An Efficient Word Segmentation Technique for Historical and Degraded Machine-Printed Documents

M. Makridis[1], N. Nikolaou[1] and B. Gatos[2]

[1]Department of Electrical and Computer Engineering,
Democritus University of Thrace, 67 100 Xanthi, Greece
{nnikol, mmakridi}@ee.duth.gr

[2]Computational Intelligence Laboratory, Institute of Informatics and Telecommunications,
National Center for Scientific Research "Demokritos", GR-153 10, Athens, Greece
http://www.iit.demokritos.gr/~bgat/, bgat@iit.demokritos.gr

Abstract

Word segmentation is a crucial step for segmentation-free document analysis systems and is used for creating an index based on word matching. In this paper, we propose a novel methodology for word segmentation in historical and degraded machine-printed documents. The proposed technique faces problems such as having text of different size, having text and non-text areas lying very near and having non-straight and warped text lines. It is based on: (i) a dynamic run length smoothing algorithm that helps grouping together homogeneous text regions, (ii) noise and punctuation marks removal as well as on obstacle detection in order to facilitate the segmentation process and (iv) a draft text line estimation procedure that guides the final word segmentation result. After testing on numerous historical and degraded machine-printed documents, it has turned out that our methodology performs better compared to current state-of-the-art word segmentation techniques for historical and degraded machine-printed documents.

1. Introduction

The segmentation process of digital documents into words can lead to a powerful description of the image content. It can lead to systems where the content of the documents is exploited efficiently through an automatic indexing and retrieval procedure. A segmentation-free approach that entails the recognition of the whole word is followed in [1][2][3] where line and word segmentation is used for creating an index based on word matching.

Previous work in word segmentation includes several approaches for machine-printed and handwritten documents. Manmatha et al. approach [4] is based on a scale space technique for word segmentation on handwritten documents of George Washington's collection. Word extraction for machine-printed documents is examined in the work of Park et al. [5]. A 3D neighborhood graph model analyses the structural information of documents and with the use of angle and distance constraints word components are identified. Park and Govindaraju [6] present a methodology that takes advantage of the spacing between words in a phrase to aid the handwritten recognition process. The determination of word breaks is made in a manner that adapts to the writing style of the individual. For the case of historical machine-printed documents, Antonacopoulos and Karatzas [7] calculate and analyze the horizontal projection profile to identify suitable spaces between words. In the work of Gatos et al. [8], word segmentation in historical machine-printed documents is based on a run length smoothing in the horizontal and vertical directions. For each direction, white runs having length less than a threshold are eliminated in order to form word segments.

In this work, we focus on the word segmentation problem in historical and degraded machine-printed documents. We face problems such as having text of different size, having text and non-text areas lying very near and having non-straight and warped text lines. The proposed technique is based on a dynamic run length smoothing algorithm that helps grouping together homogeneous text regions. It can be summarized at the following distinct steps: (i) noise and punctuation marks removal, (ii) obstacle detection, (iii) draft text line estimation and (v) word segmentation.

In the following sections, we present a detailed description of the proposed methodology, as well as experimental results that demonstrate the efficiency of the proposed methodology.

Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)
0-7695-2822-8/07 $25.00 © 2007 IEEE

2. Dynamic Run Length Smoothing

The RLSA [9] is one of the most common algorithms used in page layout analysis and segmentation techniques. It operates on binary images in a specified direction (usually horizontal or vertical) by replacing a sequence of background pixels with foreground pixels if the number of background pixels in the sequence is smaller or equal than a predefined threshold $T_{\max}$. Its purpose is to create homogenous regions in the document image.

(a) (b)

Figure 1. The two types of background sequences: (a) belonging to the same connected component and (b) belonging to different connected components.

In the proposed word segmentation technique, a modified version of the horizontal RLSA is proposed, the Dynamic Run Length Smoothing Algorithm (DRLSA), in order to overcome the drawbacks of the original algorithm, such as grouping inhomogeneous components or different slanted lines. DRLSA also works successfully with documents containing characters with variable font size. Before the application of DRLSA a connected component analysis is necessary. Two types of background pixels (white) sequences are considered. The first type concerns sequences which occur between two foreground pixels (black) which belong to the same connected component as shown in Fig. 1a. In this case, all background pixels of the sequence are replaced with foreground pixels. The second type of background pixels sequence occurs between two different connected components (see Fig. 1b). In this case, constraints are set in regard to the geometrical properties of the connected components and the replacement with foreground pixels is performed when these constraints are satisfied.

Let CC_i and CC_j be two connected components and $S(i,j)$ a sequence of background pixels between CC_i and CC_j. We define the following three metrics:

- $L(S)$: Length of the sequence $S(i,j)$, that is the number of white pixels.
- $H_R(S)$: Height ratio between CC_i and CC_j, which is defined as follows:

$$H_R(S)=\frac{\max\{h_i,h_j\}}{\min\{h_i,h_j\}} \tag{1}$$

where h_i,h_j, the heights of CC_i and CC_j, respectively.

- $O_H(S)$: The horizontal overlapping between the bounding boxes of CC_i and CC_j, which is defined by the following equation:

$$O_H(S)=\max\{Yl_i,Yl_j\}-\min\{Yr_i,Yr_j\} \tag{2}$$

where $\{Xl_i,Yl_i\}$ and $\{Xr_i,Yr_i\}$ the coordinates of the upper left and down right corner of the CC_i's bounding box. The horizontal overlapping $O_H(S)$ is demonstrated in Fig. 2.

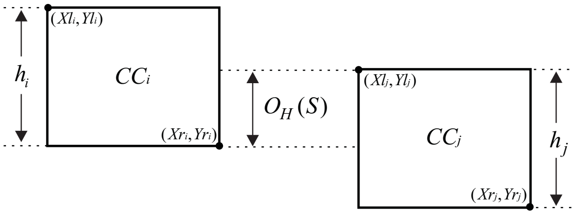


Figure 2. Graphical depiction of horizontal overlapping.

Based on the above metrics, a sequence of background pixels is replaced with foreground pixels only if:

$$(L(S)\le T_l)\wedge(H_R(S)\le T_h)\wedge(O_H(S)\ge T_o) \tag{3}$$

where T_l,T_h,T_o are predefined threshold values.

The length threshold T_l of each sequence is related to the heights of the connected components. If h_i and h_j express the heights of CC_i and CC_j then

$$T_l=a\cdot\min\{h_i,h_j\} \tag{4}$$

where a is a constant value.

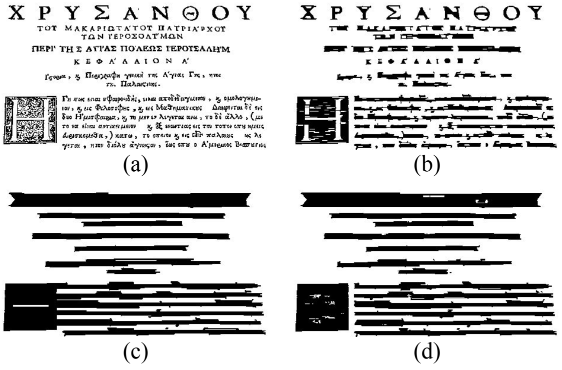
After several experimentations we have set T_h to 3.5. The horizontal overlapping threshold T_o is expressed as the percentage of coverage in regard to the component with the smallest height, that is:

$$T_o=c\cdot\min\{h_i,h_j\} \tag{5}$$

where c is set to 0.4 after experimentations. This means that at least 40% of the shortest component height must be covered in order for a link to be established.

The DRLSA in regard to the original algorithm can prevent the creation of inhomogeneous groups of components, namely to have large and small characters grouped together. Also, it has tolerance against warped or skewed text, that is components of different text lines that are close to each other at the horizontal direction is not likely to be linked even when T_l

receives large values. This happens due to the horizontal overlapping constraint. An example which shows a comparison between the original RLSA and the proposed DRLSA for text-line detection is shown in Fig. 3. The original document of Fig. 3a is processed with the RLSA where $T_{\max}$ is set to a small value equal to 7 pixels. Small size text is merged into word or sets of words but large size text is maintained into individual characters (Fig. 3b). If a larger value of $T_{\max}$ is used, for example 50 pixels, the resulted image is formed as Fig. 3c shows. Large text is correctly merged into text lines but the graphic element is linked with the small size text. In the case of Fig. 3d, in which the DRLSA is applied, all text elements were correctly formed into text lines without links with the graphic element. The value of factor a (eq. 4) used here is 5.



(a) (b)

(c) (d)

Figure 3. RLSA - DRLSA comparison: (a) original image; (b) RLSA with small threshold; (c) RLSA with large threshold and (d) DRLSA.

3. Proposed methodology

The proposed technique is divided into four consecutive stages explained in detail in this section.

3.1 Noise and punctuation marks removal

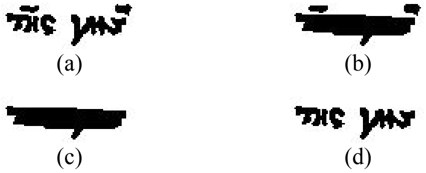
The purpose of this stage is to remove small noisy connected components and to erase punctuation marks in order to improve the structure of the background and simplify the following procedures of the technique.

First, noisy - non text elements are filtered out based on three features of the connected components and their corresponding bounding boxes: height, elongation (the ratio of the shorter to the longer side of each bounding box) and density (the ratio of the foreground pixels to the total number of pixels in each bounding box). Connected components with height shorter than 4 pixels, or density smaller than 0.08, or elongation smaller than 0.08 can not be considered as characters and they are filtered out.

The second type of filtering removes punctuation marks. It is based on the comparison of the connected components from two images, the initial document image and the resulted image after the application of the DRLSA with $a = 1.5$ (see eq. 4). Let I_1 be the original image (see Fig. 4a), I_2 the image after the application of the DRLSA (see Fig. 4b). The number of pixels P_{I_2} of each connected component $CC_i \in I_2$ is calculated, that is the number of the black pixels. In the defined area of each $CC_i \in I_2$, the sum P_{I_1} of the corresponding black pixels of I_1 is also calculated. The ratio of these two sums is taken into account as in the following equation:

$$P_R = \frac{P_{I_2}}{P_{I_1}} \le T_R \qquad (6)$$

Components which correspond to pixel size ratio smaller than T_R are removed and a new image I_3 (see Fig. 4c) is produced. Punctuation marks are removed because they are mainly isolated objects. This means that after the application of the DRLSA their size is likely to remain constant or change by a small factor contrary to text components. The final result is obtained by an AND operation between images I_1 and I_3 as shown in Fig. 4d. A proper value for T_R was experimentally found to be equal to 1.15.



(a) (b)

(c) (d)

Figure 4. Punctuation marks removal: (a) original image I_1; (b) application of DRLSA (I_2); (c) components with small pixel size ratio are removed (I_3) and (d) resulted image after the operation I_1 AND I_3.

3.2 Obstacles detection

The purpose of obstacles is to isolate different text lines and different text columns by defining regions within a document that a horizontal run length procedure is not allowed to cross. Obstacles are used in [10] where text line extraction is performed in documents of multiple text columns. In this paper, two types of obstacles are extracted. Column obstacles are extracted to define different text column spaces, while text line obstacles to locate regions between text lines which belong in the same column. After the application of noise and punctuation marks removal as well as of DRLSA with $a = 1.5$ (see eq. 4) in order to obtain a first draft text line detection, two types of obstacles are detected, column and text line obstacles.

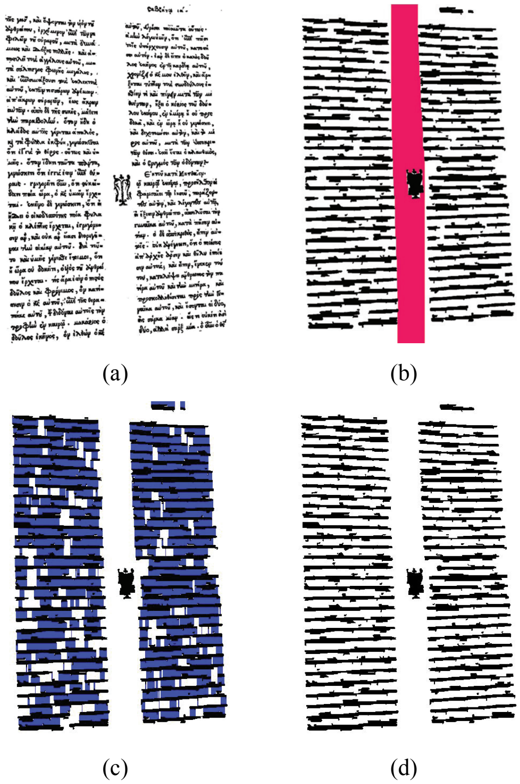


Figure 5. Example of obstacle detection and text line estimation: (a) original Image; (b) column obstacles; (c) text line obstacles and (d) draft text line estimation.

To detect column obstacles, the vertical projection profile is calculated fist. In documents with multiple columns, large valley areas exist in cases that vertical projection values are much smaller than mean value. These areas are considered as candidate obstacle areas. Also, the histogram $h_v(w)$ of vertical white run lengths is constructed and the mean value m_v of $h_v(w)$ is calculated. Line segments which correspond to white run lengths whose length is larger than $k \cdot m_v$ and spatially located in a candidate obstacle area are characterized as vertical obstacle lines. k is a constant and its value is experimentally set to 4. An example of column obstacle detection is given in Fig. 5b.

For the case of text line obstacles, line segments whose length is smaller than $M_v = \arg\max h_v(w)$ represent the text line obstacles. M_v indicates the distances between components belonging to different text lines. Fig. 5c shows an example of text line obstacles.

3.3 Draft text lines estimation

Starting from the image after noise and punctuation marks removal (see section 3.1), the algorithm forms text lines by applying DRLSA with $a = 1.5$ (see eq. 4) without crossing pixels which belong in the detected column and text line obstacles. Parameter T_l of DRLA of eq. 4 gets a large value since obstacles ensure that pixels belonging to different text lines will not be linked even if text is warped or text columns are very near. An example of text line formation is presented in Fig. 5d.

3.4 Word segmentation

The word segmentation procedure of the proposed technique is applied independently to each text line detected from the previous stage of the algorithm. All connected components of a text line L_i are first sorted according to their x coordinate and the histogram H_d of the horizontal distances between adjacent bounding boxes is constructed. A negative value for a distance of vertically overlapped bounding boxes is considered to be zero.

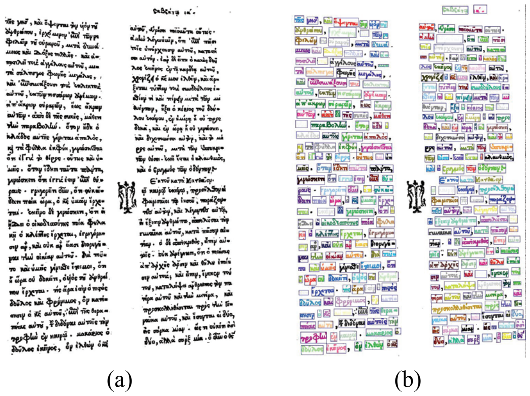


Figure 6. Example of word segmentation result.

In order to achieve a proper word segmentation of the components of the text line, adjacent connected components with distance smaller D_T are considered to belong to the same word. Threshold value D_T is defined by the following equation:

$$D_T = l + Max_v \qquad (7)$$

where Max_V represents the peak of the histogram H_d and l a constant tolerance value, experimentally set to 2.

4. Evaluation and experimental results

The proposed algorithm was tested on numerous historical and degraded machine-printed documents. An example of the achieved word segmentation results is demonstrated in Fig. 6. The algorithm performed successfully even in cases with text of different size, or with text and non-text areas lying very near, or with warped text lines. We observed some problems only in cases that the draft text line estimation failed. In order to compare with current state-of-the-art approaches, we also implemented two RLSA and projection profiles based approaches. Similar approaches have been used for the word segmentation of historical and degraded machine-printed documents in [8] and [7] respectively. We manually marked the correct word segments (ground truth) in a set of 15 historical images. The performance evaluation was based on counting the number of matches between the words detected by the algorithms and the words in the ground truth [11]. The performance was recorded in terms of detection rate and recognition accuracy, while as an overall measure we used the F-measure which is a weighted harmonic mean of detection rate and recognition accuracy [12]. As it is depicted in Table 1, the proposed algorithm outperforms the two other state-of-the-art approaches and achieves an overall evaluation measure of 75.8%.

Table1. Comparative results

	Detection rate	Recognition Accuracy	F-measure
RLSA based technique	70.4%	57.7%	63.4%
Projection Profiles based technique	69.1%	61.7%	65,2%
Proposed Technique	81.7%	70.7%	75.8%

5. Conclusions

In this paper, we propose a novel technique for word segmentation in historical and degraded machine-printed documents. The proposed technique faces problems such as having text of different size, having text and non-text areas lying very near and having non-straight and warped text lines. Comparative experimental results using several historical and degraded machine-printed documents demonstrate the efficiency of the proposed methodology. In the future work, we plan to extend our methodology for character segmentation as well as to adapt it to work with handwritten documents.

Acknowledgements

This research is carried out within the framework of the Greek Ministry of Research funded R&D project POLYTIMO [13] which aims to process and provide access to the content of valuable historical books and handwritten manuscripts.

References

[1] Y. Lu, C. Tan, H. Weihua, Fan L., “An approach to word image matching based on weighted Hausdorff distance”, *Sixth International Conference on Document Analysis and Recognition*, 2001, pp. 10-13.

[2] A. Marcolino, V. Ramos, M. Ármalo, J. C. Pinto, “Line and Word matching in old documents”, *Proceedings of the Fifth IberoAmerican Sympsium on Pattern Recognition*, 2000, pp. 123-125.

[3] H. Weihua, C. L. Tan, S. Y. Sung, Y. Xu, “Word shape recognition for image-based document retrieval”, *Int. Conference on Image Processing*, 2001, pp. 8-11.

[4] R. Manmatha, Jamie L. Rothfeder, “A scale space approach for automatically segmenting words from historical handwritten documents”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, pp. 1212-1225.

[5] H.C. Park, S.Y. Ok, Y.J. Yu, H.G. Cho, “A word extraction algorithm for machine-printed documents using a 3D neighborhood graph model”, *International Journal on Document Analysis and Recognition*, 2001, pp. 115-130.

[6] J. Park, V. Govindaraju, “Use of adaptive segmentation in handwritten phrase recognition”, *Pattern Recognition*, 2002, pp. 245-252.

[7] A. Antonacopoulos, D. Karatzas, “Semantics-Based Content Extraction in Typewritten Historical Documents”, *Eighth International Conference on Document Analysis and Recognition*, 2005, pp. 48-53.

[8] B. Gatos, T. Konidaris, K Ntzios, I. Pratikakis and S.J Perantonis, “A Segmentation-free Approach for Keyword Search in Historical Typewritten Documents”, *8th International Conference on Document Analysis and Recognition*, Seoul, Korea, 2005, pp. 54-58.

[9] F.M. Wahl, K.Y. Wong, R.G. Casey, “Block Segmentation and Text Extraction in Mixed Text/Image Documents”, *Computer Graphics and Image Processing*, 1982, pp. 375-390.

[10] T.M. Breuel, “Two geometric algorithms for layout analysis”, *Document Analysis Systems*, Princeton, NJ, 2002.

[11] A. Antonacopoulos, B. Gatos and D. Bridson, “ICDAR2005 Page Segmentation Competition”, *8th International Conference on Document Analysis and Recognition*, Seoul, Korea, 2005, pp. 75-79.

[12] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, ``Performance measures for information extraction,'' in *Proc. DARPA Broadcast News Workshop*, 1999, pp. 249-252.

[13] POLYTIMO project, http://iit.demokritos.gr/cil/Polytimo, 2007.