Parallel Genetic Algorithm for Document Image Compression Optimization

Aysha V
College of Applied Science Pattuvam (IHRD)
Kannur University
Kannur, India
aysha_v@yahoo.com

Dr Kannan Balakrishnan
Department of Computer Applications
CUSAT
COCHIN, India
mullayilkannan@gmail.com

Dr S Babu Sundar
Department of Computer Science
King Khalid University
Abha, SA
drsbsundar@gmail.com

***Abstract*—This work proposes a parallel genetic algorithm for compressing scanned document images. A fitness function is designed with Hausdorff distance which determines the terminating condition. The algorithm helps to locate the text lines. A greater compression ratio has achieved with lesser distortion.**

Keywords-Parallel genetic algorithm; compression; document image; compression ratio; fitness function.

I. INTRODUCTION

Document image processing is a very important matter in our day to day life. Even though computerization is done in many of the offices, piles of old documents are remaining to enter in the digital world. Survey and land records, documents in registration offices, patent information, valuable manuscripts in palm leaf etc are some of the examples. Document Image Processing is a shortcut for entering such documents in the digital world. Document imaging help to capture, store, search and retrieve images easily. Document Image Processing involves image segmentation, analysis, enhancement, compression, reconstruction, transmission of document image etc.

The document image is captured through a scanner and stored as jpeg image. Document image enhancement helps to get clarity for the image. Compression of document images reduces the storage space required and also reduces the transmission time for sending images through internet. The scanner level implementation of the compression algorithm makes the work easier. The compression of document image reduces the bottleneck that occurs during facsimile transmission. This work proposes a parallel genetic algorithm based method for faster compression and for achieving a better compression ratio.

II. EARLIER WORKS AND PRESENT STATUS IN DOCUMENT IMAGE COMPRESSION

Document Image Compression helps us to use less storage space and easy access of data. There are two types of Compression techniques: Lossless and Lossy. Lossy Compression techniques are applicable for ordinary digital images, because of the limitation of our eyes. Even though certain pixel portions are lost human eye can interpret the image. But this is not completely adoptable in the case of Document image compression. If some of the text portions are lost the meaning may change or the reader may be unable to deduce the meaning. Some of the conventional encoding technique for compression are Huffman coding, inter pixel redundancy, psycho visual Redundancy, JPEG, JPEG2000 compression etc.

B F Wu, C C Chin and Y L Chan [5] provide a method to compress the text plane using the pattern matching technique, called JB2. Wavelet transform and zero tree coding are used to compress the background and the text's color plane in their paper "Algorithm for compressing compound document images with large text/background overlap". In their Paper, "Binary image compression using identity mapping backpropagation neural network" [1], Murshed, Nabeel A, Bortolozzi, Flavio, Sabourin and Robert describes the compression of handwritten signatures and their reconstruction. They observed that, the lowest and highest reconstruction errors were 3.05 multiplied by 10^{-3}% and 0.01% respectively. Patrice Y Simard, Henrique S Malvar, James Rinker and Erin Renshaw proposed a system in their paper "A Foreground/ Background separation Algorithm for Image Compression" [2] known as SLIm (Segmented layered Image) for separating text and line drawing from background images, in order to compress both more effectively. This approach is different from DjVu, Tiff-FX, and MRC by being simple and fast. In the paper, "Differentiation of alphabets in handwritten texts" A

Seropian , M.Grimald, Dr N Vincent [3] the authors used fractal compression and statistical methods.

In the paper "Lossless Generalized Data Embedding", Mehmet Utku Celik, and Akmat Murat Tekalp [4], introduces a method to achieve a lossless recovery of the original image by compressing portions of the signal that are susceptible to embedding distortion and transmitting these compressed descriptions as a part of the embedded payload.

In the paper "Automatic Text Detection Using Multi Layer Color Quantization in Complex Color Images", Soo-Chang Pei and YU-Ting Chuang [6] explain an algorithm as follows: The input image is quantized to several quantized images with different number of quantized color. For each quantized image, it was put to 3D histogram analysis to find some specific colors, which are probable text candidates. Each bi-level image relative to its color candidate could be produced. By calculating some spatial features and relationships of characters, text candidates should be identified. Then combine all text candidates to single quantization layers so as to localize text region accurately.

In their paper " Text Document Authentication by Integrating inter character and word spaces water marking" [7], Hujuan Yang and Alex C Kot suggests a method which makes use of the integrated inter character and word spaces for water marking embedding. An overlapping component which is of size 3 is utilized, whereby the relationship of the left and right spaces of the character is employed for the watermark embedding. The integrity of the document can be ensured by comparing the hash value of the character components of the document before and after watermark embedding, which can be applied to other line shifting and word shifting methods as well. While the authenticity of the document can be ensured by generating the gold-like sequence, which takes the secret key of the authorized user/owner as the seed value, and it is subsequently XORed (Exclusive OR) with hash value of the character components of the document to generate the content-based watermark. The capacity of the water mark has increased compared with conventional line shifting and word shifting methods. In the International Conference on Image Processing 2004, George Paolidis, Sofia Takeridou and Christodoulos Chamzas presented a paper " Jpeg-matched Data filling of Sparse images" which proposes a non linear projection scheme for data filling that matches the baseline JPEG coder and produces good compression results and improved image quality. Meftah, Boudjelal, Debakla, M., Zaagane, M., Benyettou, A., Lezoray, Olivier "Spiking neuron network for image segmentation (2008)" [8], in this abstract they claim that Spiking Neuron Networks (SNNs) are often referred to as the 3rd generation of neural networks which have potential to solve problems related to biological stimuli. They derive their strength and interest from an accurate modeling of synaptic interactions between neurons, taking into account the time of spike emission. SNNs overcome the computational power of neural networks made of threshold or sigmoid units. Based on dynamic event driven processing, they open up new horizons for developing models with an exponential capacity of memorizing and a strong ability to fast adaptation. Moreover, SNNs add a new dimension, the temporal axis, to the representation capacity and the processing abilities of neural networks.

In the article, "Learning to Segment Document Images, K S Seshkumar, Anoop Namboodiri and C V Jawahar [9], the authors used a hierarchical frame work for document segmentation as an optimization problem. The model incorporates the dependencies between various levels of the hierarchy unlike traditional document segmentation algorithms. This framework is applied to learn the parameters of the document segmentation algorithm using optimization method like gradient descent and Q-learning. The novelty of their approach lies in learning the segmentation parameter in the absence of ground truth.

For fractal image and image sequence compression, Lucia Vences and Isaac Rudomin[10] used Genetic Algorithm to compress an image. Here Genetic Algorithms are used for randomly generated population of Local Iterated Function Systems (LIFS) , the one whose attractor is the first frame in the sequence. "Collage theorem" is considered as the foundation for the paper. There are different studies available on optimizing algorithms with the help of simple Genetic Algorithms (GA) [15] [17] [19] [22] and Parallel or Hybrid Genetic Algorithms [11] [12] [13] [14] [18] [20] [21] [23]. And studies like Probabilistic networks in image processing are also available [16]. Most of the available works are suitable for ordinary images. And many works are not suitable for document images due to the complex nature of document images. Some of the studies available are suitable for bi-level document images and they are not applicable for gray scale or color images. Genetic Algorithms are applied for binary as well gray scale images.

III. COMPRESSION OPTIMIZATION USING PARALLEL GENETIC ALGORITHM

Compression of Document images was very relevant for managing disk space for a long period. Now the storage cost is reduced and the bottle neck lies at transmission time. And due to advancement in technology it is easy to retrieve information and store files without much delay. But considering the internet traffic, it is relevant to reduce the file size for transferring information with high speed. And most of the internet service providers fix their rate for certain GB of downloadable data. In such situations, if there are efficient lossless compression algorithms available to compress document images significantly, then it will help to reduce congestion in the network. The structure of the document image vary from language to language, context to context, content to content etc. Even though it contain certain common features like line spacing, word spacing, column spacing etc. there are certain difficulties and issues in identifying regions, i.e. certain difficulties in identifying boundaries, sometimes not exact boundaries, blurred edges, incomplete edges etc. In order to compress the data with ordinary algorithms, it is difficult to deal with large two dimensional space of document images. The compression technique using genetic algorithms helps to improve the scanning procedure of document images and store the image

in a compressed format at the scanner level itself. And most of the available algorithms are for ordinary images, not for document images. So this work is an attempt to use parallel genetic algorithm for the lossless compression of document images. In 1960s John Holland proposed a method known as Genetic algorithm, which is useful and efficient when

- The search space is large, complex or poorly understood
- The domain knowledge is scarce or expert knowledge is difficult to encode to the narrow search space.
- No mathematical analysis is available.
- Traditional search methods fail.

Genetic Algorithms are a particular class of evolutionary search algorithms for global optimization. A genetic algorithm is designed by two components

- Genetic representation of the problem domain
- A fitness function to the problem domain

A. Representation of the problem

A document is scanned and the image is stored as a "*.jpeg" file. (Fig:1) The image is represented as Quad Tree by recursively applying the following method for every quadrant:

- 0 to |(m-1)/2) | and 0 to |(n-1)/2| → first quadrant
- 0 to |(m-1)/2) | and |(n-1)/2|+ to n → second quadrant
- |(m-1)/2| + 1 to m and 0 to |(n-1)/2| → third quadrant
- |(m-1)/2| + 1 to m and |(n-1)/2| +1 to n → fourth quadrant (Fig :2)

Recursion stops when desired size of leaf node is obtained. The leaf node of the quad tree is a table with gray scale values. When a leaf node of 16x16 size or 32x32 size is obtained, each matrix is considered as a ***population***. In a population a row is assigned as a ***chromosome***. A chromosome is formed with multiple or single genes. (Fig:3). When gene size increases the compression ratio increases and the quality of reconstructed image decreases.

B. Designing the fitness function

In our method, the distance between two chromosomes X and Y are computed using Hausdorff Distance (d_H):

$$\text{DISTANCE} = d_H(X, Y)$$

$$d_H(X,Y) = \max\{\sup_{x \in X} \inf_{y \in Y} d(x,y), \sup_{y \in Y} \inf_{x \in X} d(x,y)\}$$

$$\text{FITNESS} = 1 - \text{DISTANCE}^{1/2}$$

FITNESS is in the range 0 TO 1

Two chromosomes from a chosen population are compared and the fitness between the chromosomes is computed. The chromosome with highest fitness value is chosen.

C. The Method and Operations

Mating, Crossover and mutation are applied with Constrained Run Length Encoding compression algorithm [10]. In this work multiple populations are considered simultaneously. Multiple chromosomes from each population are considered for achieving massive parallelism. A quantized index table is kept for deciding encoding index of the chromosomes. This index representation reduces the size of the encoded image.

THE ALGORITHM:

1. Select n populations
2. For each population select k chromosomes (repeat the following steps 3 through 6 until all chromosome and all populations are considered.)
3. Apply a genetic operation
4. Evaluate fitness using Hausdorff distance
5. Find out the chromosomes with highest fitness.
6. Get the index value of respected chromosome from index table. Store the index.
7. End;
8. End;

IV RESULTS AND CONCLUSIONS

Parallel genetic algorithm's execution time is better than ordinary GA. (table 1).

ACKNOWLEDGMENT

This work is done as a part of our research pursuing at Department of Computer Applications, Cochin University of Science and Technology, Cochin. We extend our wholehearted gratitude to our colleagues who encouraged us in the work.

REFERENCES

[1] Murshed, Nabeel A; Bortolozzi, Flavio; Sabourin, Robert " Binary image compression using identity mapping backpropagation neural network" Proc. SPIE Vol. 3030, p. 29- 35, Applications of Artificial Neural Networks in Image Processing II, Nasser M. Nasrabadi; Aggelos K. Katsaggelos; Eds. (SPIE Homepage) March 1997

[2] Patrice Y Simard, Henrique S Malvar, James Rinker and Erin Renshaw, "A Foreground/ Background separation Algorithm for Image compression," Proc. of the Data compression connference, 2004, IEEE Computer Society , 2004.

[3] A Seropian , M.Grimald, Dr N Vincent, "Differentiation of alphabets in handwritten texts" Proc. of the 17th International Conference on Pattern Recognition- ICPR 04, IEEE Computer Society.

[4] Mehmet Utku Celik, and Akmat Murat Tekalp, "Lossless Generalized Data Embedding", IEEE Transactions on Image processing vol14, No:2 February 2005.

[5] B F Wu, C C Chin and Y L Chan " Algorithm for compressing compound document images with large text/background overlap" IEEE proc. Vis, Image Signal processing vol, 15 L, No:6 December 2004

[6] Soo-Chang Pei and YU-Ting Chuang "AutomaticText Detection Using Multi Layer Color Quantization in Complex Color Images" IEEE International Conference on Multimedia and Expo(ICME) 2004

[7] Hujuan Yang and Alex C Kot, " Text Document Authentication by Integrating inter character and word spaces

water marking", IEEE International Conference on Multimedia and Expo(ICME) 2004.

[8] http://en.scientificcommons.org

[9] K S Seshkumar, Anoop Namboodiri and C V Jawahar" Learning to Segment Document Images", Springer_Verlag, Berlin Heidel Berg 2005.

[10] Lucia Vences and Isaac Rudomin," Genetic Algorithms for Fractal Image and Image Sequence Compression", Computer Visual 1997.

[11] Sven E Eklund, " A massively parallel architecture for distributed genetic algorithms", Elsevier: Parallel Computing, 30, 2004, 647-676.

[12] Seth Bacon, " A brief Overview of Parallel Genetic Algorithms"

[13] Yong Fan, Tianzi Jiang and David J Evans," Volumetric Segmentation of Brain Images Using Parallel Genetic Algorithms", IEEE Transactions on Medical Imaging, vol 21, No:8, August 2002.

[14] Qizhi Yu, Chong Cheng Chen and Zhigeng Pan, " Parallel Genteic Algorithm on Programmable Graphics h ardware", zhejiang university, Fuzhons University, P R China.

[15] Mantas Paulinas, Andrius Usinskas, " A Survey of Genetic Algorithm Application for Image Enhancement and Segmentation",ISSN 1392-124x Information Technology and Control, 2007, vol,36, No:3.

[16] Andras Barta and Istvan Vajk,, " Document Image Analysis by Probabilistic Network and Circuit Diagram Extraction", Informatic 20, 2005, 291-301.

[17] Kazunon Otobe, Keitanaka and Masayuki hirafuji," Knowledge Acquisition on Image Processing based on Genetic Algorithms", Proc. of the IASTED International Conference Signal and Image Processing, October 28-31, 1998, Las Vegas, Nevada, USA.

[18] Enrique Alba, Fransico Luna, Antoni O J NeBro," Advances in Parallel Heterogeneous Genetic Algorithm for Continuous Optimization", Int.Appl.MNath.Con.Sui,2004,vol 4 No:3, 317-333.

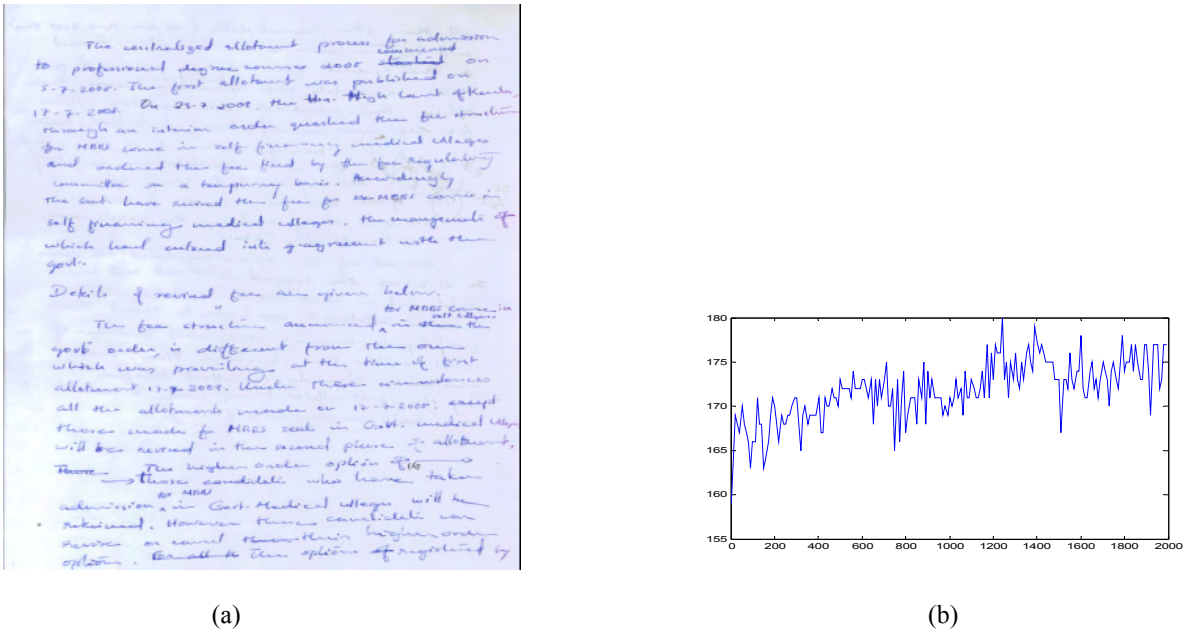
[19] Susmita Ghosh, " Incorporating Ancestors Influence in Genetic algorithms", Applied Intelligence 18, 7-25, 2003.

[20] Dudy Lim, Yew-Soon Ong, Yoochu Jim, Benhard Sendhoff, Bu-Sung Lee, " Efficient Hierarchical Parallel Genetic Algorithm using Grid Computing", Elsevier October 2006.

[21] R Nedunchelian K Koushik, N Meiyappan, V Raghu, " Dynamic task scheduling using parallel Genetic Algorithms for Hetrogeneous Distributed Computing"

[22] K.Otobe, K.Tanakan and M Hirafuji, " Image Processing and Interactive Selection with Java based on Genetic Algorithms", Computational Modelinf Lab, Japan.

[23] Erik Canta-Paz, "A survey of Parallel Genetic Alorithm", Dept of Computer Science and Genetic Algorithm Laboratory, Illinois.



(a) (b)

Figure 1. (a) original hand written image and (b) The intensity plot of the image

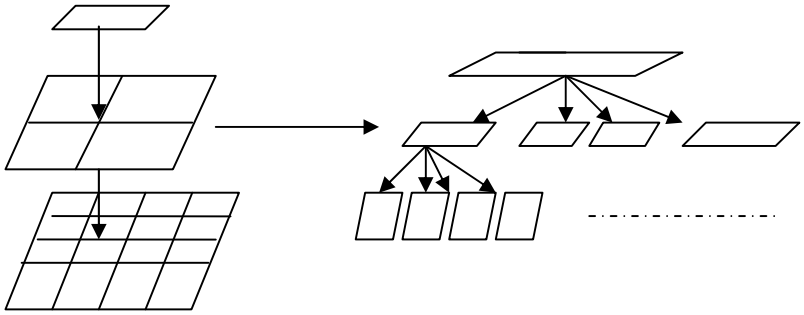


Figure 2. Quad tree generation of a document image.

42	252	232	252	54	66	47	48
0	32	253	34	34	34	34	54
22	31	35	45	32	65	45	56
43	34	38	56	12	67	67	67
56	45	36	76	23	87	56	78
76	46	37	87	34	89	45	87
88	47	76	45	23	90	65	45
90	56	65	46	35	45	78	34

(a)

42	252	232	252	54	66	47	48

(b)

42	252	232	252

(c)

Figure 3 : a) A sample population of size 8x8 b) a chromosome from the population of length 8 c) a gene of length 4

TABLE I. ANALYSIS OF THE RESULTS

Sl no	Name of image	size	PSNR dB	Compression ratio	Time in sec for ordinary GA	Time in Sec for Parallel GA
1	Doc001.jpeg	1566x1102x3	27.23	5:1	273	162
2	Doc002.jpeg	1346x1124x3	26.38	4:1	251	148
3	Doc003.jpeg	1324X1096x3	26.34	4:1	247	146
4	Doc004.jpeg	1545x1134x3	27.04	4:1	248	175
5	Doc005.jpeg	1200x1014x3	26.06	5:1	245	142
6	Doc006.jpeg	1376x1120x3	26.28	4:1	249	147
7	Doc007.jpeg	1406x1224x3	27.03	4:1	273	169
8	Doc008.jpeg	1488x1136x3	26.47	4:1	269	164