Multimedia Documents Segmentation and Classification

Cho Thida Hlaing
University of Computer Studies, Yangon
chothidahlaing@gmail.com

Abstract

Document page segmentation and classification are important parts of the document analysis process. Page segmentation is that the page is decomposed into blocks. It is also called "page decomposition" or "zoning". The goal of page classification is to label these blocks according to their contents. In this paper, multimedia documents segmentation and document's features classification is presented. Multimedia documents usually consist of a mixture of text, graphic, and image. One-scan run-length smearing algorithm with block merging is emphasized for document segmentation task. This smearing algorithm is a document page segmentation algorithm using a bottom-up approach. Document classification task is performed based on features of text, graphic and image. Separation and classification of text, graphic, and image are advantageous in reproducing, transmitting storing the multimedia document and extraction different parts of document.

1. Introduction

In the majority of document image analysis and understanding applications, page segmentation is performed as the first step. Page segmentation is the identification of areas of interest in the image of a document page. After the areas of interest have been identified in the page image, they can be classified according to their contents [1].

There are three main methods for automatically document segmentation. They are bottom-up (or data driven) method, top-down (or model driven) method and hybrid method. In the bottom-up methods, a document is segmented into small blocks such as characters, and then merges them into bigger blocks as word and text lines. The Docstrum algorithm, the run-length smearing algorithm and one-scan run-length smearing algorithm are examples of the bottom-up methods. In the top-down methods, a document is segmented from large components (high-level) to smaller, more detailed, sub-components (lower-level) [9]. Hybrid method is a mixed method of top-down and bottom-up.

In this paper, 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. Each block represents one type of media. Features-based hierarchical classification algorithm is then employed to recognize the segmented blocks.

2. Document Segmentation

The purpose of document segmentation is to segment and separate text, image, and graphic embedded in the document. A modified one-scan run -length smearing algorithm with block merging technique is utilized to segment the document. The advantage of our segmentation algorithm is that it only needs one scan in either x or y direction he resulting in the tremendous reduction of processing time.

2.1. Run-Length Smearing Algorithm (RLSA)

The run-length smearing algorithm (RLSA) works on binary images where white pixels are represented by 0's and black pixel by 1's [4].The input image must be a clean and de-skewed bitmap. Smearing algorithm replaces 0's by 1's if the number of adjacent 0's by 1's if the number of adjacent 0's is less than or equal a given constraint [3] and [10]. This algorithm's processing steps are summarized as follow:

```
Input   :  binary image
Process:
        Step1:
          (1.1) Smearing operation is performed
                row by row
          (1.2) Smearing operation is performed
                column by column
          (1.3) Combine two results from step 1.1 and
        step 1.2 by performing a logical AND
        operation between two images.
    Step2: Last, the connected component
        algorithm is applied to find blocks.
    Output: Segmented block
```

2.2. One-Scan Run--Length Smearing Algorithm

RLSA needs two scans in both horizontal and vertical smearing to achieve the segmentation goal. Scanning process is very time consuming. To avoid the double scan problem of RLSA, block merging technique is proposed to replace the second scan [6]. This algorithm's processing steps are summarized as follow:

```
Input   :  binary image
Process :
    Step1: Smearing (horizontal or vertical
        direction)
    Step2: Block merging is performed according
        according to following two conditions:
        -The length of these two blocks are
          nearly equal.
        -The distance between these two
          blocks is smaller than a
          preselected threshold.
Output: Segmented blocks
```

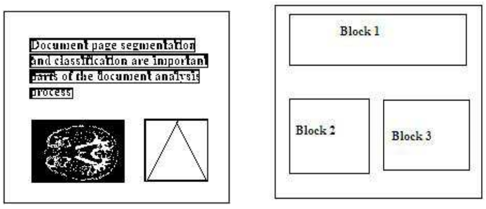
2.3. Smearing and Block Merging

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 [6]. Smearing can be performed on clean, de-skewed rectangular layout and binary images. Threshold values for smearing should be selected to minimize the following two conditions:

1. Words belonging to different columns are not connected and
2. Words belonging to different rows are not connected.

These threshold values usually depend on text size and resolution [3]. For example, let us horizontal smearing operation with threshold value 5.

Before smearing:
111110000011110000111111111
After smearing :
111110000011111111111111111


Document page segmentation
and classification are important
parts of the document analysis
process
Block 1
Block 2
Block 3

(a) **(b)**

Figure 1. (a) Document after smearing (b) Document after block merging

2.4. Segmentation Steps

```
Step1:
    For each row i
      For each column j
        Merge each black pixel to form character
        blocks using horizontal smearing threshold.
Step2:
    Merge each character block to form word
    blocks by using horizontal smearing threshold.
Step3:
    Merge each word block to form line blocks by
    using horizontal smearing threshold.
Step4:
    Merge each line block to form paragraph
    blocks by using horizontal smearing threshold.
```

Each block has its right, left, top and bottom.
.

3. Document Classification

After document segmentation, features in document are classified. In order to achieve the best possible results with OCR and storage, the contents of the document have to be examined [5]. Segmented blocks that are produced from document segmentation algorithm in Section 2.2, is fed to the media classification module. Features- based hierarchical classification algorithm is employed to classify the multimedia document into text, graphic, and image. Text can be detected according to periodic behavior. Graphic and image can be detected by using connectivity histogram.

3.1. Text Classification

Text block possesses a periodic pattern. The classification of text can be achieved according to the periodic property inherent to the text. X-profile or Y-profile can be used to show periodic pattern [7]. If Y-profile of the considered block exhibits periodic behavior as in Figure 2, this block can be declared as text block.

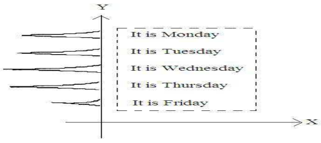
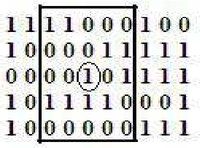


Figure 2. Periodic behavior of text block using Y-profile

3.2. Graphic and Image Classification

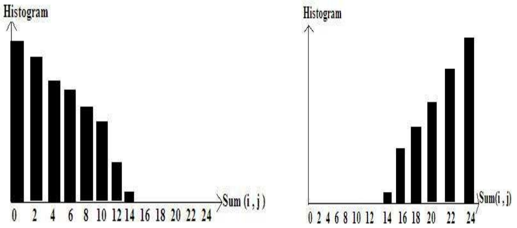
Connectivity histogram is utilized to classify non text blocks into graphic and image. The block with sparse dark pixels is defined as graphic and the block with dense dark pixels is defined as image. So, an operator evaluates the dark pixel distribution presented in the block as in Figure 3. An n x n mask centered at each dark pixel is built to compute connectivity dark pixel. In Figure 3, an 5 x 5 mask centered at a dark pixel is built. Then, calculate the number of dark pixels connected to the pixel. Each dark pixel will attain a value indicating the number of dark pixel connected to it. Last, generate the connectivity histogram by summing the number of dark pixels with the same connectivity value as in Figure 4. If the histogram is heavily distributed in the right part as in Figure 4b, then the considered block is classified as an image block. Otherwise, it is classified as a graphic block. Figure 4 illustrates the connectivity histogram of graphic and image. The x-axis represents the number of connected dark pixels and y-axis represents the number of dark pixels possess the same value in the x-axis [7] and [11].



```
1 1 1 1 0 0 0 1 0 0
1 0 0 0 0 1 1 1 1 1
0 0 0 0 1 0 1 1 1 1
1 0 1 1 1 1 0 0 0 1
1 0 0 0 0 0 0 1 1 1
```

Number of connectivity dark pixel = 9

Figure 3. The computation of connectivity dark pixel with n equaling 5



(a) (b)

Figure 4. Connectivity histogram of (a) Graphic and (b) Image

4. System Process Flow

The following figure shows the process flow of system. This system consists of four main processes. First, binarization of input document is performed because input image is binary to perform smearing operation. Binarization of input document is shown in Figure 6b. Second, horizontal smearing is performed to form line blocks by using suitable threshold. Horizontal smearing example is shown in Figure 6c.Third, block merging is performed to form paragraph blocks as in Figure 6d by using suitable threshold. Fourth, blocks in document are classified into text blocks, graphic blocks, and image blocks. In Figure 6e, text blocks are shown into blue blocks, graphic blocks are shown into green blocks and image blocks are shown into red blocks.

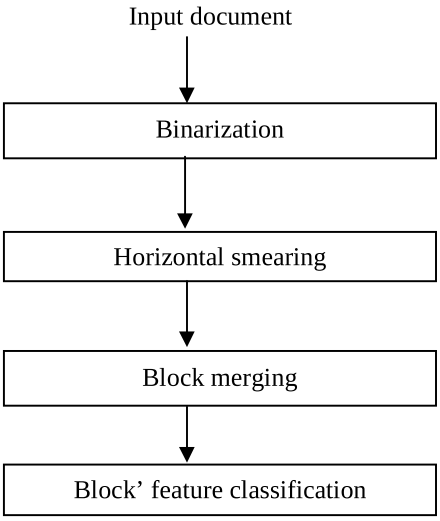


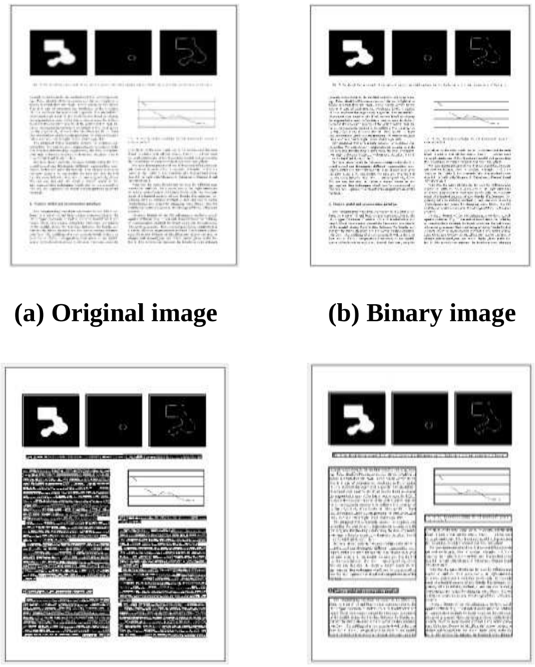
Figure 5. Process flow of system

5. System Implementation and Experimental Results

Tests are performed on simple papers with text, graphic and image contents. An example document is shown in Figure 6a. Manual global thresholding technique that uses single threshold for every image is used to convert binary document image. Shown in Figure 6b is binary document. In this paper, suitable smearing and block merging thresholds for segmentation are manually selected according to scan resolution.

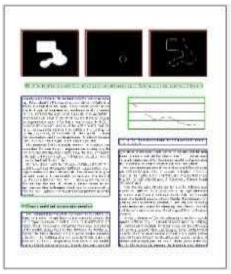
In many structured papers, when scanned at 100 dots per inch (dpi), distance between two words will be represented by nearly 0.1 inches while distance between two rows in one paragraph will be represented by nearly 0.1 inches. Words belonging to one row will be connected to be a line block and words belonging to different rows will be connected to be a paragraph block. For example, if the document is scanned at 100 dpi, threshold values for horizontal smearing and vertical block merging will be 9 or 10 or 11 or 12. In Figure 6b, input document is scanned at 96 dpi. So, 10 for horizontal smearing threshold value are used and 9 for vertical block merging threshold value are used to segment document. Small threshold value does not support to be paragraph block. This does not show text block's periodic behavior. Large threshold value connects different feature blocks. This causes classification errors.

In document's features classification, we can detect blocks that have periodic behavior as text blocks. Text paragraphs more than one line can usually be described periodic behavior. Images with sparse dark pixels (line drawings) can be detected as graphics and dark images (color pictures) with dense dark pixels can be detected as images. This method cannot classify embedded text in the graphics and images [6] and [8]. Each media block is showed by using different colors (red color for image block, green color for graphic and blue color for text block). The main purpose of system implementation is to support feature extraction.



(a) Original image **(b) Binary image**

(c) Horizontal smearing **(d) Block merging**



(e) Block's Feature Classification

Figure 6. (a) Original image, (b) Binary image, (c) Horizontal smearing to form line blocks, (d) Block merging to form paragraph blocks (e) Texts with blue blocks, graphics with green blocks, and images with red blocks.

6. Conclusion

In this paper, we have presented a multimedia document segmentation and classification method. By using this method, double scan problem of RLSA can be avoided. By using this system, physical features (text, graphic and image) embedded in documents can be easily classified and help to extract. Logical features (title, heading, characters and digits in the text body, mathematical expressions, tables, flowcharts, plots, diagrams, and logos, etc) can also be helped to extract. Text blocks produced from classification phase can be used for OCR (Optical Character Recognizer) as inputs.

7. References

[1] A.antonacopoulos and R.T. Ritchings, "Flexible Page Segmentation Using the Background", 1994.

[2] Bontee Kruatrachue, Narongchai Moongfangklang, Kritawan Siriboon, "Fast Document Segmentation Using Contour and X-Y Cut Technique", 2005.

[3] Daniel X.Le, George R. Thoma, and Harry Wechsler, "Automated Borders Detection and Adaptive Segmentation for Binary Document Images", 1996.

[4] Faisal Shafait, Daniel Keysers, and Thomas M. Breuel, "Performance Comparison of Six Algorithms for Page Segmentation", 2006.

[5] Jaakko Sauvola and Matti Pietikainen, "Page Segmentation and Classification Using Fast Feature Extraction and Connectivity Analysis", 1995.

[6] Keyur V Patel, "Partial Eight Direction Based Line Segmentation Algorithm for Epigraphical Script Images", October 3, 2009.

[7] Kuo-Chin Fan and Chin-Hwa Liu, “Segmentation and Classification of Multimedia Documents”, 1992.

[8] Oleg Okun David Doermann and Matti Pietikainen, “Page Segmentation and Zone Classification: The State of the Art”, November 1999.

[9] Premnath Dubey, “Optical Character Recognition an Overview”.

[10] Rene Baston, Karl MacMillan, and Christoph Dalitz, “PageSegmentation”, February 09, 2010.

[11] Tinku Acharya and Ajoy K. Ray,“Image Processing Principles and Applications”, 2005.