Abstract

In this paper a comprehensive survey of the state of the art lossy and lossless techniques available in the literature has been presented and the merits and pitfalls of each technique are analyzed. This study congregates the pioneer works in two dimensional (2D) compression techniques, both in pixel domain and transform domain. The evolution of compression of three dimensional (3D) medical images from 2D compression has also been discussed. Compressed medical image has to be both diagnostically lossless and less bandwidth in addition to visual quality. Region of Interest coding (ROI) which achieves diagnostic quality image with less bandwidth has been explored. In spite of proven compression technologies, only the lossless compression has been used widely around the world and the reason for the same has been investigated. In addition it also investigates several factors; why one needs to go for lossy compression.

Keywords: 3D Medical Image, Context Based Coding, DCT, DWT, Predictive Coding, VOI

1. Introduction

Many advanced medical imaging technologies like Computer Tomography (CT) and Magnetic Resonance Imaging (MRI) are 3D images and are important in biomedical field. They are multiple images or sequence of images taken at different cross section leading to very high volume of data. The huge increase in the acquisition of clinical data increases the volume of data to be stored. In addition as medical images are acquired at the cost of radiation exposure it has to be preserved for some duration of the time. Increase in volume of the data to be stored calls for compression techniques. Modern medical field like telemedicine is also enjoying the advances in communication engineering, so that patients from remote villages also can get consultation from experienced doctors in remote location. In telemedicine compression is not only required for data preservation, but also for efficient bandwidth usage for transmission.

Compression technology available in the literature is broadly classified into lossy and lossless. Lossless compression is further divided into lossless and near lossless; in lossless compression there will not be any loss of data due to compression and are reversible, near lossless compression are irreversible and does not affect data processing after decompression. Lossy compression is completely irreversible, still visually acceptable, but visual acceptance is not the main criteria in medical applications as in multimedia compression. Lossless compression is mostly based on statistical coding\(^1\), prediction principle\(^2\) and context modeling\(^3\). Statistical coding relies on statistical occurrences of frequent and rare data. It codes the frequent data with less number of digits and rare data with more number of digits. Such type of coding is more suitable for 2 dimensional (2D) text data as estimation of statistical data is more time consuming for images with high resolution. Predictive coding works on the principle of correlation and it produces best result when neighboring blocks in 2D images are similar. Context modeling uses the context provided by data that has been coded already to determine the current encoding of data.

*Author for correspondence*
Compared to statistical modeling and predictive coding, context modeling provides better compression.

Lossy coding is primarily classified into Vector Quantization (VQ) and transform coding technique and they offer high compression ratio. VQ technique is based on codebook generation after quantization. The representation of quantized values with some indices to a set of quantized level is called codebook. While VQ coding offers slightly higher compression ratio than lossless coding, the cost of memory for codebook storage and member search in codebook is more. Also the major drawback of VQ coding is that it results in blockiness due to loss of edges. Transform coding as the name signifies, it transforms the pixel domain data into transform domain mostly in frequency domain. The transform based techniques are proved to show some definite position in which the variation of harmonics and its magnitude occurs. It is also found that most of the harmonics are negligible in magnitude and hence can be ignored. Transform coding after energy compaction, is quantized and then coded statistically or predicatively.

Recently subband coding; a type of transform coding has gained attention due to its multi resolution in nature. Due to its multi resolution nature it is used for progressive transmission i.e. the main components are transmitted first and then the refined portion are transmitted. The most popular transform using wavelet is used for subband coding and in digital processing it is called Discrete Wavelet Transform (DWT). The next step after energy compaction is quantization and coding, The coding process in subband coding is primarily classified into Embedded Zero Tree Wavelet Coding (EZW), Embedded Block Coding by Optimized Truncation (EBCOT), Set Partitioned In Hierarchical Tress (SPIHT) and Set Partitioned Embedded Block (SPECK). Lossless compression can also be achieved by using the wavelet family called Integer Wavelet Transform (IWT). The rest of the paper is organized as follows: Section 2 presents different volumetric compression approaches and section 3 gives a comprehensive view of Volume of Interest Coding (VOI) for more compression. Section 4 elaborates on the compression techniques being used at present and section 5 concludes this paper.

2. Volumetric Images Compression Approaches

In general there are two design approaches followed; Statistical prediction based approaches work in pixel domain and transform based approach work in transformed signal domain typically in frequency domain. Compression factor together with visual quality is considered as the main parameter in the design for non medical application. However medical application requires clinically lossless data and hence relevant data is the main focus rather than compression factor. Progressive transmission is required when the radiologists have to determine, on the fly whether the desired image set is being transmitted or not before transmitting the entire image set. VOI coding is another technique used for compression, where the diagnostic region cannot withstand the compression.

2.1 Prediction and Contextual based Compression

Prediction and context based compression are primarily lossless compression technologies. Efficiency of lossless image compression depends on two components; statistical modeling and coding. Statistical analysis of images shows that neighboring grey pixels are highly correlated.
after the separation of color image into its components. This correlation is being used in prediction methodology, for the computation of prediction parameter. The prediction parameter is then used to compute the current pixel and then the error is computed between the original and predicted pixel. The prediction error may be further statistical coded like Huffman code to reduce the entropy and thereby compression is achieved\textsuperscript{15,16}. The general block diagram for prediction based compression with transform coding is shown in Figure 1. The more straightforward way to compute the prediction parameter is by averaging over blocks of 8 x 8 or 16 x 16. While this is simple, it does not give exact predictor for each pixel and hence the residuals may become roughly equal to the original. The predictors computed are fixed in nature and hence they are called linear predictors. Here the data cannot be compressed optimally when the input symbols to the statistical coder are not independent.

The problem with linear predictors is overcome by adaptive or context based predictors. In context modeling, the probability of the current pixel $U$ is estimated conditioned on the combination of its $m$ previously encoded pixels $x_1, \ldots, x_m$. The combination of these pixel values is called context. Hence the computed predictor changes with respect to regions like smooth grey area or edges. The adjustment of predictor parameters can be made very efficient since it is based on local information. Low entropy distributions can be obtained through larger conditioning regions or contexts. However this means larger parameters and high model cost\textsuperscript{16}. As there is no loss of data due to compression, the performance is mainly measured by the statistical modeling cost, the operational memory cost and the entropy cost.

### 2.2 Transform based Compression

Transform coding started with the opening of Discrete Cosine Transform (DCT)\textsuperscript{17}. 2D DCT has been adopted by many practical video technologies like MPEG for compression due to its low complexity. 2D DCT transforms the image or one video frame in time space domain to frequency domain, where the energy compaction occurs in few frequencies with decorrelation between frequencies. Hence only those frequencies are retained and entropy coded and others are discarded by choosing suitable quantization factor. The quantization factor for each frequency will be different and is according to the harmonic variations in the coefficient matrix\textsuperscript{18}. For 3D images, usually 3D DCT has been computed on 8 x 8 x 8 or 16 x 16 x 16 cubes rather than on multiple frames due to memory and algorithmic complexity. Cube based 3D DCT results in blocking artifacts, which destroys the edges of a medical image. However in medical images one needs to have clear detail and sharpness. Also, while there are many fast transform techniques available for 2D DCT there are not many fast 3D DCT techniques are available in the literature. Inadequate localization of frequency and hence artifact is due to insufficient number of basis functions compared to basis size (number of frequency components). Implementation block diagram of DCT based compression of three dimensional images is shown in Figure 2.

The next popular transform used for compression is Discrete Wavelet Transform (DWT) due to its reduced blocking artifact. In DWT a low pass and high pass Finite Impulse Response (FIR) filter analysis are applied on $x(n)$ and down sampled to get $l(n)$ and $h(n)$. The total number of samples in $l(n)$ and $h(n)$ is equal to $x(n)$. In synthesis stage $l(n)$ and $h(n)$ are up sampled and filtered with synthesis filters. The sequences can be further decomposed to give finer details. In a dyadic decomposition the lowest frequency band is decomposed in a recursive fashion and the number of such decomposition is referred to as dyadic levels. DWT can be extended to

![Figure 2](image-url) Banu DCT based compression implementation.
multiple dimensions by applying the transforms independently in each direction and is called as separable transform. Although non separable filters exist, it is difficult to design such filters and hence mostly separable filters are used for subband analysis or DWT. The number of decompositions performed in each dimension depends on a number of factors, such as the preferred compression rate and the size of the medical image. In general, the higher the desired compression ratio, the more times the transform is performed.

The natural block overlapping in DWT leads to smoothness along the edges. Also the number of basis function is larger than the number of basis size (number of subbands) which removes the artifacts unlike DCT. DWT not only results in energy compaction and decorrelation, but also results in self similarities i.e. if the wavelet coefficient is small in the low frequency band; it is expected to be small in the high frequency band for the same spatial location. These self similarities can be effectively used to have better compression ratio. Several compression algorithms have been proposed based on this and the popular known schemes are 2 dimensional EZW, SPIHT and SPECK. All these algorithms have their 3D version and the 3D version of EZW provides 20% to 22% decrease in compressed file size compared to their 2D counterpart. The 3D variant of SPIHT proposed by Bilgin et al. in exploits redundancies in all directions to give better compression factor.

The embedded zero tree wavelet algorithms proposed by Shapiro is the first one to address the self similarity present after DWT coding. The bits in the bit stream are generated in the order of importance by setting threshold for each pass. It can be only lossy coding, as the coefficients with larger values than the current threshold are quantized. In order to have a clear view and an idea, 2D case is reviewed here and 3D is the extension in an axial direction. As this is bitplane coding it can be stopped at any bit plane once the required bitrates are achieved. They are also called as zerotree coding and zero trees are quad tree which has insignificant coefficients with respect to the current threshold. Every coefficient in a dyadic wavelet transform is related to a set of coefficients at the next-finer level corresponding to the same spatial location in the image. A coefficient at a coarse level is called a parent, and its spatially related coefficients at the next-finer level are called as its children. This relationship can be represented by use of a tree structure, as shown in Figure 3. The relationship between the parent and the child is given by 

\[(2x,2y), (2x+1,2y), (2x,2y+1), (2x+1,2y+1)\]

![Figure 3. Banu Original Wavelet coefficients after 3 levels of decomposition](image)

Where \((x,y)\) is the coordinate of the parent. According to this relationship the coarsest coefficient has only three children (coefficient \((0,0)\)) whereas all the other coefficients except for the coefficients at the finest scale have four. The coefficients at the finest scale are childless.

The coarsest coefficient is referenced by root location and is shown in Figure 3. As seen in the Figure 3 the root has only three children whereas its child has four children. For better understanding of the compression concept, the original wavelet coefficients are shown in Figure 3. The coefficients which are less than the current threshold are called insignificant coefficients and coefficients greater than the threshold are marked with positive (+) or negative (-). The quad tree with all insignificant coefficients is marked with R and with at least one insignificant coefficient is marked with I. For clarity Figure 4 shows one from each type i.e. root location with all insignificant coefficients R and one with significant and insignificant coefficients combined I. When the number of roots become more it is obvious that we get more compression as the whole group is marked by one bit.

Set Partitioning in Hierarchical Tress (SPIHT) algorithm is an efficient method for lossy and lossless coding of natural images. Set partitioning refers to the way the quad tree is partitioned, divided the wavelet coefficients at a given threshold. Figure 4 shows how a spatial orientation tree is defined in a pyramid
constructed with recursive four subbands splitting. The ordering of the co-efficients are similar to EZW coding and they are in hierarchies. According to this relationship, the SPIHT algorithm saves many bits that specify insignificant coefficients. The only difference between EZW and SPIHT is the way in which the zero trees are coded. Moving from one threshold to other, the location of wavelet coefficients undergoes transitions. Instead of delivering a code for the symbols +,-, R and I by EZW to mark locations, the SPIHT algorithm uses states I, R, I, V, S, R, and S, V and outputs code for state-transitions such as I R -> I V, S R -> S V etc. For definition of states and coding procedure the reference 5 can be referred where a very good simple understandable procedure is available. Due to the coding of state transitions it takes one bit in the places where EZW codes with two bits.

3D EZW coding considers the coefficients in three directions for tree coding. Similar to 2D EZW, every coefficient in a dyadic wavelet transform is related to a set of coefficients at the next-finer level corresponding to the same spatial location in the image. A coefficient at a coarse level is called a parent, and its spatially related coefficients at the next-finer level are called as its children. This relationship can be represented by use of a tree structure, as shown in Figure 5. The relationship between the parent and the child is given by

\[(2x,2y,2z), (2x+1,2y,2z), (2x,2y+1,2), (2x,2y,2z+1), (2x+1,2y+1,2z), (2x+1,2y,2z+1), (2x,2y+1,2z+1)\]

Where (x,y,z) is the coordinate of the parent. The parent child relationship is in the form of cube rather than a square as in 2D coding.

Telemedicine became central part of health care for the betterment of rural Indian citizen. However poor bandwidth due to the lack of infrastructure may stop or delay the information transmission. In this case a low quality image can be delivered for the initial arrival with the patient’s interest in consideration. However the original images should be made available for subsequent download as required. Poor bandwidth should not lead to the provision of a sub-optimal image data set to a patient or a referring practitioner, where the clinical context suggests that any data loss might put the patient at risk. Progressive transmission is the one which provides such facility. DWT based compression is best suitable for progressive transmission due to the bit plane coding which results in multi resolution. Here the image quality is gradually increased. Progressive transmission

![Figure 4. Banu EZW algorithm after stage 1 with threshold 32.](image)

![Figure 5. Banu Parent child relationship in 3D EZW coding (Adopted from 4th Muri workshop presentation material).](image)
capability is preferred because it allows users to gradually recover images from low to high quality images and to stop at desired bitrate including lossless recovery.

Lossy coding using transforms coding are compared with respect to the compression ration they yield for particular PSNR. In this case DWT has been proved to provide better compression ratio than DCT and any other transform coding. Also compared with DWT except DCT all other transform codings are more costly in terms of complexity. In spite of all these compression types, medical image compression can be evaluated only with physician’s input.

3. Volume of Interest (VOI) or Region of Interest Coding

Diagnostically important regions like tumor, lesions and brain active regions cannot tolerate high compression ratio and hence radiologist prefer lossless compression for those region while outside that region can be compressed by lossy method. A hybrid technique which gives good quality image with high compression rate and lossless coding for diagnostically important region is called Region of Interest/Volume of Interest (ROI/VOI) coding. Such kind of coding in 2D image is called ROI and 3D image is called VOI coding. Contour tracing or segmentation is one of the techniques used to identify the ROI. Contour tracing is used along with some types of compression like entropy or wavelet coding for ROI/VOI coding. A biorthogonal wavelet with EZW coding is employed as a compression technique followed by medical image segmentation using an active contour to separate the diseased part in 20. There are various works available on clustering and segmentation in the literature. Recently a survey has been presented on clustering methods for segmentation and on implementing segmentation algorithms using Graphic Programming Units (GPU) by Arumugadevi et al. 21 and Erik Smistad et al. 22 respectively. Wavelet based SPIHT is not suitable for ROI coding as it does not give the explicit locations of significant pixels, instead Wavelet Difference Reduction (WDR) algorithm is used for ROI coding. In WDR 23,24 the output from the significance pass consists of the signs of significant values along with sequences of bits which concisely describe the precise locations of significant values.

The best way to explain this is to consider a simple example. Suppose that the significant values are w(2) = +34.2, w(3) = -33.5, w(9) = +48.2, w(12) = +34.2, and w(34) = +28.2. The indexes for these significant values are 2, 3, 7, 12, and 34. Rather than working with these values, WDR works with their successive differences: 2, 1, 4, 5, and 22. In this latter list, the first number is the starting index and each successive number is the number of steps needed to reach the next index. The binary expansions of these succeeding differences are (10)2, (1)2, (100)2, (101)2, and (10110)2. The Most Significant Bit (MSB) for each can be dropped as it is always 1. This bit and the signs of the significant transform values can be used as separators in the symbol stream. The ensuing symbol stream for this example is then +0 - +0 +01 - 0110. As the MSB is dropped after finding the difference between the locations of significant bits, it is called WDR algorithm.

VOI coding technique based on 3D subband block hierarchical partitioning (3D-SBHP) is introduced in 25 for 3D medical image compression that supports, a highly scalable wavelet transform based entropy coding algorithm. The authors addressed all the parameters that affect the effectiveness of VOI coding, including the size of the VOI. The authors also discussed an approach to optimize VOI decoding by assigning a decoding priority to the different wavelet coefficient bit-planes simultaneously. A number of techniques are presented in the literature to facilitate ROI coding. A method based on 3D IWT and modified EBCOT that exploits the correlation in 3 dimensions is presented in 26. However the method to obtain the VOI region is not detailed. In 27-28 ROI coding efficiency is further improved by exploiting the inherent characteristics of the medical images such as symmetry. The Figure shown below gives such kind of symmetry. This leads to further compression rate while maintain the quality of the image.

Recently there are many studies found in literature to explore the mirror symmetry in 3D medical images 28-30 in an attempt to investigate the morphological changes between healthy and pathological structures. In addition to the information of symmetry, there need to be correlation information between the symmetrical parts for better compression results. As a proof that literature is going in the right direction, in 25 the authors have presented a correlation modeling for CT images. These studies when combined effectively with the coding paradigms will result in better compression rate along with lossless coding for better diagnosis.
4. What is being Used Now and Why

A loss of clinical data either compromises or has the potential to compromise the value of an imaging examination to a patient. Also, semi-automated analysis of data sets like MRI and CT is calls for lossless compression and hence currently lossless compression is used by the wider community. A number of studies overseas have examined the issue of lossy compression in medical imaging. There are numerous examples of studies testing the acceptable limits of image compression ratios, on many different modalities. The quality assessment of ultrasound video with various techniques have been explored and it is found that, High efficiency Video Coding (HEVC) can compress the medical ultrasound videos at low bitrates without compromising on the diagnostic accuracy. General conclusion is that some levels of lossy compression are suitable for some purposes and some modalities; however there remains considerable uncertainty to decide the level and type of compression factor, for any particular examination or modality. For example, while compression of digital mammograms is not permitted in the USA by the Food and Drug Administration, the Royal College of Radiologists, and the German Radiology Society in Europe, has published acceptable lossy compression ratios of 20:1 and 15:1 respectively.

Canadian Association of Radiologists (CAR), has conducted a survey on the use of lossy compression and concluded that the use of such compression would not increase the legal responsibility of physicians, if used and implemented appropriately. However, this conclusion leaves much room for interpretation and it may lead to inconsistent application of compression.

4.1 Why Lossy Coding

Lossy coding is the need of the hour in the field of Bioinformatics and there are multiple constraints that will arise due to lossless coding and are

4.1.1 Bandwidth

Bandwidth issues in rural India may delay transmission of an uncompressed or lossless compressed image. In these cases, the radiologist should make careful use of appropriate levels of lossy image compression to speed the initial arrival of the images, with the patient’s interests foremost in their consideration.

4.1.2 Storage Costs

There is an argument that the costs of storing full data sets for every imaging examination would be prohibitively expensive, and will become more so as imaging modalities become more complex, powerful and productive. This can be considered as an ample justification for introducing lossy compression techniques. However, it is noted that, at least in the past, costs of digital data storage have fallen steadily.

4.1.3 Retention of Medical Reports

All facilities should be made for the retention of medical reports as the availability of previous reports decisively influences the interpretation of new study. This factor again increases the storage cost and hence compression must be made lossy.

5. Conclusion

The debate on lossless or lossy compression for medical image is never ending if only the theoretical research output is considered. Evidence should be created with millions of trials with lossy compression for each modal-
ity with different types of lossy compression. It should be the combined effect of medical practitioner and a bio medical engineer. Also it should be double checked with multiple medical practitioners as diagnosis changes from person to person even with the original image.

6. References

30. Yina XX, Ng BW, Ramamohanarao K, Abbott D. Tensor b-

