Term Project Report

On

A SURVEY OF CONTEMPORARY MEDICAL IMAGE COMPRESSION TECHNIQUES

Submitted by :

Y.Praveen Kumar (03010240)
Kanchan Mishra (03010242)

Department of Electronics and Communication Engineering
IIT Guwahati

1.Introduction

Advances over the past decade in many aspects of digital technology - especially devices for image acquisition, data storage, and bitmapped printing and display - have brought about many applications of digital imaging. However, these applications tend to be specialized due to their relatively high cost. One such application is the Medical Imaging.

Advanced medical imaging technologies, such as computed tomography (CT), magnetic resonance imaging (MRI) and traditional radiography performed using computed radiography (CR) and digital radiography (DR) are fundamental tools in providing more efficient and effective healthcare systems and services. The key to the proliferation of these technologies is the digital representation of images. Digital medical images have potential benefits in terms of durability ,portability and versatality. However, problems involving storage space and network bandwidth requirements arise when large volumes of images are to be stored or transmitted, as is the case with medical images. From the diagnostic imaging point of view, the challenge is how to deliver clinically critical information in the shortest time possible. A solution to this problem is through image compression.

In this Article we discuss the need for medical image compression and then we briefly summarize the various medical image compression standards. We also discuss some of the recently proposed novel medical compression techniques and finally the scope and feature reasearch in this topic is presented.

2. Image compression

Uncompressed multimedia (graphics, audio and video) data requires considerable storage capacity and transmission bandwidth. Despite rapid progress in mass-storage density, processor speeds, and digital communication system performance, demand for data storage capacity and data-transmission bandwidth continues to outstrip the capabilities of available technologies. The recent growth of data intensive multimedia-based web applications have not only sustained the need for more efficient ways to encode signals and images but have made compression of such signals central to storage and communication technology.

The figures in Table 1 show the qualitative transition from simple text to full-motion video data and the disk space, transmission bandwidth, and transmission time needed to store and transmit such uncompressed data.

Multimedia Data	**Size/Duration**	**Bits/Pixel or Bits/Sample**	**Uncompressed Size (B for bytes)**	**Transmission Bandwidth (b for bits)**	**Transmission Time (using a 28.8K Modem)**
A page of text	11" x 8.5"	Varying resolution	4-8 KB	32-64 Kb/page	1.1 - 2.2 sec
Telephone quality speech	10 sec	8 bps	80 KB	64 Kb/sec	22.2 sec
Grayscale Image	512 x 512	8 bpp	262 KB	2.1 Mb/image	1 min 13 sec
Color Image	512 x 512	24 bpp	786 KB	6.29 Mb/image	3 min 39 sec
Medical Image	2048 x 1680	12 bpp	5.16 MB	41.3 Mb/image	23 min 54 sec
SHD Image	2048 x 2048	24 bpp	12.58 MB	100 Mb/image	58 min 15 sec
Full-motion Video	640 x 480, 1 min (30 frames/sec)	24 bpp	1.66 GB	221 Mb/sec	5 days 8 hrs

Table 1 Multimedia data types and uncompressed storage space, transmission bandwidth, and transmission time required. The prefix kilo- denotes a factor of 1000 rather than 1024.

The examples above clearly illustrate the need for sufficient storage space, large transmission bandwidth, and long transmission time for image, audio, and video data. At the present state of technology, the only solution is to compress multimedia data before its storage and transmission, and decompress it at the receiver for play back. For example, with a compression ratio of 32:1, the space, bandwidth, and transmission time requirements can be reduced by a factor of 32, with acceptable quality.

2.1 Principles behind compression

A common characteristic of most images is that the neighboring pixels are correlated and therefore contain redundant information. The foremost task then is to find less correlated representation of the image. In general, three types of redundancy can be identified:
1) Spatial Redundancy or correlation between neighboring pixel values.
2) Spectral Redundancy or correlation between different color planes or spectral bands.
3) Temporal Redundancy or correlation between adjacent frames in a sequence of images.

2.2 Different classes of compression techniques

Two ways of classifying compression techniques are mentioned here.

(a) Lossless vs. Lossy compression: In lossless compression schemes, the reconstructed image, after compression, is numerically identical to the original image. However lossless compression can only a achieve a modest amount of compression. An image reconstructed following lossy compression contains degradation relative to the original. Often this is because the compression scheme completely discards redundant information. However, lossy schemes are capable of achieving much higher compression. Under normal viewing conditions, no visible loss is perceived (visually lossless).

(b) Predictive vs. Transform coding: In predictive coding, information already sent or available is used to predict future values, and the difference is coded. Since this is done in the image or spatial domain, it is relatively simple to implement and is readily adapted to local image characteristics. Differential Pulse Code Modulation (DPCM) is one particular example of predictive coding. Transform coding, on the other hand, first transforms the image from its spatial domain representation to a different type of representation using some well-known transform and then codes the transformed values (coefficients). This method provides greater data compression compared to predictive methods, although at the expense of greater computation.

2.3 Normal Compression Vs Medical Image Compression

The coding of medical images differs from the coding of standard natural images in that it is imperative that the integrity of the diagnostic information in medical images are maintained while providing a reduction in storage space and network transmission bandwidth requirements. Inevitably, the ultimate solution is through reversible compression. However, at present, the existing state-of-the-art reversible technologies cannot achieve a significant reduction in bit-rate deemed adequate for the current practical applications in biomedical imaging .

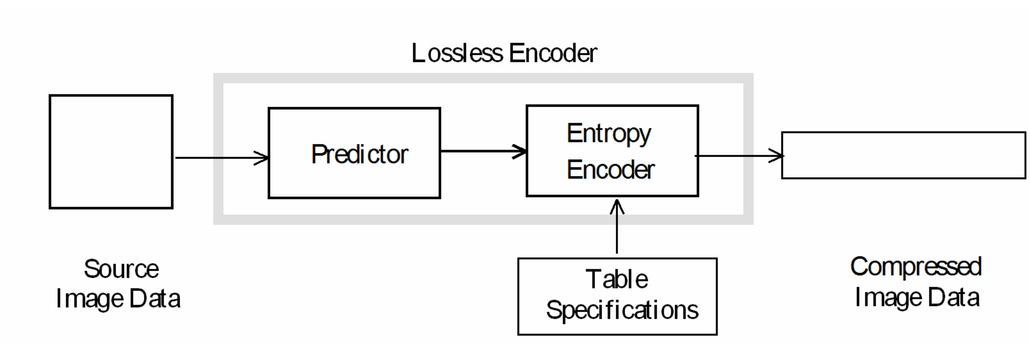
There have been numerous compression research studies examining the use of compression as applied to medical images. The papers can be categorised as focusing on just a lossless compression method, on just a lossy compression method, or focusing on both. Most have focused on lossless algorithms since the medical community has been reluctant to adopt lossy techniques owing to the legal and regulatory issues that are raised, but this situation may start to change as more lossy research is performed. Lossless image compression is typically performed in two steps, decorrelation and coding. Image decorrelation attempts to reduce the redundancy within the image. There are several common approaches that have been taken in the literature to perform this redundancy reduction step including differential pulse code modulation, hierarchical interpolation, bit-plane encoding and multiplicative autoregression. Several popular approaches for encoding are Huffman encoding, Lempel-Ziv encoding, arithmetic encoding and run-length encoding.

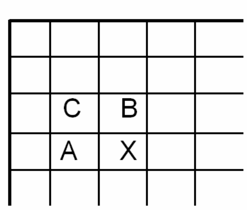
3. Medical Image Compression Standards

Emphasis is placed on those techniques that have been adopted or proposed as international standards. Particular attention is directed to the older JPEG lossless processes , the new JPEG-LS process and the lossless mode of the proposed JPEG 2000 scheme .

3.1 JPEG Predictive Lossless Standard

A predictor combines the values of up to three neighboring samples (A, B, and C) to form a prediction of the sample indicated by X in Figure. This prediction is then subtracted from the actual value of sample X, and the difference is encoded losslessly by either of the entropy coding methods - Huffman or arithmetic. Any one of the eight predictors listed in Table (under “selection-value”) can be used. Selections 1, 2, and 3 are one-dimensional predictors and selections 4, 5, 6 and 7 are two-dimensional predictors. Selection-value 0 can only be used for differential coding in the hierarchical mode of operation. The encoders can use any source image precision from 2 to 16 bits/sample, and can use any of the predictors except selection-value 0. The decoders must handle any of the sample precisions and any of the predictors. Lossless codecs typically produce around 2:1 compression for color images with moderately complex scenes.



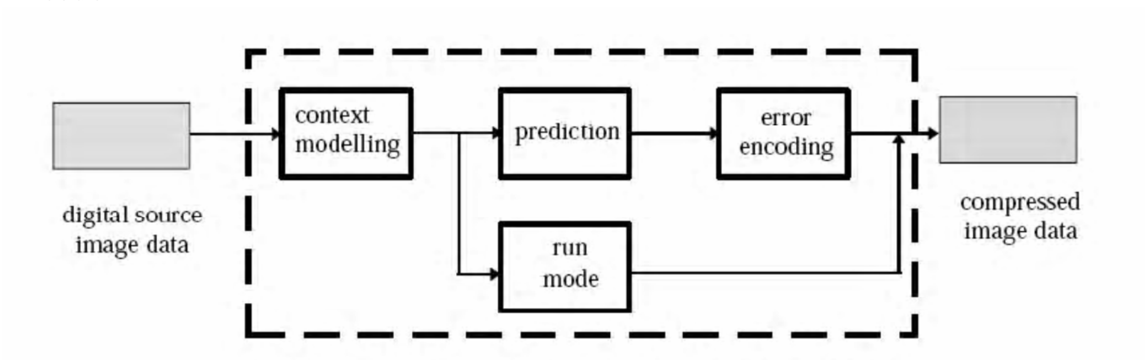


selection-value	prediction
0	no prediction
1	A
2	B
3	C
4	A+B-C
5	A+((B-C)/2)
6	B+((A-C)/2)
7	(A+B)/2

3.2 The JPEG-LS Standard

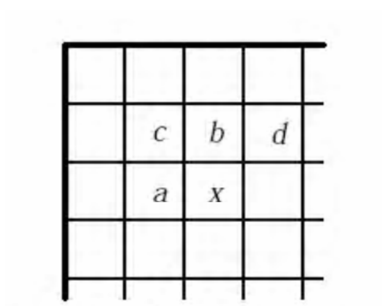
JPEG-LS is the basis for new lossless/near-lossless compression standard for compressing continuous-tone, greyscale, or colour digital still images , especially Medical Images. The standard is based on the LOCO-I algorithm (Low COmplexity LOssless COmpression for Medical Images). The algorithm uses context modeling. Context is a function of samples in the causal template used to condition the coding of the present sample. Context modeling is the procedure determining probability distribution of prediction error from the context. Each sample value is conditioned on a small number of neighbouring samples.

Encoder



Context Modeling

context is determined from four neighbourhood reconstructed samples at positions a, b, c, and d of the same component



context determines if the information in the sample x should be encoded in the regular mode (neighbours not very alike) or run mode (when neighbours are very alike).

Prediction (regular mode)

a, b, and c are used to form a prediction of the sample at position x. Prediction error is computed as the difference between the actual sample value at position x and its predicted value. This prediction error is then corrected by a context dependent term to compensate for systematic biases in prediction.

Error encoding (regular mode)

The corrected prediction error (further quantized for nearlossless coding) is then encoded using a procedure derived from Golomb coding. The Golomb coding procedures depend on the context determined by the values of the samples at positions a, b, c, and d as well as prediction errors previously encoded for the same context.

Run mode

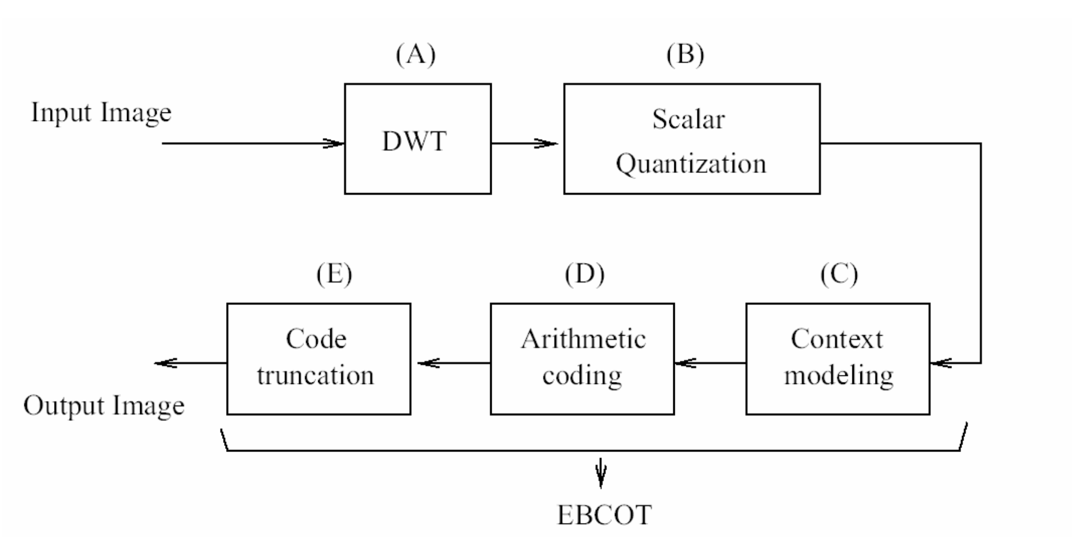
This mode is selected when reconstructed values of a,b,c and d are identical or within bounds when near-lossless coding. The mode skips prediction and error-coding. The encoder looks, starting at x, for a sequence of consecutive samples with values identical to the reconstructed value of the sample at a. The length information is encoded.

Decoding process

Encoding and decoding processes are approximately symmetrical. Decoding process is followed by a sample mapping procedure which uses the value of each decoded sample as an index to a look-up table, provided in the compressed image data. If no table is provided for a specific component the output of the sample mapping procedure is identical to the input.

3.3 Lossless JPEG 2000 Standard

JPEG2000 coding is a kind of unified lossless/lossy coding. The differences between lossless and lossy algorithms are two parts. The first part is in the implementation of discrete wavelet transform (DWT), and the second part is in the rate-control scheme.



DWT computation

DWT is carried out by the mallat decomposition of 2-channel filter banks in JPEG2000. Filters in the filter banks are classified into two types: One is an integer filter that has integer coefficients, and the other is a floating filter that has non-integer coefficients. Lossless JPEG2000 coding uses integer DWT (IWT) that is carried out by lifting schemes with integer filter and round operation.

Rate-control operation

In JPEG2000 coding, the use of two rate-control methods is allowed. One is code truncation in the EBCOT algorithm called post-quantization, and the other is pre-quantization using a scalar quantizer. Either the post-quantization or the prequantization method can be used in lossy coding. Meanwhile, no rate-control operation is required for the lossless coding.

EBCOT algorithm

EBCOT is one of the bit-plane based coding algorithms. The transformed coefficients are decomposed into bitplanes and are encoded by the MQ arithmetic coder. Then, these encoded coefficients are truncated for the rate-control. When IWT is used as DWT and pre-quantization is skipped, there is no difference between lossy coding and lossless coding until the code truncation is performed. To perform lossless coding, we have to choose IWT and skip both the pre-quantization and the post-quantization steps.

4. Some Recent and Novel Methods for Medical Image Compression

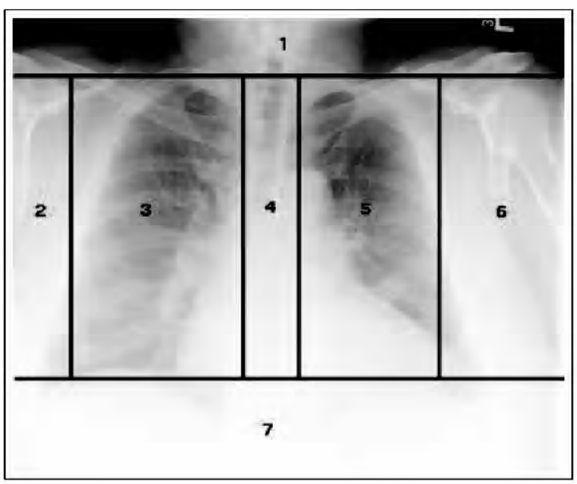
As medical/biological imaging facilities move towards complete film-less imaging, compression plays a key role. Although lossy compression techniques yield high compression rates, the medical community has been reluctant to adopt these methods, largely for legal reasons, and has instead relied on lossless compression techniques that yield low compression rates. The true goal is to maximise compression while maintaining clinical relevance and balancing legal risk. Keeping this in mind many new methods for Medical Image compression were proposed in the past two years.

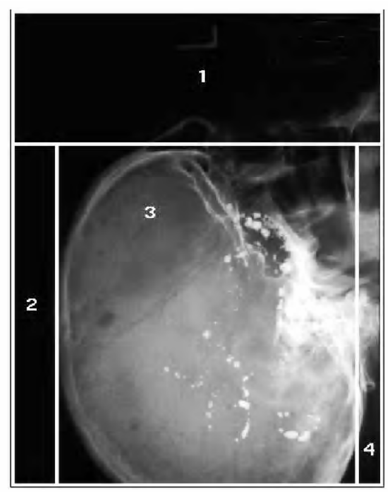
4.1 A Model Based Approach for Medical image compression

A new model-based approach to medical image compression by the use of image registration is proposed. An image that needs to be compressed will first be aligned to an image of its own type prestored in an atlas (such as the head or chest). Once a film is registered (i.e., aligned), two possibilities exist. The simpler approach is simply to read off the ‘relevant’ regions and then use lossless compression in relevant regions and lossy compression in the others. The alternative is that the new image can be subtracted from the prestored atlas image generating a residual image. This residual image will be compressed (lossless in clinically relevant regions and lossy in the others). If the alignment is done well, the residual information is minimised, thus yielding higher compression.

The regions will be defined to classify areas of the image into those that are clinically relevant and those that are not clinically relevant. These regions are stored in the atlas and have been predefined by radiologists. Depending on the need the physician may override the default regions and define new relevant regions of his own.

Lossless compression will be used in the clinically relevant regions and lossy compression will be used in areas that are not clinically relevant. Lossy compression such as JPEG, utilise a compression amount parameter that defines the amount of compression, and hence degradation, used on the image. Varying this parameter different ratios of compression can be obtained.



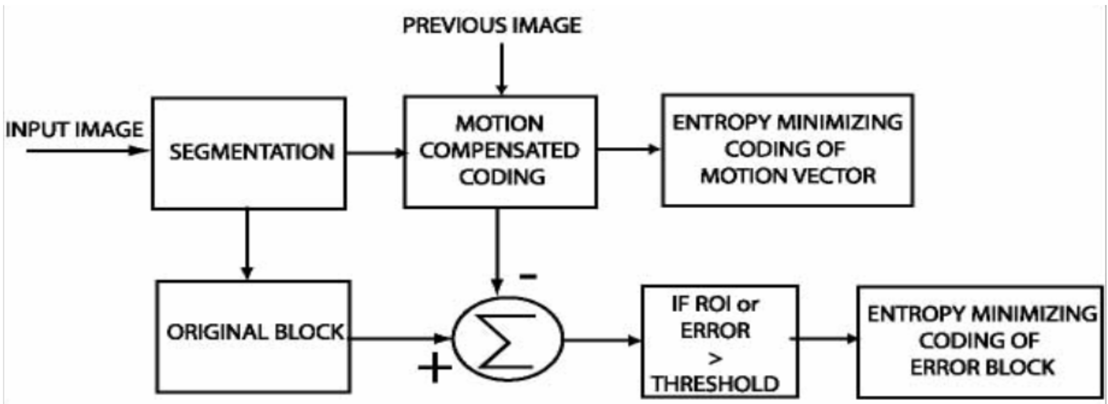


The first image is that of a partitioned chest X-ray. Regions 3 and 5 are defined to be clinically relevant. Though all other regions are clinically non-relevant region 4 is defined to be more important than the others. Hence region 3 and 5 are compressed losslessly , region 4 is compressed with lossy JPEG quality of 50 whereas all other regions are compressed ata quality of only 10. This yields compression ratio as good as 18:1 whereas 9:1 ratio is obtained if the whole image is compressed in a lossless fashion. Similarly region 3 is defined as relevant for the second image which is a partitioned skull image. The compression ratio obtained using the novel method was 3.8:1 compared to only 2.3:1 using traditional lossless compression.

This image alignment model is based on a hybrid registration technique that makes use of mutual information maximisation between two images as an initial step, followed by another methodology based on deformable modelling.

4.2 3D Medical Image Compression based on Region of Interest

The method proposed is a complete hybrid coder that uses a motion compensated coder in the overall image and an entropy minimizing, lossless coder for coding the error in the ROI (region of interest) region. The first step of an ROI based system is segmentation. Generally the image is segmented through a sequence of 3-D morphological image processing techniques. Next, motion vectors are coded for each block of the image. Finally, the error between the real image and the motion predicted image is coded for ROI blocks. Compression ratios as high as 40:1 can be achieved using this technique.



Segmentation of ROI

Segmentation algorithm relies on a 3-D extension of *mathematical morphology*, a branch of science that is built upon set theory with many application areas in image processing. It includes generation of mappings for each pixel according to the pixel's local neighborhood. Many researchers have used this technique to segment biomedical images.

ROI Based Compression Scheme

Once the ROI is segmented in each slice, a hybrid compression scheme is used for coding the images. The first slice of the volume is compressed with a lossless coder. Each slice is then coded by motion compensated coding, which also acts as a prediction filter for ROI. Finally, the difference between the real-image ROI block and the predicted-image ROI block is coded by an entropy minimizing lossless coder, e.*g.* Huffman coder.

The efficiency of the method is inversely proportional to the portion of ROI in the image. The smaller the portion of ROI in the image, the better is the resulting compression rate. In addition to its remarkable compression gain, the algorithm is accurate, since there is no degradation of diagnostic quality in ROI.

4.3 Efficient Image Compression of Medical Images Using the Wavelet Transform and Fuzzy c-means Clustering on Regions of Interest.

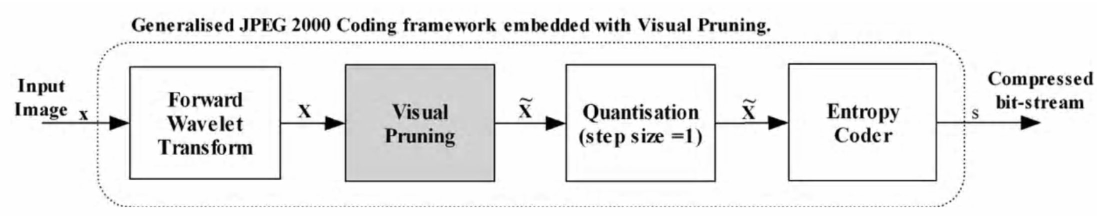
This is a novel image compression scheme, using the discrete wavelet transformation (DWT) and the fuzzy c-means clustering technique. The goal is to achieve higher compression rates by applying different compression thresholds for the wavelet coefficients of each DWT band, in terms of how they are clustered according to their absolute values. This methodology is compared to another one based on preserving texturally important image characteristics, by dividing images into regions of textural significance, employing textural descriptors as criteria and fuzzy clustering methodologies. These descriptors include cooccurrence matrices based measures.

The first compression scheme mentioned above exploits correlation characteristics of the DWT coefficients in order to assign different compression thresholds to the different distributions of correlated coefficients. First, the original image is transformed via the 2-D DWT into bands of wavelet coefficients. For a 1-level such transform 4 bands are obtained. Then, the fuzzy c-means clustering technique is applied to each such band, dividing it into two classes. The result is that we obtain two distributions of correlated coefficients for each band. The one with the larger wavelet coefficients, in terms of the magnitude of their absolute values, is considered as the important region while the other as the non-important one. Then, for each important such region of a wavelet band a lower compression ratio is applied, the same for all important regions of the wavelet domain (and equal to r1), than the one applied for the corresponding non-important regions (equal to r2). Therefore, $r1<r2$.

Concerning the second suggested compression scheme, a good measure related to second order image structure is texture .The rationale underlying the proposed compression methodology is that the significance of image regions varies in space. That is, not all image areas are important in describing the spatial probability distribution of its pixel intensities and subsequently in contributing to the visual effects of the image under consideration. A measure of such an image region significance can be derived by exploiting textural information. When the textural characteristics in an image region assume high values then, it is reasonable to suppose that the textural information content of this area is very important. Therefore, the image spatial probability distribution can be more precisely derived if a larger number of features describing it is extracted for such an area than for other ones. Thus, if a compression methodology keeps a larger number of coefficients in texturally significant regions than in the other regions then, a much better decompressed image can be finally obtained since its probability distribution can be more accurately restored.

4.4 Perceptually lossless Medical Image coding

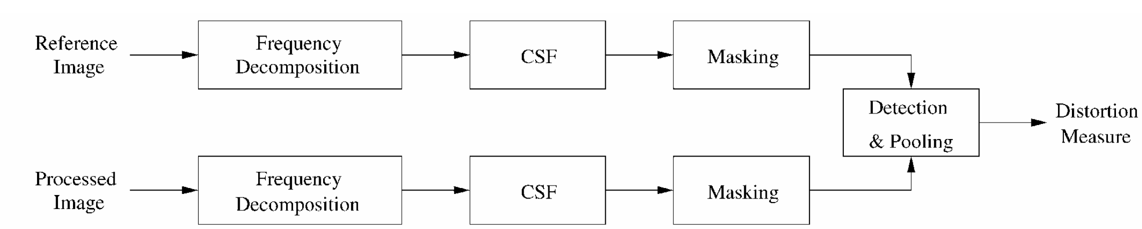
Built on the JPEG 2000 coding framework, the heart of the proposed coder is a visual pruning function, embedded with an advanced human vision model to identify and to remove visually insignificant/irrelevant information.



Vision Modeling

The HVS can be described in three parts . The first part describes the optical characteristics of the human eye with respect to its sensitivity relative to background luminance levels and varying spatio-temporal frequencies. This sensitivity is termed "contrast sensitivity", which is functionally described as the contrast sensitivity function (CSF). The second part is the visual pathway and this provides a link between the eye and the visual cortex. Finally, the third part describes the formation of images within the visual cortex. Neuron interactions in the visual cortex leads to the visual masking phenomenon. Visual masking affects a visual signal by diminishing its visibility when it is within the presence of another visual signal.

The contrast gain control (CGC) coined by Watson and Solomon serves as a vision model template implemented here. This vision model template is a unification of other earlier vision models by Teo and Heeger and by Watson and Solomon . The CGC consists of a linear transform, a masking response and a pooling and detection phase. The CGC takes two inputs, that is, a reference (original) image and a processed image.

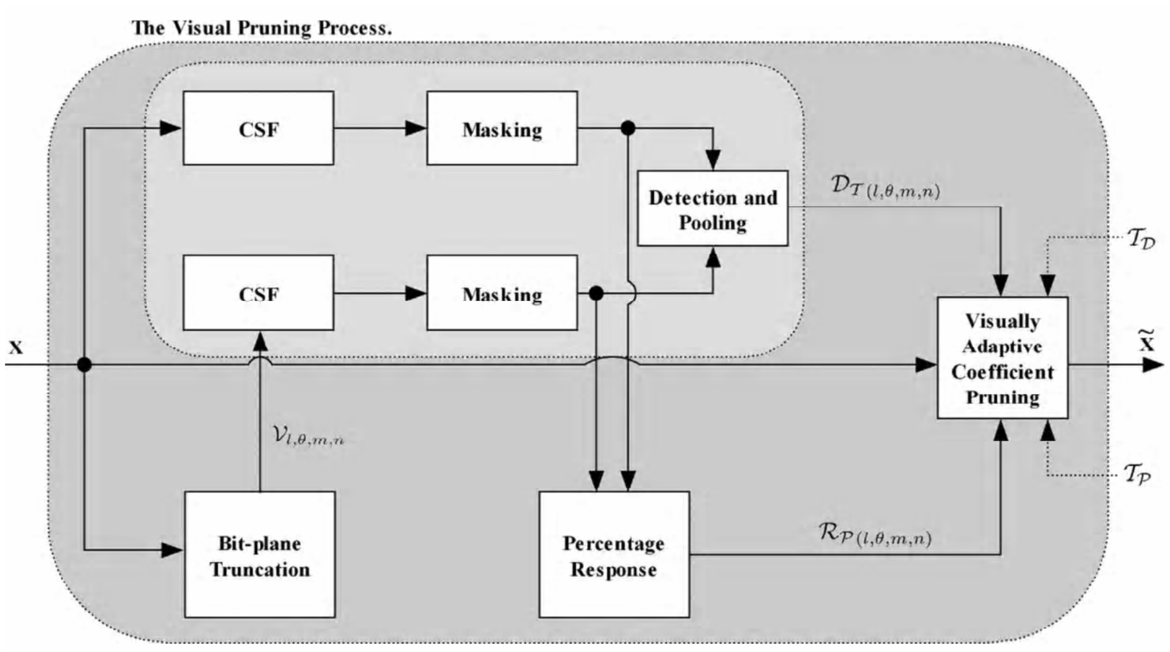


Coder Adaptation – Visual Pruning function

Approach taken here embeds the vision model into a visual pruning (VP) function. This modular approach enables the VP function to be easily adapted into other Wavelet based coding frameworks while maintaining bit-stream compliance. The VP function consists of two stages. For each frequency level, at each orientation, and at a particular location , the first stage takes in a reference coefficient, and generates a set of distorted coefficients. These distorted coefficients are generated through progressive bit-plane truncation from the least significant bit (lsb), upwards. Immediately , each distorted coefficient from the set, is compared with the reference coefficient using the vision model described previously. This generates a set of perceptual distortion measures, and a set of percentage responses. The last stage gathers the set of distortion measures, the set of percentage responses, and performs visually adaptive coefficient pruning. By comparing them to a set of predetermined JNND thresholds Td and Tp, respectively, a coefficient is truncated to a perceptually optimal bit-plane level, only when distortion measure is less than or equal to a JNND threshold, and when percentage response is less than or equal to a percentage response threshold . Thus, all transform coefficients are subjected to this perceptual filtering operation except for those in the isotropic lowpass band (LL). The values in both Td and Tp are derived from subjective experiments.

Parametrization

There are two stages to the parameterization process. The first concerns the vision model parameters, which are subjectively determined by capturing the visual nature of the images governed by the visual mechanics of the observers. The second is a set of visual thresholds Td and Tp, which are mapped to the JNND level for perceptually lossless encoding.



5. Proposed Research in the Future

The proposed research is to develop some novel compression techniques, which can be termed interframe compression and multistage compression.

Interframe compression

We use the fact that for every patient and at every image-taking session, several almost identical images are taken. The approach is to designate one of these images as a baseline image, compute the difference between it and the other images, and then losslessly compress the baseline image and the difference images. Since the difference images contain little data, the resulting compression rate is expected to be over 4.

Multistage compression

In this approach an image is first compressed at a high compression rate but with loss, and the error image is then compressed losslessly. The resulting compression is not only strictly lossless, but also expected to yield a high compression rate, especially if the lossy compression technique is good. This is because the error image will consist of zero- or small-valued elements, thus allowing for lossless compression at a high compression rate.

To evaluate the above compression techniques against traditional compression techniques, there is a need to develop image quality measures and benchmark tests, taking advantage of the contrast-sensitivity threshold of the human vision. The measures and tests will help industry assess the diagnostic quality of images that are reconstructed after compressing them using various techniques. These measures and tests are steps towards standards by which compression algorithms should be evaluated with regard to preservation of diagnostic data in Medical Images.