JOIN (Jurnal Online Informatika)
Volume 4 No. 2 | December 2019: 75-78
DOI: 10.15575/join.v4i2.330

Lossless Text Image Compression using Two Dimensional Run Length Encoding

Hilal H Huda
Telkom University, Bandung, Indonesia
hilalnuha@telkomuniversity.ac.id

Abstract-Text images are used in many types of conventional data communication where texts are not directly represented by digital character such as ASCII but represented by an image, for instance facsimile file or scanned documents. We propose a combination of Run Length Encoding (RLE) and Huffman coding for two dimensional binary image compression namely 2DRLE. Firstly, each row in an image is read sequentially. Each consecutive recurring row is kept once and the number of occurrences is stored. Secondly, the same procedure is performed column-wise to the image produced by the first stage to obtain an image without consecutive recurring row and column. The image from the last stage is then compressed using Huffman coding. The experiment shows that the 2DRLE achieves a higher compression ratio than conventional Huffman coding for image by achieving more than 8:1 of compression ratio without any distortion.

***Keywords*- Lossless compression, Huffman Encoding, Run-length Encoding, Text Image**

I. INTRODUCTION

One of the earliest activity of producing digital version of classical documents can be found in Project Gutenberg [1]. The project reached more than 50,000 of digitized documents. One of the earliest version of the digitized documents was stored as text images since the main target of the project is to digitize books. Text images are commonly used in conventional communication where texts are not directly represented by digital character such as ASCII but represented by an image. The example for this file is a facsimile file or a scanned document.

Image compression becomes a challenging research area since there are so many handheld devices that include image capturing as their feature constrained by their small storage. Text image compression itself has not gained much attention despite its importance for company or government document. Text image documents can be generalized as binary or monochrome image. There are some works related to this. GRID method has been proposed in [2] for text image lossy compression using Huffman coding and 4x4 pixels area as symbol. Block Arithmetic Coding for Image Compression (BACIC) Algorithm [3] succeeded to compress up to 16:1 compression ratio using arithmetic coding by estimating 12 previous bits with certain pattern. Class and tree-based symbol dictionary design as proposed in [4] works very well for lossless and lossy compression. Lossy compression proposed in [5] exploits key structural information to maintain the smooth contour in bi-level image. In[6], Guo et.al proposed a cost function in Bayesian framework with dictionary learning for text document compression.

In this paper, we propose a data compression technique for high resolution text images. In a high-resolution text image, one can see the printed letter as combination of horizontal and vertical straight line such that each row will have high possibility of having similarity to the next row. We wish to exploit its row-by-row and column-by-column correlation for data compression by extending the idea of Run Length Encoding (RLE). Since each row is highly correlated to the next row and text images are represented with only two colors, the similarity between row must be very high. Having row and column as symbols, each contains binary array that can be further compressed using statistical coding technique such as Huffman coding. The proposed method is further referred to as two dimensional RLE with Huffman coding (2DRLE).

The rest of this work is organized as follows. Section II describes the proposed method and dataset. The experimental results are discussed in Section III. Finally, Section IV provides the conclusions.

II. METHOD

In this section, we describe the proposed steps to achieve lossless compression for text images and the data set for evaluation.

a. Proposed Method

The proposed 2DRLE consists of three steps namely, row scanning, column scanning and entropy coding. The details of each step are given as follows:

- Row scanning
  We follow the idea of RLE [7], [8] where the each symbol is taken from row of the image instead of pixel, character, or number as symbol. Each row in the image is scanned and compared with the following row where each consecutive repeating rows will be stored as a single row. Figure 1 illustrates the process where each row will be treated as a symbol. The symbol set can be regarded as a row-wise non-repeating version of the original image. The number

Received: July 14, 2018; Revised: September 18, 2019; Received: September 25, 2019; Published: February 14, 2020

ISSN 2527-1682 (Print)
ISSN 2527-9165 (Online)

of repetition of each row is also stored for decompression process.

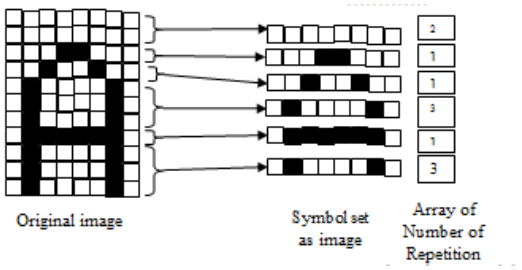


Figure 1. Row Scanning Illustration

- Column scanning
  The image without repeating row resulted from previous step is scanned column-wise and compared with the following column. Each consecutive repeating column will be represented without repetition and its number of column repetition array is stored for decompression purpose. The rest procedure is the same as row scanning but applied column-wise. Intuitively, both row and column scanning will mostly exploit repetition of the white spaces. This step is illustrated by Figure 2. The resultant of this step is a binary image without consecutive recurring row and column. This image is referred to as a symbol set since it contains the all symbols from row and column.

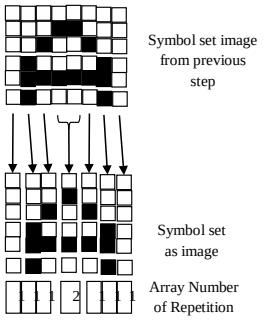


Figure 2. Column Scanning Illustration

- Huffman coding for image
  Since the symbol set is also a two dimensional binary array, we can see it as an array of symbols where the each symbol is a non-overlapping segment of the image. After obtaining number of occurrences of each symbol as illustrated by Figure 3, we can construct the code-words using Huffman coding.

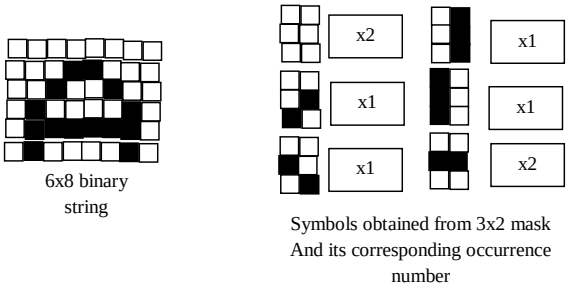


Figure 3. Symbol and its probability calculation

b. Estimated Compression Ratio

Consider a binary text image with horizontal size w pixels, vertical size h pixels, and two-color level, the image size is equal to wh bit

$$L=wh \quad (1)$$

where L denotes the number of bits in the image whereas the 2DRLE method reduces the number of bits as follows. The maximum image file size after applying RLE row and column wise is equal to

$$L=w_r\, h_r+w_r\, \lceil \log_2(w_r) \rceil+h_r\, \lceil \log_2(h_r) \rceil. \quad (2)$$

where w_r and h_r denote the width and height of the symbol set. The dimension of the symbol set is further decreased after implementing the Huffman coding.

c. Data Set

In our experiment, we use Scanned paper PBM CCITT 1 as used in [3] whose dimension is 2376x1728=4105728 bits. One can see in Figure 4 that the test image contains large amount of white space.

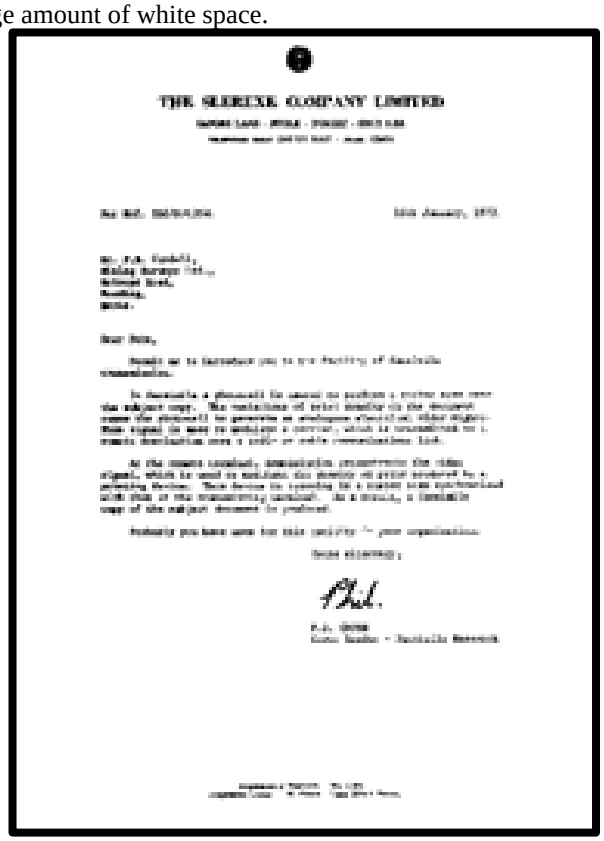


Figure 4. Image Test CCITT 1

III. RESULT AND DISCUSSION

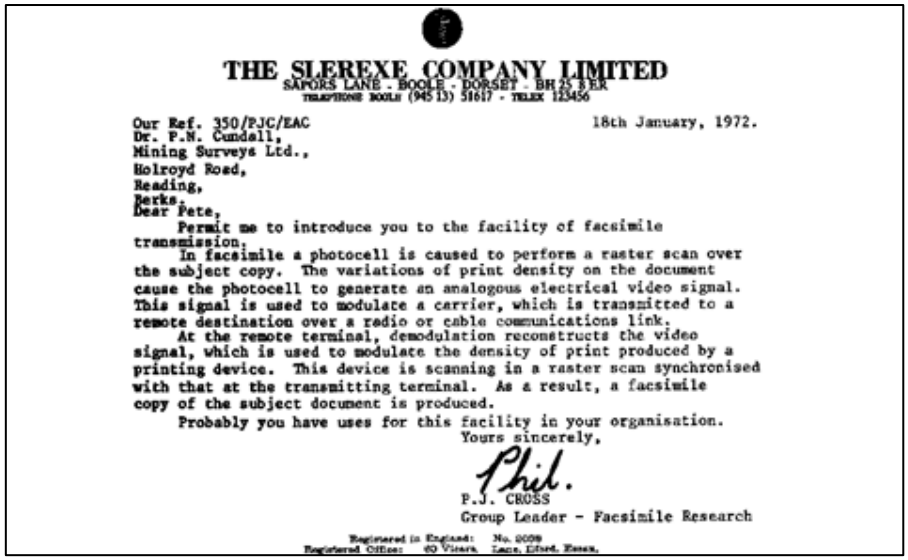
In this experiment, we apply regular Huffman coding only to test image. Original image size is 4105728 bits. The following Table shows how the mask sizes affect the compression size. The larger mask size the higher compression ratio we will obtain. However, we are constrained by the number of symbol generated. 4x4 mask will generate 2^(4×4)=2^16symbols or 65536 codewords

which is very large and computationally inefficient. The regular Huffman coding achieves the best performance at 2x4 mask with a compression ratio of 7.75:1.

Table 1. Huffman Coding with various Mask Size

Mask Size	2x1	1x2	2x2	2x4	4x2
Compr. File Size	2052864	2052864	1026432	513216	513216
Dictionary Size	27	27	440	16254	17703
Total Comp ratio	2052891 2	2052891 2	1026872 3.9983	529470 7.7544	530919 7.7332

To evaluate our proposed method, we implement the 2DRLE by applying the row and column scanning followed by Huffman coding to the test image as proposed on section II.



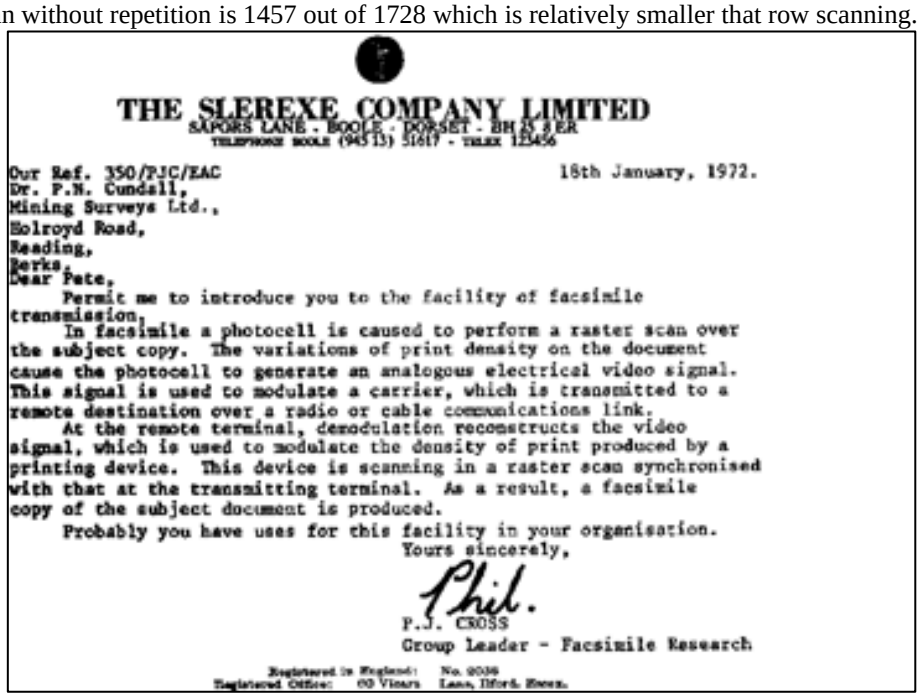
THE SLEREXE COMPANY LIMITED
SAPORS LANE · BOOLE · DORSET · BH25 8ER
TELEPHONE BOOLE (945 13) 51617 · TELEX 123456

Our Ref. 350/PJC/EAC 18th January, 1972.
Dr. P.N. Cundall,
Mining Surveys Ltd.,
Holroyd Road,
Reading,
Berks.
Dear Pete,
Permit me to introduce you to the facility of facsimile transmission.
In facsimile a photocell is caused to perform a raster scan over the subject copy. The variations of print density on the document cause the photocell to generate an analogous electrical video signal. This signal is used to modulate a carrier, which is transmitted to a remote destination over a radio or cable communications link.
At the remote terminal, demodulation reconstructs the video signal, which is used to modulate the density of print produced by a printing device. This device is scanning in a raster scan synchronised with that at the transmitting terminal. As a result, a facsimile copy of the subject document is produced.
Probably you have uses for this facility in your organisation.
Yours sincerely,
Phil.
P.J. CROSS
Group Leader - Facsimile Research

Registered in England: No. 0009
Registered Office: 60 Victoria Lane, Ilford, Essex.

Figure 5. Symbol set as original Image without row repetition only

Figure 5 shows how row scanning terminates repetition in vertical direction. Most repetition are white space but there are some non-white space or part of character included in repetition. After row scanning the number of row without repetitions is 1057 out of 2376. Figure 6 shows how column scanning partly eliminates repetition in horizontal direction. Unlike row scanning, column scanning, in our test image, only includes white space. One can see that the white space on the right side of the image is not removed. This is due to some tiny black pixel noise one the right side of the image. The number of column without repetition is 1457 out of 1728 which is relatively smaller that row scanning.



THE SLEREXE COMPANY LIMITED
SAPORS LANE · BOOLE · DORSET · BH25 8ER
TELEPHONE BOOLE (945 13) 51617 · TELEX 123456

Our Ref. 350/PJC/EAC 18th January, 1972.
Dr. P.N. Cundall,
Mining Surveys Ltd.,
Holroyd Road,
Reading,
Berks.
Dear Pete,
Permit me to introduce you to the facility of facsimile transmission.
In facsimile a photocell is caused to perform a raster scan over the subject copy. The variations of print density on the document cause the photocell to generate an analogous electrical video signal. This signal is used to modulate a carrier, which is transmitted to a remote destination over a radio or cable communications link.
At the remote terminal, demodulation reconstructs the video signal, which is used to modulate the density of print produced by a printing device. This device is scanning in a raster scan synchronised with that at the transmitting terminal. As a result, a facsimile copy of the subject document is produced.
Probably you have uses for this facility in your organisation.
Yours sincerely,
Phil.
P.J. CROSS
Group Leader - Facsimile Research

Registered in England: No. 0009
Registered Office: 60 Victoria Lane, Ilford, Essex.

Figure 6. Symbol set as original Image without row and column repetition

Finally, we can see the comparison of performance of various compression techniques on Table II for the same test image. The 2DRLE achieves a better compression ratio than traditional Huffman coding. However, our proposed method is still outperformed by BACIC technique.

Table 2. Comparison with other compression method

Coding Technique	Mask Size	Compression ratio
Huffman Coding	2x4	7.7544
Row+RLE	-	2.2376
Row+Col+RLE	-	2.6461
Row+Col+RLE+Huffman	7x1	8.82
BACIC	12	17.6

IV. CONCLUSION\

Text image can be viewed as set of horizontal and vertical lines. We can exploit that fact for text image compression. In this paper, we proposed a two dimension RLE combined with Huffman coding (2DRLE). Using the proposed scheme for eliminating row repetition, our experiment shows that there are non-white space row-repetition that is eliminated. It means that there are part of character that is compressed. Our proposed method also achieves a compression ratio up to 8.82:1 which outperforms regular Huffman coding for CCITT 1 as the test image. In the future, we wish to test our algorithm to larger size of test images to see the behavior of text in a large size font. Using various size and shape of pixel combinations as symbol is also interesting ideas to be explored for RLE and Huffman coding implementation.

V. ACKNOWLEDGEMENT

The author would like to thank to Prof. Azeddine Beghdadi from L2TI - Institut Galilée - Université Paris 13 – Sorbonne Paris for the useful discussion.

VI. REFERENCES

[1] M. Hart, “The history and philosophy of Project Gutenberg, in Project Gutenberg,” *Proj. Gutenb.*, pp. 1–11, 1992.

[2] A. Broder and M. Mitzenmacher, “Pattern-based compression of text images,” *Data Compression Conf. Proc.*, pp. 300–309, 1996.

[3] M. D. Reavy and C. G. Boncelet, “An algorithm for compression of bilevel images,” *IEEE Trans. Image Process.*, vol. 10, no. 5, pp. 669–676, May 2001.

[4] Y. Ye and P. Cosman, “Dictionary design for text image compression with JBIG2,” *IEEE Trans. Image Process.*, vol. 10, no. 6, pp. 818–828, Jun. 2001.

[5] M. Reyes, X. Zhao, D. Neuhoff and T. Pappas, “Lossy compression of bilevel images based on Markov random fields,” *IEEE Int. Conf. Image Process.*, vol. 2, 2007.

[6] J. A. and C. B. Y. Guo, C. Lu, “Model-based iterative restoration for binary document image compression with dictionary learning,” *IEEE Conf. Comput. Vis. Pattern Recognition, Honolulu,* 2017.

[7] E. L. Hauck, “Data compression using run length encoding and statistical encoding,” *U.S. Pat. Pat. 4,626,829,*.

[8] V. Watson, “Run-length encoding,” *US Pat. Pat. 10/143,542.,*.