DETERMINATION OF RUN-LENGTH SMOOTHING VALUES FOR DOCUMENT SEGMENTATION

N. Papamarkos, J. Tzortzakis and B. Gatos

Electric Circuits Analysis Laboratory
Department of Electrical & Computer Engineering
Democritus University of Thrace
67100 Xanthi, Greece
email : papamark@voreas.ee.duth.gr

Abstract

The segmentation is a pre-processing procedure in document analysis systems. In this paper, a document segmentation method is proposed which is based on run-length smoothing algorithm. Run-length based segmentation methods use heuristic techniques to determine the values of the smoothing variables. In this paper, we present an unsupervised technique, based on the distributions of the black and white run-lengths, for the calculation of proper values of the smoothing variables.

1. Introduction

In this paper, we concerned with the problem of segmentation of mixed type documents. Segmentation of documents is one of the most important procedures in Optical Character Recognition (OCR) systems. A document segmentation technique consists of many stages and its final result must be the separation of the characters [1].

Meny algorithms for segmentation of documents have been proposed. Kasturi and Trivedi [2] classify these algorithms in two basic approaches. The first approach, is a called bottom-up method, which starts by first segments the document into small blocks (marks), and then merges them into bigger blocks. The second approach is called top-down method and starts by segments the document into large blocks and then analyses them in order to achieve separation of the characters of the text blocks. In the first category belongs the method of Fletcher and Kasturi [3] which starts by first finding the connected components of an image and then separating graphics from text using the relative frequency of occurrence of components as a function of their areas. In the next step, an iterative procedure is applied to improve the initial estimation by using the Hough transform to all connected components [4-5]. This approach is quite complex and computational expensive.

In the top-down approaches belongs the projection-profile method (PPM) and methods based on run-length segmentation algorithm (RLSA). PPM is addressed with many disadvantages such as the sensitivity of the document skew. On the other hand, the RLSA is the most powerful procedure for top-down block segmentation. This technique is first introduced by Wong et al. [6] and Wahl et al. [7] and taken up again by Wang and Shihari [8]. It is a low complexity technique and can segment documents into rectangular blocks and then classified them into text, graphics or more detailed objects. Until recently, the RLSA method imposes a smoothing on the document using two parameters defined in a heuristic way (one for the vertical and one for the horizontal direction). For the block classification, additional parameters, defined also in a heuristic way, are used leading to the necessity to train the system with documents having similar fonts or other morphological characteristics. It must be noted that the method is not robust since if the assumptions made for the determination of the heuristic parameters are not satisfied, the method will fail.

In this paper, we propose a top-down unsupervised method for segmentation of digitised documents. The proposed method has two main stages. In the first stage, we determine the document's major blocks, in the second stage we proceed to the text line extraction stage. For these stages, where we use the RLSA, we propose an algorithm for automated calculation of the proper horizontal and vertical smoothing values. Specifically, the horizontal smoothing value (*hsv*) and the vertical smoothing value (*vsv*) are calculated according to the mean character length (*mcl*) and the mean text line distance (*mtld*) of the document. The values of *mcl* and *mtld* are determined using the contributions of the horizontal and vertical run-lengths.

2. RLSA

The RLSA is applicable in binary images. The basic idea of RLSA is to take advantage of the white runs existing in the horizontal and vertical directions. For each direction, RLSA

eliminates white runs whose lengths are smaller than a threshold smoothing value (*sv*).

Usually, the RLSA has three main stages. First, in the original image, RLSA is applied horizontally, row-by-row by using *hsv* and then vertically, column-by-column with *vsv*. After horizontal and vertical smoothing, we have two bit-maps, which are next combined by a logical AND operation to produce a new smoothing image. This image has small gaps that interrupt blocks of text lines. Therefore, an additional horizontal smoothing operation is performed by using a new suitable smoothing value *ahsv*.

The RLSA based segmentation is a top-down procedure, i.e., it must be applied successively to obtain separation of major blocks (paragraphs, pictures, drawing etc.), text-line blocks, word blocks and finally character blocks. For each stage the three smoothing values must be appropriate selected. According to the work of Wong et al. (1982), the smoothing values of hsv and vsv depend on "the number of pixels covering the length of long words, whereas ahsv should be set to cover a few character widths". These conditions give satisfactory results only for major block segmentation but, unfortunately, there is not have been proposed any systematic procedure for their calculation. For an image of 2024X2024 pixels, Wong et al. use as suitable values *hsv=300*, *vsv=500* and *ahsv=30*. Chauves et al. (1993), use a top-down RLSA based segmentation technique to build an office document analysis system. In this system, they give some examples of the application of the RLSA but do not propose an analytical method for the calculation of the smoothing values.

2.1 Determination of major blocks

The proposed method is based on pre-estimation of the mean character length and the mean text line distance of a document. To do this, we develop a technique that uses the horizontal and vertical distributions of white and black runs. We consider that a proper value for *hsv* must be equal to twice the *mcl*, while in the vertical direction, the proper value of *vsv* must be taken equal to the *mtld* of the document. That is

$$hsv = 2\,mc \qquad (1)$$

$$vsv = mtld \qquad (2)$$

To obtain the proper values of *mcl* and *mtld*, we examine not only the white runs but also the black runs in the vertical and horizontal directions. Specifically, initially we calculate three histograms. These histograms give the distributions of black runs in the horizontal and vertical directions and of white runs in the horizontal direction. For example, the form of these histograms for the document of Fig. 1 is given in Fig. 2.

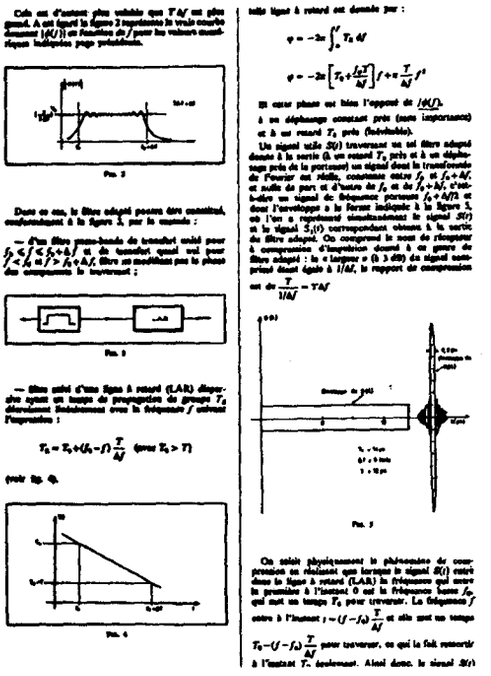


Fig. 1 Original document.

The global maximum of the histogram of horizontal black runs (*gmhbr*) constructed mainly from part of the characters. Therefore, it gives information about the mean character length. We observe that the *mcl* belongs to a region defined by multiply the *gmhbr* by two suitable coefficients m_1 and m_2. That is

$$mcl \in [int(m_1 \cdot gmhbr), intu(m_2 \cdot gmhbr)] \qquad (3)$$

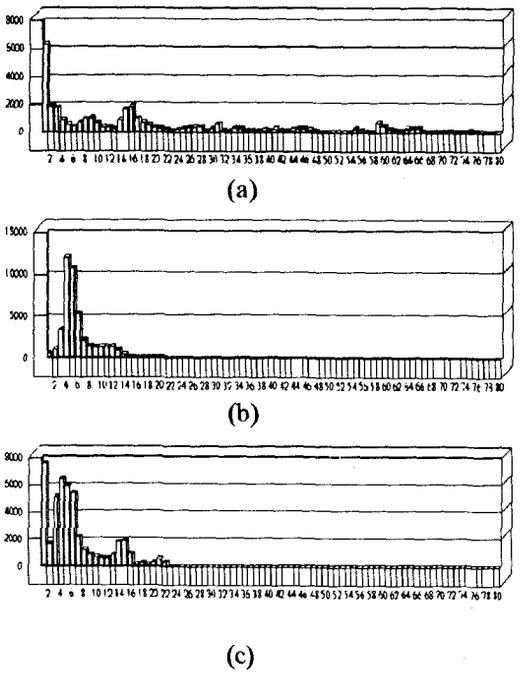
where *int(x)* is the integer part of x, and

$$intu(x) = \begin{cases} x, & \text{if } x \text{ is integer} \\ int(x)+1, & \text{otherwise} \end{cases} \qquad (4)$$

On the other hand, we know that the mean character height (*mch*) is close enough to the *mcl*. However, it is obvious that the *mch* corresponds to a local or global maximum in the histogram of vertical black runs, and moreover, we accept that this maximum belongs to the region defined by (3). Therefore, we must determine the proper values of coefficients m_1 and m_2. To do this we examine a large number of documents. In each document we multiply the *gmhbr* by m where $m \in \{3.5, 4.0, 4.5, 5.0, 5.5\}$ and for each m we compare the quantity $m \cdot gmhbr$ with the real character length (*rcl*). Making a large number of experiments we have found that

the smallest deviations are obtained for $m=4.5$. For this value the region of deviation of *rcl* is approximately [-28%,40%] that corresponds to $m_1=0.72m=3.24$ and $m_2=1.4m=6.3$. Therefore, according to (3), we accept that

$$mcl \in [int(3.24 \cdot gmhbr), intu(6.3 \cdot gmhbr)] \quad (5)$$



(a)

(b)

(c)

Fig. 2 (a) Distribution of vertical white runs. (b) Distribution of horizontal black runs.(c) Distribution of vertical black runs.

It should be noted that the above region for *mcl* has been verified with all the tested documents. For the document of Fig. 1, we have from Fig. 2(b) that $gmhbr=4$ *and* $mcl \in [12,26]$. As we can observe in Fig. 2(c), truly in this region there is a global peak which corresponds to 15. This value agrees with the real value of *mcl* and gives $hsv=30$.

We can accept that in a readable document the *mtld* is always greater than int(0.8*mcl*). This is the lower limit for *mtld*. On the other hand, the white pixels belonging in the areas between horizontal text lines give a global peak in the search area of the histogram of vertical white runs. Therefore, the upper limit of *mtld* can be large enough and it is taken equal to be 80. So, the *mtld* is determined as the position of the global histogram maximum in the range [*int*(0.8*mcl*),80].

3. Extraction of text lines

The next stage of the segmentation procedure is the extraction of text lines in each major block. Before this stage the user can select the blocks that will be passed to text line extraction stage. The procedure for text lines extraction follows a similar to the major block extraction procedure. Specifically, first we use a horizontal and a vertical run-length smoothing, apply the OR operation and after this the text lines isolated with a contour following and a minimum surrounding rectangular procedure. In this stage, we work only inside to the separated major

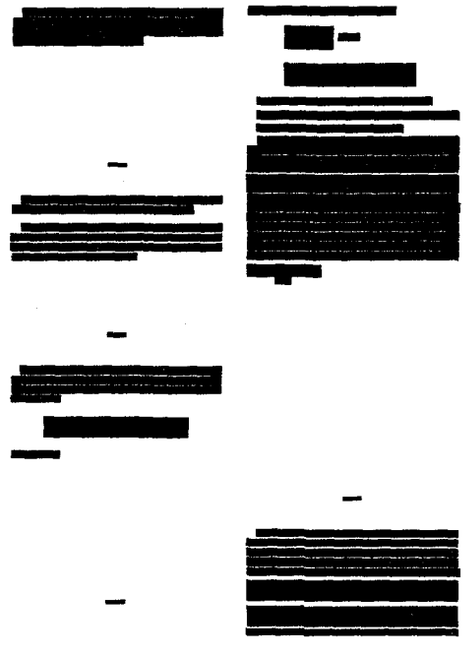


Fig. 3 Extraction of text-lines.

blocks. Therefore, it is impossible to have contact of two different blocks and hence the *hsv* can be a large enough, and we observed that in this stage *hsv must be* equal to five times the *mcl*. In contrast to *hsv*, the *vsv* must be very small in order to have not contacts of the text lines. Even zero value gives sufficient results in the most of the cases. However, with zero value of *vsv* we have problems with broken characters as i and j. This problem is more serious in the Greek characters where we have accent characters. To overcome these difficulties, and after many tests, we decide *vsv* to be equal to

$$vsv=intu(0.15mtld) \quad (6)$$

For our example we have $mtld \in [12,80]$ and $mtld=vsv=16$. We note that the same documents have been used by Chauvet et al. [9]. Comparing the major blocks we can observe that in our approach we separate the captures in contrast to the work of Chauvet et al. [9] where many captions have been embodied in drawing blocks.

For the text line extraction stage we have found that $hsv = 5 \cdot mcl = 5 \cdot 15 = 75$ and $vsv = intu(0.15mtld) = intu(0.15 \cdot 16) = 3$. Using these

values the text line extraction procedure results to the well-isolated text lines of Fig. 3.

4. Conclusions

In this paper, we propose a system for unsupervised segmentation of digitised documents. The entire system proceeds in two stages, i.e., major block smoothing stage and the text line smoothing stage. Our approach is based on RLSA, where we develop a technique for automated determination of the proper smoothing values. The smoothing values are determined by examination of the contribution of the horizontal and vertical run-lengths in the document.

Experimental results shows that the proposed segmentation method is applicable in any mixed type document and gives high quality results.

5. References

1. L. O'Gorman. and R. Kasturi. *Document Image Analysis,* IEEE Comp. Society Press, 1995.
2. R. Kasturi and M.M. Trivedi, "Image Analysis Aplplications", Marcel Dekker, N.York, 1990.
3. L. A. Fletcher and R. Kasturi, "A robust algorithm for text string separation from mixed text/graphics images," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 10, No. 6, pp. 910-918, November 1988.
4. R. C. Gonzalez and P. Wintz, "Digital Image Processing", 2nd ed., Addison-Wesley, N.York, 1987.
5. C. Strouthopoulos, N. Papamarkos and C. Chamzas, "Identification of text-only areas in mixed type documents", Proc. of *IEEE Workshop on Nonlinear Signal and Image Processing*. Greece, pp. 162-165, 1995.
6. K.Y. Wong, R.G. Casey and F.M. Wahl, "Document analysis system," IBM J. Res. Devel., Vol. 26, No. 6, pp. 647-656, 1982.
7. F. M. Wahl, K. Y. Wong and R. G. Casey, "Block Segmentation and Text Extraction in Mixed Text/Image Documents," Computer Graphics and Image Processing, Vol. 20, pp. 375-390, 1982.
8. D. Wang and S.N. Shihari, "Classification of newspaper image blocks using texture analysis," Computer Vision Graphics and Image Processing, Vol. 47, pp. 327-352, 1989.
9. P. Chauvet, J. Lopez-Krahe, E. Taflin and H. Maitre, "A system for an intelligent office document analysis, recognition and description," Signal Processing, Vol. 32, pp. 161-190, 1993.