Segmentation of Handwritten and Printed Arabic Documents

Ghazouani Fethi, IFN[1], ENIT, Tunis, Tunisia Email: gfethi@yahoo.fr

and

Maddouri Mondher FST, Tunis, Tunisia Email: mondher.maddouri@fst.rnu.tn

Maddouri Snoussi Samia, ENIT, Tunis, Tunisia Email: samia_maddouri@yahoo.f

El Abed Haikel, Email: elabed@tu-bs.de [1]

Volker Margner, Email: Maergner@ifn.ing.tu-bs.de [1]

[1]Institute for Communications Technology, Braunschweig, Germany

***Abstract*—on this paper, we proposed a new text line segmentation of handwritten and typewriting Arabic document images that uses the Outer Isothetic Cover (OIC) algorithm of a digital object. In the first step, we use this method to segment the composed document into text blocs. In the second step, for each text bloc we will extract the text lines. Finally, line text will be segmented into words or into pieces of Arabic word (PAWs).**

The first results obtained in the current stage of the proposed method over a dozen texts are encouraging. We have also tested this method on documents written in Latin scripts.

***Keywords*—handwritten and modern document; text line segmentation; document image; pieces of Arabic words**

I. INTRODUCTION

The first step in the automatic document recognition is the segmentation of the text image into text line. The objectives of this step is to assign each component of the text to the appropriate line; to make it possible to prepare the data for further processing such as normalization, word segmentation and feature extraction.

The segmentation of handwritten text is complicated by the variation of the distance interline and the undulation of baselines generate different orientations of the text. The characters in two lines of text may touch or overlap. This considerably complicates the segmentation line. In Arabic script, these situations are frequently due to the presence of ascending and descending characters. The massive presence of diacritical symbols often generates false lines.

Most work on the segmentation of a page in line is based on a decomposition of the image into connected components. After the separation of lines, we focus on the separation of words for each line, then segmenting each word in pieces (parts) of words.

In the framework of this article, we focus essentially on image segmentation of Arabic documents into blocks of text and lines. Then we apply our method to the segmentation of Latin documents.

In the following, we present some technique applied to the segmentation of documents into text lines. Then we present our approach to the segmentation of Arabic handwritings and printed documents into blocks, text lines and words or parts of words.

The results of our segmentation method are shown subsequently, by tests on historical and modern documents.

Finally we end our article with a conclusion and perspectives that show a possible extension of the proposed approach.

II. RELATED WORKS

Several works have been proposed for the segmentation of documents. For example, Bennasri and al. have proposed a method to extract lines of text Arabic script, using the projection [2]: First the document is divided into multiple columns to correct the problem of sinuosity. Then, the starting points of all lines are detected using the minimum partial projection of the profile. Then, a contour tracking part of each line is carried out: first in the direction of writing, then in the opposite direction.

Nicolaou et al. [3] proposed a technique to segment the lines of Latin manuscripts using tracers (axes) minima. These minima are estimated using the vertical projection histogram. This method was tested on a sub-database of ICDAR 2007 consists of 20 documents containing 476 lines and 80 documents containing 1771 lines. This technique achieved an extraction rate that is equal 98.6%.

Another approach to the extraction lines Arabic manuscripts of ancient texts was proposed by Zahour and al. in [4]. Initially, the document is divided into columns of equal size. Then, each column of the document is segmented into three types of text blocks: small blocks which generally represent the diacritical symbols, medium blocks that correspond to body text and large

blocks reflect the overlap of words between adjacent lines.

There are also methods of segmentation using the Hough transformed. This technique is widely used for extraction of text lines [5]. For example in [6], the Hough transform is used with a method of grouping connected components. For this, the connected components are extracted and then the contours and edges of these components are detected.

Louloudis et al. proposed in [7], a technique for the extraction of lines and words of ancient Greek manuscripts. The Hough transform is applied to connected components using the centroids of rectangles encompassing their points as voters. These rectangles are estimated by calculating the average size of characters in the document. The proposed system was tested on the basis of documents ICDAR 2007 which is divided into 80 documents containing 1773 lines and 40 ancient manuscripts containing 1095 lines. The extraction rate of lines is 97%.

Another method used for segmentation is the snake or the contour. With this technique Bukhari et al. proposed a method for extracting parameterized snake lines of handwritten documents [8]. The proposed system was tested on the basis of documents ICDAR 2007 which is divided into 80 documents containing 1770 lines. The extraction rate of line is 96.3%.

Du et al. used the Mumford-Shah model [10] for the extraction of the lines of Latin manuscripts [9]. The proposed system was tested on 100 Chinese documents, 96documents and 100 Indian Korean documents. The extraction rate of lines is 98% for Chinese documents, 98% for Indian documents and 96% for Korean documents.

A new method proposed by Vasant Manohar et al. in [11], this method involves grouping text lines segmented by a set of methods for segmentation of handwritten texts line in an undirected graph. The graph nodes correspond to connected components and the edge connecting pairs of connected components.

III. PROPOSED WORK

The proposed method realizes the segmentation of handwritten and/or printed text lines, into words and into pieces of words. It is based on the algorithm for the construction of the isothetic covers of a digital object [1].

We thought to find a new segmentation method of document images. We started by construct the Outer Isethetic Cover (OIC) of documents. So, we made a change to this algorithm in order to segment a document into blocks of text, if the document is composed of text blocks, each block is then segmented into a set of lines then each line will be segmented into words and/or pieces of words.

A. Construction of the Outer Isothetic Cover (OIC)

In order to segment to segment document, we construct the outer isothetic covers of the corresponding document image after its binarization. To do this, we impose an isothetic set of grid size g on the binarized image.

Let Q_1, Q_2, Q_3, and Q_4 be the four quadrants incident at a grid point $p(i, j)$, as shown in Fig. 1. The grid point p is decided to be a vertex depending on how many of the quadrants have object containment. Interestingly, there arise $2^4 = 16$ different arrangements considering object containments of these four quadrants, which can be reduced to five cases. Let Cq ($q = 0, 1, \cdots, 4$) denote the case of all the arrangements for which q out of 4 squares are occupied by the object. If p belongs to Case C_1, then it is a 90^0 vertex of the isothetic polygon; and if it is a 270^0 vertex, it belongs to C_3. For Case C_2, if the diagonal quadrants are occupied, then p is considered as a 270^0 vertex; otherwise, p is a nonvertex grid point lying on some edge of the polygon.

For case C_0, p is just an ordinary grid point lying outside the polygon, whereas, for case C_4, p is a grid point lying inside the polygon [1].

So, in order to draw to draw polygon correspond of text blocks, of text lines and of multiple words. We have modified the algorithm TIPS [1]. With a proper grid size g, each polygon is constructed.

First, the document is binarized. Then the grid points **are traversed in the raw-major order until a 90° vertex** ('start vertex') is found. Subsequent grid points are classified, marked as 'visited', and the direction is determined from each such grid point until the start vertex is reached. This completes the outer isothetic cover corresponding to an object (text blocks, text lines or word). The procedure is iterated over the remaining set of unvisited grid points until the next 90° vertex is found, which subsequently derives the polygon corresponding to another object. Finally, all the grid points are visited and the algorithm reports the vertex sequences of all the isothetic polygons corresponding to the text blocks, to text lines or words in the input document.

Setting the grid size *g*: In order that each isothetic polygon corresponds to each object and hence results in a sequence of vertices, specifying an appropriate grid size is necessary. So, for each case of segmentation, the grid size g is chosen after a set of tests on a set of document images.

B. Segmentation of the document in text blocks

A document can be composed by one or more paragraphs (or blocks of text). These text blocks can be

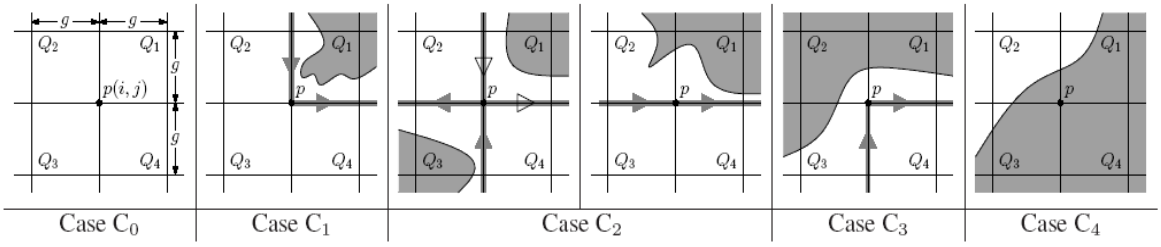


Figure 1. Five combinatorial cases (16 subcases) [12].

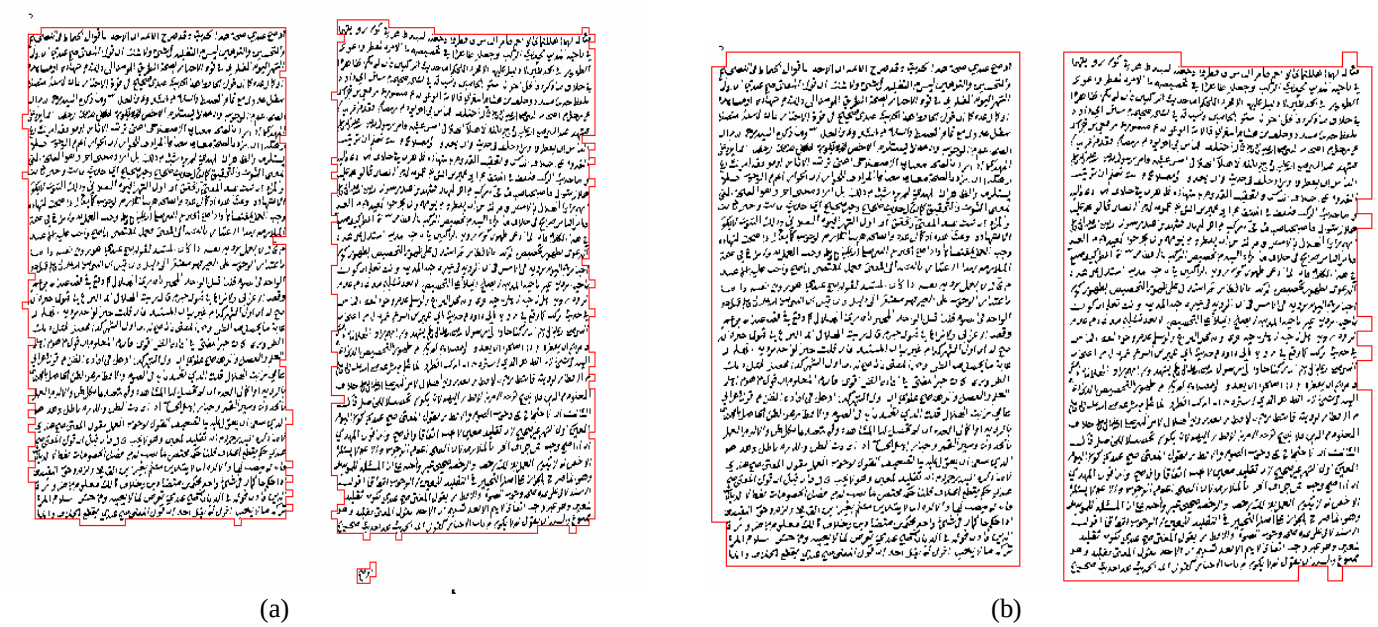
arranged in parallel horizontally or vertically (Figure 2).

In this case, we thought to divide the document into parts of texts designed to simplify further processing i.e. segmentation of the text lines.

The extraction of these text blocks is made by varying the size of the grid *g*, the more the size *g* is large, the more between them. Contrary to the Latin script, a line can be segmented into word or into characters.

IV. EXPREMENTS AND RESULTS

In order to evaluate the results of our approach, we have



(a) (b)

Figure 2. Segmentation of handwritten Arabic document into text block (a) $g = 16$ (b) $g = 13$.

the results are better. The figure 2 shows the results for document segmentation into blocks of text by changing the grid size *g* for handwritten Arabic texts.

C. Segmentation of (block) text in lines

A text or text block can be segmented into text lines. By changing the grid size *g*, we have applied our approach to segment a text into texts line. At the difference to segment a document to text blocks, the same algorithm is applied with the operator of mathematical morphology (the closure) in order to obtain the entire polygon line. The results are shown in Fig. 4.

D. Segmentation of line into words and/or into pieces of word :

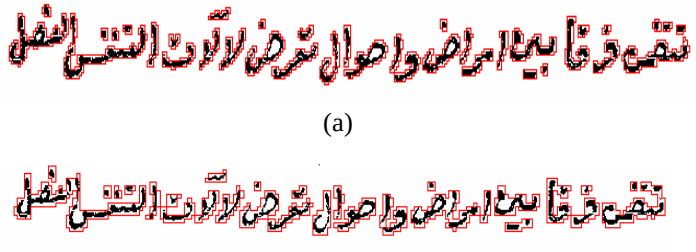
Then we have used the algorithm to extract words from line. A line can be segmented into words and/or parts of word Arabic manuscript (printed respectively). This is because the Arabic writing is recursive. The word can be composed by parts of words (Pieces of Arabic Words (PAWs)) and sometimes there is enough space

tested this algorithm on a variety of documents of two types of scripts: handwriting and printed. Handwritten Arabic documents composed by text blocks were segmented into blocks; the result is shown in Figure 1. As for the handwritten script, we have applied our method on documents printed Arabic and Latin;

The first database is a collection of 200 forms written by 200 different native writers. The writers were asked to write a paragraph of an Arabic text including up to 10 sentences. There were no restrictions for the writing. This database is an extension of the standard benchmarking IfN/ENIT database. The second collection of Arabic handwritten documents includes scans from historical documents collected during a research project in the IfN. The printed Latin text is from the Google Books (version 7, August 2007)

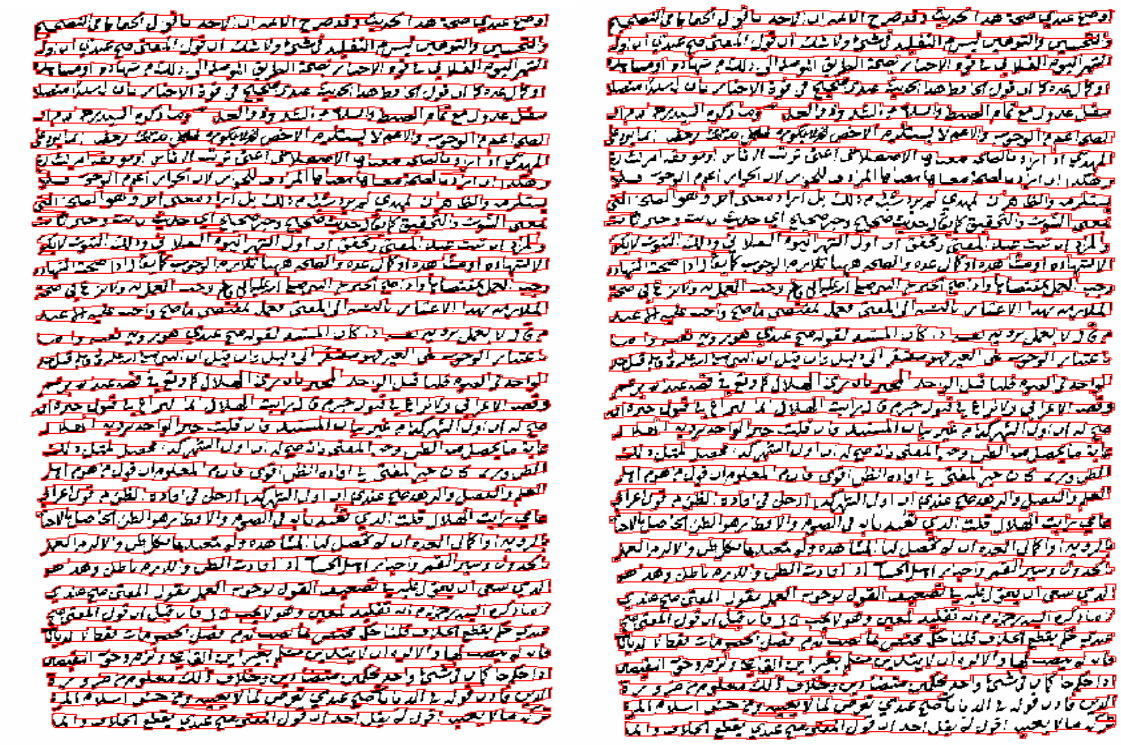
Then we tried to do the segmentation text blocks into text lines. The result of the segmentation of handwritten Arabic text is shown below. In the same way, we are showing in the following figures the result of such segmentation for printed Arabic and Latin script:

The last step of our method is to extract the connected components from the line. The result of handwritten Arabic text line segmentation into words or parts of words has been shown in Fig. 3. In the same step we are applied our method on a printed text line Arabic and Latin.



(a) (b)

Figure 3. Segmentation of handwritten text line Arabic into word or into pieces of words: (a) $g = 2$ (b) $g = 4$



(a) (b)

Figure 4. Text lines segmentation (a) $g = 1$ (b) $g = 2$

V. CONCLUSION

In this paper we presented some techniques of segmentation methods. Then we proposed a new segmentation method for document images handwritten and printed script. The idea of this method is inspired from the algorithm of construction of isothetic covers of a digital object [1]. So we have shown that the results of such segmentation depend of the variation of the grid size g. Then, to segment composed documents into text blocks, we used a large value of *g*. And to extract the line text from blocks and the words or pieces of words from text line, we have reduced the grid size *g*.

The results of our method are preferment for proper images document, especially for the type of printed texts document. This because in this type of script, the characters of two lines can neither touches nor overlaps. Instead of the handwritten, these situations exist frequently, which will sometimes give incorrect results.

REFERENCES

[1] A. Biswas, P. Bhowmick, B.B. Bhattacharya, Construction of isothetic covers of a digital object: A combinatorial approach 2010.

[2] Bennasri, A., Zahour, A. et Taconet, B. (1999). Extraction des lignes d'un texte manuscrit arabe. *Vision Interface'99*.

[3] Nicolaou, A. et Gatos, B. (2009). Handwritten text line segmentation by shredding text into its lines. *International Conference on Document Analysis and Recognition*.

[4] Zahour, A., Likforman-Sulem, L., Boussellaa, W. et Taconet, B. (2007). Text line segmentation of historical arabic documents. *In 9th Int.Conf. on Document Analysis and Recognition.*

[5] Duda, R. O. et Hart, P. E. (1972). Use of the hough transformation to detect lines and curves in pictures. *Commun. ACM*.

[6] Malleron, V., Eglin, V., Emptoz, H., Dord-Crouslé, S. et Régnier, P. (2009). Text lines and snippets extraction for 19th century handwriting documents layout analysis. *International Conference on Document Analysis and Recognition*.

[7] Malleron, V., Eglin, V., Emptoz, H., Dord-Crouslé, S. et Régnier, P. (2009). Text lines and snippets extraction for 19th century handwriting documents layout analysis. *International Conference on Document Analysis and Recognition*.

[8] Bukhari, S. S., Shafait, F. et Breuel, T. M. (2009). Scriptindependent handwritten textlines segmentation using active contours. *In ICDAR09.*

[9] Du, X., Pan, W. et Bui, T. D. (2009). Text line segmentation in handwritten documents using mumford-shah model. *Pattern Recognition.*

[10] Mumford, D. et Shah, J. (1989). Optimal approximation by piecewise smooth functional and associated variational problems. *Commun. Pure Appl. Math.*

[11] Vasant Manohar, Shiv N. Vitaladevuni, Huaigu Cao, Rohit Prasad, and Prem NatarajanGraph Clustering-based Ensemble Method for Handwritten Text Line Segmentation. ICDAR 2011.

[12] Aisharjya,Sakar et al. Word Segmentation and Baseline Detection in Handwritten Documents Using Isothetic Covers. *International Conference on Frontiers in Handwriting Recognition* 2010