DIGITAL IMAGING BASICS

Image Compression Techniques for Medical Diagnostic Imaging Systems

Majid Rabbani and Paul W. Jones

As a result of recent technological advances, a significant (and increasing) number of the images in a typical radiology department are represented in digital form. This includes images that have been captured using digital imaging modalities such as computed tomography and magnetic resonance imaging as well as images that have been digitally converted from film originals as in computed radiography. To provide economical storage of such images and to allow for their efficient transmission over various networks, it becomes necessary to consider digital image compression techniques. In this article, the fundamental concepts of several popular reversible and nonreversible image compression schemes are reviewed. In addition, the performance of the schemes in terms of compression ratio and reconstructed image quality as applied to medical diagnostic images is investigated.

Copyright © 1991 by Eastman Kodak Company

KEY WORDS: image compression, predictive coding, transform coding, progressive transmission

I. THE NEED FOR IMAGE COMPRESSION

THANKS TO recent advances in digital image capture and storage technology, the use of digital imaging modalities in radiology has increased dramatically in the past decade. Currently, about 15% of the radiographs taken in a hospital are generated directly in digital form. Prominent examples include computed tomography (CT), magnetic resonance imaging (MRI), ultrasound (US), nuclear medicine (NM), and digital subtraction angiography (DSA). Furthermore, conventional chest or skeletal x-rays, which constitute another 70% of the radiographs, can be converted into digital signals for subsequent processing in computed radiography (CR). The reason for this interest in digital images is clear: representing images in digital form facilitates image transfer as well as

allowing for visual diagnostic information to be manipulated in useful and novel ways.

Despite these advantages, there is one potential problem with digital images, namely, the large number of bits required for their representation. As an example, consider the conversion of a 14 × 17-in chest radiograph into digital form. Assuming the radiologist would examine the image at a typical viewing distance of 14 in, spatial frequencies as high as 7 cycles/mm may need to be captured (based on data for the human visual system contrast sensitivity function).¹ This translates into a sampling rate of 70 μm per picture element or pixel, and the resulting image contains approximately 5000 × 6000 pixels. Since each pixel in a medical image is usually represented by at least 10 bits (corresponding to 1024 brightness levels), the total number of bits for a single image of this type is approximately 300 mega(10⁶) bits.

Currently, magnetic tapes are the most widely used storage media in hospitals, but a typical magnetic tape can only store about 3 images of this size. As a result, some hospitals use optical disks as their primary storage media. A typical setup might consist of 100, 14-in optical disks in a jukebox that can store a total of several tera(10¹²) bits of data. However, the digital

From the Electronic Imaging Research Laboratories, Eastman Kodak Company, Rochester, NY.

Address reprint requests to Dr Majid Rabbani, Electronic Imaging Research Laboratories, Eastman Kodak Company, Rochester, NY 14650-1816.

*Copyright © 1991 by Eastman Kodak Company
0897-1889/91/0402-0003\$03.00/0*

storage of the large number of radiographs that are captured every day in a typical hospital would soon tax even this technology to its limit. Proposed picture archiving and communications systems² (PACS) envision such an all-digital environment, but for PACS to be an economically viable alternative to film, the average number of bits required to represent an image must be reduced by using efficient image compression algorithms without, of course, compromising the diagnostic information.

Image compression can also benefit a PACS environment in the area of image communications. In a local area network, workstations are close to the picture archives, and direct communication using large bandwidth channels is feasible. Newly emerging fiber optics technology (fiber distributed data interchange [FDDI]) provides communication at a bandwidth of 100 Mbits/sec. However, even at an optimistic 20% throughput rate for a shared network, the transmission of the digitized chest image in our example would take 15 sec, whereas in many applications it is desirable to access the image in less than 2 sec. This problem of efficient transmission is further aggravated in wide area networks, which include low-bandwidth channels such as telephone lines for long-distance switched communication. For example, the transmission of a much lower resolution image, such as a 512×512 , 12-bit CT image, would take almost a minute using an Integrated Services Digital Network network with a 64 kilobits/second transmission rate.

II. Types of Image Compression

The discussion in the preceding section clearly illustrates the need for image compression in a PACS environment. There are many approaches to image compression, but they can be categorized into two fundamental groups: reversible and nonreversible.

In reversible compression (also known as bit-preserving or lossless), the reconstructed image after compression is numerically identical to the original image on a pixel-by-pixel basis. Obviously, reversible compression is ideally desired since no information is compromised. However, only modest compression ratios (generally in the range of 1.5:1 to 4.0:1

depending on the original image characteristics) are possible with reversible compression.*

In nonreversible compression (also known as lossy), the reconstructed image contains degradations relative to the original. As a result, much higher compression ratios can be achieved as compared to reversible compression. In general, more compression is obtained at the expense of more distortion. It is important to note that these degradations may or may not be visually apparent. In fact, the term visually lossless has been used to characterize lossy compression schemes that result in no visible loss under normal viewing conditions. Unfortunately, this is a subjective definition and extreme caution should be taken in its interpretation. It is conceivable that an algorithm that is visually lossless under certain viewing conditions, eg, a 19-in video monitor viewed at a distance of 4 ft, could result in visible degradations under more stringent conditions, eg, a 14×17 -in secondary image printed on film. A related term is information preserving, which has been used by the American College of Radiology and National Electrical Manufacturers Association (ACR/NEMA). According to the ACR/NEMA data compression standard report,³ a compression scheme is information preserving if the resulting image retains all of the "significant information" of the original image. The standard does not define what is meant by significant information and has left its definition to the user.

The general framework for any compression scheme is shown in Fig 1. It includes three components: image decomposition or transformation, quantization, and symbol encoding. The relative importance of each component varies from one technique to another, and not all components are necessarily included in a particular technique. The image decomposition or transformation is usually a reversible operation and is performed to reduce the dynamic range of the signal, to eliminate redundant information, or, in general, to provide a representation

*Compression ratio is defined as the ratio of the number of bits used for the original image to the number of bits for the compressed image. Higher compression ratios indicate more efficient compression.

Fig 1. General compression framework.

that is more amenable to efficient coding. The primary difference between reversible and nonreversible techniques is the inclusion of the next stage, quantization, in nonreversible techniques. By quantizing the data, the number of possible output symbols is reduced. The type and degree of quantization has a large impact on the bit rate and quality of a nonreversible scheme. The final stage, symbol encoding, may be as simple as using fixed-length binary code-words to represent the symbols resulting from the decomposition or quantization stages, or it might use a variable-length code as a means of achieving rates close to the fundamental information-theoretic limits.

There is significant debate concerning the use of nonreversible compression schemes for medical diagnostic images. The level of coding distortion that can be accepted for clinical diagnosis purposes has not yet been well-defined, and given the wide variety of resolutions and image characteristics resulting from the different imaging modalities, a thorough clinical evaluation of tolerable distortion for different classes of images is required. As a result, it is often argued that nonreversible compression in any form is not acceptable in a PACS environment as it might compromise diagnostic accuracy. Despite this argument, there are at least two reasons why lossy compression schemes are of interest in medical imaging.

First, it is sometimes desirable to use a

nonreversible compression scheme in conjunction with a reversible scheme. This results in two types of data being stored for each image. The first type corresponds to a relatively low quality image generated by a nonreversible compression scheme operating at a high compression ratio such as 40:1. The second type of data represents the difference on a pixel-by-pixel basis between the lossy reconstructed image and the original image. This difference, or residual, image is compressed by a reversible scheme prior to storage. The memory required to store both the lossy image and the lossless difference image is still less than the original uncompressed image. This approach finds use in a situation where the radiologist is using a low-bandwidth wide area network to access remotely stored pictures of a patient with multiple diagnostic studies and several images per study. In such a case, it may be necessary to browse through a few images before the particular image(s) of interest can be located. The highly compressed image can serve as a guide for the initial selection process, resulting in great savings in time and bandwidth. In many cases, the radiologist might even be able to extract the required information from this highly compressed image. If further information is needed, the difference image can then be transmitted to provide a completely lossless reconstruction. This idea of transmitting partial image information in stages is termed progressive transmission (PT). We will examine this concept in more detail in a later section.

The second argument for using lossy compression is based on the observation that in some cases the degradation caused by the compression is less than the noise introduced by the other system components, eg, film noise, digital scanner noise, etc. In such systems, the use of compression jeopardizes diagnostic accuracy no more than any other component of the system.

In the following sections, several popular compression techniques are briefly reviewed. In section III, a reversible compression algorithm termed lossless differential pulse code modulation (DPCM) is discussed. A lossy compression algorithm based on the discrete cosine transform (DCT) is then discussed in section IV. Finally, in section V, progressive transmission

Fig. 2. DPCM predictor configuration.

techniques such as bit plane (BP) encoding, the reduced-difference pyramid, and the S-transform are examined. The performance of the various schemes is compared using examples from medical diagnostic imaging.

III. REVERSIBLE COMPRESSION TECHNIQUES: LOSSLESS PREDICTIVE CODING

The values of adjacent pixels in an image are often similar, ie, they are highly correlated. This property is exploited in predictive coding techniques where the value of a given pixel is predicted based on the values of the surrounding pixels. Predictive coding can be approached in a number of ways, but the most common is DPCM.⁴ This section will focus on lossless DPCM, but DPCM can also be operated as a lossy technique.

In DPCM, an image is typically encoded one pixel at a time across a raster scan line as shown in Fig 2 for two consecutive raster lines (the scanning direction is from left to right). The pixel to be encoded is labeled x_m , and it is

assumed all pixels scanned prior to x_m (shown in the figure as A, B, C, etc) have already been encoded and transmitted. As a result, both the transmitter and the receiver can access these previous values to form a prediction (estimate) for x_m . Using m previous pixel values (typically, $m \leq 4$), a linear prediction, denoted by $\hat{x}_m$, can be formed as:

$$\hat{x}_m = \sum_{i=0}^{m-1} \alpha_i x_i, \quad \text{Equation (1)}$$

where the α_i 's are predictor coefficients (weighting factors) specified by the user and the x_i 's denote the previous pixels. Although an optimum set of predictor coefficients can be determined for each image, simple fixed predictors such as $\hat{x}_m = A$, $\hat{x}_m = 0.5A + 0.5B$, or $\hat{x}_m = A + B - C$, are often used in practical implementations. The prediction is rounded to the nearest integer and clipped to the range $[0, 2^n - 1]$ for an n -bit image. A differential (error) image, e_m , is then constructed as the difference between the rounded prediction and the actual pixel value, ie,

$$e_m = x_m - \hat{x}_m. \quad \text{Equation (2)}$$

The motivation for this process is that the differential image typically has a greatly reduced variance and significantly less redundancy than the original image and can thus be encoded more efficiently. Referring to the block diagram in Fig 1, the construction of the differential image from the original image corresponds to the first stage, where the image is transformed to make it more suitable for coding.[†]

As an example, consider the 512×512 , 12-bit CT image shown in Fig 3. A portion of the

Fig 3. Original CT image.

[†]In a lossy DPCM scheme, the differential image is quantized to a limited number of values in order to allow for greater compression at the expense of errors in the reconstruction.

Fig 4. Histogram of original CT image.

Fig 5. Histogram of CT differential image.

histogram of the original image is shown in Fig 4, and Fig 5 shows a portion of the histogram of the corresponding differential image using the predictor $\hat{x}_m = 0.5A + 0.5B$. Note that the values in the differential image are represented by 13 bits (-4095 to $+4095$) since each is the difference between two 12-bit values. Unlike the original image histogram, the histogram of the differential image is highly peaked around zero. This is because the prediction is generally close to the actual value resulting in a small predic-

tion error. The error is significant only around edges or areas of sudden change that constitute a small percentage of the total pixels in the image. Different images yield differential image histograms with a similar shape but differing widths. The standard deviation of the CT image in Fig 3 is $\sigma = 525.12$ while the standard deviation of the differential image is only $\sigma_e = 31.47$; the variance is reduced by a factor of 278.

If the differential image is encoded with fixed-length codewords, no compression would

be achieved. In fact, there would be an expansion in the amount of data since each pixel in the differential image needs to be represented by $n + 1$ bits rather than the original n bits. However, by utilizing variable-length codes, the average number of bits required to represent each pixel can be substantially reduced. The idea behind variable-length coding is simple: shorter codewords are assigned to the more probable difference values (symbols), and longer codewords are used to represent the less probable symbols. The net result is a reduction in average bit rate compared to fixed-length codes. Huffman suggested a means of constructing uniquely decodable variable-length codes (known as Huffman codes) with the smallest average codeword length among all possible codebooks.^{5,6}

Table 1 illustrates a portion of the Huffman codebook used to encode the 12-bit CT image in the previous example. The codebook was developed based on the differential image histogram shown in Fig 5. As expected, small differences are coded with codewords much shorter than 13 bits, whereas differences with a magnitude larger than 128 (which comprise less than 1% of the total number of difference values) are coded with 20 bits (consisting of a 7-bit prefix plus 13 bits for the actual difference value).[‡]

Using this code on the CT differential image results in an average bit rate of 4.90 bits/pixel, which corresponds to a compression ratio of 2.45:1. This experiment was repeated for a set of 24 CT images with the same resolution and bit depth. In each case, a local Huffman codebook, ie, a codebook matched to the statistics of the differential image being encoded, was used. The compression ratios varied from 1.45:1 to 3.23:1, with an average of 2.13:1. Additionally, a set of 24, 512×512 , 12-bit MRI images was encoded using local Huffman codes. The average compression ratio for the set was 2.00:1. This is slightly lower than the CT results because of higher noise levels in the MRI images. In general, noise, due to its unpredictable nature, can significantly affect the compression ratio in any lossless scheme.

The local Huffman codebooks used in the preceding example are optimal in that they are

Table 1. DPCM Huffman Codewords for the CT Image

Difference Value	Probability	Bits	Codeword
-5	0.0314	5	01110
-4	0.0361	5	11001
-3	0.0404	5	11101
-2	0.0447	4	0000
-1	0.0474	4	0011
0	0.2710	2	10
1	0.0466	4	0010
2	0.0430	5	11111
3	0.0395	5	11100
4	0.0340	5	01111
5	0.0300	5	01100
Mag > 128	0.0088	20	1100001 (prefix) + 13 bits

designed specifically for each image, but unfortunately a local codebook requires two passes through the image: an initial pass to calculate the differential image histogram and construct the codebook, and a second pass to encode the difference values. One way of eliminating the need for a two-pass algorithm is to construct a global Huffman codebook, ie, a codebook based on an average histogram. For a set of similar images, eg, all CT images or all MRI images, the differential image histogram is fairly stable and a global Huffman code usually results only in a small loss in compression efficiency.

To assess the effect of global Huffman coding on the compression ratio, the following experiment was performed. First, a global Huffman code was developed based on the average of the 24 CT images and was used to encode that set. Then, a global Huffman code based on the average of the 24 MRI images was used to encode the MRI set. Finally, a global codebook based on the average of all 48 images was used to encode the complete set. The average compression ratios for these three cases are summarized in Table 2, along with the results for the local Huffman codebooks. As expected, switching from local codebooks to a global codebook (either CT only or MRI only) results in a slight loss in compression efficiency. Interestingly, the use of a single CT + MRI global codebook does almost as well as the separate CT or MRI global codebooks. This confirms the utility of global Huffman coding in situations where the extra complexity associated with local Huffman coding is not justified.

[‡]This codebook is actually a modified Huffman code.⁷

Table 2. Lossless Coding Compression Ratios

Technique	CT	MRI
DPCM		
Local Huffman Code	2.13	2.00
Global Huffman code (CT)	2.01	—
Global Huffman code (MRI)	—	1.95
Global Huffman code (CT + MRI)	1.97	1.94
Bit plane encoding (arithmetic coding)	2.01	1.72
Reduced-difference pyramid (local Huffman code)	1.93	1.74
S-transform pyramid (local Huffman code)	1.99	1.80

IV NONREVERSIBLE COMPRESSION TECHNIQUES: DISCRETE COSINE TRANSFORM

A general transform coding scheme involves subdividing an $N \times N$ image into smaller $n \times n$ (eg, $n = 8$ or 16) blocks and performing a reversible orthogonal transformation on each block.⁸ For each $n \times n$ block, the transformation produces an $n \times n$ set of transform coefficients. The goal of this process is to decorrelate the original signal, resulting in the energy being redistributed among only a small set of transform coefficients. In this way, many coefficients can be discarded after quantization and prior to encoding. A block diagram of a basic transform coding scheme is shown in Fig 6.

A transformation can be viewed as a decomposition of the original image into a set of basis images. In the case of the Fourier transform (or the general family of sinusoidal transforms), the basis images consist of sines and/or cosines with different spatial frequencies, and each transform coefficient is proportional to the fraction of energy in the original image that corresponds to that particular frequency. As a result, it is natural to also interpret the transform as a spectral or frequency decomposition of the original image.

A practical transform for the purpose of image compression should have a strong decorrelating effect on typical imagery, should preferably consist of image-independent basis functions, and should have a fast implementation. Since its introduction in 1974, the discrete cosine transform (DCT) has by far become the most popular transform for image compression applications.⁹ The forward two-dimensional (2D) DCT of an $n \times n$ block of pixels is defined as^{10,11}:

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

Equation (3)

where $f(j, k)$ represents the original pixel values in the block and $F(u, v)$ represents the transform coefficients at spatial frequencies u and v . To recover the original pixel values from the transform coefficients, the inverse 2D DCT is used:

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

Equation (4)

where

$$C(w) = \begin{cases} \frac{1}{\sqrt{2}} & \text{for } w = 0 \\ 1 & \text{for } w = 1, 2, \dots, n-1. \end{cases}$$

Equation (5)

Transmitter

Receiver

Fig 6. Transform coding block diagram.

The DCT is separable, ie, the transform kernel can be decomposed into two one-dimensional (1D) kernels specifying separate horizontal and vertical operations. Thus, an $n \times n$ block of pixels can be transformed by first performing a 1D transform along each row of the block followed by another 1D transform along each column of the resulting output.

The effect of DCT on a typical image block is best illustrated through an example. Consider the following 8×8 block of pixel values taken from the 12-bit CT image in Fig 3.

$$f(j, k) = \begin{bmatrix} 743 & 966 & 1059 & 1041 & 984 & 943 & 922 & 923 \\ 932 & 1043 & 1037 & 994 & 947 & 922 & 915 & 941 \\ 1039 & 1032 & 990 & 948 & 926 & 924 & 943 & 977 \\ 1033 & 989 & 937 & 913 & 918 & 941 & 979 & 1016 \\ 998 & 945 & 919 & 910 & 929 & 966 & 1004 & 1034 \\ 970 & 932 & 927 & 939 & 969 & 1008 & 1035 & 1047 \\ 940 & 925 & 934 & 964 & 1001 & 1022 & 1047 & 1054 \\ 953 & 945 & 967 & 999 & 1030 & 1044 & 1052 & 1050 \end{bmatrix}.$$

The transformed block, using Equation (3), is given by:

$$F(u, v) = \begin{bmatrix} 1946.13 & -26.63 & 7.96 & -13.96 & -5.38 & -4.15 & 2.72 & -2.80 \\ -23.19 & 48.32 & -13.33 & -25.10 & -11.26 & -9.33 & -1.30 & -1.41 \\ 11.98 & -24.35 & -55.82 & -20.50 & -8.98 & -3.89 & -2.98 & -0.74 \\ -15.01 & -34.94 & -22.40 & -8.48 & -3.87 & -1.15 & 0.53 & -0.66 \\ -4.75 & -15.16 & -11.67 & -6.01 & -1.75 & 0.03 & -0.24 & 0.69 \\ -5.31 & -10.72 & -5.42 & -0.36 & 0.42 & 0.43 & 1.57 & 0.75 \\ -1.45 & -3.93 & -3.48 & -0.19 & 0.20 & -0.83 & 1.19 & 0.47 \\ -0.97 & -3.61 & -1.32 & -0.24 & 0.31 & 0.65 & 0.46 & 0.73 \end{bmatrix}.$$

Because of the scaling used in Equation (3), the top-left transform coefficient is twice the average brightness (DC value) of the original image block. The horizontal frequency increases from left to right, and the vertical frequency increases from top to bottom. It can be seen that the majority of the block energy is concentrated in the low frequency coefficients. This result is typical since most image blocks contain only small variations, corresponding to the lower spatial frequencies.

It is important to realize that the transform operation by itself does not achieve any compression; it only makes the data more suitable for compression. It is the subsequent quantization

of the transform coefficients and their efficient encoding that results in compression. The level of compression usually depends on the degree of quantization performed; coarser quantization results in higher compression but also introduces more degradation. In general, reconstruction quality can be traded for bit rate through the adjustment of the quantizer.

There are various strategies for the selection, quantization, and encoding of the transform coefficients.¹² In one scheme, known as zonal sampling, only those transform coefficients that are located in a prespecified zone in the transform block (usually the low frequencies) are retained and all the other coefficients are set to zero. Each retained coefficient is then quantized and encoded with a fixed-length codeword. The structure of the quantizer and the number of bits used to encode each quantized coefficient can be a function of the location of the coefficient within the block. For example, a few lowest frequency coefficients may be encoded with 7 or 8 bits, while the other retained coefficients are encoded with less bits. An advantage of the zonal sampling scheme is that it has a fixed bit rate (and compression ratio) for all images, a desirable feature in many applications.

A major drawback of fixed-rate schemes is that the reconstructed images generally do not exhibit the same subjective quality. This is because the frequency content varies depending on the level of detail and noise present in an image. While compressing every image by the same ratio might result in excellent reconstruction quality for some images, it may prove inadequate for others. For example, in the zonal sample selection strategy, if a certain block contains a lot of high-frequency detail, the coefficients outside the retention zone would contain significant energy and their omission results in a large reconstruction error. To assure a constant reconstruction quality, the compression ratio should be allowed to vary for different images. A means of achieving this in a DCT-based scheme is to include all the coefficients in the quantization process. A different quantizer can be used for each coefficient in order to control the amount of the distortion for that coefficient. A motivation for this is the varying perceptual importance of each coefficient as a

function of frequency. Usually, a finer quantizer (producing less distortion) is used to quantize the low frequency coefficients, while a coarser quantizer can be used for the higher frequencies. To simplify the implementation, the various quantizers can be designed so that they only differ by a scaling factor. A single quantizer can then be used for all the coefficients provided that the coefficients are scaled (normalized) prior to quantization to achieve the desired step size. At the decompression stage, each quantized coefficient is denormalized prior to the inverse DCT operation.

Following our example, consider a simple implementation in which each coefficient is divided by a constant scaling (normalization) factor of $D = 16$ and then uniformly quantized by being rounded to the nearest integer. This results in the following array:

$$F^*(u, v) = \begin{bmatrix} 122 & -2 & 0 & -1 & 0 & 0 & 0 & 0 \\ -1 & 3 & -1 & -2 & -1 & -1 & 0 & 0 \\ 1 & -2 & -3 & -1 & -1 & 0 & 0 & 0 \\ -1 & -2 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

As expected, the normalization and quantization process produces many zero-valued coefficients. Unlike the zonal sample selection technique, the position of the nonzero quantized coefficients in each block is not known beforehand, and this information, along with the actual amplitude of each nonzero coefficient, has to be encoded. One popular approach is to use a variable-length code to encode the amplitude of each nonzero coefficient in combination with its relative position from the previously encoded nonzero coefficient. For more details on this technique, the reader is referred to the draft proposal on the world standards for the compression of still, continuous-tone, color images.¹³

To continue with our example, the block of pixels is reconstructed by first denormalizing

the coefficients (each nonzero coefficient is multiplied by $D = 16$), and then applying the inverse DCT. The error introduced in the values of the original image block due to compression is given by

$$e(j, k) = \begin{bmatrix} -2 & -6 & -15 & -3 & -5 & -9 & -26 & -29 \\ 29 & 13 & -17 & -2 & -8 & 0 & 1 & 1 \\ 10 & -9 & -8 & -8 & -11 & 2 & 1 & 6 \\ 18 & 11 & 6 & -22 & -33 & -6 & 1 & 1 \\ 3 & 9 & 1 & -27 & -18 & 23 & 26 & 18 \\ -2 & 13 & 5 & -11 & 11 & 38 & 7 & -2 \\ 1 & 20 & 16 & 1 & 8 & 6 & -19 & -12 \\ -7 & 29 & 20 & -1 & -7 & -23 & -21 & 21 \end{bmatrix}.$$

where $e(j, k) = f(j, k) - \hat{f}(j, k)$, and $\hat{f}(j, k)$ denotes the reconstructed image. Figure 7 shows the 12-bit CT image of Fig 3, which has been compressed by a factor of 25:1 using the above DCT scheme. The errors resulting from the compression process are illustrated in Fig 8. The error image, g , is generated using

$$g(i, j) = 8[f(i, j) - \hat{f}(i, j)] + 2048 \quad (\text{Equation 6})$$

where the factor of 8 is included to make the errors more visible, and the constant value of

Fig 7. CT image compressed with DCT at 25:1.

Fig 8. Error image for CT image compressed with DCT at 25:1.

2048 has been added to make the errors positive and suitable for printing.

The subjective differences between the original image and this reconstructed image are very subtle even when viewed under stringent viewing conditions. In general, there exists no suitable quantitative measure, which relates well to subjective quality. A commonly used measure is the root-mean-squared error (RMSE) which is defined as

$$\text{RMSE} = \sqrt{\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N [f(i, j) - \hat{f}(i, j)]^2},$$

(Equation 7)

and represents the standard deviation of the error image. The RMSE corresponding to the reconstructed image of Fig 7 is 8.20 and is insignificant compared to the dynamic range of the original image (0 to 4095). For some applications, the CT image in our example could be compressed to an even greater extent (in excess of 50:1) while still maintaining good recognizability. Such compression ratios can be achieved by increasing the normalization value D , which controls the quantizer step size.

V PROGRESSIVE TRANSMISSION TECHNIQUES

In PT, partial image information is transmitted in stages, and at each stage, an approxima-

tion to the original image is reconstructed at the receiver. The reconstructed images progressively improve as more information is received. The transmission can be stopped if an intermediate version of the image is satisfactory or if the image is found to be of no interest. PT is motivated by the need to transmit images over low bandwidth channels (relative to the amount of data), eg, telephone lines, particularly in those cases where quick recognition is important or total transmission time may be limited. An example of this scenario is telebrowsing, where a remote image data base is searched for a particular image or images. Of course, conventional compression algorithms such as DPCM will decrease the transmission time, but they generally require the majority of the image to be reconstructed to achieve recognition and permit only a single reconstructed image of fixed quality.

An important aspect of PT is the need for image hierarchies, ie, a way to organize the image data in order of importance. Each level of the hierarchy corresponds to a reconstructed image at a particular resolution or level of quality. It is convenient to classify image hierarchies into fixed-resolution hierarchies or variable-resolution hierarchies. In fixed-resolution hierarchies, the reconstructed image is inherently the same size as the original image, and the value at any particular pixel location is refined as one moves from level to level. In variable-resolution hierarchies, the images corresponding to the levels of the hierarchy vary in spatial resolution. This approach naturally results in a pyramid structure, where the base of the pyramid represents the full resolution image, ie, the original image, and as one moves up the pyramid, the images decrease in spatial resolution and size. For example, if the levels of the hierarchy are formed by repeatedly averaging over 2×2 blocks, then the pyramid representation would appear as in Fig 9. For display purposes in a PT scheme, it may be desirable to interpolate or replicate the reduced size images to a larger size for improved recognizability. In this section three PT schemes are considered: bit plane encoding, which is a fixed-resolution PT scheme, and the S-transform and reduced-difference pyramids, which represent variable-resolution hierarchies.

Fig 9. Image pyramid structure.

Fixed-Resolution Hierarchies: Bit Plane Encoding

Consider an $N \times N$ image in which each pixel value is represented by k bits. By selecting a single bit from the same position in the binary representation of each pixel value, an $N \times N$ binary image (consisting of the values '0' or '1') called a bit plane can be formed.¹⁴ For example, we can select the most significant bit (MSB) of each pixel value to generate an $N \times N$ binary image representing the MSB plane. Repeating this process for the other bit positions, the original image can be decomposed into a set of k , $N \times N$ bit planes (numbered 0 for the least significant bit (LSB) plane through $k - 1$ for the MSB plane). Progressive transmission can be achieved by transmitting the bit planes in a sequence, starting with the MSB plane and ending with the LSB plane. The image reconstructed from the MSB plane is a binary image, and additional gray levels are added as more bit planes are received. A lossless reconstruction is possible if all bit planes are used. Since each bit plane is a binary image, lossless binary compression techniques can be applied to reduce the bit rate.¹⁵ The amount of compression is usually largest for the MSB and decreases as one moves through the bit plane sequence. This property allows for significant compression if the transmission is terminated after receiving just the first few bit planes.

Bit plane encoding algorithms typically take advantage of the existence of large uniform areas in each plane to achieve high compression. The simplest method of encoding binary data with frequent clusters of 0's and 1's is runlength coding.¹⁶ Both 1D and 2D runlength coding techniques are routinely used for the encoding of binary text and documents in the field of facsimile transmission.¹⁷ In runlength coding, a new message set is constructed based

on the runs of 0's and 1's. A run is merely the number of times a value (either 0 or 1) occurs consecutively along a given direction, typically along a scan line. Variable-length codes, such as global Huffman codes, can be used to encode the various runlengths. In our experiments, we used a more elaborate technique based on adaptive arithmetic coding, which is part of the proposed world standards for the compression of bilevel images.¹⁸ Although the details of the technique are beyond the scope of this article, it is representative of current technology available for the compression of binary images.

The results of bit plane encoding the set of CT and MRI test images are summarized in Table 3. In the absence of any compression, each bit plane would be represented by 1 bit/pixel. The table entries represent averages over the 24 images within each set. As expected, the first few significant bit planes are highly compressible. In fact, it takes an average of only 1.18 bits/pixel to encode the first 6 MSB planes of a CT image. This is a better than a 5:1 compression ratio with reference to the 6 bits that it takes to represent those planes in uncompressed form. For the set of MRI images, the corresponding average bit rate is even less at 1.05 bits/pixel. These results are important since the first few MSB planes usually contain the majority of the structural information of the image. Figure 10 is a good illustration of this claim. It shows the CT image of Fig 3 reconstructed from the first three MSB bit planes (at 0.18 bit/pixel), and Fig 11 shows a reconstruction from the first six MSB planes (at 0.80 bit/pixel). Finally, the last entry of Table 3

Table 3. Bit Plane Results for the CT and MRI Images

Bit Plane Number	CT Average Bits/Pixel	MRI Average Bits/Pixel
BP 11 (MSB)	0.010	0.005
10	0.092	0.025
9	0.127	0.077
8	0.189	0.148
7	0.289	0.263
6	0.472	0.532
5	0.674	0.925
4	0.766	1.000
3	0.818	1.000
2	0.840	1.000
1	0.843	1.000
0 (LSB)	0.844	1.000
Total	5.964	6.975

Fig 10. First three bit plane of CT image.

shows the resulting bit rates when all the bit planes are summed to obtain a lossless reconstruction of the image. The corresponding compression ratios are summarized in Table 2. Although the bit plane encoding compression ratios may not be as high as lossless DPCM, the differences are usually insignificant and the added PT capability makes bit plane encoding worthy of consideration.

Variable-Resolution Hierarchies

In its simplest form, a variable-resolution hierarchy can be generated by repeatedly sub-

Fig 11. First six bit planes of CT image.

sampling the original image data. For example, the original $N \times N$ image could be subsampled by a factor of two in both dimensions, and then the resulting $N/2 \times N/2$ image could again be subsampled, and so on, until an image consisting of a single pixel is obtained. Unfortunately, subsampling can introduce aliasing that becomes more pronounced at the higher levels of the pyramid, but more importantly, the pyramid created by subsampling cannot be compressed efficiently. Forming a mean pyramid by repeatedly averaging over 2×2 blocks alleviates the aliasing problem to some extent since the averaging process acts as a lowpass filter. However, the problem of efficient storage of the pyramid still remains.

Referring to Fig 9, each pixel value x at a given level in the mean pyramid needs to be resolved into four new values (x_1, x_2, x_3 , and x_4) in order to reconstruct the image at the next level of the pyramid. The reconstruction process at each level thus utilizes the pixel values from the previous level in conjunction with any additional information needed to determine the new values. There are several techniques that achieve compression by the efficient characterization and storage of this additional information.

In one technique, known as the reduced-difference pyramid,¹⁹ the pyramid is constructed by computing the following four values at each level:

$$\begin{aligned} x &= 1/4(x_1 + x_2 + x_3 + x_4), \\ d_1 &= (x_1 - x_2), \\ d_2 &= (x_2 - x_3), \\ d_3 &= (x_3 - x_4). \end{aligned} \quad (\text{Equation 8})$$

The value of x is simply the average of the pixel values in the 2×2 block and is used to form the image at the next level of the pyramid. An example of a mean pyramid formed by this process is shown in Fig 12 for the CT image used previously. The d_i 's represent difference values between adjacent pixels and are stored as the additional information needed to reconstruct the x_i values from x . It can be shown that at each reconstruction stage, the x_i values can be losslessly reconstructed if k bits are used to represent x and $k + 1$ bits are used to represent the d_i 's, where k is number of bits/pixel in the original image. If fixed-length codes are used to

Fig 12. Mean pyramid for CT image.

encode the d_i 's, the storage of the pyramid would result in a slight expansion compared to the original image. However, as in lossless DPCM, the histogram of the difference values is highly peaked around zero and variable-length coding can be used to advantage. For example, applying a local Huffman code to the difference values for the set of 24 CT images results in an average compression ratio of 1.93:1. Similarly, applying a local Huffman code to the difference values for the set of 24 MRI images results in an average compression ratio of 1.74:1. These results are summarized in Table 2. The effect of using a global Huffman code rather than local Huffman codes on the difference values can be extrapolated from the DPCM results in Table 2 since the encoded values have similar characteristics in both cases.

Another approach that also yields a mean pyramid is the S-transform.^{3,20,21} The following equations are used to construct the pyramid at each stage:

$$x = \frac{1}{4}(x_1 + x_2 + x_3 + x_4),$$

$$d_1 = \frac{1}{2}(x_1 + x_2 - x_3 - x_4),$$

$$d_2 = \frac{1}{2}(x_1 - x_2 - x_3 + x_4),$$

$$d_3 = (x_1 - x_2 + x_3 - x_4).$$

Equation (9)

As before, x is the mean over a 2×2 block and is used to form the image at the next level of the pyramid. The d_i difference values represent the additional information needed to resolve x into x_1, x_2, x_3 , and x_4 . It can be shown that by retaining only k bits for x , $k + 1$ bits for d_1 and d_2 , and $k + 2$ bits for d_3 , a lossless reconstruction of the x_i values can be obtained. The S-transform is slightly more efficient than the reduced-difference pyramid at the expense of an increased number of computations. Encoding the difference values with local Huffman codebooks yields average compression ratios of 1.99:1 for the 24 CT images and 1.80:1 for the 24 MRI images. These results are summarized in Table 2. Again, the impact of using global Huffman codes on the difference values can be extrapolated.

From Table 2, it can be seen that both the reduced-difference pyramid and the S-transform pyramid provide compression ratios that are competitive with those achieved from bit plane encoding. As a result, the choice of a particular progressive transmission technique can be made on the basis of implementation ease and usefulness of intermediate images in the progressive reconstruction.

CONCLUSION

The field of digital image compression encompasses a wide range of techniques in order to achieve the goal of reducing the number of bits required to represent an image. These techniques differ in reconstructed image quality, complexity, compression efficiency, and special features such as progressive transmission. In this article, some of the basic concepts and techniques that are particularly useful for the

compression of medical diagnostic images have been examined. For more information on these and other compression techniques, the reader is referred to the list of review articles and books presented at the end of this paper.

ACKNOWLEDGMENT

The authors would like to acknowledge K. Vizy, J. Birkelund, S. Daly, P. Melnychuck, and E. Harding, all of Eastman Kodak Company, Rochester, NY, for their valuable contributions to this manuscript.

REFERENCES

1. Barten PGJ: The SQRI method: A new method for the evaluation of visible resolution on a display. *Proc Society for Information Display* 28(3):253-262, 1987
2. Duerinckx AJ (ed): *Proc Society of Photo-optical Instrumentation Engineers Picture Archiving and Communication Systems (PACS) for Medical Applications*, 318, 1982
3. American College of Radiology/National Electrical Manufacturers Association Standards Publication for Data Compression Standards, NEMA Publication PS-2, Washington, DC, 1989
4. Musmann HG: Predictive image coding, in Pratt WK (ed): *Image Transmission Techniques*, *Adv Electronics Electron Phys* (suppl 12). Orlando, FL, Academic Press, 1979, pp 73-112
5. Huffman DA: A method for the construction of minimum redundancy codes. *Proc Institute of Radio Engineers* 40:1098-1101, 1952
6. Abramson N: *Information Theory and Coding*. New York, NY, McGraw-Hill, 1963, pp 77-81
7. Hankamer M: A modified Huffman procedure with reduced memory requirement. *Institute of Electrical and Electronics Engineers Trans Commun COM-27*: 1979, pp 930-932
8. Clarke RJ: *Transform Coding of Images*. London, Academic Press, 1985
9. Ahmed N, Natarajan T, Rao KR: Discrete cosine transform. *Institute of Electrical and Electronics Engineers Trans Computers C-23*(1): 1974, pp 90-93
10. Chen WH, Smith CH: Adaptive coding of monochrome and color images. *Institute of Electrical and Electronics Engineers Trans Commun COM-25*(11): 1977, pp 1285-1292
11. Chen WH, Pratt WK: Scene adaptive coder. *Institute of Electrical and Electronics Engineers Trans Commun COM-32*(3): 1984, pp 224-232
12. Habibi A: Survey of adaptive image coding techniques. *Institute of Electrical and Electronics Engineers Trans Commun COM-25*(11): 1977, pp 1275-1284
13. Pennebaker W: JPEG Technical Specification, Revision 8. Working Document No. JTC1/SC2/WG10/JPEG-8-R8, Aug 1990
14. Schwartz JW, Barker RC: Bit-plane encoding: A technique for source encoding. *Institute of Electrical and Electronics Engineers Trans Aerospace Elec Systems AES-2*(4): 1966, pp 385-392
15. Kunt M, Johnsen O: Block coding of graphics: A tutorial review. *Proc Institute of Electrical and Electronics Engineers* 68(11): 1980, pp 770-786
16. Huang TS: Coding of two-tone images. *Institute of Electrical and Electronics Engineers Trans Commun COM-25*(11): 1977, pp 1405-1424
17. Hunter R, Robinson AH: International digital facsimile standards. *Proc Institute of Electrical and Electronics Engineers* 68(11): 1980, pp 854-867
18. Langdon Jr GG, Rissanen J: Compression of black-white images with arithmetic coding. *Institute of Electrical and Electronics Engineers Trans Commun COM-29*(6): 1981, pp 858-867
19. Wang L, Goldberg M: Reduced-difference pyramid: A data structure for progressive image transmission. *Opt Eng* 28(7):708-716, 1989
20. Lux P: Redundancy reduction in radiographic pictures. *Opt Acta* 24(4):349-365, 1977
21. Blume H, Fand A: Reversible and irreversible image data compression using the S-transform and Lempel-Ziv coding. *Proc Society of Photo-optical Instrumentation Engineers Medical Imaging III Image Capture and Display*, 1091, 1989, pp 2-18

REVIEW ARTICLES AND BOOKS

1. Rabbani M, Jones PW: *Digital Image Compression Techniques*. Society of Photo-optical Instrumentation Engineers Tutorial Series, Vol TT7. Bellingham, WA, 1991
2. Jain AK: *Fundamentals of Digital Image Processing*. Englewood Cliffs, NJ, Prentice-Hall, 1989
3. Netravali AN, Haskell BG: *Digital Pictures: Representation and Compression*. New York, NY, Plenum Press, 1988
4. Jayant NS, Noll P: *Digital Coding of Waveforms*. Englewood Cliffs, NJ, Prentice-Hall, 1984
5. Rosenfeld A, Kak AC: *Digital Picture Processing*, vol 1. Orlando, FL, Academic Press, 1982
6. Pratt WK: *Digital Image Processing*. New York, NY, Wiley-Interscience, 1978
7. Gonzalez RC, Wintz PA: *Digital Image Processing*. Reading, MA, Addison-Wesley, 1977
8. Roos P, Viergever A, van Dijke MCA, et al: Reversible intraframe compression of medical images. *Institute of Electrical and Electronics Engineers Trans Medical Imag* 7(4):328-336, 1988
9. Jain AK: Image data compression: A review. *Proc Institute of Electrical and Electronics Engineers* 69(3):349-389, 1981
10. Netravali AN, Limb JO: Picture coding: A review. *Proc Institute of Electrical and Electronics Engineers* 68(3): 366-406, 1980