Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Statistical Analysis of Difference image for Absolutely Lossless Compression of Medical Images

Robina Asraf (*Student MemberIEEE)* , Muhammad Akbar (*Member IEEE)*

Noman Jafri (Member IEEE)

***Abstract*— Absolutely lossless compression technique has been recently proposed for medical image compression, It is a hybrid of lossless and lossy compression for medical images using difference image[1]. It is a multi step process in which we have used lossy coding followed by difference image coding. The combination of two compression ratios is resultant. In this paper we analyze statistics associated with difference image. The main objective of this work is to exploit some statistical measures for a difference image to compress it losslessly. Difference image statistics plays a vital role in our technique to get maximum compression ratios. Proposed scheme is simple, computationally economical and can achieve higher compression ratios than existing standard lossless compression techniques and also meets the legal requirement of medical image archiving.**

I. INTRODUCTION

Despite of rapid growth in digital communication systems, mass storage devices and processor speed, requirement for data storage capacity and transmission band width continue to exceed the capability of available technologies. The science of obtaining a compact representation of an image while maintaining all the necessary information is referred to as image compression.

Image compression can be broadly classified into two types, namely Lossless compression and Lossy compression. Lossless compression method is also referred to as reversible, entropy or noiseless coding as it enables complete recovery of the original image from the compressed data. The main draw back of this method is low compression ratio, especially in the case of images, the compression ratio could be as low as 2:1. Its main application is in medical images, where any loss can affect diagnostic accuracy. Our proposed method is an effort to obtain compression ratios higher than obtained by existing standards, without any loss of data. A hybrid of Lossless-Lossy compression is proposed in which difference image statistic plays pivotal role.

This paper is arranged such that brief discussion on hybrid compression scheme is given then brief revision of lossless compression techniques used is presented. Next we define some statistical measures used to assess compression capability. In section V experiments and results are presented. In section VI we conclude our work with future prospects.

Robina Ashraf is a research student at College of Signals, National University of Sciences and Technology Pakistan. Dr. Muhammad Akbar is her supervisor and Dr. Noman Jafri is her co-supervisor. E mail address is robina0321@yahoo.com

II. HYBRID COMPRESSION SCHEME

Original image is compressed with loss then it is decompressed. The difference of original and retrieved image is taken and is compressed with lossless coding. The overall compression will be achieved by combining, both lossy and lossless compressed data. At decoder side reverse procedure is applied to obtain the reconstructed image. Absolute difference of original and reconstructed images resulted in zero matrix which means absolutely lossless compression. Any lossy/lossless compression techniques could be used. We have used NNVQ (Neural Network Vector Quantizer) as lossy compressor and Huffman coding for lossless compression [1]. Fig.1 shows the flow of events for our proposed scheme.

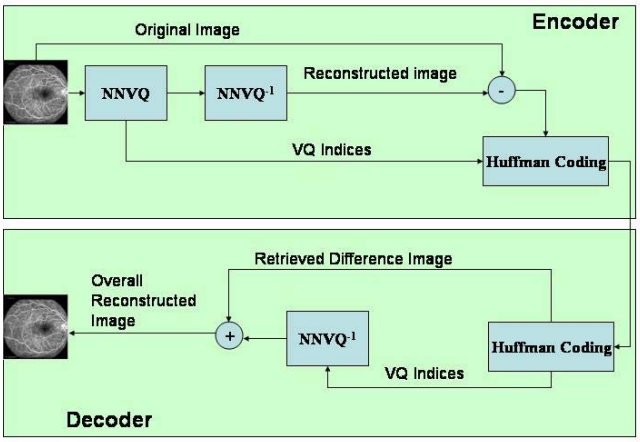


Fig. 1. Hybrid Compression Scheme for medical image

If A is compression ratio for lossy compressed image and B is compression ratio for Difference image then overall compression ratio achieved will be:

$$CR_{overall} = A * B \Big/ A + B$$

Overall compression ratio is calculated for X-ray Lung as A= 40, B=14, C= (40*14/54) =10.37. In this paper it is to be explained that how could we get B=14, in lossless scenario.

III. LOSSLESS COMPRESSION

The goal of lossless image compression is to generate an absolutely equivalent but shorter representation than the

original image. This is an important requirement for medial imaging domains, where not only high quality is in demand, but unaltered archiving is a legal requirement. In Huffman coding short code words are assigned to those input blocks with high probabilities and long code words to the ones with low probabilities. A Huffman code is designed by merging together the two *least probable* characters, and repeating this process until there is only one character remaining. A code tree is thus generated and the Huffman code is obtained from the labeling of the tree.

Repeated occurrence of the same character is called a *run*. Number of repetitions is called the *length* of the run. Run of any length is represented by three characters. For example eeeeeeetnnnnnnnn @e7t@n8. Coding of binary images is further simplified because repeat character is not required, only alternate repetition lengths are given, hence more compression.

Other lossless coding schemes such as Lempel Ziv (LZ) coding, Context Adaptive Lossless Image Compression (CALIC) or Arithmetic coding, etc. are described in [2].

IV. STATISTICAL MEASURES

To evaluate the compression capability of difference image we have used the following measures:

a. Non zero data

To prove that difference matrix contain less data to be compressed. We describe a measure *nnz (number of nonzero elements in a matrix).*

b. Max value in difference image

It is very important information that tells about the number of bits to represent the value.

c. Entropy

Let X be image with M X N size and 8 bit /pixel. Thus, X(i,j) represents the pixel at location (i,j).
then entropy is given as

$$H(X) = \frac{1}{MN} \sum_{i=0}^{255} p_{ij} \log_2 p_{ij}$$

Where p_{ij} represents the probability of pixel X(i,j) occurrence[3]. This is a measure of average information of a source. More information associated with source more will be its entropy.

d. Redundancy

Coefficient of redundancy is defined as

$$D = 1 - H(X) / H_{\max}$$

where H_{max} is maximum possible value for H(X) (8bits for gray scale images).

e. Probability distribution

If the source symbols are equally probable, its entropy maximizes and source provides maximum possible average information per source symbol. With distribution concentrated on some symbols entropy decreases, redundancy increases and hence more compression possible. For example: out of 200 source symbols for probability distribution 20/200, 40/200,60/200,80/200 entropy will be 1.8464 ,for distribution,120/200,80/200,0,0 entropy will be 0.9707 and for probability distribution,180/200,20/200 entropy will be 0.4690. This shows that probability distribution is vital for compression capability analysis.

f. Expectation and Variance

For a continuous variable x with distribution f(x), expected value can be defined as

$$E[X] = \int_{-\infty}^{\infty} x f(x) d$$

As long as defining integral is convergent. For discrete variable it can be re-expressed as sums. It is center of mass of probability distribution called mean and denoted by μ_x .The Kth central moment is moment of centered random variable (x-μ),whose mean is zero.

$$\mu_k = E(X - \mu)^k f(x) dx$$

The variance measures the spread of the probability mass about the mean [4].

V. EXPERIMENTS AND RESULTS

Experiments are done on a large number of test images. Only six are included here for demonstration purposes. These include X-ray lung, retinal image, Spinal MRI, , MRI brain, CT pancreatic, CT Spine, Pregnancy ultra sound [5],[6],[7]. All images are divided into 4x4 blocks, each block is treated as a vector of 16 elements and preprocessed to train.

The experimental work is mostly done in MATLAB [8]. The original images are gray scale images with 8 bits per pixel. As the test data is similar to the training data the quality of reconstructed images through NNVQ are comparable to lossless compression. That's why the difference images contain very little data. Original images and difference of these with NNVQ reconstructed images are shown in fig.2 and fig.3 respectively.

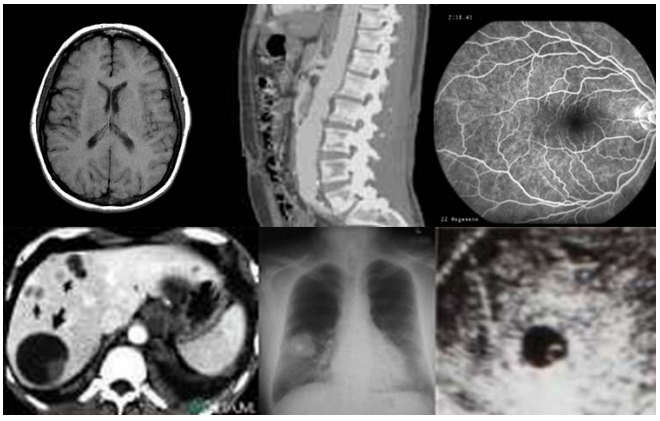


Fig.2. Original Images

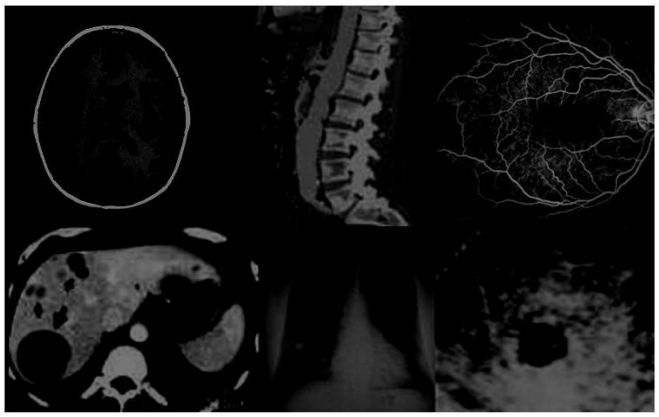


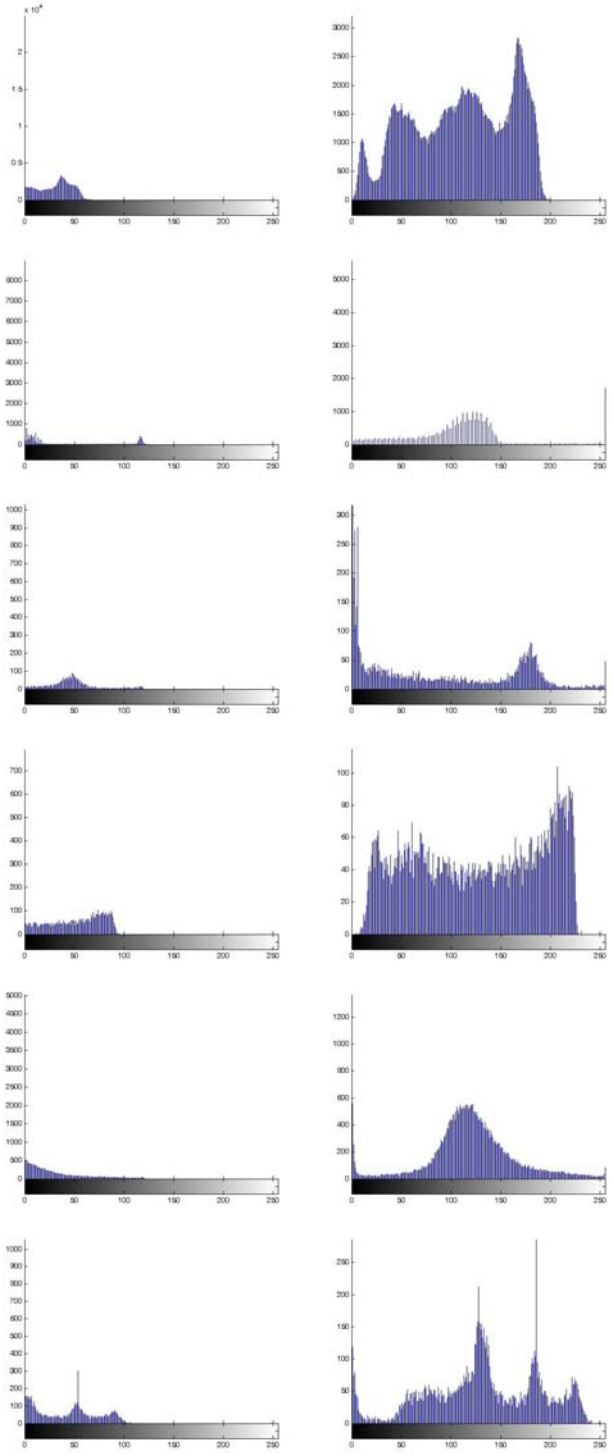
Fig.3. Difference Images

Now we treat all difference images. First we find total number of pixels in original images then compare these with number of non zero elements in difference images, It is found that in all test images more than 50% data comes out to be zero which resulted in long strings of zeros or concentration of probability distributions towards black. Here run length coding is used to compress these long strings of zeros efficiently. Another parameter is maximum value which is 255 in original images but for all difference images it comes around or less than 128 which means one bit decrease to represent each symbol. Dynamic range of gray levels in each of the difference image is reduced from 255 to about 128 or less. Here we get another step towards compression of file by representing difference image with lesser bits than original image.Table.1 shows the nonzero data and max value for difference images in test set.

TABLE.1

Difference Image	Total bytes	Nonzero bytes	Ratio % nnz/Total	Maximum Value
MRI Brain	65536	8077	12.32	123
CT Spine	12870	6145	47.74	114
Retinal Image	46656	14567	31.22	121
Xray Lung	264000	104870	39.72	67
Pregnancy Ultrasound	10230	4224	41.29	95
CT Pancriase	8910	2342	26.28	120

Next we found entropy and probability distribution for each of the difference image. In fig.4 comparisons of probability distribution of original and difference images are given.
In original image, all the gray levels from 0 to 256 contribute in distribution. As we know that for equiprobable source symbols, entropy is maximum. While for difference images all of probability distributions are concentrated towards black, entropy becomes less than original image. Average length of code is dependant on entropy, when entropy decreases average length of code also decreases that's how



(a) (b)

Fig.4 (a) Histogram of difference images
(b) Histogram of original images

We achieve high compression ratios for a difference image in lossless mode. Table.2 gives entropy and variance for original and difference image. Variances shown are calculated by using sum of variance matrix. Comparison of

variance matrix for original and difference images in test data is plotted as graphs (respectively) and shown in fig.5.

TABLE.2

Images	Entropy		Variance(sum)	
	Original	Difference	Original	Difference
MRI Brain	14.702	11.084	11597	4184
CT Spine	13.308	11.492	3352	1767
Retinal Image	15.859	14.085	16987	10473
Xray Lung	12.008	10.192	21905	7420
Pregnancy Ultra Sound	13.109	11.837	6706	3280
CT Pancriase	12.081	10.120	6259	1987

Fig.5 shows that variance of difference images is less than original in all cases hence compression capability is more.

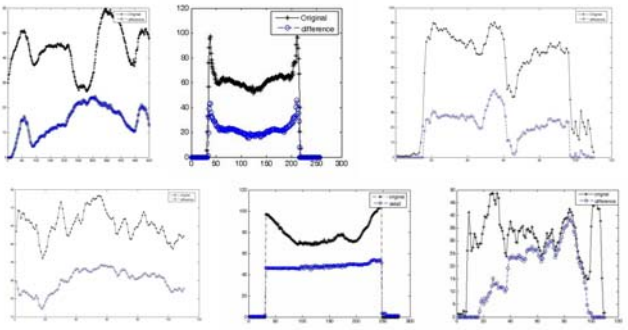


Fig.5. variances Original image against difference image

Table.3 shows overall results for hybrid compression scheme and compare it with JPEG2000 Loss Less.
'A' is compression ratio between bytes of original and lossy (NNVQ) compressed images. 'B' is compression ratio between bytes of difference image and its lossless compressed version. '$CR_{overall}$' is overall compression ratio. While JPEG2000 is applied directly on original images.

TABLE.3.

Images	Proposed Technique			JPEG2000 LOSSLESS
	A	B	$CR_{overall}$	
MRIBrain	40	16.6	11.73	1.73
CT Spine	40	8.56	7.05	1.48
Retina	40	11.55	8.96	1.54
Xray lung	40	14.02	10.37	1.78
Pregnancy	40	10.924	8.58	1.65
CT Pancrias	40	9.416	7.62	1.63

For NNVQ compression ratio (approx~40) is chosen for all images. Difference images are compressed in three steps:

- A type of run length encoding to suppress only long strings of zeros.
- Lesser bits representation due to squeezed dynamic range.
- Huffman coding of code is then produced.

Decoding is reverse order process of these steps.

VI. CONCLUSIONS

By all this analysis we have concluded that difference image contain less data, it can be represented with lesser number of bits, also its probability distribution, entropy and variance support compression. Hybrid compression technique, in which difference image is used, can achieve higher compression ratios as compared to standard lossless compression techniques.
Particularly Table.3 shows that lossless compression ratio achieved by JPEG2000 (1~2) is far less than proposed technique (5~10). Future work can be done for generalization and standardization of proposed technique for a variety of medical image modalities. Experiments can be done on different hybrids of lossless and lossy compression techniques. Another pair that we have tried is JPEG and Arithmetic coding.

ACKNOWLEDGMENT

I acknowledge the financial support by Higher Education Commission (HEC) Pakistan for carrying out this research.

REFERENCES

[1] Robina Ashraf, Muhammad Akbar "Absolutely Lossless Compression of Medical Images", proceeding of 27th Annual IEEE EMBS Conference Sep. 2005 Shanghai China.

[2] Khalid Sayood, Introduction to Data Compression. Morgan Kaufmann 2nd edition 2000.

[3] Hazem Munawer Al-Otum, "Qualitative and quantitative image quality assessment of vector quantization, JPEG and JPEG2000 compressed images", Journal of Electronic Imaging. vol. 12(3) pp 511-521 July 2003.

[4] Edward R. Dougherty, Random Processes for Image and Signal Processing. SPIE 1999.

[5] http://medmuseum.com/museum/

[6] http://www.med.harvard.edu/AANLIB/cases/

[7] http://www.radiotherapy.com/Simulation/

[8] Rafael C. Gonzalez, Richard E. Woods, Steven L.Eddins, "Digital Image Processing Using MATLAB." Pearson Prentice Hall 2004.