

# Large Scale Implementation of Optimal Image compression Algorithms

Helen Chou, Feng Chen, Daniel Valentino\*, Lu Huang\*, John Villasenor

Department of Electrical Engineering  
\*Department of Radiological Sciences  
University of California, Los Angeles,  
Los Angeles, California, 90095

## ABSTRACT

Despite over a decade of research and development, medical image compression has not yet been widely implemented on clinical picture archiving and communication systems (PACS). We have developed a prototype interface which incorporates both lossless and lossy compression into a browsing system that enables the efficient use of network and storage resources. Such a system allows an user to quickly browse through a large set of image icons created from lossy compression and selectively retrieve the original images for diagnosis from the optical disk that contains losslessly compressed image data. For lossless compression, we implemented modality specific techniques which combines preprocessing, adaptive prediction and entropy coding, giving a compression improvement of 20% over JPEG predictors. The lossy compression algorithm consists of subsampling followed by wavelet transform coding and achieves compressed CR images of sufficient quality for browsing at a compression ratio of about 2000:1.

## 1.INTRODUCTION

### 1.1 UCLA PACS and radiology

A significant trend in the practice of radiology is the increasing use of high volume imaging modalities such as spiral computed tomography (spiral CT), echo-planar magnetic resonance imaging (fast MRI), high resolution computed radiography (CR), digital subtraction angiography (DSA), and potentially digital mammography.

The UCLA Picture Archiving and Communication System currently serves a large portion of diagnostic radiology, three intensive care units, and several clinical sections. All CR, CT, and MR studies are acquired, archived, and automatically distributed to diagnostic and review workstations. Over 1 terabyte of image data was archived in 1994 alone, and the total image data archived to date exceeds 3.5 terabytes. On average, 500 megabytes of image data are distributed per day to each workstation, and some high volume sections of archive receive over 1.5 gigabytes per day. As we move to a fully digital radiology department, the volume of archived data will increase rapidly. Using the current trends in utilization of imaging modalities and number of procedures performed at UCLA, we estimate that the department will generate over 5 terabytes of image data per year (excluding angiography). With such high volume of data, two main problems can impede the performance of PACS: 1) the data acquired will be difficult to archive efficiently and cost effectively, and 2) reviewing images from the vast archive can be extremely time consuming and difficult through the existing PACS network.

Therefore, we have implemented a prototype image browsing system on the UCLA PACS which combines two data compression schemes and a quick search icon database to save network and storage resources. Such a system will greatly facilitate the clinicians in performing various tasks related to medical images. The compression algorithms that we have used are in general well known, and include adaptive linear predictive coding for the lossless coding and subband decompositions for lossy coding.

## 2. METHODS

### 2.1 Lossless compression

Lossless coding of medical image data is performed using a two-stage approach. In the preprocessing stage, the outer edges of the digital image template are zeroed. In the compression stage, linear predictive coding is performed using a recursively computable predictor similar to that used in the JPEG lossless coding standard. The preprocessing stage is modality-specific, and improves the compression by eliminating data that are certain to contain no diagnostically relevant information. For example, the standard digital CR template of 2048 x 2048 pixels usually contains a border along the outer edge into which no image data is mapped during the digitization process. The size of this region varies with study, but typically comprises 10 - 15% of the total template area and can be easily identified using simple automatic search techniques. CT images have circular support but are digitized to a 512 x 512 template, leaving about 20% of the template unused. Often the data values in these "black" areas are not identically zero because of small perturbations associated with noise in the imaging or reconstruction process. Setting the data in these regions to zero aids significantly in the compression process by improving the performance of the entropy coding that is performed in subsequent processing.

After preprocessing is complete, the image is compressed using a block-based adaptive lossless linear prediction scheme. Since medical images are highly nonstationary, the image is first partitioned into blocks of size  $M \times M$ , with the prediction coefficients then determined locally based on the image statistics within each block. The prediction is based on the standard autoregressive random field model with a DC bias at the output, meaning that the value of each pixel is modelled as a linear combination of neighboring pixels as described in general form by the equation:

$$y(m, n) = \sum_k \sum_l a(k, l) y(m - k, n - l) + a_0 + e(m, n)$$

where  $a(k, l)$  is an array of 2D prediction coefficients,  $a_0$  is the constant bias for each block of the image, and  $e(m, n)$  is the prediction error sequence. The region of support of the coefficients  $a(k, l)$  determines the recursive computability and the computational complexity of the predictor. For each  $M \times M$  block, the values of  $a(k, l)$  are obtained by solving the set of normal equations constructed using values of the locally measured 2D covariance function.

One of the fundamental tradeoffs of this approach is the size of the region  $M \times M$ . Clearly, if  $M$  is too small, the savings in bandwidth due to better prediction enabled by a more highly localized model will be offset by the increased bandwidth needed to transmit the prediction coefficients for each of the image blocks. In practice, one wants to use the maximum

possible region  $M \times M$  over which the image data can be assumed to be approximately stationary. This approach of viewing a fundamentally non-stationary random field as locally stationary has been applied previously in the field of speech processing, where researchers have discussed the concept of quasi-stationarity<sup>1</sup>. A general, higher dimensional formulation of quasi-stationarity appropriate to images also exists<sup>3</sup>. A quasi-stationary two-dimensional random field (i.e. image) is one in which the statistical characteristics may vary with location, but at every location there exists a locally appropriate, stationary covariance function. The size of the region over which this covariance function approximates (to within some epsilon) the true covariance is a measure of the interval of stationarity, or correlation length.

The correlation length for medical images of course depends strongly on the modality concerned. To investigate this further, we analyzed medical images in the context of non-stationary random fields, leading to the following conclusions: First, medical images are quasi-stationary, but with correlation lengths that vary strongly with location. For example, many of the larger structures in chest studies are very homogenous, and a single covariance model is accurate to better than 10% over regions encompassing many hundreds of pixels. Interpixel correlations in these areas are consistently over 0.9. As might be expected, structures that are small in size and show greater detail have much smaller correlation lengths (under ten pixels); interpixel correlations in these portions of the image are often below 0.5. Consistent with their higher resolution, CT and MR display more rapidly varying statistics than do lower resolution modalities such as PET and CR. The second conclusion is that the assumption of isotropicity within an image plane is in generally reasonable. For a particular location of a particular image, of course, strong angular dependence can occur, but on average the locally appropriate covariance functions show no directional preference.

In the interest of implementational simplicity, we chose the size of the image blocks to be  $M \times M = 30 \times 30$  for both MR and CR<sup>4</sup>. This number is slightly higher than the typical correlation lengths associated with these images. This is reasonable given that one ideally wants pick the value for  $M$  beyond which the penalty due to the failure of the stationarity assumption just offsets the bandwidth savings that resulted from choosing the larger  $M$ . Given sufficient computational resources, it is also possible to implement a more sophisticated approach in which statistically homogenous regions of varying size and shape are automatically identified and used for prediction. The performance of such a scheme would certainly be superior to the block-based algorithm that we have implemented, but the degree of the potential improvement is uncertain and the implementation would be quite complex. The mask we used in the prediction consists of pixels offset by  $(-1,0)$ ,  $(0,-1)$ , and  $(-1,-1)$  relative to the pixel being coded, and has the same support as the mask used in the JPEG lossless coding standard. The prediction errors resulting from the block-adaptive prediction generally have low variance, and can be effectively compressed using entropy coding. We have chosen arithmetic coding over Huffman coding because no training of a codeword table is required and also for the fact that arithmetic coding can represent a codeword with non-integer number of bits, which gives better performance overall.

The compression results from the adaptive 2D predictive scheme in combination with entropy coding and preprocessing are very encouraging in that by utilizing the local image statistics and modality specific information, the performance of the 2D covariance predictor is

consistently superior than the general compression schemes available. For CR images, where the image statistics are fairly consistent over the range of pixels within our  $M \times M$  region segmentations, the 2D covariance predictor performs over 88% better than the UNIX Compress and GNU-Zip, and is superior over the best JPEG predictor by about 25%. Also, within our expectations, the more statistically varying MR images are more difficult to compress, and since no preprocessing is done at this point of development, the 2D covariance predictor gave only about 45% improvement over the UNIX Compress and GNU-Zip, and about 10% better than the best JPEG predictor. We ran the lossless compression algorithm on a significant number of images on the archive to show the following data:

**Table 1: Lossless Compression Results for the adaptive 2D covariance algorithm**

Modality	Number of images tested	Average image size (bytes)	Compression ratio
CR (Computed Radiograph)	20	8390656	5.023
MR (Magnetic Resonance image)	41	2273162	2.468

The table below shows a comparison of the commonly used lossless algorithms with the above data compiled for CR images:

**Table 2: Comparing with other lossless compression methods (CR images)**

Other methods size: 8390656 bytes	Number of images tested	Compression Ratio
Best JPEG predictor	20	4.38
UNIX Compress (Lempel-Ziv)	20	2.665
GNU Zip Compress	20	2.660

As the tables indicate, better compression is achieved for the adaptive covariance predictor at the cost of computational complexity as reflected in the longer compression/decompression times. However, the speed can be improved by further optimizing the algorithm in terms of implementing it onto a hardware FPGA or by reducing the redundancy of multiplies in the software implementation. Of course, the entropy coder, which takes up approximately 37% of the total run-time, can also be modified to the suit characteristics of the compression algorithms, thereby reducing the speed even further. Nevertheless, the adaptive covariance compression scheme serve as an excellent software data compression solution to the fast growing PACS archive. With the numbers acquired above, it is estimated that the implementation of the lossless compression algorithm will reduce existing stored data by about 70%.

## 2.2 Lossy compression

Due to the large volume of medical image data an extremely high compression ratio is needed in order to fit all the data onto a local magnetic diskette. For example, the UCLA hospital currently performs approximately 200,000 CR procedures per year, and each procedure usually generates two 8 megabyte images. In order to fit all data acquired over a four-year period into a 6GB disk array, a compression ratio of over 2000:1 is required. While such a compression ratio would be impossibly high for intraframe coded images at the resolutions associated with non-medical images, the high resolution and bit depth for CR images leaves significant room for high compression. We developed a lossy compression scheme consisting of subsampling, subband decomposition, adaptive scalar quantization, run-length coding and arithmetic coding which is used to compress 8-megabyte CR images by approximately 2000:1. The goal of this compression is to produce an image which retains sufficient information to enable browsing but is small enough to be easily transmitted over low-bandwidth links.

A CR image is typically 2048 x 2048 pixels with 10 or 12 bits of intensity data for each pixel. The intensity of each pixel is stored using 2 bytes, leading to approximately 8 megabytes of storage space for each image. Since CR images are usually highly correlated and contain mostly low-frequency components, smoothing/subsampling by a factor of 8 in each dimension is first performed to reduce the dimensions of the images to 256x256. In addition, the intensity resolution is reduced from 12 bits to 8 bits. The combination of these two steps gives an initial reduction in data volume of 128:1 with relatively minor degradation of most image features.

Once the reduced-resolution version of the image is created, compression is performed using an algorithm including transform coding and adaptive scalar quantization. The transform constitutes one of the most important elements of a compression algorithm, affecting both the image quality and the implementation complexity. To explore the relative performance of difference transforms, we performed implementations using the block discrete cosine transform (DCT) employed in the JPEG<sup>4</sup> and MPEG<sup>5</sup> compression standards. The DCT performs well when low compression ratios are used, but at high compression ratios only the lowest-frequency DCT coefficients from each 8-pel by 8-pel block are retained, resulting in severe blocking artifacts in the reconstructed image. Wavelet transforms offer an alternative to the DCT and can lead to very efficient image compression if the right filters are used. The principle behind the wavelet transform, as elaborated in a number of recent papers<sup>6</sup>, is to hierarchically decompose an input signal into a series of successively lower resolution reference signals and their associated detail signals. At each level, the reference signal and detail signal (or signals in the multidimensional case) contain the information needed to reconstruct the reference signal at the next higher resolution level. Efficient image coding is enabled by allocating bandwidth according to the relative importance of information in the reference and detail signals, and then applying scalar or vector quantization to the transformed data values<sup>7,8</sup>. We have studied the performance of both block-based and full-frame DCT versus wavelet coding and concluded that a well-optimized wavelet-based compression algorithm generally gave a higher quality image (in terms of both perceptual appearance and quantitative error measures) at the same compression ratio. A further advantage of the wavelet transform is that it can be implemented

using very low computational complexity if the correct filters are used<sup>9</sup>. After the transform is performed, scalar quantization is used to quantize the coefficients in each subband. The step size for each subband is chosen adaptively based on the statistics of each subband to minimize the overall distortion due to the quantization. After quantization, the subband coefficients contain a large number of zeros, therefore runlength coding is used to eliminate the redundancy among these coefficients. Finally, arithmetic coding is performed on the run-length coded coefficients to further reduce the data size. An additional compression of 15-20:1 is achieved after the subband coding stage, which combined with subsampling, gives an overall compression of approximately 2000:1 (Figure 2).

### 2.3 Browsing database

Because of the high compression ratio obtained using the lossy compression approach described above, we are able to store icons of images acquired within the last few years onto a local magnetic disk. The next step is to find an efficient method to organize and manage the huge number of iconified images. To achieve high-performance image browsing, we structured the browsing system in a client/server architecture by implementing the image browsing database using Sybase SQL (Structured Query Language) server. The iconified images and their associated attributes are stored in the relational image browsing database where multiple end users can retrieve specific sets of iconified images qualified by these attributes. The set of attributes referencing the image icons includes patient medical record number, patient name and other demographic information such as imaging modality, imaging protocol, imaging date and time, file identification number, the logical file name for the location on the archive, and the compression method used along with the related parameters. Those attributes that identify the set of image icons uniquely are retrieved from the PACS image database at the beginning of image compression process. Additional attributes for the image icons are stored as pointers to the physical locations of the images on the icon database and on the optical archives. Since the image icons may be accessed through the image icon name attribute, two database reads are required for retrieving a set of image icons. Access to the image browsing database is through SQL. When an user selects an image icon, indicating his/her desire to view the image in full resolution, the browsing system will initiate image retrieval from the PACS image archive for the desired original image.

### 2.4 Browsing interface

The browsing interface is a graphical user interface(GUI) which efficiently controls the accessing of the icons and the losslessly compressed archival images. Through the interface, multiple-end users can search the icon database for a series of candidate images based on search parameters delineated in the last section. These images are reconstructed and displayed in a reduced resolution format, in what we call the damnified form. If the image file queried has more than one frame, then all of the frames will be displayed in the iconified form in addition to any other chosen set of images. If the number of icon images exceed the display interface limit, the images can be scrolled onto or off the screen. The user can subsequently browse for other icons or choose a specific image to be shown in full resolution. The chosen image is retrieved and reconstructed from the optical archive and displayed in full resolution. In the case where the image has more than one frame, only the selected frame(s) with the relevant information will be

retrieved from the optical disks. This browsing system targets the fact that direct retrieval from the optical archives through the PACS network for each candidate image is extremely time-consuming, and often the images retrieved contains little relevant information. For example, a MR or CT scan usually contains 10 to 20 frames, only of a few of which contain diagnostic information. But if these images are first accessed through the browsing system in the iconified format, the subset of frames or image files that are of potential diagnostic interest can immediately be identified. To facilitate this, the lossless compression scheme uses intraframe compression, so that a specific frame of an image file can be retrieved and reconstructed separately without having to transmit and decompress the entire image file, which can exceed 10 megabytes in size.

### 3. CONCLUSION

The image browsing system makes efficient use of the storage and network resources by incorporating both lossless and lossy compression schemes and a browsing database locating at a local magnetic diskette. As the results have shown, the lossless compression scheme outperforms the general compression schemes by targeting to the local characteristics of the image as well as to the imaging modality. This enables a saving of over 70% in storage space on the optical disk archive. The lossy compression scheme achieves high compression ratio and acceptable reviewing quality by using subsampling and wavelet transform coding. The compressed CR icons, which are about 4 kilobytes each in size, can then be most efficiently stored in the icon database. Due to the high compression ratio, the browsing database can contain a few year's worth of acquired images in compressed format, thereby enabling the browsing system to have an extensive set of image icons on record. The browsing interface provides a user-friendly front end to this entire system, and contains a number of essential search functionalities. As the browsing system is incorporated into the PACS, we anticipate faster and more efficient data archiving and retrieval for the physicians.

The next generation PACS must provide an information infrastructure that can support advanced clinical applications that display and process greater and greater volumes of data. Such applications include intelligent indexing systems, natural language queries for decision support, and virtual reality for planning and evaluating therapeutic procedures, and potentially teleconsultation and teleintervention<sup>10</sup>. These applications will in turn require even higher bandwidth and greater use of compression technology.

### 4. REFERENCES

1. M. G. Amin, "Time-frequency Spectrum Analysis and Estimation for Non-Stationary Random Processes," in Time-frequency Signal Analysis Methods and Applications, B. Boashash, Ed., Longman Cheshire: Melbourne, 1992.
2. W. Martin and P. Flandrin, "Wigner-Ville spectral analysis of non-stationary processes," IEEE Trans. ASSP, vol. 33, pp. 1461-1470, 1985.
3. S.M. Rytov, Y.A. Kraviso, and V.I. Tatarskii, Principles of Statistical Radiophysics, Berlin: Springer-Verlag, 1978.
4. G.K. Wallace, "The JPEG Still Picture Compression Standard," Comm. ACM, vol. 34, pp. 30-45, 1991.
5. D. Le Gall, "MPEG: A video compression standard for multimedia applications,"

Comm. ACM, vol. 34, pp. 46-58, 1991.

6. I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Common. Pure Appl. Math.*, vol. XLI, pp. 909-996, 1988.

7. M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform" *IEEE Trans. on Image Proc.*, vol. 1, pp. 205-220, 1992.

8. P. Desarte, B. Macq, and D. Slock, "Signal-adapted multiresolution transform for image coding," *IEEE Trans. Information Theory*, vol. 38, pp. 897-903, 1992.

9. J. D. Villasenor, B. Belzer, J. Liao, "Wavelet Filter Evaluation for Image Compression," to appear in *IEEE Transactions on Image Processing*, August, 1995.

10. D. J. Valentino et al., "Building a Dependable Next Generation PACS," *SPIE Medical Imaging 1995: PACS Design and Evaluation*.

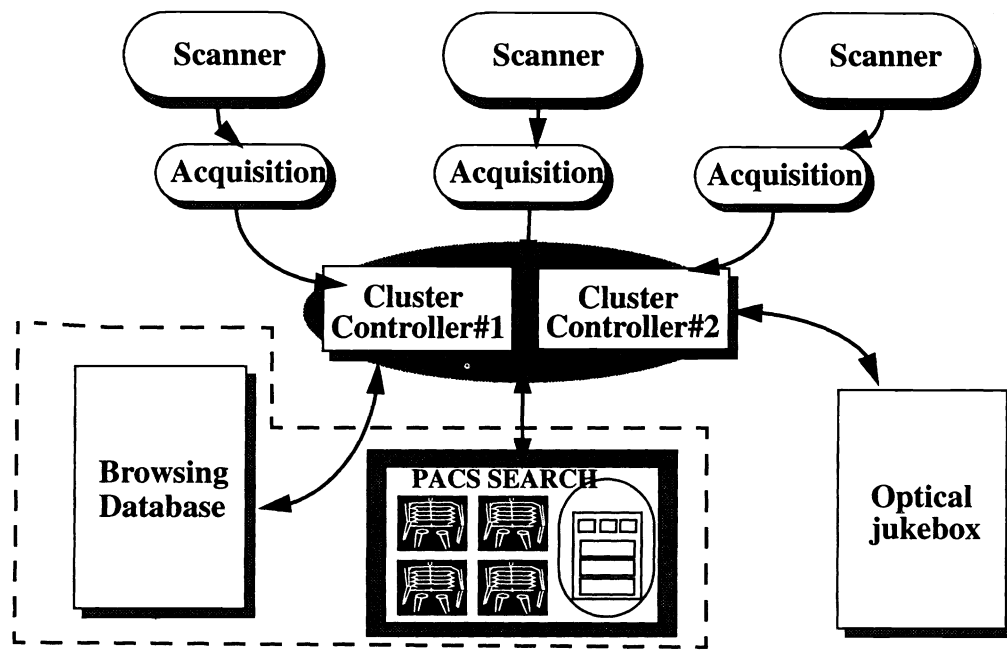


Figure 1: Overview of PACS browser with the rest of the network

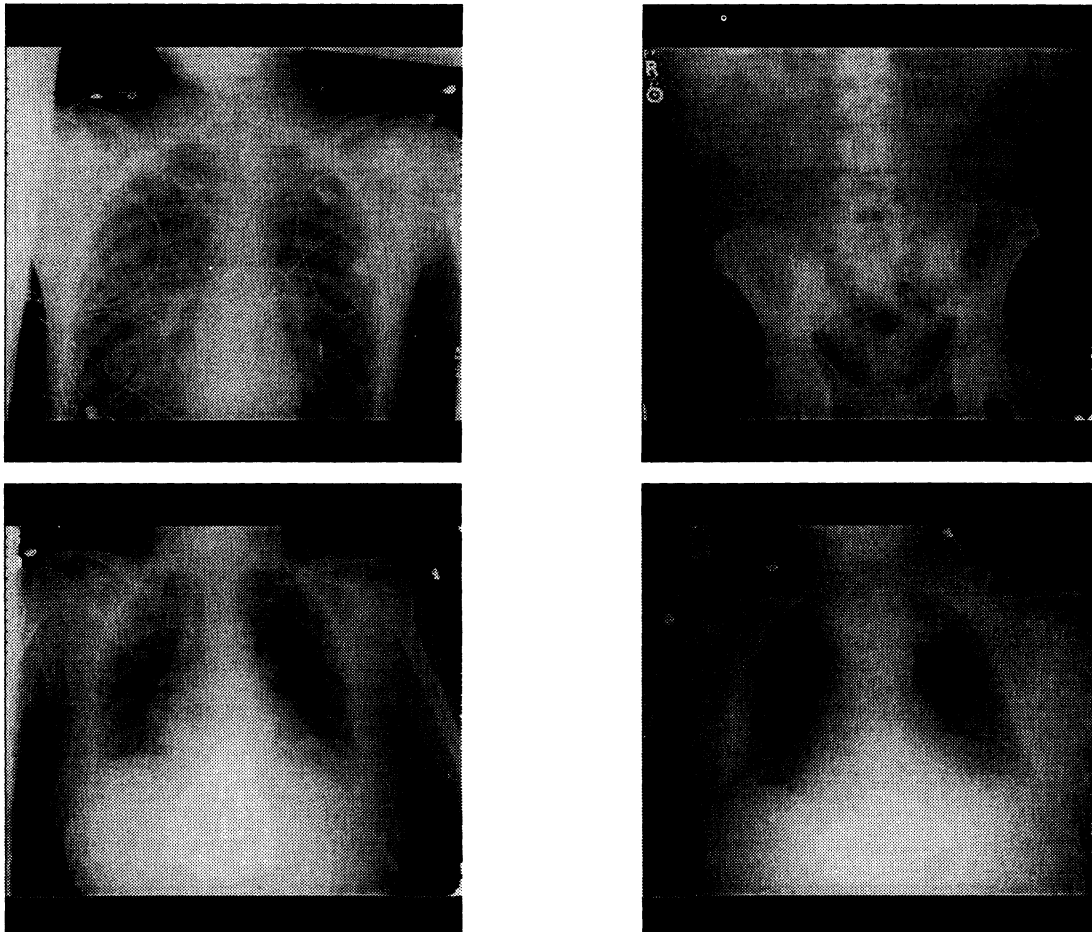


Figure 2: Sample CR images compressed at 2000:1