CHAPTER 2

SURVEY ON MEDICAL IMAGE COMPRESSION ALGORITHMS

2.1 INTRODUCTION

Development of compression schemes for medical data is a rapidly evolving field with growing applications in the health care services like teleconsultation, teleradiology, e-health, telemedicine and statistical medical data analysis. Compression techniques reduce file storage size and transmission time by removing the redundancy and irrelevancy present in the data. Redundancy reduction aims at removing duplication from the signal source. Irrelevancy reduction omits parts of the signal that will not be noticed by the signal receiver, namely Human Visual System. In general there are three types of redundancies such as, spatial redundancy or correlation between neighboring pixel values, spectral redundancy or correlation between different spectral bands and temporal redundancy or correlation between adjacent frames in a sequence of images. Compression algorithms take advantage of one or more redundancies, to reduce the number of bits required to store and transmit the image data without any appreciable loss of useful information.

2.2 CLASSIFICATION OF COMPRESSION ALGORITHMS

Images can be compressed using lossy or lossless techniques. Lossless compression (Das & Burgett 1993, Ramabadran & Chen 1992,
Kuduvalli & Rangayyan 1992) is reversible compression where the reconstructed image is exactly the same as the input image, and obtains compression ratio in the range 2:1 to 3:1 or greater, depending on the content. Lossy compression (Chen & Ramabadran 1994, Wu 2002) is irreversible compression technique which obtains compression ratio around 50:1 to even 100:1 depending on the compress quality level and the image content. Here reconstructed image are not the same as the input image, but with imperceptible degradation. Though high compression is needed for volumetric medical data, clinicians are reluctant to accept lossy compression, as any information loss or error caused by compression process could affect their diagnostic decisions and also legal challenge may arise. So, medical community relies on lossless or near lossless compression methods. In the past lot of algorithms has been proposed exclusively for compression of medical images. Though many methodologies are available for compression, we have discussed in this chapter only three major kinds viz., Transform based compression, Region of Interest based compression and Neural Network based compression along with the survey on the work reported so far on medical images with these techniques.

2.2.1 Transform Based Compression

The basic frame work in which this transform based compression works is as shown in Figure 2.1. Most of the compression algorithm developed falls under this category.

![Figure 2.1 Transform based compression scheme](image-url)
This kind depends on transforms that convert time domain signals into frequency domain signals, so as to give a sparse representation for the image. Here the spatially distributed energy is concentrated in fewer coefficients which are quantized and encoded. The quantizer which is non invertible block introduces the compression. Major used transformations are Discrete Cosine Transform and Discrete Wavelet Transform.

2.2.1.1 Discrete Cosine Transform

Transform coding became popular mainly due to the introduction of the Discrete Cosine Transform—an efficient transform with high computational efficiency and compression performance that is close to the performance of optimal Karhunen Loeve Transform (Jain 1979). This fact has made the DCT favourable for still image and video coding. The widest used commercial product that is used for the still image coding scheme, the ISO/ITU-T standard Joint Photographic Experts Group created in late 1980’s is based on DCT (Christopoulos et al 2000). Among several modes of JPEG, baseline is the most popular. Here the image is divided into 8x8 blocks, DCT is computed for each and every block, the resulting blocks are individually quantized by a standard quantization table, zigzag scanned and Huffman coded. This has the advantage of simple computation, but at lower rates suffers from blocking artifacts due to coarse quantization of coefficients. Basic 2D-DCT function and its inverse functions are shown in Equations (2.1) and (2.2).

\[
F(u,v) = \frac{4C(u)C(v)}{n\times n} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} f(i,j) \cos \left( \frac{(2i+1)u\pi}{2n} \right) \cos \left( \frac{(2j+1)v\pi}{2n} \right) \quad (2.1)
\]

\[
f(i,j) = \frac{4C(u)C(v)}{n\times n} \sum_{u=0}^{n-1} \sum_{v=0}^{n-1} F(u,v) \cos \left( \frac{(2i+1)u\pi}{2n} \right) \cos \left( \frac{(2j+1)v\pi}{2n} \right) \quad (2.2)
\]
where \( C(x) = \begin{cases} \frac{1}{\sqrt{2}}, & x = 0 \\ 1, & \text{otherwise} \end{cases} \)

A 3D compression concept for a 2D still image was proposed by Tai et al (2000). In this method the image is divided into 8x8 non overlapping subblocks, 2D DCT is performed on every subblock and activity value \( \varepsilon \) is calculated by Equation (2.3) using six coefficients as shown in Figure 2.2 for each subblock.

\[
\varepsilon = a^2 + b^2 + c^2 + d^2 + e^2 + f^2
\]  

(2.3)

Depending upon the predefined threshold values the blocks are classified into four different classes and the class information is stored in classmap. After classification, those subblocks with the same class are grouped together to form the 3-D cuboid according to the row major scan of class-map. At the decoder side, the cuboid structure can be reconstructed by the class-map which has already been transmitted. The 3D DCT is performed on the cuboids and coefficients are uniformly quantized to produce the quantized coefficients. A refined quantizer is used to quantize the dc coefficients. As for ac coefficients, a coarse quantizer is adopted to increase the compression ratio. The quantized coefficients are scanned in a 2-D zigzag pattern and Huffman encoded. This strategy not only raises the compression ratio but also alleviates blocking effects and outperforms the technique by Wu & Tai (1998) and JPEG in terms of bitrates and PSNR values.
In the work by Wu & Tai (2001), all the 8x8 blocks of the image were DCT transformed. A translation function gathers the same frequency components from all individual spectrum and reorganizes the spectrum to yield multiresolution spectrum. Most of the bands in high frequency domain are all low amplitude coefficients, which are regarded as insignificant bands. Those high frequency bands contain only zero after quantization. By discarding zero blocks (8x8 blocks which contain only zeroes) compression ratio will be raised compared to block based transform coding schemes. As for the other bands, they have the properties of similarity that is exploited to reduce the bit rate further. For each significant band a best match band is found by computing the errors with other subbands. The band which produces less error becomes the best match band. The best match data are zigzag scanned and are lossless arithmetic coded. With experimental results reported for various kinds of medical images such as sonogram, angiogram and x-ray, its performance was better than the JPEG and offers best decoded quality under the same bit rate. Compared to the scheme in Tai et al (2000), this method has the benefit of a smaller computational burden.

Medical image compression by sampling DCT coefficients was presented by Wu (2002). The original image in spatial domain is divided into
non overlapped sub images (16x16) and is converted in to spectral domain using DCT. The spectrum for still image is a two dimensional signal. The 2D signal is converted in to 1D signal by zigzag scanning, which possesses the property of compacting energy to low frequency regions. A sampling algorithm which provides adaptive sampling is applied to the 1D signal. The significant samples are selected from the 1D signal such that connecting two significant samples by linear interpolation constitutes an approximation to the original fragment. If the distortion formed by the line of two significant samples and the original waveform exceeds a given constant value specified by the user, then the samples forming the area are stored as non redundant samples. As to the other samples those that can be approximately predicted are taken as redundant samples, which are not necessary to be stored or transmitted. The significant samples selected here are used to represent the original signal and the other samples are discarded. The line segments between two significant samples are entropy coded by Huffman codes. The encoder blocks are shown in Figure 2.3. On the decoder side, a linear segment to connect two consecutive significant samples is used to reconstruct the decoded signal sequences.

![Figure 2.3 Encoder system](image)

**Figure 2.3 Encoder system**

Before adaptive sampling a classification of the sequences is performed because importance of every subblock inside the medical image is not equivalent. Therefore, the allowable sampling distortion area for different regions can be adopted adaptively as well to achieve a higher compression
ratio. This strategy achieves lower bit rates for medical images including sonogram, X-ray, CT and angiogram with higher PSNR values compared to JPEG and algorithms by Wu & Tai (1998) and Tai et al (2000).

Another work in (Chen & Tai 2004) adopts DCT to perform subband decomposition, followed by modified SPIHT data organization and entropy coding. The input image is subjected to 8x8 DCT. A translational function is applied on each block which categorizes the blocks into four (LL, LH, HL and HH) subbands. The coefficients from LL, LH, HL and HH subbands across all 8x8 DCT blocks are grouped respectively to form entire image in the frequency domain as shown in the Figure 2.4.

![Figure 2.4 Subband decomposition and data organization on the whole image](image)

In the second level, the grouped LL subbands are transformed to spatial domain by 4x4 IDCT. The 8x8 DCT is again applied to the spatial domain elements, which results in LL, LH, HL and HH subbands. SPIHT algorithm is applied which reduces redundancy of the same level subbands and the coefficients are entropy coded. For a given bit rate, this method yields higher PSNR than original SPIHT and JPEG 2000.
2.2.1.2 Discrete Wavelet Transform

Wavelet coding is proving to be a very effective technique for medical image compression, giving significantly better results than the JPEG standard algorithm which is based on DCT (Manduca 1995, Saipetch et al 1995). The standard steps in such compression are to perform the Discrete Wavelet Transform (DWT), quantize the resulting wavelet coefficients and losslessly encode the quantized coefficients. These coefficients are usually encoded in raster-scan order, although common variations are to encode each sub-block in a raster-scan order separately or to perform vector quantization within the various sub-blocks.

A good encoding scheme for wavelet coefficients, termed Embedded Zerotree Wavelet coding (EZW), was described by Shapiro (1993). Some of the ideas underlying EZW have been significantly modified and enhanced by Said and Pearlman (1996). Their approach, termed Set Partitioning in Hierarchical Trees (SPIHT), yields significantly better compression than conventional wavelet compression with similar computational complexity, and represents the state-of-the-art in general-purpose image compression. In (Manduca & Said 1997) the wavelet transformed medical images were coded with SPIHT and the results were compared to conventional wavelet compression, where the wavelet coefficients are quantized, run length and Huffman coded and to JPEG compression. For all compression ratios tried, the RMS differences between the original and compressed images were significantly smaller for the SPIHT algorithm than for the standard wavelet algorithm and JPEG.

In (Tohumoglu 1999), the well known Shapiro’s EZW algorithm for image coding was modified. This is designed to optimize the combination
of zerotree coding and Huffman coding. It is a two iteration EZW which combines two or more iterations of the original algorithm into one, comparing the coefficient values simultaneously to two different thresholds. This gives low bit rates for a given SNR over traditional EZW and JPEG and works at higher speed.

A diagnostically lossless medical image compression method by Qi et al (2000) describes wavelet transform modulus maxima and convex hull algorithm to extract the convex hull (many MRI and CT medical images have a convex hull that bounds the foreground of the image) which is all that is used for diagnosis that contains all diagnostic information of the original image (foreground). UNIX utilities compress and pack and lossless JPEG were used to compress the original image and denoised image of the original image size whose pixels outside convex hull region are set to zero. Compressing a denoised image improved CR (Compression ratio) by approximately 1.5 times when experimented with compress, pack and lossless JPEG.

In the Layered Set Partitioning in Hierarchical Trees (LSPIHT) (Hwang et al 2003) for scalable transmission systems, the encoded bit streams are divided into a number of layers for transmission and reconstruction. The LSPIHT algorithm can be viewed as a sequence of operations using the SPIHT algorithm with one operation for each layer. Starting from the base layer, the LSPIHT encodes one layer at a time until the design of the top layer is completed. In LSPIHT, the CR and resolution associated with each layer can be pre-specified before encoding. This layered image transmission through LSPIHT outperformed SPIHT based simulcast system.
As the proposed algorithms which will be discussed in chapter 3 and 4 are based on wavelet transform, a brief discussion on the various wavelet decompositions and popular wavelet based encoders for images which were the motivation for developing new algorithm for three dimensional medical data is done at this point.

**Wavelet Based Decomposition**

Decomposition of images in to subbands can be done through two dimensional filters or using separable transforms that can be implemented using one dimensional filters that operates on all rows first and then on all columns (Sayood 2006). Most of the image compression algorithms use the later one for image decomposition. Figure 2.5 shows decomposition of image of size $N \times M$. The image is filtered through two level filter bank structure with low pass filter $H_0$ and high pass filter $H_1$. In the first level we filter all rows and down sample the output of the filters to get two subbands (lowpass and highpass) of size $N \times \frac{M}{2}$. The two bands are subsequently subjected to column processing to give four subbands of size $\frac{N}{2} \times \frac{M}{2}$. The band obtained by low pass filtering both the rows and columns is indicated as LL subband of the image, the one obtained by low pass filtering the rows and high pass filtering the columns is referred as HL subband, similarly subimage obtained by high pass filtering the rows and low pass filtering the columns is referred as LH and the subimage obtained by high pass filtering both the rows and columns is known as HH image. LL, HL, LH and HH subbands obtained by first level decomposition are shown in Figure 2.6.
Figure 2.5 One level decomposition of image into four subbands

Figure 2.6 Subbands obtained from first level of decomposition

These subbands can be further subjected to filtering and sub sampling until we obtain a desired subband structure suitable for the encoder designed. The wavelet filters used for filtering is independent of the decomposition method. DWT can use any set of wavelet filters and decompose the image in any way. Few popular ways of applying discrete wavelet transform to the image so as to partition it into several subbands are
- **Standard Decomposition**

In this method wavelet transform of every row of the image is computed, which results in a transformed image where the first column contains averages and all the other columns contain differences. The standard algorithm then computes wavelet transform of every column. This result in one average value at the top left corner and rest of the top row containing averages of differences with all other coefficient values transformed into differences (Salomon 2004). Figure 2.7 shows the standard decomposition of the image.

![Standard Wavelet Decomposition](image)

**Figure 2.7 Standard wavelet decomposition**

- **Pyramid Decomposition**

In this method the wavelet transform is applied alternatively between rows and columns. Here, all the rows are filtered first which yields low pass band i.e., the averages in the left half and high pass band i.e., the differences in the right half. The next step is to filter all the columns which results in averages in top-left quadrant of the image and differences in all other bands. This process is again repeated only on the top-left quadrant sub
image. Again and again the process is repeated only on the resulting top-left sub image until only one average is left at the top-left corner and all other pixel values have been reduced to differences. This decomposition is depicted in Figure 2.8.

![Pyramid image wavelet decomposition](image)

**Figure 2.8 Pyramid image wavelet decomposition**

- **Line Wavelet Decomposition**

  Here, wavelet transform is applied to all the rows first, which splits the image into L1 and H1. Subband L1 alone is again split into L2 and H2. This process is repeated until only one left most band is average and all other bands are differences. The wavelet transform is then applied recursively to the leftmost column, resulting in one smooth coefficient at the left-top corner of the coefficient matrix. This method is shown in Figure 2.9. As this technique exploits only correlation among the rows, its implementation is simple and execution is fast, about twice that of the standard decomposition.
Figure 2.9 Line wavelet decomposition

- **Laplacian Pyramid**

  In Laplacian pyramid decomposition (Burt & Adelson 1983), the image is low pass filtered and the upsampled version of the lowpass image is subtracted from the original image. This process can be iterated by decomposing the coarse version repeatedly. The image is partitioned into a Gaussian pyramid (lowpass subbands) and a Laplacian pyramid that consists of detail coefficients (Highpass subbands). Only the Laplacian pyramid is required to reconstruct the image. The transformed image is bigger than the original image in this case.

- **Quincunx Decomposition**

  Quincunx decomposition is illustrated in Figure 2.10. Here decomposition proceeds level by level and decomposes subband $L_i$ of level $i$ into subbands $H_{i+1}$ and $L_{i+1}$ of level $i+1$. It is efficient and computationally simple. On average, it achieves more than four times the energy compaction of the line method.
Uniform Decomposition

This method is also known as wavelet packet transformation which is illustrated in Figure 2.11. When the image is subjected to wavelet transformation it forms four subbands and in the next level decomposition each and every subband is again split into four subbands resulting in sixteen subbands. When the process is repeated for \( l \) times, then it results in \( 2^l \times 2^l \) subbands. The computational cost of every level of decomposition is very high because it computes \( n^2 \) coefficients for every level of decomposition, where \( n \) is the side length of the image (Wong & Kuo 1993). In spite of high average reconstruction qualities, the perceptual quality of the image starts degrading at lower bit ratios than for the other decomposition methods because the support for a single coefficient is global. Its removal has the effect of blurring the reconstructed image.
Adaptive Wavelet Packet Decomposition

To avoid the heavy computation in uniform decomposition, in adaptive wavelet packet decomposition the subbands that do not contribute to energy compaction are exempted from decomposition. This leads to subbands of different sizes. The main problem in this type of decomposition is finding an algorithm that will determine which subband splits can be skipped. Two such decompositions are shown in Figure 2.12.
Wavelet Based Encoders

Attracted by the data structure of the wavelet transform, many algorithms for data compression were proposed by researchers in the past based on wavelet transform. We discuss in this section only three benchmark encoders.

Set Partitioning in Hierarchical Trees (SPIHT)

Set Partitioning in Hierarchical Trees encoder was developed by Said & Pearlman (1996). It is an embedded coder that refines the ideas presented in Shapiro’s (1993) Embedded Zero tree Wavelet (EZW) coder. In this algorithm the image is subjected to any appropriate subband transformation such as discrete wavelet transform. After transformation the image is represented by an indexed set of transformed coefficients $C_{i,j}$, located at pixel position $(i,j)$ in the transformed image. The transformed image is said to exhibit a hierarchical pyramidal structure defined by the level of decomposition, with the topmost level being the root. The finest coefficients lie at the bottom level of the pyramid while the coarsest coefficients lie at the top level. Between the subbands there is a spatial self similarity and the coefficients are better magnitude ordered if we move downward in the pyramid following the same spatial orientation. A tree structure called spatial orientation tree, defines the spatial relationship on the hierarchical pyramid. Figure 2.13 shows how the spatial orientation tree is defined in a pyramid constructed with recursive four band splitting. Each node of the tree corresponds to a pixel and is identified by the pixel coordinate. Its direct descendants correspond to the pixels of the same spatial orientation in the next finer level of the pyramid. The tree is defined in such a way that each node has either no offspring or four offspring, which always
form a group of $2 \times 2$ adjacent pixels. In the Figure 2.13 arrows are oriented from the parent node to its offspring. The pixels in the highest level of the pyramid are the tree roots which are grouped in $2 \times 2$ adjacent pixels. The coefficient in the top left corner of the array does not have any offsprings.

The following four types of sets, which are sets of coordinates of the coefficients, are used to explain the algorithm.

- $O(i,j)$: This is the set of coordinates of the offsprings of the wavelet coefficient at location $(i,j)$. As each node can either have four offsprings or none, the size of $O(i,j)$ is either zero or four.

- $D(i,j)$: This is the set of all descendants of the coefficient at location $(i,j)$. Descendants include the offsprings, the offsprings of the offsprings and so on.

- $H$: Set of coordinates of all spatial orientation tree roots (nodes in the highest pyramid level)

- $L(i,j)$: This is the set of coordinates of all the descendants of the coefficient at location $(i,j)$ except for the immediate offsprings of the coefficient at location $(i,j)$. i.e., $L(i,j)=D(i,j)-O(i,j)$

A wavelet coefficient at location $(i,j)$ in the pyramid representation has four direct descendants at locations, $O(i,j)=\{ (2i,2j), (2i,2j+1), (2i+1,2j), (2i+1, 2j+1) \}$. 
The algorithm makes use of three lists: the List of Insignificant Pixels (LIP), the List of Significant Pixels (LSP) and the List of Insignificant Sets (LIS). The LSP and LIS lists contain the coordinates of coefficients and LIS contains coordinates of the roots of sets of type $D$ or $L$. The coder starts with computation of threshold value $T=2^{\frac{\log_2 C_{\text{max}}}{2}}$, where $C_{\text{max}}$ is the value of largest coefficient to be encoded. At the start of the algorithm, LIP list is initialized with the set $H$. Those elements of $H$ that have descendants are placed in LIS as type $D$ entries. The LSP list is initially empty. Then the coordinates of LIP and LIS are examined for significance (A coefficient is said to be significant if it is greater than the threshold value and a set $D(i,j)$ or $L(i,j)$ is said to be significant if any coefficient in the set has a magnitude greater than the threshold). The insignificant coefficients are retained in LIP or transferred from LIS to LIP. The significant coefficients are transferred to LSP. After locating and recording all the significant pixels for the given threshold, the threshold is reduced by a factor of two and the process repeats. By the end of each stage, all coefficients that have been found to be significant will have their most significant bits recorded. As passes occur further (refinement pass) more precision is added to the value stored for each pixel. In this manner, the SPIHT algorithm performs a rough sorting of pixel
values by magnitude and records the values one bit at a time. It is the separation of bit planes that makes SPIHT, an embedded coder.

One of the main features of the proposed coding method is that the ordering data is not explicitly transmitted. Instead, it is based on the fact that the execution path of any algorithm is defined by the results of the comparisons on its branching points. So, if the encoder and decoder have the same sorting algorithm, then the decoder can duplicate the encoder’s execution path if it receives the results of the magnitude comparisons, and the ordering information can be recovered from the execution path. The complete algorithm with pseudo code is available in Said & Pearlman (1996).

- **Set Partitioned Embedded bloCK coder (SPECK)**

The Set Partitioned Embedded bloCK coder (SPECK) (Pearlman et al 2004) algorithm has its roots primarily in the ideas developed in the SPIHT and EBCOT (Taubman 2000) image coding algorithms. It is different from some of the above mentioned schemes in that it does not use trees which span and exploit the similarity across different subbands. Rather, it makes use of sets in the form of blocks.

Similar to SPIHT, a set $T$ of pixels is significant with respect to $n$ if
\[
\max_{(i,j) \in T} \{|c_{i,j}|\} \geq 2^n
\]
otherwise it is insignificant. The SPECK algorithm makes use of rectangular regions of image. These regions or sets are referred as sets of type $S$, which can be of varying size. The size of a set is the cardinality $C$ of the set, i.e., the number of elements in the set. During the course of the algorithm, other types of sets are also formed known as sets of type $I$. These sets are formed by chopping off a small square region from the top left portion of a large square region as shown in Figure 2.14. Set $I$ is decomposed into $S$ sets in a prescribed way, so as to progress through the transformed
image from coarser to finer resolution subbands. The coding part of SPECK always takes place on the $S$ sets. The motivation behind this is to exploit the clustering of energy found in transformed images and concentrate on those areas of the set which have high energy. This ensures that pixels with high information content are coded first.

![Figure 2.14 Partitioning of image into sets $S$ and $I$](image)

The algorithm maintains two lists: LIS – List of Insignificant Sets and LSP – List of Significant Pixels. The former contains sets of type $S$ of varying sizes which have not yet been found significant against a threshold $n$ while the latter contains those pixels which have been tested significant against $n$. It starts by partitioning the image into two sets: set $S$ which is the root of the pyramid, (In the transformed image, among the hierarchical pyramidal structure of subbands at different levels of its decomposition, the topmost band is the root of the pyramid) and set $I$ which is left of the image after taking out the root. Set $S$ is added to the LIS and is processed by testing it for significance against the threshold $n = n_{\text{max}}$. If not significant it stays in the LIS.

If $S$ is significant, it is quadrisected, i.e., partitioned into four subsets $O(S)$, each having size approximately one-fourth the size of the parent set $S$. Each of these offspring sets $O(S)$ is tested for significance for the same $n$ and, if significant, is quadrisected once more. If not significant, it is added
to the LIS. Each significant subset is, in turn treated as a set of type $S$ and processed recursively, until pixel level is reached where the pixels that are significant in the original set $S$ are located and thereby coded. The pixels/sets that are found insignificant during this selection process are added to LIS to be tested later against the next lower threshold. After all current sets of type $S$ have been tested against $n$, the set $I$ is processed next by testing it against the same threshold $n$. If it is found to be significant, it is octave band partitioned as shown in Figure 2.15. The new set $I$ formed by this partitioning process is now reduced in size.

![Figure 2.15 Partitioning of set I](image)

In this way, regions that are likely to contain significant pixels are grouped into relatively smaller sets and processed first, while regions that are likely to contain insignificant pixels are grouped into a large set. Once all the sets have been processed for a particular threshold $n$, the refinement pass is initiated which refines the quantization of the pixels in the LSP. Once this is done, the threshold is lowered and the sequence of sorting and refinement passes is repeated for sets in the LIS against this lower threshold. This process is repeated until the desired rate or till last threshold, corresponding to $n=0$. The complete algorithm with pseudo code is available in Pearlman et al (2004),
JPEG 2000

JPEG 2000, is popular compression standard which works based on wavelet transform. First step of the JPEG 2000 encoder is to transform the color components (if it is a color image) by means of either a Reversible Component Transform (RCT) or an Irreversible Component Transform (ICT). Each transformed component is then compressed separately. The component transform is preceded by a DC level shifting (if the pixels have unsigned values). The RCT is a decorrelating transform. Each transformed component of the image is partitioned into rectangular, nonoverlapping regions called tiles. Tiles may have any size, up to the size of the entire image. Each tile is compressed individually. The main reason for having tiles is to enable the user to decode parts of the image (region of interest). The decoder can identify each tile in the bitstream and decompress just those pixels included in the tile. A tile is compressed in four main steps. The first step is to compute wavelet transform. Two wavelet transforms are specified by the standard. They are the (9/7) floating point wavelet (irreversible) and the (5/3) integer wavelet (reversible). Either transform allows for progressive transmission, but only the integer transform can produce lossless compression. L level of decomposition is done, where L is a parameter determined by the encoder. In step two, the wavelet coefficients are quantized. Each subband can have a different quantization step size. The quantization step size may be determined iteratively in order to achieve a target bit rate or in order to achieve a predetermined level of image quality. Step three uses the MQ coder to arithmetically encode the wavelet coefficients. The EBCOT algorithm has been adopted for the encoding step. The principle of EBCOT is to divide each subbands into blocks (termed codeblocks) that are coded individually. The bits resulting from coding several code blocks become a packet and the packets are the components of the bitstream. The last step is to construct bitstream. This step places the packets, as well as many markers, in the
bitstream. The markers can be used by the decoder to skip certain areas of the bitstreams and to reach certain points quickly. Using markers the decoder can decode certain code blocks before others, thereby displaying certain regions of the image before other regions. The bitstream is organized in layers, where each layer contains higher resolution image information. Thus decoding the image layer by layer is a natural way to achieve progressive image transmission and decompression.

JPEG 2000 performs better than the original JPEG, especially for images where very low bit rate or very high image quality are required. For lossless or near lossless compression, JPEG 2000 offers only modest improvement over JPEG.

2.2.2 ROI Based Compression

In medical images, diagnostically useful information is mostly gathered and occupies a small area in the image. This feature can be utilized to compress an image. To useful areas (which is called as regions-of-interest, i.e. ROI), low compression is applied or even without compression to keep the quality of image high in this area. To other areas, compression is applied to give high density compression. The compressed image is kept with useful information as well as small size. This section provides ROI coding techniques proposed in the past applied for medical images.

The principle of the general scaling method, which is a component of JPEG 2000 part II is to scale (shift) transform coefficients so that the bits associated with the ROI are placed in higher bit planes than the bits associated with the background. During encoding process, the most significant ROI bit planes are first placed in the bit stream before background bit planes of the image. Thus, the ROI will be decoded, or refined, before the rest of the image. Regardless of the scaling, a full decoding of the bit stream results in a
reconstruction of the whole image with the highest fidelity available. If the bit stream is truncated, or the encoding process is terminated before the whole image is fully encoded, the ROI will be of higher quality than the rest of the image.

This technique is slightly modified in MAXSHIFT method (Bradley & Stentiford 2002), which is described in part I of JPEG 2000. In this method the wavelet coefficients corresponding to region of interest are scaled up by the encoder such that the minimum coefficient belonging to the ROI is larger than the maximum coefficient of the background (non-ROI area) and also the background coefficients are scaled down to values less than one. This allows the decoder to decode the ROI first without any need for ROI mask generation or information about shape of the ROI from encoder. The advantage of MAXSHIFT method over scaling method is that there is no need to transmit the ROI shape information as it is implied and decoder is as simple as non-ROI capable decoder. Figure 2.16 illustrates the scaling and MAXSHIFT method.

![Diagram of image compression methods](image.png)

**Figure 2.16** (a) Full Image Compression (b) ROI scaling based method (c) MAXSHIFT method
A lossy to lossless ROI compression scheme was proposed by Liu et al (2002a), where input images are segmented into the foreground and background, and a chain code-based shape coding scheme (Liu et al 2002b) is used to code the ROI's shape information. Critically sampled shape adaptive Integer Wavelet Transforms (IWTs) (Minami et al 2001) are performed on the foreground and background image separately. Coding of these coefficients is done through Set Partitioning in Hierarchical Trees (SPIHT) with EBCOT’s (Embedded Block Coding with Optimized Truncation) context model for bit plane coding. Finally, the shape-coding bit stream, the foreground bit stream, and the background bit stream are combined into a single bit stream. Experiments conducted on chromosome images proved the efficiency of the proposed scheme both in the context of lossy and lossless image compression.

Another ROI coding was proposed by Tasdoken & Cuhadar (2003), based on Region-Based Integer Wavelet Transform (RB-IWT). Here lifting scheme based IWT (Dewitte & Cornelis 1997) is applied to the partitioned image and modified SPIHT is used for encoding the arbitrary shape regions. This scheme was claimed to perform better than SPIHT based ROI coding Object-Based extension of the Set Partitioned Embedded Block (OB-SPECK) (Islam & Pearlman 1999) coding algorithm.

Employing Region Based Discrete Wavelet Transform (RBDWT) (Li & Li 2000), Penedo et al (2003) proposed two ROI coding methods. One method applied an Object-Based extension of the SPIHT (OB-SPIHT) coding algorithm and other method applies Object Based extension of the Set Partitioned Embedded bloCK (OB-SPECK) coding algorithm. With application to digital mammography, it is claimed to exhibit much higher quality than SPIHT and JPEG 2000. Dilmaghani et al (2003) brought about ROI coding through Embedded Zerotree Wavelet (EZW) algorithm on coefficients of ROI in wavelet domain. They also developed criteria for
optimum bit allocation for subbands of different resolution. Vector quantization based methods for ROI coding technique were proposed by Nasrabadi & King (1988) and Cziho et al (1998). In VQ, the vectors of input image are compared to the elements of a codebook called the code vectors or code words, and only the index of the nearest code vector is transmitted. Size of codebook and code word determines the quality of reconstructed image.

ROI segmentation algorithms suffered complexity and large execution time, so methods based on neural networks were introduced (Buller et al 1996, Chen et al 2002). The method presented in Chen et al (2002) was a semi-automatic method of ROI segmentation because the centre of ROI was determined manually. In Abdou & Tayel (2008), an automatic segmentation of ROI via an Artificial Neural Network and an Introduced Difference Fuzzy Model (IDFM) (Tayel & Abdou 2006) was presented. Here the non-ROI is coded with Fast and reduced bit Embedded Zerotree Wavelet algorithm (FEZW) (Tayel & Abdou 2005) and the ROI is coded with the same algorithm but with higher refinement level. The block diagram of the algorithm is as shown in Figure 2.17.

![Figure 2.17 Automatic bichannel compression technique for medical images](image-url)
The main advantage of this algorithm is that it automatically segments the ROI and reduces complexity when compared to others that used both lossy and lossless coding within the same medical image.

2.2.3 Neural Network Based Compression

Image Compression using Artificial Neural Networks is a topic where research is being carried out in various directions towards achieving a generalized and economical network. The algorithms discussed above are based on fixed transforms where as neural network compression use adaptive techniques. Advantages of Neural Networks (NN) include robustness under noisy conditions or incomplete data and simple decoding (Dony & Haykin 1995). Various research works are directed towards achieving quick convergence of the network without loss of quality of the restored image. Feedforward Networks using Back propagation Algorithm adopting the method of steepest descent for error minimization is popular and widely adopted for image compression. The general parameters deciding the performance of Back Propagation Neural Network Algorithm includes the mode of learning, information content, activation function, target values, input normalization, initialization, learning rate and momentum factors (Otair & Salameh 2004, Otair & Salameh 2005). Medical images contain a number of distinct gray levels with narrow difference with their neighborhood pixels. This makes back propagation algorithm to slow down in converging. The compression achieved is also not high.

To overcome these drawbacks a new approach using cumulative distribution function is proposed in (Anna Durai & Anna Saro 2008). Computational complexity is involved in compression of raw pixels of an image in spatial domain or the mathematically transformed coefficients in frequency domain using Artificial Neural Networks. If the gray levels of the pixels in an image and their neighbors are mapped in such a way that the
difference in the gray levels of the neighbor with the pixel is minimum, then compression ratio as well as the convergence of the network can be improved. To achieve this, the Cumulative Distribution Function (Gonzalez & Woods 2008) is estimated for the image and it is used to map the image pixels. When the mapped image pixels (preprocessed) are used as input, the Back propagation Neural Network yields high compression ratio as well as it converges quickly. There will not be any loss in data in the preprocessing and hence the finer details in the image are preserved in the reconstructed image. The convergence time, PSNR and compression ratio for BPNN has been improved by this approach.

There exist hundreds of modalities of medical images and each modality has hundreds of sub classes for different organs and different sizes, in such a situation it is difficult to generalize a neural network for all modalities. To tackle this problem, having a prior knowledge about nature and size of acquisition for a single type of medical image, a flag byte was proposed in (Ashraf & Akbar 2006) which is automatically set by the image size and some features. The number of modalities determines the number of bits in the flag. This flag byte is then used to select a particular compression architecture configuration. For codebook design, Kohonen’s self organizing feature map (Laha et al 2004) method is applied, which provides good VQ codebooks leading to better quality reconstructed images as compared to LBG (Linde-Buzo-Gray) algorithm (Linde et al 1980). Once trained and code book is ready it is transmitted to the receiver and then afterwards for any subsequent use it is assumed that receiver know the code book. The only overhead is the flag byte which depends on how many types of medical images are to be treated, prior knowledge. This algorithm is defined clearly in Figure 2.18.
Decoding process is only a lookup table. Decoder simplicity is vital in medical images where we encode once and decode many times for diagnosis and discussions. The results of this method have shown that compression ratios are better than JPEG for same PSNR.

The major drawbacks of compression using NN includes slow training, moderate compression ratios and quality of reconstructed image is highly dependent on training data.

2.3 FEATURES OF COMPRESSION ALGORITHMS

In the previous section, medical image compression techniques based upon DCT, DWT, Region of Interest and Neural Networks were presented. In all these approaches compression performance is influenced by many factors like transform selection, encoder design, segmentation algorithm, network selection, training algorithm etc., So general features of these approaches are summarized using Table 2.1.
Table 2.1 Features of compression techniques based upon DCT, DWT, ROI and NN

<table>
<thead>
<tr>
<th>Compression Technique based on</th>
<th>Features</th>
</tr>
</thead>
</table>
| Discrete Cosine Transform    | * High computational Efficiency.  
* Achieves compact representation for highly correlated signals i.e., closed to the performance of optimal KLT.  
* No full frame processing.  
* Good compression performance.  
* Suffers from blocking artifacts.  
* Used in many coding systems such as JPEG, MPEG and H.26X due to high energy compaction property in the frequency domain. |
| Discrete Wavelet Transform   | * High compression ratio compared to DCT approach.  
* No blurring of images.  
* Multiresolution analysis.  
* Compact energy spectrum.  
* Improves visual quality.  
* High computational complexity.  
* Used in JPEG 2000 Scheme. |
| Region of Interest           | * Can be transform based or non-transform based.  
* Preserves image quality in diagnostically critical regions.  
* High compression ratio with good quality in ROI.  
* Algorithm is dependent on the kind of image and ailment.  
* Part of JPEG 2000. |
| Neural Networks               | * Predictive approach.  
* Superior over traditional methods when dealing with noisy or incomplete data.  
* Quality of reconstruction image is highly dependent on training data.  
* Test data should be similar to training data.  
* Non transformed approach.  
* Simple decoding.  
* No progressive enhancement. |
2.4 CONCLUSION

In this chapter, a review of various coding techniques for medical images that have exploited the unique features of medical images is presented. The techniques are classified into three categories and their basic features were discussed. Among the various approaches discussed, the approaches that were felt promising to attack the problems stated in chapter 1 of this thesis are DWT and ROI based approaches. So under wavelet based compression, various wavelet decomposition methods for efficient representation of images prior to compression and three popular wavelet based encoders were discussed.

Though there are many techniques proposed with unique characteristics, research has to be done to develop efficient and simple techniques that will produce high quality reconstructed images with high CR particularly for three dimensional medical data and enable their use in portable and mobile devices, which have limited computing power. One such developed technique is discussed in the next chapter.