Page Segmentation and Document Layout Analysis for Scanned Image by using Smearing Algorithm

Nay Win Htun , Lin Min Ko
naywinhtunmec@gmail.com, linminko@gmail.com
Computer University (Maubin)

Abstract

This paper presents a feature-based system which utilizes domain knowledge to segment and classify scanned image documents. Documents usually consists of a mixture of text and image. Text block possesses an interesting property that the x-profile or y-profile of text block is a periodic pattern. Image block possesses generate the connectivity histogram by summing the number of dark pixels with the same connectivity value. Initially, one-scan run-length smearing algorithm (RLSA) with block merging is proposed to segment the document. After segmentation process, the next task is to classify the segmented block. The classification task is then performed based on the rules induced from the features or primitives associated with each document. In this system, proper use of domain knowledge is proved to be effective in accelerating the segmentation speed and decreasing the classification error.

Keywords: one-scan run-length smearing, block merging, connectivity histogram, text block, image block.

1. Introduction

In a world changing technology, it can be reassuring to know that something remains constant; that is the need for newer and better technology. While traditional text can no longer satisfy the users, the emerging of multimedia fills this hole in good time. As more users catch on to the pleasure of splendid multimedia document, they've got tired of the plain text-only document. It's apparent that the evolution of multimedia technologies helps propel the multimedia system to a decided edge over its rival (traditional text-only system). In this paper, we will study and develop a state-of-art multimedia technology that is the segmentation and classification of the document.

In a document, the information revealed by the document is composed of text and image. The RLSA algorithm for automatic separation and classification of text and image is advantageous in reproducing, transmitting, and storing the document. With the separation and classification tasks done, the next task is to recognize the separated text, compress the separated image. With all these done, the storage space and transmission time will be greatly reduced as the processed results need much less storage that the original forms.

The emergence of multimedia research is not very long. In 1982, Wong[5] proposed a document analysis system to process technical document. He developed a constrained run length smearing algorithm to segment the document into several blocks and classify these blocks according to their geometrical properties. In 1984, Nagy [3] proposed another document segmentation algorithm called recursive x-y cut algorithm to do the task. Toyoda[4] presented a study about the segmentation of Japanese newspaper by adopting the synthetic method to analyze the structure of Japanese newspaper. As to the block classification, Srihari[9] proposed a texture analysis based algorithm to classify text and image. However, this algorithm cannot classify horizontal line and vertical line. Recently, Stephen[10] presented a newspaper reading system by utilizing four filters to classify segmented blocks into different types of document. The evaluation of the page segmentation algorithms was done on the University of Washington III (UW-III) database [12]. The database consists of 1600 English document images with manually edited ground-truth of entity bounding boxes. These bounding boxes enclose text and non-text zones, text-lines and words. We used the 978 images that correspond to the UW-I dataset pages.

Thus, we will present a document processing system which can automatically separate and classify the text and image embedded in the document and analyze the structure of each separated document. A modified one-scan run-length smearing algorithm with block merging techniques is first utilized to segment the document. The advantage of our segmentation algorithm is that it only needs one scan in either x or y direction resulting in the tremendous reduction of processing time.

In this paper, we divide four stages for the system. The typical stages have the following structures: the first stage of this system is preprocessing stage, which includes binarization process, noise removal process and skew correction process. The second stage is document segmentation stage by using RLSA algorithm with block merging techniques. The third stage is document classification stage, which includes classification of text and image. The last stage is

the implementation stage of page segmentation and document layout analysis for scanned image.

2. Preprocessing

In this system, the preprocessing stage in document understanding primarily involves the following processes: binarization process, removal of noise and correction of skew that is introduced in original document image during scanning /acquisition. We look into each of these problems and present commonly used techniques in this section. Some of these preprocessing algorithms could also be implemented as part of other modules during layout and structure analysis. However, pre-processing algorithms that are more specific to other modules of document understanding, such as character recognition are not studied here. Such methods might include binarization of a gray-level image and scaling of characters to a standard size.

2.1 Binarization

Initially, in this system, binarization is the process of converting a gray scale image (0 to 255 pixel values) into binary image (0 to 1 pixel values) by selecting a global threshold that separates the foreground from background. Each pixel is compared with the threshold and if it is greater than the threshold it is made 1 or else 0. Here a threshold of 128 is chosen as shown in Figure 1.

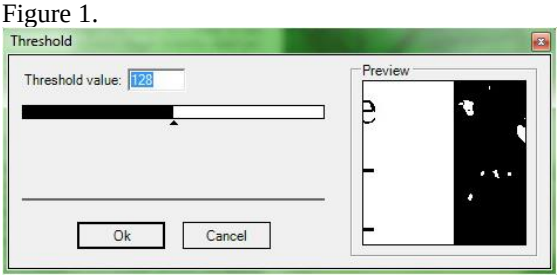

Figure 1: Binarization process for threshold value

$$Y(x,y) = 1 \quad \text{if } X(x,y) > T \qquad (1)$$
$$Y(x,y) = 0 \quad \text{if } X(x,y) <= T$$

We calculate the threshold values by using Equation (1) on the original document. Thus, we let that X is original document and Y is binarized document and x , y are the pixel values of the original document.

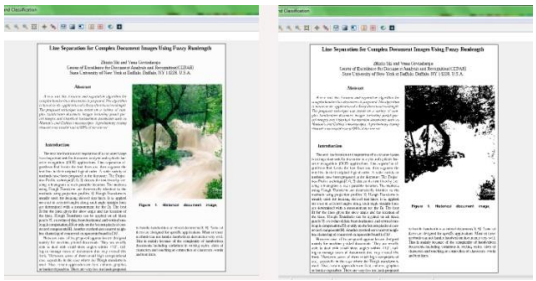

Figure 2: Original Document and Binary Document

2.2 Noise Removal

Noise is a common problem in most of the image understanding problems. These involve white noise that is introduced by interferences in the sensors and amplification circuits of the digitization mechanism (scanner, camera).The common sources of noise include white noise, salt and pepper noise, quantization artifacts, etc. The noise introduced during scanning or due to page quality has to be cleared before further processing. For this the document is scanned for noise using a moving 5 x 5 window. If all non-zero pixels are confined to the central 3 x 3 section, all those pixels are set to 0.

2.3 Skew Correction

Skew is introduced in a document image when a document is scanned or imaged at an angle with respect to the reference axes. The problem of skew correction plays an important role in the effectiveness of many document analysis algorithms, such as text line estimation, region boundary detection, etc. For example, algorithms based on projection profiles assume an axis-aligned scan. The primary challenge in skew correction is the estimation of the exact skew angle of a document image. This noise removal process and skew correction process works if these process is require for the document.

3. Document Segmentation

The purpose of document segmentation is to segment the document into several blocks with each block representing one type of document. The method used in segmenting for a document is the one-scan run-length smearing algorithm with block merging. The Run-Length Smearing algorithm (RLSA) operates on the document where any two nonadjacent segments with distance smaller than a preselected threshold are smeared and treated as a merged segment. The algorithm is first applied row by row and then column by column to obtain two intermediate results. The two intermediate results are then combined by performing a logical AND operation to generate the final result. Last, the connected component algorithm is applied to find the blocks. Our proposed document segmentation algorithm first performs the smearing operation either in the horizontal direction or vertical direction. The block merging technique is then applied to form the blocks. The following sub-sections are the detail descriptions of our proposed document segmentation algorithm.

3.1 Run-Length Smearing Algorithm

After preprocessing stage, the run-length smearing algorithm (RLSA) works on this binary image where white pixels are represented by 0's and black pixels by 1's. The algorithm transforms a binary sequence x into y according to the following rules:

(i) 0's in x are changed to 1's in y if the number of adjacent 0's is less than or equal to a predefined threshold C.

(ii) 1's in x are unchanged in y.

These steps have the effect of linking together neighboring black areas that are separated by less than C pixels. The RLSA is applied row-wise to the document using a threshold Ch, and column-wise using threshold Cv , yielding two distinct bitmaps. These two bitmaps are combined in a logical AND operation. Additional horizontal smearing is done using a smaller threshold, Cs, to obtain the final bitmap. Then, connected component analysis is performed on this bitmap, and using threshold C21 and C22 on the mean run of black pixel in a connected component and block height, connected components are classified into text and non-text zones. Smearing is an operation to connect two nonadjacent segments into one merged segment if the distance between these two segments is smaller than a threshold. The smearing operation is performed in either horizontal or vertical direction depending on the arrangement of this binary document. Suppose that the smearing operation is performed in the horizontal direction one line at a time, two segments having distance less than a threshold will be merged into one segment. To illustrate the smearing operation for this binary image, we consider the following example with threshold value being selected as 4.

If (X(x,y) = 0)

$$Y(x,y) = 0 \quad \text{if } X(A) > T \qquad (2)$$
$$Y(x,y) = 1 \quad \text{if } X(A) <= T$$

We let that A is number of adjacent 0 , T is threshold value , X is before smearing stage , Y is after smearing stage and x,y are the pixel values of the document and T is threshold value.

before smearing

1 1 1 1 1 1 1 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 1 1

after smearing

1 1 1 1 1 1 1 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1

In this binary image, with the smearing operation being performed, each segment can be further merged with the segments in the previous row to form a larger segment according to the connectivity property. Suppose the one-scan run-length smearing operation is performed in the horizontal direction, two segment s_1 and s_2 are merged if the projections of these two segments in the vertical direction overlap.

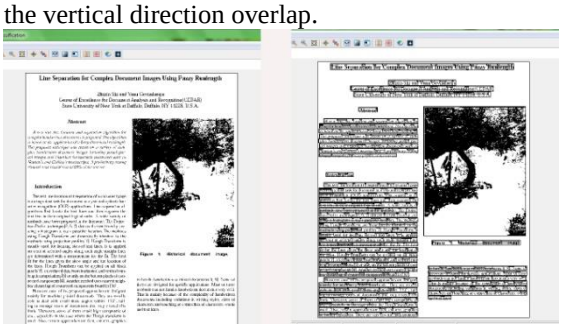

Figure 3: Binary Document and Viewing Document after Employing Run-Length Smearing Algorithm

3.2 Block Merging

In this section, run-length smearing algorithm needs two scans in both the horizontal and vertical directions to achieve the segmentation goal. As we know, the scanning process is very time-consuming. To avoid the double scan problem, we propose the block merging technique to replace the second scan. The principles of block merging are quite simple. After using RLSA algorithm, two blocks are merged if they satisfy the following conditions on this viewing document after employing RLSA algorithm as shown in Figure 4.

(i) The length of these two blocks are nearly equal.

(ii) The distance between these two blocks is smaller than a preselected threshold.

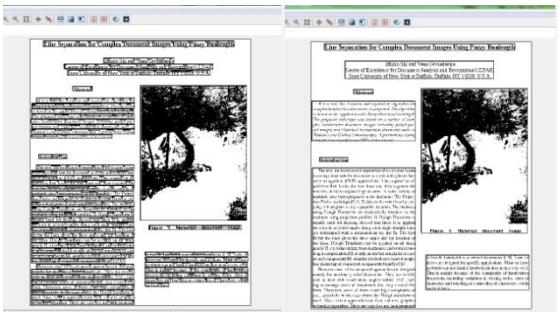

Figure 4: Document after Employing Run-Length Smearing Algorithm and Viewing Document after Employing Block Merging

4. Document Classification

With the document segmentation being done, the next task is to classify the segmented block. In this section, we will discuss how to classify the segmented block. The proposed solution is a feature-based system that is the classification is made based on the features associated with the block.

(i) Classification of Text: Text block possesses an interesting property that is the x-profile or y-profile of text block is a periodic pattern. After segmentation process, the classification of text can be achieved according to the periodic property

inherent to the text. The text classification process, the y-profile of the considered block exhibits periodic phenomenon as shown in Figure 5, therefore we can declare the considered block as a text block.

Figure 5: Periodic Behavior of Text Block

(ii) Classification of Image: To achieve the classification goal, we develop an operator to evaluate the dark pixel distribution presented in the block. Bellows are the description of the proposed method. First, an *n* x *n* mask centered at each dark pixel is built. Then, calculate the number of dark pixels connected to it. Each dark pixel will attain a value indicating the number of dark pixel connected to it. An example illustrating the computation of connected dark pixel with *n* equaling 3 is shown in Figure 6. Last, generate the connectivity histogram by summing the number of dark pixels with the same connectivity value. If the histogram is heavily distributed in the right part as shown in Figure 7, then the considered block is classified as an image block.

Figure 6: The Computation of Connectivity Dark Pixel with n equaling 3.

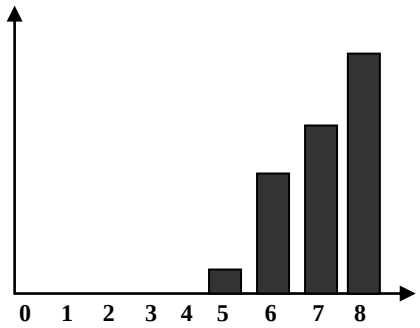

Figure 7: Connectivity Histogram of Image

5. Implementation of the System

In this paper, we present a document processing system which automatically separates and classifies the text and image embedded in the document and analyzes the structure of each separated document is shown in Figure8. The document image obtained by the scanning of a hard copy magazine document. The system can be accept five file formats type are .jpg , .png , .tif , .bmp and .gif . Preprocessing stage consists of three steps: binarization, noise removal and skew correction. Binarization is the process of converting a gray scale image into binary image. Noise removal introduced during scanning or due to page quality has to be cleared before further processing. Skew is introduced in a document image when a document is scanned or imaged at an angle with respect to the reference axes.

After preprocessing stage, the next task is to segment the document. Run-Length Smearing Algorithm (RLSA) operates on the document where any two nonadjacent segments with distance smaller than a preselected threshold are smeared and treated as a merged segment, to obtain the blocks. The block merging technique is then applied to form the blocks. Two blocks are merged if the length of these two blocks is nearly equal and the distance between these two blocks is smaller than a preselected threshold.

Before segmentation stage, the next task is to classify the segmented block. The text classification process, the y-profile of the considered block exhibits periodic phenomenon, therefore we can declare the considered block is a text block and the connectivity histogram is heavily distributed in the right part, then the considered block is classified as an image block. In the classified output image consists of two blocks: text block and image block. Text block represent color is blue and image block represent color is red in Figure 9.

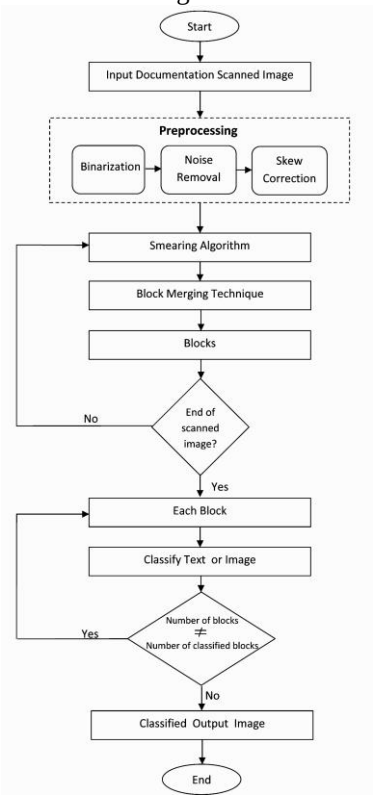

Figure 8: Overview of the System

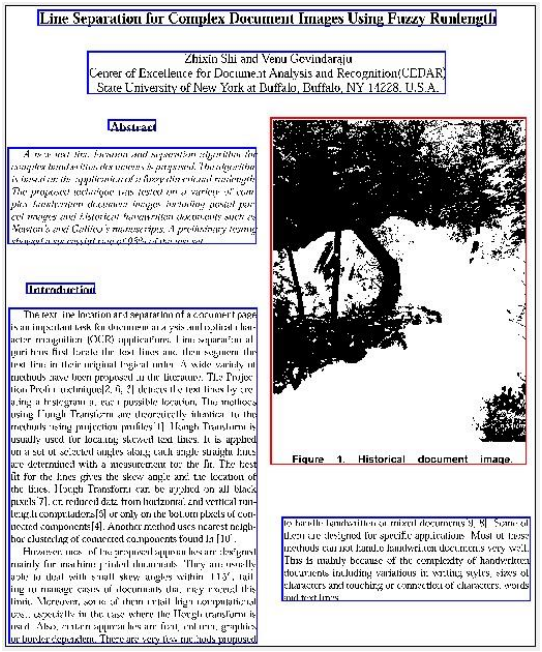

Figure 9: Output Image

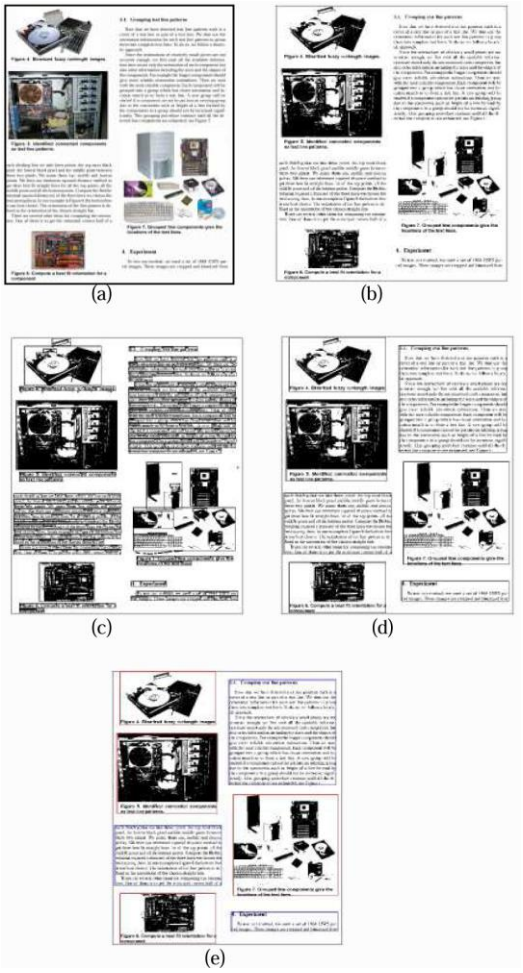

Figure 10: (a) Original Document, (b) Binary Document, (c) Document after Employing Run-Length Smearing Algorithm, (d) Document after Employing Block Merging , (e) Output Document

6. Experimental Results

This section presents some experimental results illustrating the segmentation and classification of document. Shown in Figure 2 is the original document and binary document. The document with text and image being segmented and classified are shown in Figure 2 and 3 respectively. Figure 10 shows another example illustrating the segmentation and classification of document.

7. Conclusion

In this paper, we have presented a method for the segmentation and classification of documents. This system works via one-scan run-length smearing algorithm and block merging, the document is segmented into several blocks with each block representing one type of documents. A feature-based classification algorithm is then employed to recognize the segmented block. This paper concludes with some experimental results illustrating the computer segmentation and classification of documents.

References

[1] Cattoni, R., Coianiz, T., Messelodi, S., Modena, C.M.: Geometric layout analysis techniques for document image understanding: a review. Technical report, IRST,Trento, Italy (1998)

[2] Das, A.K., Saha, S.K., Chanda, B.: An empirical measure of the performance of a document image segmentation algorithm. IJDAR 4 (2002) 183–190

[3] G. Nagy and S. Seth, " Hierarchical representation of optically scanned documents", Proc. Of 7th Int. Conf. on Pattern Recognition, pp. 347-349, 1984.

[4] J. Toyoda, Y. Noguchi, and Y. Nishimura, " Study of extracting Japanese newspaper article", Proc. Of 6th Int. Conf. on Pattern Recognition, pp. 1113-1115, 1982.

[5] K.Y. Wong, R.G. Casey, and F.M. Wahl, "Document analysis system", IBM J. Res. Develop, vol. 6, pp. 642-656 , November 1982.

[6] Kanai, J., Nartker, T.A., Rice, S.V., Nagy, G.: Performance metrics for document understanding systems. In: Proc. ICDAR, Tsukuba, Japan (1993) 424–427

[7] Liang, J., Phillips, I.T., Haralick, R.M.: Performance evaluation of document structure extraction algorithms. CVIU 84 (2001) 144–159

[8] Mao, S., Rosenfeld, A., Kanungo, T.: Document structure analysis algorithms: a literature survey. Proc. SPIE Electronic Imaging 5010 (2003) 197–207

[9] S.N. Srihari and G.W. Zack, " Document image analysis", Proc. of 8th Int. Conf. on Pattern Recognition, pp. 434-436, 1986.

[10] W.L. Stephen and S.N. Srihari, " Reading newspaper text", Proc. Of 10th Int. Conf. on Pattern Recognition , pp. 703-705, 1990.

[11] Yanikoglu, B.A., Vincent, L.: Ground-truthing and benchmarking document page segmentation. In: Proc. ICDAR, Montreal, Canada (1995) 601–604

[12] Guyon, I., Haralick, R.M., Hull, J.J., Phillips, I.T.: Data sets for OCR and document image understanding research. In: Handbook of character recognition and document image analysis, World Scientific, (1997) 779–799