Pattern Recognition Letters 19 (1998) 585-593

# The Centroid method for compressing sets of similar images

**Kosmas Karadimitriou [1] , John M. Tyler [2]**

*Department of Computer Science, Louisiana State University, Baton Rouge, LA 70803*

Received 8 August 1997; revised 25 November 1997; accepted 25 February 1998

## Abstract

Similar images are images with common features, similar pixel distributions, and similar edge distributions. Fields such as medical imaging or satellite imaging often need to store large collections of similar images. In a set of similar images the image similarities represent patterns that consistently appear across all images; this results in "set redundancy". This paper presents the Centroid method that extracts and uses these similarity patterns to reduce set redundancy and achieve higher lossless compression in sets of similar images. Experimental results with a medical image database demonstrate that the Centroid method can deliver significantly improved image compression. 

*Keywords*: Image compression; Lossless compression; Medical image compression; Set redundancy; Similar images.

## 1. Introduction

Research in data compression has grown rapidly the last 50 years producing a large number of lossless and lossy compression methods (for some excellent reviews see Netravali and Limb, 1980; Jain, 1981; Bassiouni, 1985; Rabbani and Jones, 1991; Wong et al., 1995). Data compression is possible because of data redundancies. In gray-scale digital images three basic data redundancies can be identified and reduced: the coding redundancy, the interpixel (or spatial) redundancy, and the psychovisual redundancy. Most data compression methods are based on the same principles and on the same theoretical compression model. Fig. 1 depicts the lossless compression model and Fig. 2 shows the more general lossy model (the lossless model can be derived from the lossy one by omitting the "Quantization" step). Pixel mapping reduces the interpixel redundancy. Quantization reduces the psychovisual redundancy. The final step of symbol encoding reduces the coding redundancy.

For individual images, this general compression model is sufficient. However, sets of similar images contain additional redundancy due to the existence of common information that appears as similar patterns across these images. Compression methods based on the current compression model only eliminate intra-image redundancy, but not this type of inter-image redundancy. The term set redundancy has been introduced by Karadimitriou (1996) to describe the inter-image redundancy. It has been shown that identifying the common patterns in sets of similar images and using them to reduce set redundancy can significantly improve compression (Karadimitriou, 1996; Karadimitriou and Tyler, 1996, 1997).

[1] Corresponding author. E-mail: kosmas@bit.csc.lsu.edu
[2] E-mail: tyler@bit.csc.lsu.edu

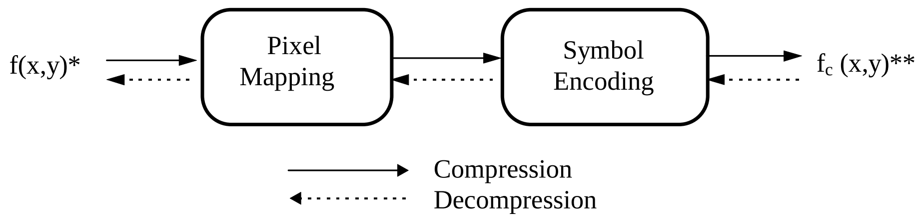


Fig. 1. Lossless compression model. * f (x,y) = original image; ** $f_c$ (x,y) = compressed image.


Fig. 2. Lossy compression model. * f (x,y) = original image; ** $f_c$ (x,y) = compressed image.

Karadimitriou (1996) proposed the Enhanced Compression Model as a more appropriate model for compressing sets of similar images. This model includes an additional step for set redundancy reduction and it is described in Section 2. Methods that achieve set redundancy reduction are referred to as SRC (Set Redundancy Compression) methods. Two SRC methods are the Min-Max Differential method (Karadimitriou and Tyler, 1996) and the Min-Max Predictive method (Karadimitriou and Tyler, 1997). In this paper a third SRC method is described, the Centroid method (Section 3). This method creates an "average image" to capture the common patterns that appear in a set of similar images. One of the best application areas for SRC methods is medical imaging. Medical image databases usually store similar images; therefore, they contain large amounts of set redundancy. Section 4 presents the application of the Centroid method on a set of CT brain scans, and Section 5 contains some concluding remarks.

## 2. The enhanced compression model

The Enhanced Compression Model is an extension of the basic compression model and includes an additional step, the set mapping step (Fig. 3). Set mapping reduces the set redundancy from a set of similar images and it can be realized in different ways. One way is to implement it as an $N$-dimensional transform by translating the origin of the $N$-dimensional coordinate system, where $N$ is the number of pixels in a given image. This $N$-dimensional transform can reduce the dynamic range of the pixel values in a set of similar images. This results in fewer bits per pixel required to store these values. The following example addresses these concepts.

First, assume that every pixel value in an $N$-pixel image represents a coordinate value in the $N$-dimensional space. Consequently, every $N$-pixel image defines a unique point in this space. To visualize these concepts, consider the case of small 1x2 "images". Every such image contains only 2 pixels and can be represented by a unique point in the 2-dimensional space. The values of the 2 pixels are used to define the coordinates $X$ and $Y$. This is a 2-pixel image:

| | |
|---|---|
| 45 | 203 |

The corresponding point in the 2-dimensional space is presented in Fig. 4. A set of random 2-pixel images could result in a scatter plot similar to Fig. 5. As Fig. 5 shows, the range of values is {0-255} for both $X$ and $Y$ coordinates; therefore, 8 bits are needed to represent each value. Now, consider a set of 2-pixel images that are *similar* to each other. In this set, suppose that the first pixel's value in every image is in the range of {40-60} and the second pixel's value is in the range of {135-165}. The scatter plot for such a set is presented in Fig. 6. As we see, all the points are clustered together in the same area. In this case, we can translate the coordi-

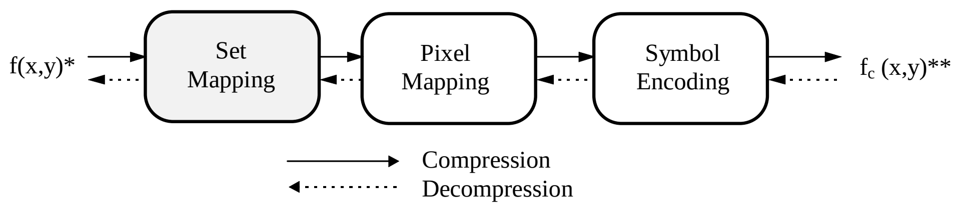


Fig. 3. The enhanced lossless compression model. * f (x,y) = original image; ** $f_c$ (x,y) = compressed image.

nate axes so that the points get closer to the origin. Fig. 7 shows this coordinate system transform.

After translating the origin, the maximum coordinate values in this cluster of points become $X_{max} = 20$ and $Y_{max} = 30$. Now, only 5 bits are sufficient to represent these new coordinates (versus 8 before), each of which corresponds to a value of a pixel in the transformed images. Of greater importance is the fact that translating the origin caused the variance of the pixel values inside every image to decrease; that is, both pixels in every image have values in the range of {0-30}, compared with the original range of {40-165}. This represents a decrease of 76% in the variance and results in a reduction of the image entropy. For example, if it were possible to reduce the range of pixel values to {0-7}, then the images could be encoded with only 3 bits/pixel. Note that this is lossless compression, because it is always possible to perform an inverse axes shift and recover the original images. In addition, this compression is based solely on the properties of the original set of images, and not on the internal properties of the individual images. Therefore, additional compression can be achieved by using any of the existing compression methods to further compress individually every image in this set.

This example with 2-pixel images demonstrated how "set mapping" can be implemented as a coordinate axes translation in the 2-dimensional space. The same procedure can be extended to *N*-dimensional space for compressing sets of images with *N* pixels per image. Note that if the origin is translated to the point where the pixel values of the images become minimum but not negative, then this translation is equivalent to creating the "minimum image" from the set and replacing every image by its difference from this "minimum image". Analogously, the "maximum image" could be used instead of the "minimum image". In this case, the origin would be translated to the point where all pixel values become negative, but with minimum absolute values (Fig. 8). Finally, the origin could be translated to the centroid of the cluster (Fig. 9) which corresponds to the "average image" of the image set. In this case, the origin translation is equivalent to creating the "average image" and then replacing every image in the set with its differences from this "average image".

Creating and using the "minimum", "maximum", or "average" images for set mapping was

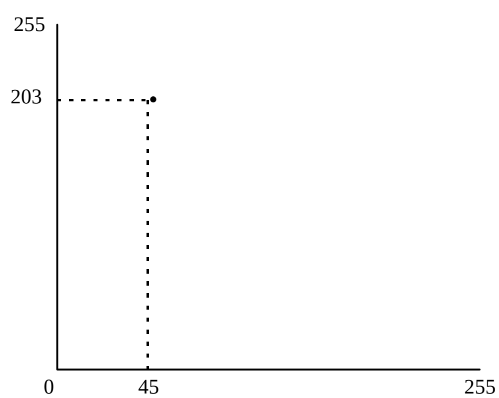


Fig. 4. Point defined by a 2-pixel image

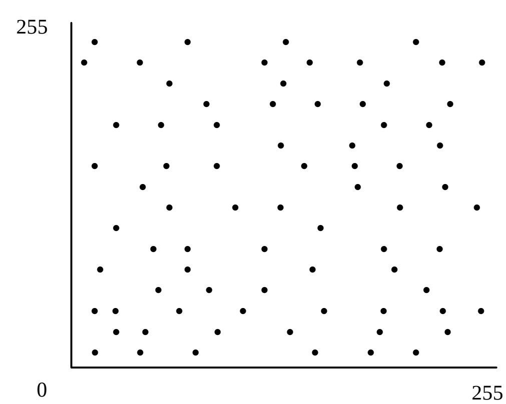


Fig. 5. Set of random 2-pixel images

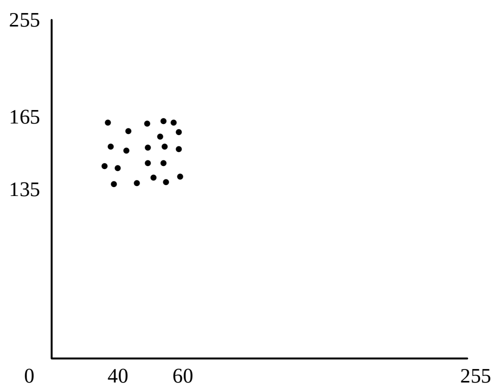


Fig. 6. Set of similar 2-pixel images

the starting point for developing practical methods based on the Enhanced Compression Model. The Min-Max Differential (Karadimitriou and Tyler, 1996) and the Min-Max Predictive (Karadimitriou and Tyler, 1997) methods use the concept of "minimum" and "maximum" images. In the next section the Centroid method is described, which uses the concept of the "average" image.

## 3. The "Centroid" method

In general, predictive methods are very useful for image compression. These methods create a prediction for the value of every pixel; then the encoder

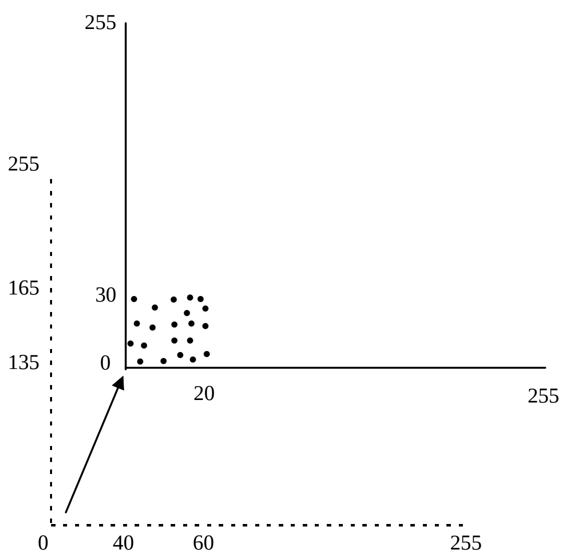


Fig. 7. Translating the origin closer to the cluster

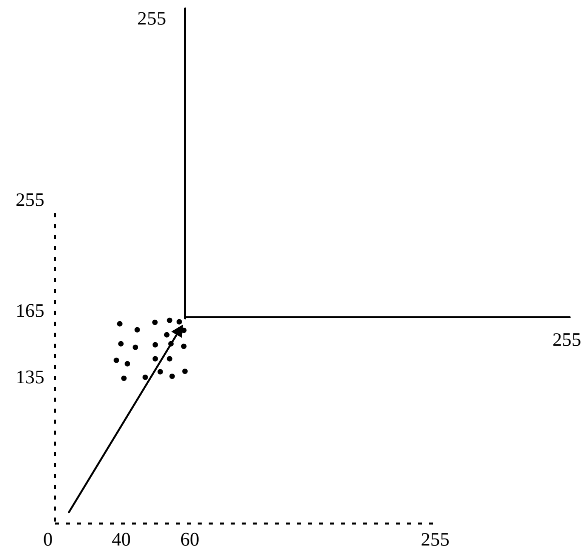


Fig. 8. Translating the origin to the "maximum point" of the cluster, or "maximum image"

stores only the error between this prediction and the pixel value. If the prediction scheme is good, then the errors are very small and have a Laplacian distribution with most of the values very close to zero. The "average image" from a set of similar images can be used to predict the pixel values in

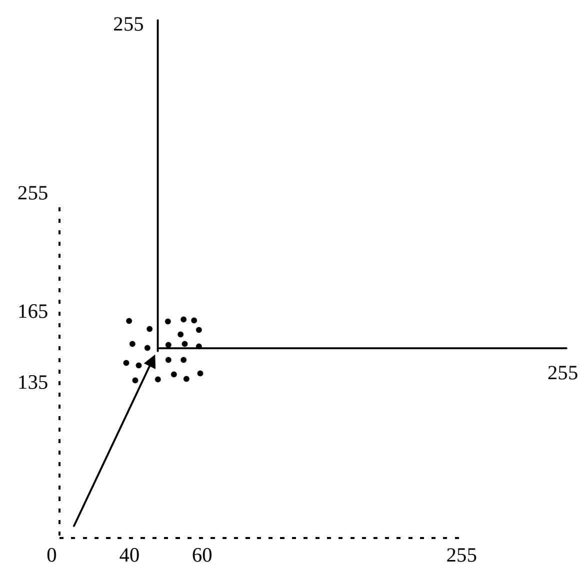


Fig. 9. Translating the origin to the centroid point, or "average image"

each image of the set. A simple model for predicting the pixel value at position $i$ in image $j$ is

$$F_{i,j} = m_i \tag{1}$$

where $F_{i,j}$ is the predicted value and $m_i$ is the average value from position $i$ across all images. This prediction scheme results from intuition rather than mathematical development. However, a formal analysis can produce more sophisticated and more accurate models. The mathematical development of a better prediction scheme that uses more effectively the "average image" is presented next.

Let us assume we have a set of $K$ images, with $N$ pixels per image. We define $x_{i,j}$ to be the pixel value of pixel $P_i$ from image $j$. For natural images, a model that can describe this value is

$$x_{i,j} = x_{i-1,j} + r_{i,j} + s_{i,j} \tag{2}$$

where $x_{i-1,j}$ is the value of the previous pixel, $r_{i,j}$ is a random error (independent and normally distributed with zero mean value), and $s_{i,j}$ is the difference of the two pixels due to the local change in intensity values (slope). Taking the sum in Eq. (2) for pixel position $i$ across all $K$ images and dividing by $K$, results in

$$\frac{\sum_{j=1}^{K} x_{i,j}}{K} = \frac{\sum_{j=1}^{K} x_{i-1,j}}{K} + \frac{\sum_{j=1}^{K} r_{i,j}}{K} + \frac{\sum_{j=1}^{K} s_{i,j}}{K} \ . \tag{3}$$

Since $r_{i,j}$ represents a zero-mean residual, we have

$$\sum_{j=1}^{K} r_{i,j} \cong 0 \ .$$

Also, the average value $m_i$ of pixel $i$ across all images is

$$m_i = \frac{\sum_{j=1}^{K} x_{i,j}}{K} \ .$$

Therefore Eq. (3) becomes

$$m_i = m_{i-1} + \frac{\sum_{j=1}^{K} s_{i,j}}{K} \ . \tag{4}$$

Subtracting Eq. (4) from Eq. (2) yields

$$x_{i,j} - m_i = x_{i-1,j} + r_{i,j} + s_{i,j} - m_{i-1} - \frac{\sum_{j=1}^{K} s_{i,j}}{K}$$

thus,

$$x_{i,j} = m_i + \varepsilon_{i,j} \tag{5}$$

where

$$\varepsilon_{i,j} = x_{i-1,j} - m_{i-1} + r_{i,j} + s_{i,j} - \frac{\sum_{j=1}^{K} s_{i,j}}{K} \ . \tag{6}$$

The term $\varepsilon_{i,j}$ can be considered as an "error term" that denotes the difference of pixel value $x_{i,j}$ from the mean value $m_i$ at pixel position $i$. From Eq. (5) we have

$$x_{i-1,j} - m_{i-1} = \varepsilon_{i-1,j} \ . \tag{7}$$

Because the images are similar, the slope difference $s_{i,j}$ at pixel position $i$ of image $j$ will be very close to the average slope difference at the same pixel position $i$ across all images. Therefore,

$$s_{i,j} - \frac{\sum_{j=1}^{K} s_{i,j}}{K} \cong 0 \ . \tag{8}$$

By replacing Eq. (7) and Eq. (8) at Eq. (6) we get

$$\varepsilon_{i,j} = \varepsilon_{i-1,j} + r_{i,j} \ . \tag{9}$$

Eq. (5) and (9) define the model for predicting a pixel value in a set of similar images. In this model, the error term is serially correlated or autocorrelated. Forecasting with autocorrelated error terms has been studied extensively (e.g., Neter et al. (1989)). Linear regression with one independent variable and first-order autoregressive error is defined as

$$x_{i,j} = \beta_0 + \beta_1 \cdot m_i + \varepsilon_{i,j}, \quad \varepsilon_{i,j} = \rho \cdot \varepsilon_{i-1,j} + r_{i,j} \tag{10}$$

where $x_{i,j}$ is the dependent variable, $m_i$ is the independent variable, $\beta_0$ and $\beta_1$ are the regression parameters, $\rho$ is the autocorrelation parameter, and $r_{i,j}$ is an independent and normally distributed random error term with zero mean value. The model defined by Eq. (5) and (9) is the same as the linear regression model defined by Eq. (10), with

$$\beta_0 = 0 \ , \quad \beta_1 = 1.0 \ , \quad \text{and} \quad \rho = 1.0 \tag{11}$$

From statistical theory, a good prediction scheme for the model of Eq. (10) is

$$F_{i+1,j} = \beta_0 + \beta_1 \cdot m_{i+1} + \rho \cdot e_{i,j}$$

where $F_{i+1,j}$ is the forecast for the (unknown) value $x_{i+1,j}$, and $e_{i,j}$ is defined as

$$e_{i,j} = x_{i,j} - \left(\beta_0 + \beta_1 \cdot m_i\right) .$$

Using Eq. (11) these equations are simplified to

$$F_{i+1,j} = m_{i+1} + e_{i,j}, \qquad e_{i,j} = x_{i,j} - m_i . \tag{12}$$

These two equations define a model for predicting the value of a pixel using the average value for that pixel position plus a correction term. This is the basis of the Centroid method. The experimental results that are presented in the next section show that this model is more accurate and provides better compression than the simpler model of Eq. (1).

## 4. Experimental results

### *4.1. The application of the Centroid method to medical images*

Medical imaging is an area in which the Enhanced Compression Model can be used very effectively. The reason is that medical images always depict parts of the same subject: the human body. Moreover, the standard procedures used in radiology result in images very similar to one another. For example, for every chest X-ray the position of the patient, the orientation of the imaging device, and the parameters used are standard. A collection of 1,000 chest X-rays is significantly more self-correlated, in the statistical sense, than a collection of random images. The same is true for any other collection of medical images that are grouped together by modality and type of exam (for example, brain axial MRI scans, liver CT scans, etc). Consequently, medical image databases store sets of images with high similarity that contain large amounts of set redundancy. This redundancy appears as common patterns across the images which can be used by the Centroid method to improve compression.

In our experiments we used a small image database with 51 CT brain images collected at M.D. Anderson Cancer Center in Houston, Texas. The images were gray-level, with resolution 512x512 pixels, and were scaled to 8 bits/pixel. They were randomly selected from patients of both sexes, various ages, and a variety of pathological conditions. From this test database, a set of images with high similarity was formed, containing 10 images (Fig. 10); this set size was chosen to ensure high similarity among the images. The remaining images were considered to belong in other sets. In this small test database, the size of those other sets was not large enough for testing purposes. However, a real-world medical image database containing hundreds or thousands of images can be partitioned in a number of sets with reasonably large sizes; in that case the Centroid method can be efficiently implemented in every set.

To facilitate the implementation of the Centroid method, these images were registered into a standard position and size. This was done with a semi-automatic procedure, in which three landmark points were identified in every image and then the images were rotated, translated, and scaled so that these three landmark points would coincide with standard

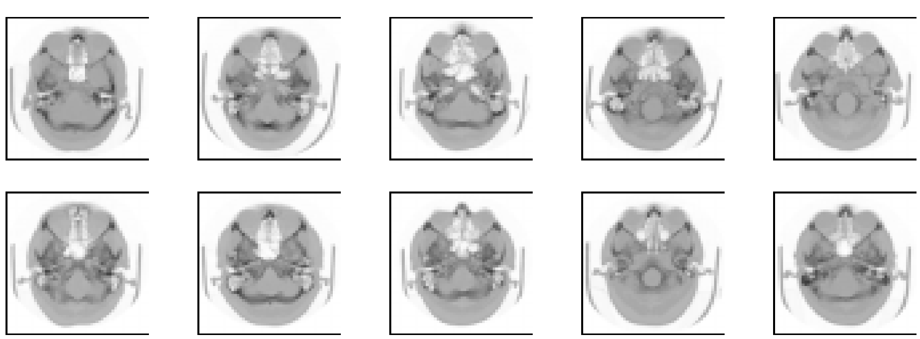

Fig. 10. The CT images used in the experiments

positions. This registration is useful in creating the "average" image. Note, however, that there is no need to store the registered versions of the original images. Once the registration parameters (rotation, translation, and scaling) for every image are known, these parameters can be used to perform the inverse registration on the "average" image, instead of using the registered versions of the original images. Consequently, there is no change in the original images.

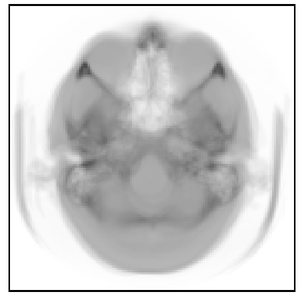

Fig. 11. The "average" CT brain image

### 4.2. Experimental procedure and test results

According to the Enhanced Compression Model, the Centroid method can be combined with standard compression techniques to improve their compression ratios. In our experiments the Centroid method was tested in combination with three well-known compression techniques: Huffman encoding, Arithmetic coding, and Lempel-Ziv compression. The first two are entropy-based techniques; the third is a dictionary-based compression scheme. Huffman encoding (Huffman, 1952; Knuth, 1985) is one of the oldest and most widely used compression methods. Arithmetic coding (Langdon, 1984; Witten et al., 1987; Moffat et al., 1995) is slower and more difficult to implement; however, it outperforms the Huffman method in compression ratios. Lempel-Ziv compression (Ziv and Lempel, 1977, 1978) is a powerful technique with many variations; in this research the LZW version (Welch, 1984) was used. All experiments were performed under the Unix operating system which includes standard implementations of Huffman and LZW methods (commands "pack" and "compress", respectively). For Arithmetic coding the publicly available software based on Moffat et al. (1995) was used.

As it has been mentioned in the previous sections, the main idea of the Centroid method is to use the "average" image as a predictor. The "average" image for our test set of CT images is shown in Fig. 11. A simple approach, suggested by Eq. (1), is to subtract this "average" image from each image to be compressed and store only the differences. However, the Centroid method is based on the more sophisticated model described in Eq. (12). Both approaches were implemented and tested to demonstrate their performance. The first approach (taking the differences from the "average") is referred to as the Differential method, and the results are presented in Table 1. The second approach is the Centroid method, and the results are presented in Table 2. In these two tables, the compression ratio is defined as

$$C = \frac{original\ image\ size}{compressed\ image\ size}$$

and the improvement in compression is defined as

$$improvement = \frac{C_{SRC} - C}{C} \cdot 100\%$$

where $C$ is the compression ratio achieved when using a regular compression method only, and $C_{SRC}$ is the compression ratio achieved when combining *SRC* with that regular compression method.

As shown in Tables 1 and 2, the compression is improved when the images are pre-processed with the "average" image. The improvement from the use of the Centroid method is 18% for Lempel-Ziv, 59% for Arithmetic, and 90% for Huffman compression. The simpler Differential approach also achieves some improvement for Arithmetic and Huffman encoding (5% and 17% respectively) but not as much as the Centroid method. These results verify the validity of the Enhanced Compression Model and demonstrate the use of the Centroid method for improving compression in sets of similar images. It is important to emphasize that the Centroid method is totally reversible, therefore the overall compression is lossless.

The performance of the Centroid method was also compared with the performance of two previously proposed SRC methods, the Min-Max Differential (MMD) method (Karadimitriou and Tyler, 1996) and

Table 1
Experimental results from taking the differences from the average

| | Average size (Kbytes) | Lossless compression ratio | Comments |
|---|---|---|---|
| Original image | 262 | N/A | |
| Huffman compressed | 190 | 1.379 : 1 | |
| Differential + Huffman | 162 | 1.617 : 1 | 17% improvement |
| Arithmetic compressed | 154 | 1.701 : 1 | |
| Differential + Arithmetic | 146 | 1.795 : 1 | 5% improvement |
| Lempel-Ziv compressed | 107 | 2.449 : 1 | |
| Differential + Lempel-Ziv | 107 | 2.449 : 1 | 0% improvement |

Table 2
Experimental results from the Centroid method

| | Average size (Kbytes) | Lossless compression ratio | Comments |
|---|---|---|---|
| Original image | 262 | N/A | |
| Huffman compressed | 190 | 1.379 : 1 | |
| Centroid + Huffman | 100 | 2.620 : 1 | 90% improvement |
| Arithmetic compressed | 154 | 1.701 : 1 | |
| Centroid + Arithmetic | 97 | 2.701 : 1 | 59% improvement |
| Lempel-Ziv compressed | 107 | 2.449 : 1 | |
| Centroid + Lempel-Ziv | 91 | 2.879 : 1 | 18% improvement |

the Min-Max Predictive (MMP) method (Karadimitriou and Tyler, 1997). Table 3 presents a summary of the results from testing the three methods on the same set of CT images. As this table shows, the Centroid method produces a larger compression improvement than MMD, but smaller than the MMP method. However, the Centroid method offers a distinct advantage over MMP: new images can be inserted in the set and compressed without recomputing the "average" image. In contrast, MMP requires recomputing the "minimum" and "maximum" images for the set, as well as recompressing all images. Thus, the Centroid method offers greater flexibility for maintaining dynamic image sets, whereas the MMP method offers better compression for storing fixed image sets. Clearly, the choice is application-dependent.

Table 3
Summary of compression improvements (%) using SRC methods

| | Centroid | MMD | MMP |
|---|---|---|---|
| Huffman | 90 | 48 | 129 |
| Arithmetic | 59 | 28 | 93 |
| Lempel-Ziv | 18 | 13 | 37 |

Finally, a note on the computational time requirements. An image file of 262 Kbytes can be preprocessed by the Centroid method in less than a second on a Sun SPARC 20 workstation, which is approximately as fast as the MMD and MMP methods. The additional time required by all SRC methods is a small tradeoff considering the improvement in compression produced.

## 5. Conclusion

Image compression is possible because of the existence of different types of redundancy in digital images. Current compression methods usually target the intra-image redundancies; however, sets of similar images contain significant amounts of inter-image redundancy, the set redundancy. Set redundancy can be used to improve compression in sets of similar images. This paper describes the Enhanced Compression Model that extends the current theoretical compression model by including a set mapping operation. Set mapping identifies and uses the common patterns that appear in similar images to reduce set redundancy. The Centroid method was developed as a practical technique for performing the set mapping operation.

One of the best application areas for the Enhanced Compression Model is medical imaging. Our tests with CT brain scans showed that when medical images are pre-processed with the Centroid method, lossless compression ratios are improved by as much as two-fold.

There are also many other application areas where the Enhanced Compression Model can be implemented. For example, satellite image databases often contain sets of images taken over the same geographical areas, and under similar weather or lighting conditions. Therefore, they contain set redundancy that can be reduced by the Centroid method. Future research will explore implementation of SRC methods in these fields. Also the use of lossy methods will be investigated as an alternative to the use of the lossless techniques for the Enhanced Compression Model.

## References

Bassiouni, M.A., 1985. Data compression in scientific and statistical databases. IEEE Trans. Software Eng. 11, 1047-1058.

Huffman, D.A., 1952. A method for the construction of minimum redundancy codes. Proc. IRE 40, 1098-1101.

Jain, A.K., 1981. Image data compression: a review. Proc. IEEE 69, 349-389.

Karadimitriou, K., 1996. Set Redundancy, the Enhanced Compression Model, and methods for compressing sets of similar images. Ph.D. Dissertation, Louisiana State University, Baton Rouge, LA.

Karadimitriou, K., Tyler, J.M., 1996. The Min-Max Differential Method for large-scale storage and compression of medical images. In: Proc. of Annual Molecular Biology and Biotechnology Conference, Baton Rouge, LA.

Karadimitriou, K., Tyler, J.M., 1997. Min-Max Compression Methods for medical image databases. ACM SIGMOD Record 26, 47-52.

Knuth, D.E., 1985. Dynamic Huffman coding. J. Algorithms 6, 163-180.

Langdon, G.G., 1984. An introduction to arithmetic coding. IBM J. Res. Develop. 28, 135-149.

Moffat, A., Neal, R., Witten, I.H., 1995. Arithmetic coding revisited. In: Proc. IEEE Data Compression Conference, Snowbird, UT.

Neter, J., Wasserman, W., Kutner, M.H., 1989. Applied Linear Regression Models. Irwin, Burr Rigde, IL.

Netravali, A.N., Limb, J.O., 1980. Picture coding: a review. Proc. IEEE 68, 366-406.

Rabbani M., Jones, P.W., 1991. Digital Image Compression Techniques. SPIE Press.

Welch, T.A., 1984. A technique for high performance data compression. IEEE Comput. 17, 8-19.

Witten, I.H., Neal, R.M., Cleary, J.G., 1987. Arithmetic Coding for data compression. Comm. ACM 30, 520-540.

Wong S., Zaremba, L., Gooden, D., Huang H.K., 1995. Radiologic image compression - a review. Proc. IEEE 83, 194-218.

Ziv J., Lempel A., 1977. A universal algorithm for sequential data compression. IEEE Trans. Inform. Theory 23, 337-343.

Ziv J., Lempel A., 1978. Compression of individual sequences via variable-rate coding. IEEE Trans. Inform. Theory 24, 530-536.