down to 0.55 bpp, the radiologists believed that image usefulness degraded at considerably higher bit rates. As such, the radiologist's subjective opinion of an image's usefulness for diagnosis may not necessarily be a good predictor of measurement accuracy.

4. CONCLUSIONS

Three radiologists measured the four primary blood vessels on MR chest images that were compressed with predictive PTSVQ. We determined that down to 0.55 bpp these measurements differed only at the noise level from measurements made on the original images. This analysis was performed using the t-test and the Wilcoxon test with both independent and personal gold standards. Furthermore, it was determined that a radiologist's opinion of an image's usefulness for a measurement task may not necessarily correspond with the actual utility of the image for the measurement task.

5. REFERENCES

[1] P. Cosman, C. Tseng, R. Gray, R. Olshen, L. Moses, H. Davidson, C. Bergin, and E. Riskin, "Tree-structured vector quantization of CT chest scans: Image quality and diagnostic accuracy," *IEEE Trans. Medical Imaging*, vol. 12, pp. 727–739, 1993.

[2] J. Sayre, D. R. Aberle, M. I. Boechat, T. R. Hall, H. K. Huang, B. K. Ho, P. Kashfian, and G. Rahbar, "Effect of data compression on diagnostic accuracy in digital hand and chest radiography," in *Proceedings of Medical Imaging VI: Image Capture, Formatting, and Display*, vol. 1653, pp. 232–240, SPIE, Feb. 1992.

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

[4] J. Lee, S. Sagel, and R. Stanley, *Computed Body Tomography with MRI correlation*, vol. 2. New York: Raven Press, 1989. Second Edition.

[5] D. Stark and J. WG Bradley, *Magnetic Resonance Imaging*. St. Louis: Mosby-Year Book Inc., 1992. Second Edition.

[6] R. Brandenburg, V. Fuster, E. Giuliani, and D. McGoon, *Cardiology: Fundamentals and Practice*. Chicago: Year Book Medical Publishers, Inc., 1987.

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

[8] C. Tseng, S. Perlmutter, P. Cosman, K. Li, C. Bergin, R. Olshen, and R.M.Gray, "Effect of tree-structured vector quantization on the accuracy of vessel measurements in mr chest scans," *IEEE Trans. Medical Imaging*. Submitted for possible publication.

[9] G. W. Snedecor and W. G. Cochran, *Statistical Methods*. Ames, Iowa: Iowa State University Press, 1989.

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

[11] R. Miller, Jr., *Simultaneous Statistical Inference*. New York: Springer-Verlag, 1981. Second Edition.