A MODIFIED RLSA TECHNIQUE FOR AUTOMATICALLY READING TRADEMARKS

Dr. Ghazali Sulong
Zaidah Ibrahim
Faculty of Computer Science and Information System
University Technology Malaysia
Skudai, Johor Darul Takzim

ABSTRACT

This paper proposes a technique for automatically reading trademarks that have interflow of mixed-mode contents of text and nontext. The design is based on the known knowledge of the structure of the trademarks. First of all, block segmentation using modified Run Length Smoothing Algorithm (RLSA) decomposes the trademark image into single-mode blocks. Then, group classification classifies the blocks into various groups followed by classification of text and nontext block. Then, extraction process of the text block proceeds using the average length of the runs of the white pixels. The experimental results show that the proposed technique is capable to perform text extraction for the trademarks that satisfy the requirements.

Keywords: block segmentation, classification, RLSA, text extraction, trademark

ABSTRAK

Kertas kerja ini mencadangkan satu kaedah untuk mengekstrak teks secara otomatik dari logo yang mengandungi beranika jenis dan saiz teks dan bukan teks. Kaedah ini bergantung kepada maklumat yang diperolehi terlebih dahulu mengenai struktur logo. Mula-mula, blok segmentasi menggunakan teknik Run Length Smoothing Algorithm (RLSA) yang telah diubah suai, mencantumkan imej logo kepada blok-blok yang sejenis. Kemudian, teknik klasifikasi kumpulan mengklasifikasikan blok-blok tersebut kepada jenis-jenis logo yang telah ditentukan dan diikuti pula oleh proses mengekstrak teks menggunakan teknik purata pixel putih dalam sesuatu aliran. Rumusan eksperimen menunjukkan bahawa teknik yang dicadangkan ini berjaya mengekstrak teks dari beranika jenis logo yang menetapi syarat-syarat yang telah ditetapkan.

1.0 INTRODUCTION

Today, there is a growing trend in mostly every office towards the use of electronic input and storage. There is urgent need to convert their large volumes of paper document databases to an electronically searchable and efficiently stored digital form.

The data of interest in any given document like trademarks usually consists of a mixture of text and nontext which must be separated for subsequent processing. Automatic segmentation is necessary since manual identification and manipulation of text and nontext consumes a lot of labor costs and data volumes. In this paper, we briefly describe an algorithm to extract the text from a mixed-mode trademark and we assume a priori knowledge of the document structure in the design. Then, character segmentation technique followed by character recognition technique will be discussed. The last section indicates the experimental results of this research.

2.0 TRADEMARKS

A trademark or logo uniquely identifies a company or an organization. At the moment, there exists more than 200,000 trademarks that register at the Domestic and Consumer Affair. When a new company registers, its' trademark needs to be compared with the existing ones. So far, this process is done manually. Thus, to overcome this problem, an automatic recognition of trademark is needed. But, before recognition can be performed, the mixed-mode trademark needs to be segmented into one of text and nontext . Then only the extracted text can be further processed for recognition

A trademark may consists of any kinds of text and nontext. The text part may consists of the Roman alphabets, uppercase or lowercase letters, with multifont and sizes. The nontext part may be in the form of simple lines and figures. A trademark can be categorized into two different categories based on content, which are :

1. text or nontext only
2. combination of text and nontext

The second group can be subdivided into three other subgroups, that are :

a. text and nontext that are nicely separated from each other
b. text that overlaps or touches the nontext

c. text that is located within closed borders

The appendix shows some examples of trademarks based on the different groups mentioned above. This research only focusses on the second cgroup of the trademarks and they are referred as Group 1, Group 2, and Group 3, respectively.

3.0 OVERVIEW OF DESIGN

A functional block diagram illustrating the main functions of the text retrieval system for a trademark is shown in figure 1.

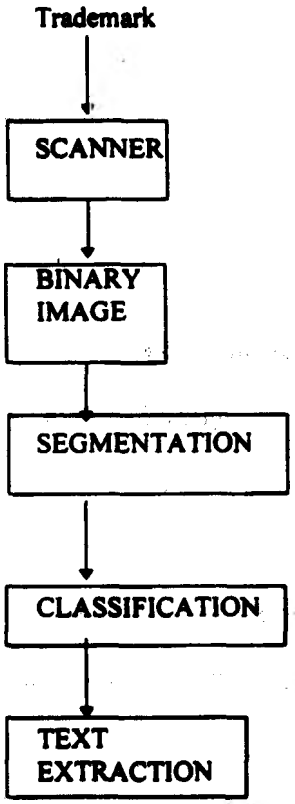

Figure 1 Functional block diagram for text retrieval system for trademarks

The trademark is digitized by a scanner and thresholded into binary images [Shaaban, 96]. Then, the binary image is smeared to form various blocks under block segmentation. After that, the blocks are being classified into different groups followed by text extraction. In order for the system to perform successfully, the trademark should satisfy the following requirements :

1. the text and nontext should be in black pixel with white background
2. the text should have the same straight base line
3. the text should consists of at least two characters

3.1 SEGMENTATION

Block segmentation decomposes a document image into rectangular blocks each of which includes one of text or nontext. Block segmentation techniques can be categorized as being either top-down or bottom-up. In top-down technique, large regions are recursively segmented into subregions, like Run Length Smoothing Algorithm (RLSA) [Wahl, F.M., et.al., 82, Wong, K.Y., et.al., 82]. In bottom-up technique, pixels are initially grouped together as connected components and progressively merged into larger regions like Connected Component Analysis (CCA) [Fletcher, L.A., et.al., 88]. The top-down technique is fast and very effective for processing documents that have a specific format. The bottom-up technique is more resistant to noise and skew within a document but time consuming. This paper chooses a modified RLSA for block segmentation as we know the nature of the document.

3.1.1 MODIFIED RLSA

This system receives a "clean" binary image which is free of noise and not skewed. To segment the image into regions for further processing, we begin with smearing the image using modified RLSA.

Modified RLSA is applied to a binary sequence in which white pixels are represented by 0's and black pixels by 1's. A set of adjacent 0's or 1's is called a run. The algorithm converts a binary input sequence f into an output sequence g by using a simple rule : the 0's in f are changed to 1's in g if the run length of 0's i.e. the number of adjacent 0's is less than or equal to a predefined threshold c. For instance, if the threshold c is 2 and the input f is :

f: 0010001110100110000

then the output g is :

g: 1110001111111110000

Modified RLSA consists of the following three steps :

1. a horizontal smearing is applied to the original image by a predefined threshold *c1*
2. a vertical smearing is applied to the original image by a predefined threshold *c2*
3. the smearing results of steps 1 and 2 are combined by a logical AND operator

Different values for *c1*, c2 and c3 produce different types of RLSA images. Very small *c1* may cause rectangular blocks to occur around individual characters. Slightly larger values for *c1* may close up individual characters with the nontext part. Similar comment is applied to the values of *c2* and *c3*. Thus, the threshold values should be carefully chosen to obtain the desired result

The trademark is being scanned at 200 dpi and after a few experiments, 50 has been chosen to be a good choice for c1 and 10 for *c2*.

3.2 BLOCK/GROUP CLASSIFICATION

Block classification classifies the blocks after modified RLSA process into one of text or nontext. Most existing block classification techniques are based on the discrimination of statistical local or global features. A two-dimensional plane consisting of mean value of the block height versus run length of the block mean black pixel is determined to classify document blocks into text and nontext [Wong, K.Y., et.al., 82]. A rule-based classification technique uses the features like height, aspect ratio, density and perimeter [Fisher, J.L., et.al., 90]. A robust block classification technique based on clustering rule which is the mean vertical and horizontal transition of white to black pixels in a block [Shih, F.Y., et.al., 92]. These techniques are based on a page layout and there is no overlapping or touching of text and nontext.

On the other hand, trademark consists of only one or two small blocks of black pixels and the text and nontext may be overlapping or touching each other. Thus, to differentiate between a text and nontext block, first of all the blocks need to be classified into various groups as mentioned earlier.

Group classification is based on edge detection. Edge-based shape representation involves segmenting the shape boundaries into simple pieces and then constructing descriptions of both the individual pieces and the relationships between pieces. But in this case, only one edge is used. The left edge is deteted by scanning the binary image

from left to right, referred as *Left*. Then, group classification is performed according to the following algorithm :

```
Begin
If Left = 1 Then
        Find topwidth, middlewidth, bottomwidth
        If middlewidth > topwidth Then
                Group 2
        Else
                Group 3
Else
        Group 1
End IF
End
```

Topwidth, middlewidth and bottomwidth represent the values of the width of the top, middle and bottom of the blocks produced after modified RLSA process. These measurements are made by adding up the black runs on the top, middle and bottom of the blocks. This measurement is made possible since for Group 2, the text is always longer in width compared to the nontext.

3.3 TEXT EXTRACTION

3.3.1 Group 1

To differentiate between a text block and a nontext block in Group 1, the average length of the white runs is calculated for each row in the block, referred as *avgdist*. Then, all the *avgdist* in the same block is being averaged again, referred as *bigavg*.

A text block should have the same values of *avgdist* for each row as *bigavg*. However, in certain cases, there occurs a slight difference in some values of the *avgdist* in a text block. For instance, the length of white runs between the characters K and J is very much different compared to the length of white runs between the characters B and D. Notice that, the length of the white runs between characters K and J is approximately twice the length of the white runs between characeters B and D. Thus, the *avgdist* for each row should be approximately 50% larger or smaller than the *bigavg*.

For example, the following values are the *avgdist* for 2 unknown blocks.

Block 1 Block 2

row 1	4	4
row 2	3	1
row 3	2	3
row 4	4	1
row 5	5	2
bigavg = 4		*bigavg* = 2
2 <= {4, 3, 2, 4, 5} <= 6		1 <= {4,1,3,1,2} <> 3

From these values, we can conclude that block 1 is a text block and block 2 is a nontext block.

There are a few cases where the nontext block consists of only black pixels, with not a single white pixel. In this case, we have to add another statement in the algorithm which indicates that if the *bigavg* is equal to zero, then it is a nontext block.

3.3.2 Group 2

From the width measurement, the text block can be determined but the text block also consists of the nontext part that overlap or touches it. This is done by performing a horizontal cut on the block based on the height of the left and the right edges of the block with the greater width. The nontext part can be deleted in the character recognition process which is not being discussed here.

3.3.3 Group 3

Group 3 consists of borders around the desired image and these borders need to be extracted or deleted first. Once the borders have been deleted, the remaining images will have to go through the same process as Group 1 for text extraction.

Usually, the borders are in the form of rectangle or circle. To delete them, the left and right edges need to be detected. Then, line neighborhood density-based technique is used. If its' direct neightbor on the right (for the left edge) and on the left (for the right edge) equals to black, change it to white. This process continues for every pixel on both edges. As a result, the whole borders have been deleted.

4.0 EXPERIMENTAL RESULTS

Several different trademarks from the different groups mentioned earlier have been tested. Satisfactory results are only obtained if the trademarks satisfy the system requirements. See the appendix for some samples of the tests. The first page shows the results of modified RLSA. The consecutive pages show sample tests for trademarks from Group 1, Group 2 and Group 3, respectively. On the left is the original image while on the right is the extracted text.

5.0 CONCLUSION

A text extraction technique has been proposed for automatically reading trademarks. Block segmentation is based on a modified RLSA while group classification is based on the physical structure of the trademarks. Text extraction is based on the average length of the runs of the white pixels. Satisfactorily results have been obtained but more improvements can be added to the system to make it more flexible. More research is being done on this including character recognition for the extracted text from Group 2.

BIOGRAFI

Dr. Ghazali Sulong adalah seorang Prof. Madya di Fakulti Sains Komputer dan Sistem Maklumat. Beliau sedang aktif menjalankan penyelidikan dalam bidang Image Processing And Pattern Recognition termasuk Handwritten Text Recognition dan Teleradiology.

Zaidah Ibrahim adalah seorang pelajar Sarjana Lanjutan di Fakulti Sains Komputer dan Sistem Maklumat. Beliau memperolehi Sarjana dalam jurusan Sains Komputer dari Northrop University, Los Angeles pada tahun 1988. Beliau pernah berkhidmat sebagai pensyarah di Pusat Pendidikan Persediaan, Institut Teknologi MARA, Shah Alam.

6.0 REFERENCES

[Chen 93] Chen, C.H. (1993). **Handbook Of Pattern Recognition And Computer Vision**, World Scientific, Singapore.

[Ellman 90] Ellman, D.G. (1990). **A Review Of Segmentation And Contextual Analysis Techniques For Text Recognition**. Pattern Recognition, Vol. 23, No. 3/4, pp. 337-346.

[Esposito et. al. 90] Exposito, F., Malerba, D. Semeraro, G. Annese, E., and Scafuro, G, **An Experimental Page Layout Recognition System for Office Document Automatic Classification :an Integrated Approach for Inducive Generalization**. Proc. IEEE, pp. 557-562.

[Fisher et. al 90] Fisher, J.L., Hinds, S.C. and Dámato, D.P, **A Rule-Based System for Document Image Segmentation**. Proc. IEEE, pp. 567 - 572.

[Fletcher et. al. 88] Fletcher, L.A. and Kasturi, R, **A Robust Algorithm For Text String Separation From Mixed Text/Graphics Images**. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 10, No. 6, November, pp. 910 - 918.

[Houle et. al. 91] Houle, Gilles and Eom, Kie-Bum, **On the Use Of A Priori Knowledge To Character Recognition**. Proc. IEEE, pp. 1415 - 1420.

[Hung 85] Hung, S.H.Y. **On the Straightness Of Digital Arcs**. IEEE Trans. Pattern Anal. Machine Intelligence, Vol. PAMI-7, No. 2, March 1985.

[Koplowitz, et. al. 95] Koplowitz, J. and Plante, S, **Corner Detection For Chain Coded Curves**. Pattern Recognition, Vol. 28, No. 6, pp. 843-852.

[Lebourgeois et. al. 92] Lebourgeois, F., Bublinski, Z. and Emptoz, H, **A Fast And Efficient Method For Extracting Text Paragraphs And Graphics From Unconstrained Documents**. Proc. IEEE, pp. 258 - 267.

[Lu 95] Lu, Yi. **Machine Printed Character Segmentation - an Overview**. Pattern Recognition, Vol. 28, No. 1, pp. 67 - 80.

[Shih et. al. 92] Shih, Y., Chen, S.S., Hung, D.D. and Ng, P.A, **A Document Segmentation, Classification And Recognition System**. Proc. IEEE, pp. 258 - 267.

[Schurman 92] Schurman. **Document Analysis - From Pixels to Contents**. Proc. IEEE, Vol. 80, No. 7, July, pp. 1101 - 1119.

[Syaaban 96] Handwritten Text Recognition. Unpublished Technical Report.

[Wahl et. al. 82] Wahl, F.M., Wong, K.Y. and Casey, R.G, **Block Segmentation And Text Extraction In Mixed Text/Image Documents**. Computer Graphics and Image Processing 20, pp. 375 - 390.

[Wong et. al. 82] Wong, K.Y., Casey, R.G. and Wahl, F.M. **Document Analysis System**. IBM J. Res. Develop, Vol. 26, Nov, pp. 647 - 656.

[Wu 82] Wu, Li-De. **On The Chain code Of A Line**. IEEE Trans. Pattern Anal. Machine Intelligence, Vol. PAMI-4, No. 3, May 1982.

[Young et. al. 86] Young, T.Y. and Fu, King-Sun, **Handbook Of Pattern Recognition And Image Processing**. Academic Press Inc., San Diego.

[Zlatopolsky 94] Zlatopolsky, A.A. Automated Document Segmentation. Pattern Recognition Letters 15, pp. 699 - 704.