Chapter 2 Digital Image Compression

2.1 Data Compression and Data Redundancy

Data compression is defined as the process of encoding data using a representation that reduces the overall size of data. This reduction is possible when the original dataset contains some type of redundancy. Digital image compression is a field that studies methods for reducing the total number of bits required to represent an image. This can be achieved by eliminating various types of redundancy that exist in the pixel values. In general, three basic redundancies exist in digital images that follow.

Psycho-visual Redundancy: It is a redundancy corresponding to different sensitivities to all image signals by human eyes. Therefore, eliminating some less relative important information in our visual processing may be acceptable.

Inter-pixel Redundancy: It is a redundancy corresponding to statistical dependencies among pixels, especially between neighboring pixels.

Coding Redundancy: The uncompressed image usually is coded with each pixel by a fixed length. For example, an image with 256 gray scales is represented by an array of 8-bit integers. Using some variable length code schemes such as Huffman coding and arithmetic coding may produce compression.

There are different methods to deal with different kinds of aforementioned redundancies. As a result, an image compressor often uses a multi-step algorithm to reduce these redundancies.

2.2 Compression Methods

During the past two decades, various compression methods have been developed to address major challenges faced by digital imaging.[3—10] These compression methods

can be classified broadly into lossy or lossless compression. Lossy compression can achieve a high compression ratio, 50:1 or higher, since it allows some acceptable degradation. Yet it cannot completely recover the original data. On the other hand, lossless compression can completely recover the original data but this reduces the compression ratio to around 2:1.

In medical applications, lossless compression has been a requirement because it facilitates accurate diagnosis due to no degradation on the original image. Furthermore, there exist several legal and regulatory issues that favor lossless compression in medical applications.[11]

2.2.1 Lossy Compression Methods

Generally most lossy compressors (Figure 2.1) are three-step algorithms, each of which is in accordance with three kinds of redundancy mentioned above.

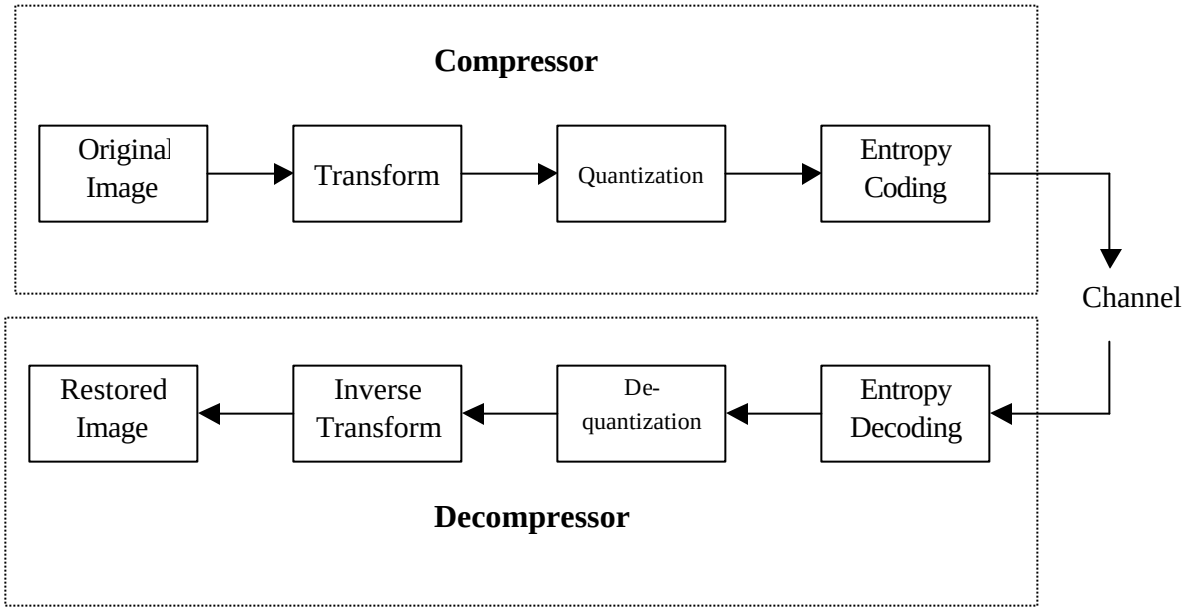

Figure 2.1 Lossy image compression

The first stage is a transform to eliminate the inter-pixel redundancy to pack information efficiently. Then a quantizer is applied to remove psycho-visual

redundancy to represent the packed information with as few bits as possible. The quantized bits are then efficiently encoded to get more compression from the coding redundancy.

2.2.1.1 Quantization

Quantization is a many-to-one mapping that replaces a set of values with only one representative value. Scalar and vector quantization are two basic types of quantization. SQ (scalar quantization) performs many-to-one mapping on each value. VQ (vector quantization) replaces each block of input pixels with the index of a vector in the codebook, which is close to the input vector by using some closeness measurements. The decoder simply receives each index and looks up the corresponding vector in the codebook.

Shannon[12] first showed that VQ would result in a lower bit rate than SQ. But VQ suffers from a lack of generality, since the codebook must be trained on some set of initial images. As a result, the design of the codebook will directly affect the bit rate and distortion of the compression.

Riskin et. al.[5] presented variable-rate VQ design and applied it to MR images. Cosman et. al.[13] used similar methods to compress CT and MR chest scans. Xuan et. al.[14] also used similar VQ techniques to compress mammograms and brain MRI.

2.2.1.2 Transform Coding

Transform coding is a general scheme for lossy image compression. It uses a reversible and linear transform to decorrelate the original image into a set of coefficients in transform domain. The coefficients are then quantized and coded sequentially in transform domain.

Numerous transforms are used in a variety of applications. The discrete KLT (Karhunen-Loeve transform), which is based on the Hotelling transform, is optimal with its information packing properties, but usually not practical since it is difficult to compute. [15,16] The DFT (discrete Fourier transform) and DCT (discrete cosine transform) approximate the energy-packing efficiency of the KLT, and have more efficient implementation. In practice, DCT is used by most practical transform systems since the DFT coefficients require twice the storage space of the DCT coefficients.

Block Transform Coding

In order to simplify the computations, block transform coding exploits correlation of the pixels within a number of small blocks that divide the original image. As a result, each block is transformed, quantized and coded separately. This technique, using square 8*8 pixel blocks and the DCT followed by Huffman or arithmetic coding, is utilized in the ISO JPEG (joint photographic expert group) draft international standard for image compression.[17-19] The disadvantage of this scheme is the blocking (or tiling) artifacts appear at high compression ratios.

Since the adoption of the JPEG standard, the algorithm has been the subject of considerable research. Collins et. al.[20] studied the effects of a 10:1 lossy image compression scheme based on JPEG, with modifications to reduce the blocking artifacts. Baskurt et. al.[21] used an algorithm similar to JPEG to compress mammograms with a bit rate as low as 0.27 bpp (bits per pixel) while retaining detection ability of pathologies by radiologists. Kostas et. al.[22] used JPEG modified for use with 12-bit images and custom quantization tables to compress mammograms and chest radiographs.

Moreover, the ISO JPEG committee is currently developing a new still-image compression standard called JPEG-2000 for delivery to the marketplace by the end of the year 2000. The new JPEG-2000 standard is based upon wavelet decompositions

combined with more powerful quantization and encoding strategies such as embedded quantization and context-based arithmetic. It provides the potential for numerous advantages over the existing JPEG standard. Performance gains include improved compression efficiency at low bit rates for large images, while new functionalities include multi-resolution representation, scalability and embedded bit stream architecture, lossy to lossless progression, ROI (region of interest) coding, and a rich file format.[23]

Full-Frame Transform Coding

To avoid the artifacts generated by block transforms, full-frame methods, in which the transform is applied to the whole image as a single block, have been investigated in medical imaging research.[24-26] The tradeoff is the increased computational requirements and the appearance of ringing artifacts (a periodic pattern due to the quantization of high frequencies).

Subband coding is one example among full-frame methods. It will produce a number of sub-images with specific properties such as a smoothed version of the original plus a set of images with the horizontal, vertical, and diagonal edges that are missing from the smoothed version according to different frequencies.[27-29] Rompelman[30] applied subband coding to compress 12-bit CT images at rates of 0.75 bpp and 0.625 bpp without significantly affecting diagnostic quality.

Recently, much research has been devoted to the DWT (discrete wavelet transform) for subband coding of images. DWT is a hierarchical subband decomposition particularly suited to image compression.[31] Many different wavelet functions can be applied to different applications. In general, more complicated wavelet functions provide better performance. The wavelet transform can avoid the blocking artifacts presented in block transform methods and allow easy progressive coding due to its multiresolution nature.

Bramble et. al. [32] used full-frame Fourier transform compression on 12 bpp digitized hand radiographs at average rates from about 0.75 bpp to 0.1 bpp with no significant degradation in diagnostic quality involving the detection of pathology characterized by a lack of sharpness in a bone edge. However, Cook et. al.[33] investigated the effects of full-frame DCT compression on low-contrast detection of chest lesions and found significant degradation at rates of about 0.75 bpp. These results illustrate that both imaging modality and the task play an important role in determining achievable compression.

2.2.2 Lossless Compression Methods

Lossless compressors (Figure 2.2) are usually two-step algorithms. The first step transforms the original image to some other format in which the inter-pixel redundancy is reduced. The second step uses an entropy encoder to remove the coding redundancy. The lossless decompressor is a perfect inverse process of the lossless compressor.

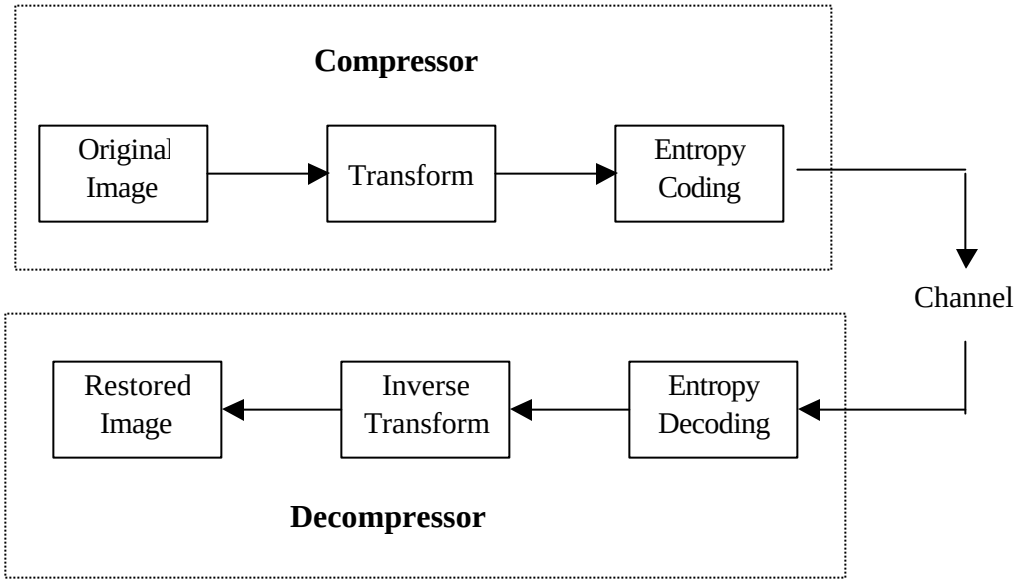

Figure 2.2 Lossless image compression

Typically, medical images can be compressed losslessly to about 50% of their original size. Boncelet et. al.[34] investigated the use of three entropy coding methods for lossless compression with an application to digitized radiographs and found that a bit rate of about 4 to 5 bpp was best. Tavakoli[35, 36] applied various lossless coding techniques to MR images and reported a compression down to about 5 to 6 bpp, with LZ (Lempel-Ziv) coding achieving the best results.

Lossless compression works best with decorrelated data. Roose et. al.[5, 37] investigated prediction, linear transformation, and multiresolution methods for decorrelating medical image data before coding them. The compression result was 3:1 and less than 2:1 for angiograms and MRI respectively. Kuduvalli and Rangayyan[6] studied similar techniques and found linear prediction and interpolation techniques gave the best results with similar compression ratios.

Here, we summarize the lossless compression methods into four categories.

2.2.2.1 Run Length Coding

Run length coding replaces data by a (length, value) pair, where "value" is the repeated value and "length" is the number of repetitions. This technique is especially successful in compressing bi-level images since the occurrence of a long run of a value is rare in ordinary gray-scale images. A solution to this is to decompose the gray-scale image into bit planes and compress every bit-plane separately. Efficient run-length coding method[38] is one of the variations of run length coding.

2.2.2.2 Lossless Predictive Coding

Lossless predictive coding predicts the value of each pixel by using the values of its neighboring pixels. Therefore, every pixel is encoded with a prediction error rather than its original value. Typically, the errors are much smaller compared with the original value so that fewer bits are required to store them.

DPCM (differential pulse code modulation) is a predictive coding based lossless image compression method. It is also the base for lossless JPEG compression. A variation of the lossless predictive coding is the adaptive prediction that splits the image into blocks and computes the prediction coefficients independently for each block to achieve high prediction performance. It can also be combined with other methods to get a hybrid coding algorithm with higher performance.[14, 39]

2.2.2.3 Entropy Coding

Entropy represents the minimum size of dataset necessary to convey a particular amount of information. Huffman coding, LZ (Lempel-Ziv) coding and arithmetic coding are the commonly used entropy coding schemes.

Huffman coding utilizes a variable length code in which short code words are assigned to more common values or symbols in the data, and longer code words are assigned to less frequently occurring values. Modified Huffman coding[40] and dynamic Huffman coding[41]are two examples among many variations of Huffman's technique.

LZ coding replaces repeated substrings in the input data with references to earlier instances of the strings. It often refers to two different approaches to dictionary-based compression: the LZ77[42] and the LZ78[43]. LZ77 utilizes a sliding window to search for the substrings encountered before and then substitutes them by the (position, length) pair to point back to the existing substring. LZ78 dynamically constructs a dictionary from the input file and then replaces the substrings by the index in the dictionary. Several compression methods, among which LZW (Lempel-Ziv-Welch)[44] is one of the most well known methods, have been developed based on these ideas. Variations of LZ coding are used in the Unix utilities Compress and Gzip.

Arithmetic coding[45] represents a message as some finite intervals between 0 and 1 on the real number line. Basically, it divides the intervals between 0 and 1 into a number of smaller intervals corresponding to the probabilities of the message's symbols. Then the first input symbol selects an interval, which is further divided into smaller intervals. The next input symbol selects one of these intervals, and the procedure is repeated. As a result, the selected interval narrows with every symbol, and in the end, any number inside the final interval can be used to represent the message. That is to say, each bit in the output code refines the precision of the value of the input code in the interval. A variation of arithmetic coding is the Q-coder[46], developed by IBM in the late 1980's. Two references are provided for the latest Q-coder variation.[47, 48]

2.2.2.4 Multiresolution Coding

HINT (hierarchical interpolation)[5, 37] is a multiresolution coding scheme based on sub-samplings. It starts with a low-resolution version of the original image, and interpolates the pixel values to successively generate higher resolutions. The errors between the interpolation values and the real values are stored, along with the initial low-resolution image. Compression is achieved since both the low-resolution image and the error values can be stored with fewer bits than the original image.

Laplacian Pyramid[49] is another multiresolution image compression method developed by Burt and Adelson. It successively constructs lower resolution versions of the original image by down sampling so that the number of pixels decreases by a factor of two at each scale. The differences between successive resolution versions together with the lowest resolution image are stored and utilized to perfectly reconstruct the original image. But it cannot achieve a high compression ratio because the number of data values is increased by 4/3 of the original image size.

In general, the image is reversibly transformed into a group of different resolution sub-images in multiresolution coding. Usually, it reduces the entropy of the image.

Some kinds of tree representation could be used to get more compression by exploiting the tree structure of the multiresolution methods.[50]

2.3 Measurements for Compression Methods

2.3.1 Measurements for Lossy Compression Methods

Lossy compression methods result in some loss of quality in the compressed images. It is a tradeoff between image distortion and the compression ratio. Some distortion measurements are often used to quantify the quality of the reconstructed image as well as the compression ratio (the ratio of the size of the original image to the size of the compressed image). The commonly used objective distortion measurements, which are derived from statistical terms, are the RMSE (root mean square error), the NMSE (normalized mean square error) and the PSNR (peak signal-to-noise ratio).

These measurements are defined as follows:

$$RMSE = \sqrt{\frac{1}{N*M}\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}[f(i,j)-f'(i,j)]^2}$$

$$NMSE = \sum_{i=0}^{N-1}\sum_{j=0}^{M-1}[f(i,j)-f'(i,j)]^2 \Big/ [\sum_{i=0}^{N-1}\sum_{j=0}^{M-1} f(i,j)]^2$$

$$PSNR = 20*\log_{10}(\frac{255}{RMSE})$$

where the images have $N*M$ pixels (8 bits per pixel), $f(i,j)$ represents the original image, and $f'(i,j)$ represents the reconstructed image after compression-decompression.

Since the images are for human viewing, it leads to subjective measurements based on subjective comparisons to tell how "good" the decoded image looks to a human viewer. Sometimes, application quality can be used as a measure to classify the

usefulness of the decoded image for a particular task such as clinical diagnosis in medical images and meteorological prediction in satellite images and so on.

When comparing two lossy coding methods, we may either compare the qualities of images reconstructed at a constant bit rate, or, equivalently, we may compare the bit rates used in two constructions with the same quality, if it is accomplishable.

2.3.2 Measurements for Lossless Compression Methods

Lossless compression methods result in no loss in the compressed images so that it can perfectly restore the original images when applying a reversible process. The frequently used measurement in lossless compression is the compression ratio. This measurement can be misleading, since it depends on the data storage format and sampling density. For instance, medical images containing 12 bits of useful information per pixel are often stored using 16 bpp.

A better measurement of compression is the bit rate due to its independence of the data storage format. A bit rate measures the average number of bits used to represent each pixel of the image in a compressed form. Bit rates are measured in bpp, where a lower bit rate corresponds to a greater amount of compression.

2.4 Summary

Digital image compression has been the focus of a large amount of research in recent years. As a result, data compression methods grow as new algorithms or variations of the already existing ones are introduced. All these digital image compression methods are concerned with minimization of the amount of information used to represent an image. They are based on the same principles and on the same theoretical compression model, which effectively reduces three types of redundancy, such as psycho-visual, inter-pixel and coding, inherited in gray-level images.

However, a 3-D medical image set contains an additional type of redundancy, which is not often addressed by the current compression methods. Several methods[51-58] that utilize dependencies in all three dimensions have been proposed. Some of these methods[51, 52, 54, 55] used the 3-D DWT in a lossy compression scheme, whereas others[53, 57] used predictive coding in a lossless scheme. In the latest paper,[58] 3D CB-EZW (context-based embedded zerotree wavelet) algorithm was proposed to efficiently encode 3-D image data by the exploitation of the dependencies in all dimensions, while enabling lossy and lossless decompression from the same bit stream.

In this proposal, we first introduce a new type of redundancy existing among pixels values in all three dimensions from a new point of view and its basic characteristics. Secondly, we propose a novel lossless compression method based on integer wavelet transforms, embedded zerotree and predictive coding to reduce this special redundancy to gain more compression. Thirdly, we expand the proposed compression method to the application of the telemedicine to support the transmission of the ROI without any diagnostic information loss and the simple diagnosis of certain disease such as multiple sclerosis in MR brain images.