Content-Based Compression of Mammograms for Telecommunication and Archiving

Brad Grinstead[1*], Hamed Sari-Sarraf[1], Shaun Gleason[2], and Sunanda Mitra[1]
[1]Texas Tech University
[2]Oak Ridge National Laboratory

Abstract

The concept of content-based image compression (CBIC) has far reaching effects in the areas of archiving and telecommunications. The purpose of this paper is to present some pilot study results from the application of CBIC to mammography. Unlike traditional compression approaches, CBIC first analyzes the content of the data before compression takes place. In this approach, prior to compression, the data is preprocessed and is segmented into two non-overlapping regions: (1) focus-of-attention regions (FARs) that contain the "important" segments of the data, and (2) background regions. Subsequently, the former regions are compressed using a lossless compression technique (maintaining fidelity), while the latter regions are compressed with the aid of a lossy technique (attaining large reductions in data). The intended result is an optimal balance between data reduction and data fidelity. In this case, compression ratios 5-6 times greater than that of lossless compression alone can be reached while preserving the important information.

1. Background

The technological advances in digital information transmission and archiving, as significant as they have been in the recent years, cannot rival the perpetual growth in the amount of produced data. In the coming years, this will especially be true in the medical field, where there is a unified movement toward the digital age. While this movement is expected to enhance patient care, it faces particularly difficult challenges in the areas of data transmission and archiving for telemedicine applications. As a specific example, consider that over the next 5-10 years there will be a movement toward full-field, digital mammography, which is primarily motivated by the improved image quality over the traditional film-based approach.

In this work, we introduce a novel approach to information compression, namely CBIC, which aims to achieve an optimal balance between data reduction and data fidelity. Unlike existing techniques [1-3], in this approach, prior to compression, the data is preprocessed and is segmented into two non-overlapping regions: (1) FARs that contain the "important" segments of the data, and (2) background regions. Subsequently, the former regions are compressed using a lossless compression technique (maintaining fidelity), while the latter regions are compressed with the aid of a lossy technique (attaining large reductions in data).

2. Segmentation of digitized mammograms for CBIC

In screening mammography the "important" image information sought by radiologists includes masses, microcalcifications, ducts, and the breast boundary. These regions often constitute less than 15% of the mammogram. Therefore, a successful segmentation of these events from the background regions (i.e., regions corresponding to connective tissue *plus* the non-breast regions of the image) will make this a prime target for CBIC. Recall that after this

[*] Corresponding Author – bgrinstead@ttu.edu

segmentation step, a combination of lossy (for background regions) and lossless (for FARs) compression techniques can be employed to achieve large reductions in data while maintaining the original image quality. Our approach to accomplishing this segmentation step employs fractal encoding.

The primary motivation for pursuing a fractal-based approach is that fractal encoding and iterated function systems, in general, are ideal for characterizing the cloud-like texture that represents the normal background tissue in mammograms. This scheme can therefore be used to flag all structures that appear to be different from the normal background tissue. Fractal encoding exploits the property of *partitioned self-similarity* of images [4]. This means that instead of being formed of copies of its whole self, as is the case in exact self-similarity, the image is composed of transformed parts of itself. In computing the coefficients of this transformation or map, it is assumed that, because of the notion of partitioned self-similarity, each subregion of the image can be described (in the sense of minimizing a dissimilarity metric) in terms of another subregion. The former subregion belongs to the range pool, $\boldsymbol{R}$, while the latter belongs to the domain pool, $\boldsymbol{D}$, of the map. If a given subregion in $\boldsymbol{D}$ cannot be mapped to any region in $\boldsymbol{R}$ (i.e., their measure of dissimilarity is above a specified threshold), then $\boldsymbol{R}$ is further partitioned into smaller subregions. This process continues recursively until either a similar subregion from $\boldsymbol{D}$ is found or a specified maximum level of partitioning, $\boldsymbol{L_{max}}$, is reached.

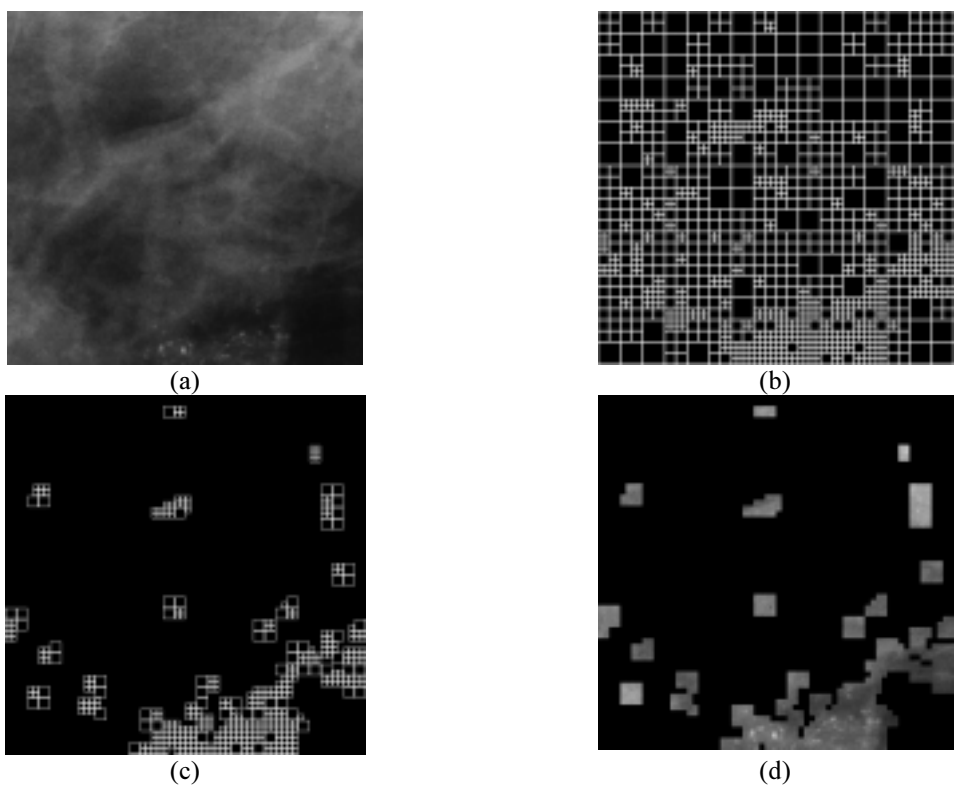

Figure 1. (a) A subregion of a digitized mammogram with clustered microcalcifications in the lower portion of the image. (b) Quadtree partitioning as a result of fractal encoding. (c) Those subregions and their 8-neighbors in (b) that never satisfied the similarity condition. (d) Input mammogram segmented into FARs and background regions (black areas).

In the first of two pilot studies conducted thus far [5,6], it was shown that during the fractal encoding process, for subregions in $\boldsymbol{R}$ that contain "important" information, $\boldsymbol{L_{max}}$ is reached. The reason is that the visual appearance of such events (e.g., microcalcifications) is different from that of the coexisting background structures. Therefore, subregions in $\boldsymbol{D}$ that can be mapped to those

areas in the image with microcalcifications are expected to be nonexistent [Figures. 1(a) - (b)]. Subregions for which L_{max} is reached along with their 8-neighbors make up the FARs, Figures 1(c), and (d). This hypothesis was further tested through a second pilot study [7], the results of which are summarized next.

Eighty, 12-bit, 50 micron/pixel mammographic images were obtained from the University of Chicago's database and were used to carry out this pilot study. These included 35 abnormal and 45 normal cases. The images were collected from 4 different hospitals. Cases containing biopsy-proven clustered microcalcifications were chosen randomly and the cluster locations were marked by experienced radiologists at each site. To mitigate problems caused by global nonuniformities, each mammogram was divided into 512x512, nonoverlapping subimages. Each subimage was then processed independently. Subsequent analysis of the data showed that on average FARs constituted only 17% of the input data but, at the same time, these FARs contained 92% of the "important" information (in this case, microcalcifications) in the data.

3. Compression Strategies and preliminary results

Notice the irregular shapes of the FARs in Figure 1(d). In order to decrease the difficulty in encoding such irregular shapes, bounding boxes can be placed around the FARs. Using this method to identify the location of the FARs makes the encoding process easier, at the cost of lowering the overall compression ratio because the bounding boxes are usually larger than the FARs they encompass. The preliminary results presented in this paper were obtained with bounding boxes manually placed around the FARs. Once the locations of the FARs have been determined, there are several compression strategies that can be employed.

First, knowing that the FARs comprise, on average, about 17% of the image, a compression ratio of approximately 5:1 can be achieved by simply setting the background to zero, leaving only the FARs. Using a standard lossless coding technique on the result could bring that ratio up to 15:1. This, however, does not allow the background structures – which are required by radiologists for context – to be seen.

The second strategy is to use a combination of lossy and lossless coding to preserve the contextual structures while preserving the image fidelity in the areas of interest. By using a standard lossy compression [8-10] on the background areas, the contextual structures can be retained. The lossless coded FARs can then be superimposed over the lossy compressed background, preserving the fidelity of the areas of interest. The cost for retaining these structures comes out of the compression ratio.

When performing lossy compression on the background, there are a few directions that can be taken. The first is to take the original subimage [Figure 2(a)] and segment out the FARs as shown in Figure 2(b). Then, lossy compression is performed on the resultant image. However, this is not a very good solution because certain transform-based lossy compression techniques (e.g. Embedded Zerotree Wavelets [9]) cause ringing to appear around hard edges and transitions as seen in Figure 2(c). Figure 2(d) shows the artifacts that can occur when lossless encoding is performed on the FARs and the resulting images are added together. However, if it can be accepted that the overhead of encoding the *entire* subimage lossily is trivial compared to the amount of data generated by the lossless coding of the FARs, then lossless coding can simply be superimposed over the entire lossy-compressed subimage, achieving the results shown in Figure 2(e). The results presented in this paper are based on this last approach.

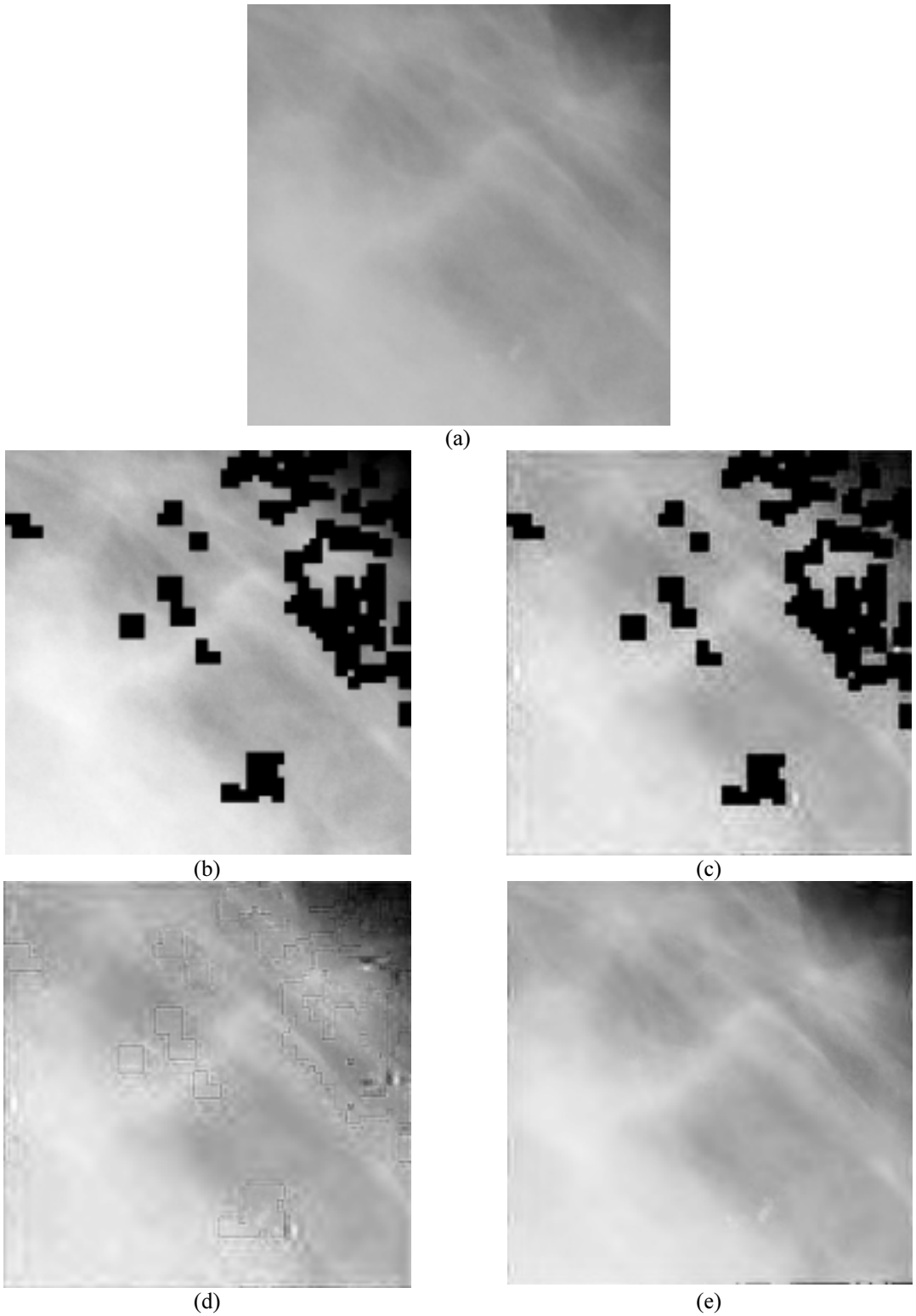

Figure 2. (a) A subimage of a digitized mammogram. (b) Subimage with the FARs removed. (c) Results of lossy-coding on (b) the segmented image. (d) The losslessly encoded FARs added to (c) the lossy-coded segmented image. (e) Results of lossily compressing the entire subimage and then superimposing the losslessly encoded FARs.

For the preliminary results presented in this paper, twelve 512 x 512 subimages of digitized mammograms were processed. The FARs resulting from the fractal encoding process were manually encased in bounding boxes for compression. The information contained in the bounding boxes was then losslessly compressed, while the entire subimage was compressed with a lossy technique. Table I shows the preliminary results of combining the two. Notice that the smaller percentage of the image contained in the FARs, the higher the compression ratio.

Table I. Compression results on 512 x 512 subimages of the breast tissue

512 x 512 Subimage	% of Image Contained in FARs	% of Image Contained in Bounding Boxes	Compression Ratio with the Background Compressed at 80:1
1	2.93%	3.00%	34.21
2	2.93%	4.00%	30.42
3	3.74%	4.70%	27.27
4	6.24%	7.32%	20.22
5	6.54%	8.50%	18.03
6	6.72%	8.00%	18.72
7	8.96%	12.00%	13.54
8	9.40%	11.00%	14.48
9	11.01%	14.15%	11.93
10	12.45%	15.00%	11.01
11	13.01%	15.00%	11.01
12	21.80%	22.00%	7.81
Average	8.81%	10.39%	18.22
Min	2.93%	3.00%	7.81
Max	21.80%	22.00%	34.21

The expression for estimating the compression on a single subimage is:

$$CR = \left(\frac{B}{LLCR} + \frac{1}{LCR} \right)^{-1} \tag{1}$$

where B is the percentage of the image contained in the bounding boxes encompassing the FARs, $LLCR$ is the lossless compression ratio used on the FARs, and LCR is the lossy compression ratio used on the background.

It can be seen from the data in Table I that the percentage of the image included in the bounding boxes is, on average, 1.2 times greater than that included in the FARs. Also gathered from the Table I trials is the fact that the average compression of the lossless coding was approximately 2:1. Using this knowledge and equation (1), some estimates of the compression of entire mammograms can be made. The FARs for twelve mammograms were generated and estimated compression ratios were determined. The average compression of an entire mammogram was found to be about 14:1. The minimum and maximum compression ratios were 6.38:1 and 18.70:1 respectively. These preliminary estimations based on the data in Table I show that the results of CBIC are, on average, about 7 times higher compression than lossless compression alone. In addition, the information in the areas of interest is guaranteed to be free of distortion, which would not be the case if lossy compression alone was applied to the mammogram.

4. Conclusion

Content Based Image Compression (CBIC) can greatly reduce the amount of data necessary to archive and transmit large databases of mammography information. By using a quadtree fractal encoding scheme to generate Focus of Attention Regions (FARs), a combination of lossy and lossless techniques can be used to provide compression 4 to 6 times that of standard lossless compression, while preserving the fidelity of the areas of interest.

While the preliminary results gathered in this initial study are quite good, there is some room for improvement. By using smaller bounding boxes, the amount of data added to irregular shapes that must be losslessly compressed can be minimized, raising the compression ratio. Also, other forms of lossless encoding which have yet to be explored may lead to increased performance.

In addition to increasing the compression ratio, further work needs to be done in automating the CBIC process. This can be accomplished by automatic generation of the bounding boxes. Also, the compression process is currently being done in separate stages, with the lossless and lossy algorithms in separate programs. This can be accomplished by combining the stages into a single process.

References

1. M. G. Strintzis, "A Review of Compression Methods for Medical Images in PACS," *Int. J. Med. Inf.* **52**(1-3), 1998, pp. 159-165.
2. H. P. Chan, et al., "Image Compression in Digital Mammography: Effects on Computerized Detection of Subtle Microcalcifications," *Med. Phys.* **23**(8), 1996, pp. 1324-1336.
3. R. M. Gray, et al., "Evaluating Quality and Utility in Digital Mammography," *IEEE Int. Conf. on Image Proc*, October 1996, pp. 5-8.
4. Y. Fisher, "Fractal image compression with quadtrees," Fractal Compression: Theory and Application to Digital Images, Y. Fisher, ed., Springer Verlag, New York, 1994, pp. 55-77.
5. H. Sari-Sarraf, et al., "A Novel Approach to Computer-Aided Diagnosis of Mammographic Images," *3rd IEEE Workshop on Applications of Comp. Vision*, December 1996.
6. H. Sari-Sarraf, et al, "Reducing Data and False Detections in Computer-Aided Diagnosis of Microcalcifications," *82nd Annual Meeting of the Radiological Society of North America*, December 1996.
7. H. Sari-Sarraf, et al., "Front-End Data Reduction in Computer-Aided Diagnosis of Mammograms: A Pilot Study," *SPIE's Medical Imaging Conf.*, February 1999.
8. S. Mitra, et al., "High Fidelity Adaptive Vector Quantization at Very Low Bit Rates for Progressive Transmission of Radiographic Images," *J. Electronic Imaging* **8**(1), 1999, pp. 23-35.
9. J. Shapiro, "Embedded Image Coding Using Zerotrees of Wavelet Coefficients," *Transactions on Signal Processing*, **41**(12), December 1993, pp. 3445-3462.
10. A. Said and W. Perlman, "A New, Fast, and Efficient Image Codec Based on Set Partitioning in Hierarchical Trees," *IEEE Transactions on Circuits and Systems for Video Technology*, **6**(3), June 1996, pp. 243-249.