Line And Word Segmentation of Handwritten Documents

G. Louloudis[1], B. Gatos[2], I. Pratikakis[2], C. Halatsis[1]

[1]Department of Informatics and Telecommunications,
University of Athens, Greece
http://www.di.uoa.gr
louloud@mm.di.uoa.gr,
halatsis@di.uoa.gr

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

Abstract

In this paper, we present a segmentation methodology of a handwritten document in its distinct entities namely text lines and words. Text line segmentation is achieved making use of the Hough Transform on a subset of the connected components of the document image. Also, a post-processing step includes the correction of possible false alarms, the creation of text lines that Hough Transform failed to create and finally the efficient separation of vertically connected characters using a novel method. Word segmentation is treated as a two class problem. The distances between adjacent overlapped components in a text line are calculated and each of these is categorized either as an inter-word or an intra-word distance after the comparison with a threshold. The performance of the proposed methodology is based on a consistent and concrete evaluation technique that relies on the comparison between the text line segmentation result and the corresponding ground truth annotation as well as the word segmentation result and the corresponding ground truth annotation.

Keywords: Document Analysis, Handwritten Documents, Hough Transform, Text Line Segmentation, Word Segmentation.

1. Introduction

Segmentation of a document image into its basic entities namely text lines and words, is a critical stage towards handwritten document recognition. The difficulties that arise in handwritten documents make the segmentation procedure a challenging task. There are many problems encountered in the segmentation procedure. For the text line segmentation procedure these include the difference in the skew angle between lines on the page or even along the same text line, overlapping words and adjacent text lines touching. Furthermore, the frequent appearance of accents in many languages (e.g. French, Greek) makes the text line segmentation a challenging task. For the word segmentation process, difficulties that arise include the appearance of slant in the text line, the existence of punctuation marks along the text line and the non-uniform spacing of words.

In this paper, we present a segmentation methodology of a handwritten document in its distinct entities namely text lines and words. The main novelties of the proposed approach consist of (i) the extension of a previously published work for text line segmentation [8] taking into account an improved methodology for the separation of vertically connected text lines and (ii) a new word segmentation technique based on an efficient distinction of inter-word and intra-word distances.

The paper is organized as follows: in Section 2, the related work is described. In Section 3 the methodology to segment text lines is detailed. Section 4 deals with the word segmentation method. In Section 5, we present the experimental results and, finally, Section 6 describes conclusions and future work.

2. Related Work

A wide variety of text line detection methods for handwritten documents has been reported in the literature. There are mainly three basic categories that these text line detection methods fall in. Methods lying in the first category make use of the Hough transform ([7, 8, 14]). In these methods, by starting from some points of the initial image, the lines that fit best to these points are extracted. The points considered in the Hough transform are usually either the gravity centers [7] or minima points [14] of the connected components. In [8] a block based Hough transform approach is applied taking into account the gravity centers of parts of connected components, which are called blocks. Methods lying in the second category make use of projections ([1, 2, 3, 11]). In [2, 3], the methodology divides the document image into vertical strips and in these strips the horizontal projections are calculated. The resulting projections are combined in order to extract the final text lines. In [1, 10], the histogram of the pixels' intensities at each scan line is calculated. The

produced bins are smoothed and the corresponding valleys are identified. These valleys indicate the space between the lines of the text. Finally, the third category deals with methods that use a kind of smearing [16, 18]. In [16], a fuzzy runlength is used to segment lines. This measure is calculated for every pixel on the initial image and describes how far one can see when standing at a pixel along horizontal direction. By applying this measure, a new grayscale image is created which is binarized and the lines of text are extracted from the new image.

Some methods that do not lie in the previous categories are [13, 17]. In [13], the text line extraction problem is seen from an Artificial Intelligence perspective. The aim is to cluster the connected components of the document into homogeneous sets that correspond to the text lines of the document. Shi et al. [17], make use of the Adaptive Local Connectivity Map. The input to the method is a grayscale image. A new image is calculated by summing the intensities of each pixel's neighbors in the horizontal direction. Since the new image is also a grayscale image, a thresholding technique is applied and the connected components are grouped into location maps by using a grouping method.

For word segmentation there exist two distinct tendencies. In the first, after taking as input a text line image, the connected components are calculated. The distances between adjacent connected components are measured using a metric such as the Euclidean distance, the bounding box distance or the convex hull metric [10, 12, 15]. Finally, a threshold is defined which is used to classify the calculated distances as either inter-word or inter-characters gaps.

In the second [9], the word segmentation problem is considered as a text line recognition task, adapted to the characteristics of segmentation. That is, at a certain position of a text line, it has to be decided whether the considered position belongs to a letter of a word, or to a space between two words. For this purpose, three different recognizers based on Hidden Markov Models are designed, and results of writer- dependent as well as writer-independent experiments are reported in the paper.

All the above techniques do not deal successfully with the separation of vertically connected text lines which is a crucial aspect towards word recognition. For the word segmentation problem, there is an inefficient discrimination between inter-word and intra-word distances. These reasons motivated us to present a novel text line and word segmentation methodology in order to solve the above problems.

3. Text line segmentation

The proposed methodology for text line segmentation in handwritten document images deals with the following challenges: (i) each text line that appears in the document may have an arbitrary skew angle and converse skew angle along the text line; (ii) text lines may have different skew directions; (iii) accents may be cited either above or below the text line and (iv) parts of neighboring text lines may be connected.

The text line segmentation methodology is based on [8] and includes three steps: (i) pre-processing, (ii) Hough transform mapping and (iii) post-processing.

3.1. Pre-processing

The pre-processing step consists of three stages. In the first stage the connected components of the binary image are extracted since the input is a binarized image. Then, the average character height AH for the whole document image is calculated. We assume that the average character height equals to the average character width AW. The final stage includes the partitioning of the connected components domain into three sub-domains which are denoted as "Subset 1", "Subset 2" and "Subset3". These sub-domains are treated in a different manner by the methodology.

"Subset 1" contains all components which correspond to the majority of the characters with size which satisfies the following constraints:

$$(0.5 * AH \le H < 3 * AH) \text{ AND } (0.5 * AW \le W) \quad (1)$$

where H, W denote the component's height and width, respectively, and AH, AW denote the average character height and the average character width, respectively.

"Subset 2" contains all large connected components. Large components are either capital letters or characters from adjacent text lines touching. The size of these components is described by the following equation:

$$H \ge 3 * AH \quad (2)$$

Finally, "Subset 3" should contain characters as accents, punctuation marks and small characters. The equation describing this set is:

$$\begin{gathered} ((H < 3 * AH) \ AND \ (0.5 * AW > W)) \\ \text{OR} \\ ((H < 0.5 * AH) \ AND \ (0.5 * AW < W)) \end{gathered} \quad (3)$$

3.2. Hough transform mapping

In this stage, the Hough transform takes into consideration a sub-domain (denoted as "Subset 1") of the connected components of the image.

In our approach, instead of having only one representative point for every connected component, a partitioning is applied for each connected component lying in "Subset 1", so as to have more representative points voting in the Hough domain. This is accomplished by partitioning every connected component of the above set to equally- sized blocks. The width of each block is defined

by the average character width *AW*. An example is shown in Fig. 1. After the creation of blocks, we calculate the gravity center of the connected component contained in each block. The set of all this points contributes to the Hough transform.

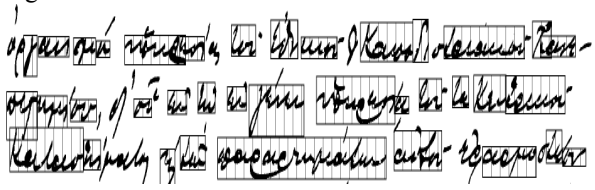


Figure 1. An example showing the connected components partitioning to blocks of width *AW*. All connected components not placed in a bounding box correspond to either "Subset 2" or "Subset 3".

The Hough transform is a line to point transformation from the Cartesian space to the Polar coordinate space. A line in the Cartesian coordinate space is described by the equation:

$$x\cos(\theta) + y\sin(\theta) = p \tag{4}$$

It is easily observed that the line in the Cartesian space is represented by a point in the Polar coordinate space whose coordinates are p and θ. Every point in the subset that was created above corresponds to a set of cells in the accumulator array of the (p,θ) domain. To construct the Hough domain the resolution along θ direction was set to 1 degree letting θ take values in the range 85 to 95 degrees and the resolution along p direction was set to 0.2**AH* (as in [7]).

After the computation of the accumulator array we proceed to the following procedure: We detect the cell (ρ_i,θ_i) having the maximum contribution and we assign to the text line (ρ_i,θ_i) all points that vote in the area $(\rho_i-5,\theta_i)..(\rho_i+5,\theta_i)$. To decide whether a connected component belongs to a text line, at least half of the points representing the corresponding blocks must be assigned to this area. After the assignment of a connected component to a text line, all votes that correspond to this particular connected component are removed from the Hough transform accumulator array. This procedure is repeated until the cell (ρ_i,θ_i) having the maximum contribution contains less than n_1 votes in order to avoid false alarms. During the evolution of the procedure, the dominant skew angle of currently detected lines is calculated. In the case that the cell (ρ_i,θ_i) having a maximum contribution less than n_2 ($n_2>n_1$), an additional constraint is applied upon which, a text line is valid only if the corresponding skew angle of the line deviates from the dominant skew angle less than 2°.

3.3. Post-processing

The post-processing procedure consists of two stages. At the first stage, (i) a merging technique over the result of the Hough transform is applied to correct some false alarms and (ii) connected components of "Subset 1" that were not clustered to any line are checked to see whether they create a new line that the Hough transform did not reveal. After the creation of the final set of lines, components lying in "Subset 3" as well as the unclassified components of "Subset 1" are grouped to the closest line.

The second stage deals with components lying in sub-domain "Subset 2". All components of this subset mainly belong to n detected text lines ($n>1$). Our methodology for splitting these components consists of the following steps:

STEP 1: Calculate y_i, which are the average y values of the intersection of detected line i and the connected component's bounding box ($i=1..n$) (see Fig. 2,3,4(a)).

STEP 2: Exclude from the procedure the last line n if the following condition is not satisfied:

$$\sum_{\substack{x=x_s \\ y=y_n-(y_n-y_{n-1})/10}}^{\substack{x_e \\ y_e}} I(x,y) \Big/ \sum_{\substack{x=x_s \\ y=y_{n-1}}}^{\substack{x_e \\ y_e}} I(x,y) > 0.08 \tag{5}$$

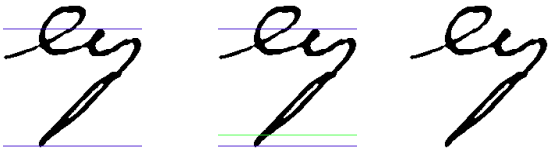
where (x_s,y_s), (x_e,y_e) are the coordinates of the bounding box of the component (see Fig. 2(b), 2(c)) and I the image of the component (value 1 for foreground and 0 for background pixels). Eq. (5) verifies that the component area near line n is due to a vertical character merging and not due to a long character descender from text line n-1 (see Fig. 2(c)).

STEP 3: For every line i, $i=1..n$-1, we define zones Z_i according to the following constraint:

$$y_i + \frac{y_{i+1}-y_i}{2} < y < y_{i+1} \tag{6}$$

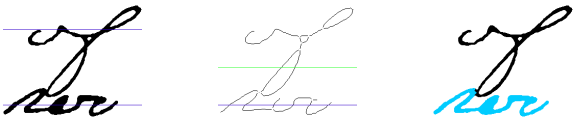
Then, we compute the skeleton of the connected component, detect all junction points and remove them from the skeleton if they lie inside zone Z_i. If no junction point exists in the segmentation zone Z_i we remove all skeleton points on the center of the zone (Fig. 3(b), 4(b))

STEP 4: For every zone z_i flag with id 1 the skeleton parts that intersect with line i. All other parts are flagged with id 2. Finally, in each zone z_i separation of the initial connected component into different segments is accomplished by assigning to a pixel the id of the closest skeleton pixel (Fig.3(c),4(c)).



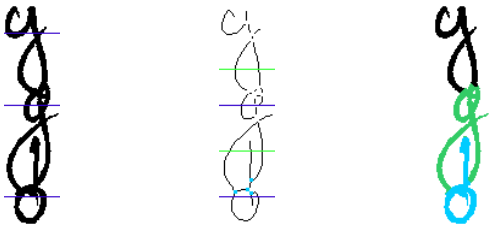
(a) (b) (c)

Figure 2: (a) the lines y_i , (b) area under green line checked in eq. 5 (c) final result.



(a) (b) (c)

Figure 3: (a) the lines y_i , (b) zones (starting from dotted line to solid line) (c) final result.



(a) (b) (c)

Figure 4: (a) the lines y_i, (b) zones (starting from dotted line to solid line) (c) final result.

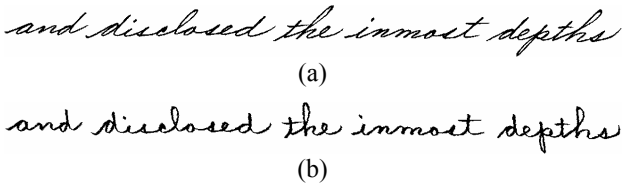
4. Word segmentation

The word segmentation procedure is divided into two steps. The first step deals with the computation of the distances of adjacent components in the text line image and the second step is concerned with the classification of the previously computed distances as either inter-word distances or inter-character distances.

4.1. Distances computation

In order to calculate the distances in the text line image, a pre-processing procedure is applied. The pre-processing procedure concerns the computation and correction of the dominant slant angle of the text line image (Fig. 5). The correction of the slant angle avoids the underestimation of the distance. Gap metrics for Roman-style text lines regard the gap as the space between two connected components (CC's), but we define the gap as the space between two overlapped components (OC's), where an OC is defined as a set of CC's whose projection profiles in the vertical direction are connected (see Fig. 6). The computation of the gap metric is considered not on the connected components but on the overlapped components.

The gap metric we used in the methodology is the Euclidean distance. The Euclidean distance between two adjacent overlapped components is defined as the minimum Euclidean distance among the Euclidean distances of all pairs of points of the two adjacent overlapped components (Fig. 7).



(a)

(b)

Figure 5: (a) a text line image and (b) the same text line after the slant correction.

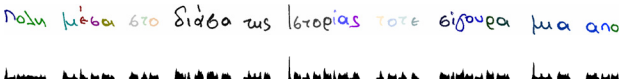


Figure 6: The overlapped components of the text line image are defined based on the projection profile.

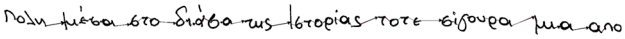


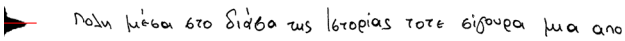
Figure 7: The arrows define the Euclidean distances between adjacent overlapped components.

4.2. Gap discrimination

For the gap classification we define a global threshold in the image. To compute this threshold we calculate the black to white transitions in every scanline of the text line image (Fig. 8). We focus on the scanline with the maximum number of black to white transitions. In this particular scanline we calculate and store all the lengths of the white runs and sort them in a top down order. Finally, we use the median of the sorted list in the line threshold equation which is defined as:

$$LT = 1.8 * M_v \tag{7}$$

where M_v is the median value of the sorted list. The applied weighting is after experimental work.



(a) (b)

Figure 8: (a) the black to white transitions in every scanline. Notice the maximum defined by the red line. (b) the text line image.

In order to define the global threshold we calculate the average of the temporary threshold along all text lines of the document image.

5. Experimental Results

We tested our methodology on a set of 102 images. This set includes images from the test set of the ICDAR 2007 Handwriting Segmentation Contest [6] as well as some images that were taken from the Historical Archives of University of Athens, Greece. We further considered historical handwritten documents which have several degradations and poor image quality. For all images, we have manually created the corresponding ground-truth in terms of text lines and words. The total number of text lines was 2240 while the corresponding number for words was 16983.

To check the effectiveness of our method we based on the evaluation methodology described in [6]. We also implemented two state of the art techniques for the sake of comparison. These techniques are the projection profiles [1] and the run length smearing algorithm [18]. The comparative results for the word segmentation in terms of detection rate (DR), recognition accuracy (RA) and FM ((2*DR*RA)/(DR+RA)) are shown in table 1. It is worth pointing out that our methodology outperforms the other two approaches achieving a detection rate of 90,4% and a recognition accuracy of 90,6%. We also took an experiment where the input to the word segmentation module was not the result of the text line segmentation. Instead, the input we gave was the ground-truth of the text lines which means that the input to the module was the perfect text lines. The results we obtained in terms of detection rate and recognition accuracy were 91,7% and 92% respectively.

Most of the errors encountered in the word segmentation phase are due to the non uniform spacing between characters of the same word image and between adjacent words.

An image example for the proposed methodology is shown in Figure 9. The results of the projection profile and the Rlsa methodology are shown in Figures 10 and 11 respectively.

Figure 9: An image example showing the final result of the proposed methodology.

Figure 10: An image example showing the final result of the projection profile methodology.

Figure 11: An image example showing the final result of the Rlsa methodology.

Table 1. Comparative experimental results.

Methodology	Detection Rate	Recognition Accuracy	FM
Proposed Methodology	90,4%	90,6%	90,5%
Projection Profile	65,7%	52,5%	58,4%
RLSA	76,4%	35%	48%

6. Conclusions and Future Work

In this paper, we present a segmentation methodology of a handwritten document in its distinct entities namely text lines and words. The main novelties of the proposed approach consist of (i) the extension of a previously published work for text line segmentation [8] taking into account an improved methodology for the separation of vertically connected text lines and (ii) a new word segmentation technique based on an efficient distinction of inter-word and intra-word distances.

Future work concerns the comparison of the proposed methodology with newer methodologies of the literature. Another issue is to try to use the result of the word segmentation method as a feedback in the text line process in order to further improve its results.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement n° 215064 (project IMPACT) and by the Greek Secretariat for Research and Development under the PENED 2003 framework.

References

[1] Esra Ataer, Pinar Duygulu, "Retrieval of Ottoman Documents", *Proceedings of 8th ACM SIGMM International Workshop on Multimedia Information Retrieval* , October 26-27, 2006, Santa Barbara, CA, USA.

[2] M. Arivazhagan, H. Srinivasan, and S. N. Srihari, "A Statistical Approach to Handwritten Line Segmentation", in *Document Recognition and Retrieval XIV, Proceedings of SPIE*, San Jose, CA, February 2007, pp. 6500T-1-11.

[3] Elisabetta Bruzzone, Meri Cristina Coffetti, "An Algorithm for Extracting Cursive Text Lines", *Proceedings of the Fifth International Conference on Document Analysis and Recognition,* Bangalore, India, 1999, pp. 749.

[4] Fu Chang, Chun-Jen Chen, Chi-Jen Lu, "A Linear-Time Component-Labeling Algorithm Using Contour Tracing Technique", *Computer Vision and Image Understanding, Vol. 93, No.2,* February 2004, pp. 206-220.

[5] 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 (ICDAR'05)*, Seoul, Korea, August 2005.

[6] B. Gatos, A. Antonacopoulos and N. Stamatopoulos, "ICDAR2007 Handwriting Segmentation Contest", *9th International Conference on Document Analysis and Recognition (ICDAR'07)*, Curitiba, Brazil, September 2007, pp. 1284-1288.

[7] Laurence Likforman-Sulem, Anahid Hanimyan, Claudie Faure, "A Hough Based Algorithm for Extracting Text Lines in Handwritten Documents", *Proceedings of the Third International Conference on Document Analysis and Recognition,* Montreal, Canada, 1995, pp. 774-777.

[8] G. Louloudis, K. Halatsis, B. Gatos, I. Pratikakis, "A Block-Based Hough Transform Mapping for Text Line Detection in Handwritten Documents", *10th International Workshop on Frontiers in Handwriting Recognition (IWFHR 2006),* La Baule, France, October 2006, pp. 515-520.

[9] F. Luthy, T. Varga, H. Bunke, "Using Hidden Markov Models as a Tool for Handwritten Text Line Segmentation", *Ninth International Conference on Document Analysis and Recognition,* Curitiba, Brazil, 2007, pp. 8-12.

[10] U. Mahadevan, R. C. Nagabushnam, "Gap metrics for word separation in handwritten lines", *Third International Conference on Document Analysis and Recognition,* Montreal, Canada, 1995, pp. 124-127.

[11] R. Manmatha, J. L. Rothfeder, "A Scale Space Approach for Automatically Segmenting Words from Historical Handwritten Documents", *IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.27, No.8,* August 2005, pp. 1212-1225.

[12] U. V. Marti, H. Bunke, "Text Line Segmentation and Word Recognition in a System for General Writer Independent Handwriting Recognition", *Sixth International Conference on Document Analysis and Recognition,* Seattle, WA, USA, 2001, pp. 159-163.

[13] S. Nicolas, T. Paquet, L. Heutte, "Text Line Segmentation in Handwritten Document Using a Production System", *Proceedings of the 9th IWFHR,* Tokyo, Japan, 2004, pp. 245-250.

[14] Y. Pu and Z. Shi, " A Natural Learning Algorithm Based on Hough Transform for Text Lines Extraction in Handwritten Documents", *Proceedings of the 6 International Workshop on Frontiers in Handwriting Recognition*, Taejon, Korea,1998, pp. 637-646.

[15] G. Seni, E. Cohen, "External Word Segmentation of Off-line Handwritten Text Lines", *Pattern Recognition,* 27(1): 41-52,1994.

[16] Z. Shi, V. Govindaraju, "Line Separation for Complex Document Images Using Fuzzy Runlength", *First International Workshop on Document Image Analysis for Libraries,* 2004, pp. 306.

[17] Z. Shi, S. Setlur, and V. Govindaraju, "Text Extraction from Gray Scale Historical Document Images Using Adaptive Local Connectivity Map", *Eighth International Conference on Document Analysis and Recognition,* Seoul, Korea, 2005, pp. 794-798.

[18] Wahl, F.M., Wong, K.Y., Casey R.G.: "Block Segmentation and Text Extraction in Mixed Text/Image Documents" Computer Graphics and Image Processing, 20 (1982) 375-390.