Comparative Study of Two Segmentation Methods of Handwritten Arabic Text
MM-OIC and HT-MM

Fethi Ghazouani[1], Samia Snoussi Maddouri[2,3] and Fadoua Bouafif Samoud[2]
[1]*LIPAH - FST, University of Tunis El Manar, Tunis, Tunisia*
[2]*National Engineering School of Tunis (ENIT), University of Tunis El Manar, Tunis, Tunisia*
[3]*Hnayka Community Faculty, Taibah University, Medina, Saudi Arabia*

Keywords: Arabic Handwriting Documents, Segmentation, Mathematical Morphology, Outer Isothetic Cover, Hough Transform.

Abstract: We present in this paper a comparative study of two segmentation methods of handwritten Arabic text. The first method is a combination of the Mathematical Morphology (MM) and the algorithm of construction of the Outer Isothetic Cover of a digital object (OIC) named MM-OIC. The second method uses the Hough Transform (HT) and MM to segment the handwriting Arabic script called HT-MM. These methods are applied in two levels of segmentation: text lines and Pieces of Words. The two proposed methods are evaluated and compared to a set of documents selected from three databases: IFN/ENIT-database (17 documents), BSB (16 documents) and KSU (30 documents) online databases. The average rate line segmentation of MM-OIC is 75%, and of HT-MM is 45%. The average rate of PAW segmentation acheive 89% for the MM-OIC and 70% for the HT-MM method. The efficiency of the MM-OIC method is explained by the fact that this method can extract the approximate form of writing, and sometimes it can exceed some problems that are related to the Arabic script such as the overlapping lines and diacratical symbols.

1 INTRODUCTION

The first step in the automatic document recognition is the segmentation of the text image into text lines. This prepares the data for further processing steps like normalization, word segmentation and features extraction.

The segmentation of handwritten text is complicated by the variation of the interline distance and by the baselines undulation that often generates different orientations of the text. The characters in two adjacent lines may touch (Figure 1(a)) or overlap (Figure 1(b) and 1(c)). This considerably complicates the text lines segmentation. In Arabic script, these situations frequently exist because of the presence of ascendant and/or descendant characters. On the other hand, the massive presence of diacritical symbols in Arabic script often generates false lines (Figure 1(d) and 1(e)).

In the literature, most works of page segmentation in lines are based on the decomposition of the image content into connected components. For Arabic handwriting, the extraction of pieces of words seems to be easier than the detection of words. Several segmentation methods for lines extraction have been presented such as: projection method ((Bennasri et al., 1999), (Nicolaou and Gatos, 2009)), k-means algorithm (Zahour et al., 2007), Hough transform method ((Bouafif

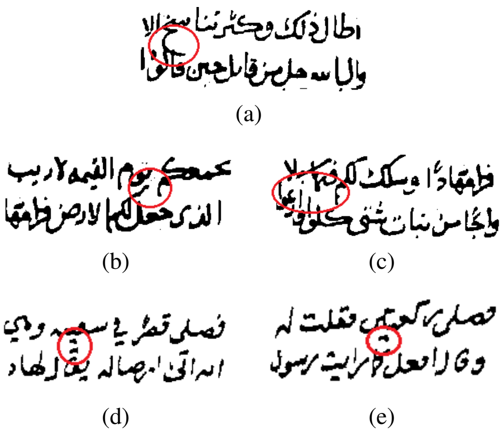

(a)
(b)
(c)
(d)
(e)

Figure 1: Examples of existing situations in Arabic script. (a) touch lines. (b,c) lines overlap. (d,e) presence of diacritical symbols between lines.

Ghazouani F., Snoussi Maddouri S. and Bouafif Samoud F..
Comparative Study of Two Segmentation Methods of Handwritten Arabic Text - MM-OIC and HT-MM.
DOI: 10.5220/0004764205280535
In *Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods* (ICPRAM-2014), pages 528-535
ISBN: 978-989-758-018-5

et al., 2006), (Malleron et al., 2009)) and snake technique or active contour (Bukhari et al., 2009). On the same, some works are proposed for the segmentation in pieces of words for instance: contour technique (Snoussi, 2003), labeling-recognition (Abdulkader,) and projection methods (Sarfaz et al., 2003).

In this framework, two segmentation methods of handwritten Arabic documents are proposed. The First method is based on a combination between Mathematical Morphology (MM) and Outer Isothetic Cover (OIC). The second method is a combination between Hough Transform (HT) and Mathematical Morphology (MM) operators. Finally, a comparative study of the two methods on the obtained results is done. We end our paper with a conclusion and perspectives that show possible extensions of this work.

2 MM-OIC SEGMENTATION METHOD

The MM-OIC method is a combination of the mathematical morphology and the algorithm of construction of the outer isothetic cover of a digital object. This method is used for the segmentation of the handwritten Arabic text into lines and into sub words. To extract lines from Arabic text, firstly, we have applied morphologic operators to connect components belonging to the same line. Then we have applied the algorithm of the construction of OIC on the MM's result in order to delimit zones of each line.

2.1 Connect Adjacent Component: MM

The Arabic script is generally represented in the form of pieces or parts of words called Pieces of Arabic Words (PAWs) (Miled et al., 1998). Each PAW represents a connected component. For this specific reason we thought about the application of morphological operators for connecting small adjacent components belonging to the same line. In this case, the latest is seen as a single object. The fact of having a single object allows to facilitate the extraction of the line after the application of the OIC algorithm (subsection 2.2). In this step we chose the closing as a basic morphologic filter to connect components. The structuring element is a rectangular form with length 1 and width 25. The closing operation allows us to connect horizontally the different components which are situated at a distance less than the structuring element. The structuring element is chosen after a learning phase on the treated images documents. The set of the used documents to choose the size of the structuring element consists of 25 images and selected from three different databases. Some of them are written by different scripters, containing curved or skewed lines and sometimes with variable sizes of the script. We have also chosen the size of the structuring element such that after the application of the MM, the adjacent lines can not be touching and overlapped.

Indeed, the choice of a length equal to 1 is explained by the fact that in the set of selected document, there is presence of texts with a very small interline and that sometimes the distance between certain characters, especially the ascendants and the descendants of two adjacent lines of text is very small. So, increasing the length value of the structuring element can lead to a touch between writing. Figure 2 illustrates an example of this situation. The distance between the two green characters of the two successive lines is too small (Figure 2(a)). Figure 2(b) shows the result of the MM after application of a structuring element with length = 1, it is clear that the components of each row are connected together. Contrariwise, in Figure 2 (c), the components of the two different lines are connected in a single object (instead of two objects) and are touched (red circle) after increasing the length of the structuring element (length = 4).

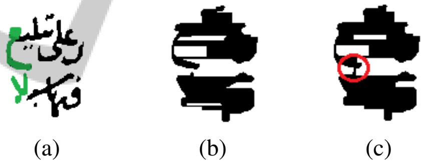


Figure 2: Results of the MM with two different values of length of the structuring element. (a) original image. (b) Applying of a structuring element with a length = 1. (c) Applying of a structuring element with a length = 4.

Secondly, in the Arabic script the inter-words space is often variable. Then, for connecting horizontally those components, we must choose an adequate width value. Two situations can exist with decreasing and increasing the width value. Decreasing this value, generally, allows only the connection of the very nearest components. In this case, we can have several big "sub-connected" objects instead of a single connected object. On the other hand, the ascendant and descendant characters often overlap and the distance between a descendant (respectively ascendant) of a line and another ascendant (respectively descendant) of the next line (respectively the previous line), in the horizontal sense, is smaller. Increase the width value allows these specific characters to be touched. Figure 3 presents an example of the second case. In Figure 3(a), the distance (in the horizontal sense) between the descendant character (green color) of the first line and the ascendant character (green color) of the second line is smaller. The correct result of the MM is obtained after applying of a structuring element with a width value = 25 (Figure 3(b)). Contrariwise, in Figure 3(c), the components of the two different lines are connected in a single object (instead of two objects) and are touched (red circle) after increasing the width of the structuring element (width = 30).

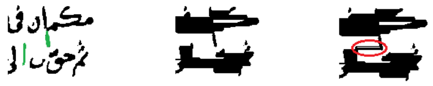

(a) (b) (c)

Figure 3: Results of the MM with two different values of width of the structuring element. (a) original image. (b) Applying of a structuring element with a width = 25. (c) Applying of a structuring element with a width = 30.

So, after an experimental phase, the best chosen width value of the structuring element is 25.
Figure 4 illustrates the result of concatenation of connected components belonging to the same line. Figure 4(a) presents the initial image and Figure 4(b) shows the morphologic operator effect.

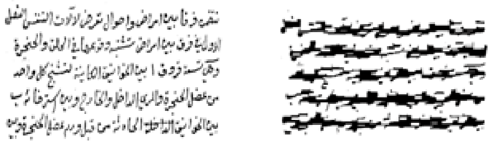

(a) (b)

Figure 4: Mathematical morphology applied to connect different components belonging to the same line. (a) original image. (b) connected component concatenation result.

2.2 Line Extraction: Construction of OIC

In order to extract lines in a handwritten Arabic document, we construct the outer isothetic cover of the corresponding document after applying the MM by the use the algorithm given by (Biswas et al., 2010).

The isothetic cover of a digital object specifies a simple representation of the object and provides approximate information about its structural content and geometric characteristics. When, the cover ”tightly” encloses the object, it is said to be an outer isothetic cover (OIC). And when the cover inscribes the object, it is considered as an inner isothetic cover (IIC) (Biswas et al., 2010). The outer isothetic cover is defined by a set of isothetic polygons, having their edges lying on the grid lines, such that the effective area corresponding to the object is minimized (Biswas et al., 2010). The Figure 5 shows an example of the set of outer polygons for different handwritten Arabic lines. The outer isothetic covers are obtained for grid size g = 2.

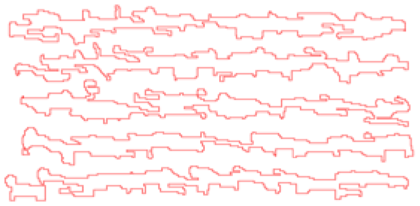
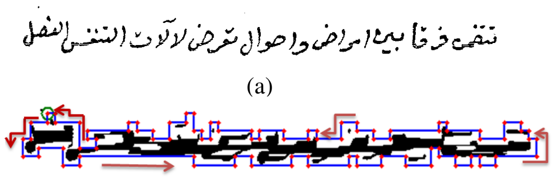


Figure 5: Set of polygons representing handwritten Arabic lines corresponding to results of MM of Figure 4(b).

Then, to draw multiple polygons corresponding to different lines, we modify the algorithm given by (Biswas et al., 2010). This idea was used by Sarkar for words segmentation of handwritten Bangla documents (Sakar et al., 2010). So, the algorithm is applied on the result of MM on handwritten Arabic document. With a proper grid size each polygon corresponds to a single line in the initial document.

In fact, on the result image of MM the grid points are traversed in the raw-major order until a 90° vertex (start vertex) is found ((Biswas et al., 2010), (Sakar et al., 2010)). Subsequent grid points are classified, marked as ”visited”, and the direction is determined from each grid point to the start vertex. Finally, the OIC is constructed when the start vertex is reached again. Figure 6 presents an example of construction of OIC for the handwritten Arabic text line after application of MM. In this figure each vertex is represented by a red point, the start vertex is surrounded (green circle) and the OIC of a line is represented with blue color.


(a)

(b)

Figure 6: Construction of the OIC of handwritten Arabic lines. (a) original image, (b) result of OIC algorithm.

With this process we extract text lines from a document image. In Figure 7 we show a simple example of text lines extraction from handwritten Arabic document. Figure 7(a) presents the initial image. Figure 7(b) illustrates the result of the MM applied to the initial image. Figure 7(c) shows the OICs construction of different objects with an appropriate grid size g (g = 2), which has been defined after a number of experimental tests on a set of 15 documents, on the result of

MM: the green polygons generally represent diacritical symbols that can be identified by their perimeters, while each large polygon represents a text line. The text line extraction result is shown in Figure 7(d).

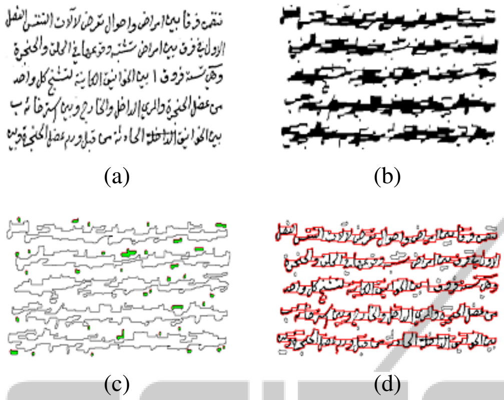

(a) (b) (c) (d)

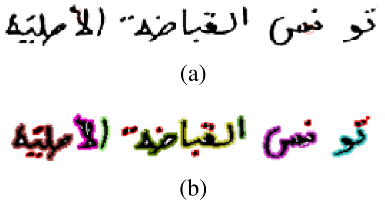
Figure 7: Text lines extraction with MM-OIC method. (a) original image, (b) MM result, (c) OIC algoritm result, (d) text lines extraction.

2.3 PAWs Segmentation

In (Snoussi, 2003) a PAW is described by a closed contour which is neither a diacritic point nor a loop. For this case, we focus on the separation of connected components of each extracted line. We apply the algorithm of the construction of the OIC, without the application of the MM, with a grid size g = 1 on the result image of line extraction step, we attain the segmentation of the line into PAWs. In this case the choice of the value of the grid size is not arbitrary. Because, with the smallest value of g (g = 1), the OIC algorithm allows the construction of the exact shape of each detected component. The set of extracted polygons contains some small polygons corresponding generally to diacritic points or punctuation mark and identified by their perimeters. Figure 8 illustrates an example of a handwritten Arabic line segmented into PAWs. Figure 8(a) presents a handwritten Arabic line. The PAWs segmentation result is shown in Figure 8(b). In this figure, each PAW is represented by a colored polygon. The red polygons represent diacritic points and punctuation marks.

3 HT-MM METHOD

The HT-MM segmentation method is a combination between the Hough Transform (HT) and Mathematical Morphology (MM). This method is applied on the handwritten Arabic script in order to detect text line and PAWs also.


(a)

(b)

Figure 8: Segmentation of handwritten Arabic line in PAWs. (a) Handwritten Arabic line. (b) Segmentation result.

3.1 Text Line Detected

The HT (Duda and Hart, 1972) is applied to the binarized edge map to generate its Hough image. For this purpose the parameters of the Hough transform are tuned in such a way that the lines are extracted as a set of connected PAWs (Bouafif et al., 2012).

In this stage the MM is applied in order to connect PAWs and charchters. Indeed, the closing filter is one of the basic tools of MM; it permits the connection of adjacent components that have a separate distance that is lower than the structuring element. The application of a closing filter, using a rectangular form of the structuring element of length 10 and width 2 on the result of HT, enabling us to construct continuous lines. This step is applied three times to connect the different components existing on the same horizontal line. The structuring element is chosen after a learning phase on the processed documents images with the same steps explained as before. A labeling stage is then applied in order to associate a label color to each connected component. Each color represents handwritten text line. Figure 9 presents the different steps of text lines extraction of a handwritten Arabic document. The initial image is shown in Figure 9(a). Figure 9(b) illustrates all lines detected after the HT application. Figure 9(c) shows the effect of MM and the labeling stage applied to the HT result. Each components color represents a text line in Figure 9(d).

3.2 PAWs Segmentation

Proposed PAW segmentation method is composed by two stages: Detection of baselines and extraction of PAW's.

3.2.1 Detection of the Baseline of the Text Lines

Baseline is an artificial line composed by a sequence of aligned pixel that connects the maximum black pixels of the characters in the text line. Different baseline extraction methods are abound in the literature (Snoussi et al., 2008). The proposed baseline detection method is applied to the text binary image with-

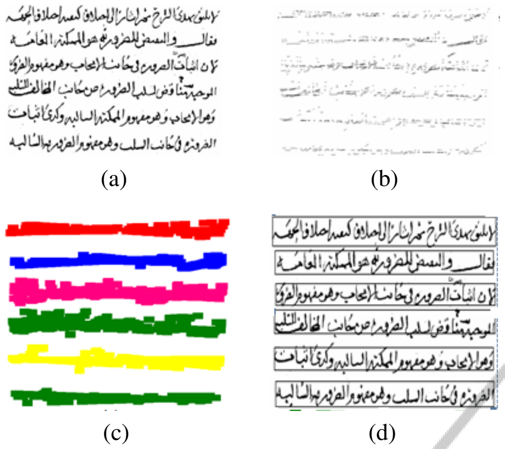


Figure 9: Different steps for text line detection.

out slant correction. It is based on the HT in order to detect the median baseline. In fact the maxima of the accumulator present the median baseline of word processing. The horizontal projection stage is applied in the Hough space in order to extract the lower and upper lines. These two lines divide the word into three parts: (1) Ascender and upper diacritic points above the upper baseline; (2) Descender and lower diacritic points under the lower baseline and (3) the main content of the word between the two baselines. Figure 10 illustrates the three baselines extracted by HT and horizontal projection. Upper and lower baselines are presented in blue color and median baseline in red color.

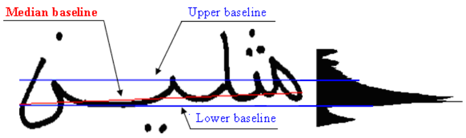


Figure 10: Median, upper and lower baseline detected from an Arabic handwritten word.

3.2.2 PAW Detection

Handwritten Arabic words can have some discontinuities due to the pen up and the binarization stage. According to these discontinuities we can detect more than the real number of a PAW in a given word. The application of morphological filter can connect some parts of PAW. A labeling stage is then applied in order to associate a label color to each connected component. The labeled component can be an isolated character or a token diacritic or characters set. The extraction of the PAW needs to eliminate the components existing below the lower baseline and above the upper baseline.

Figure 11 illustrates the different steps of PAW detection. Figure 11(a) presents the original image, where a discontinuity is surrounding in red color. The smoothing and labeling stage is presented in Figure 11(b), where the correction of discontinuity is surrounded in blue color and baselines are presented in red color. The diacritic points localized above the upper baseline are eliminated. Figure 11(c) shows two extracted PAW's.

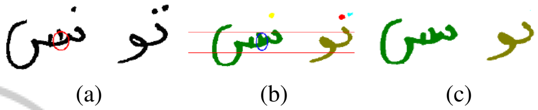


Figure 11: PAWs detection.

4 EVALUATION AND COMPARISON

The evaluation method allows a comparison between the coordinate's components automatically extracted from the segmented image and the real coordinate's components of the original text image.

4.1 Test and Results

The two segmentation methods are tested on a set of handwritten Arabic documents selected from three different databases. For the line extraction method we select a sub-set of documents divided into 30 documents extracted from the KSU[1] database containing 470 lines, 17 images selected from IFN/ENIT-databases containing 162 lines and 16 documents extracted from the BSB[2] databases containing 191 lines. To evaluate the segmentation of handwritten line into PAWs, we select 30 lines from the KSU-database composed of 922 PAWs, 30 lines from IFN/ENIT composed of 772 PAWs and 11 lines extracted from BSB images containing 300 PAWs. In Figure 12 we show the correct extraction rate of the two methods for lines segmentation.

We can see in Figure 12 that the MM-OIC method extraction rate is greater than for HT-MM. We also notice that the two methods have given the best extraction rate on the IFN/ENIT database. This is explained by the fact that the documents of this database are simple and contain lines with a large interline space. The low results of the HT-MM segmentation method is due to the choice of structural element for connecting the components. In fact, the HT-MM

[1] http://ksu.edu.sa/ar/research/manuscripts-makhtota
[2] http://www.bsb-muenchen.de/

method uses as structural element to connect horizontally different components. This enables successive lines to touch or to overlap for the cases of documents containing lines with a relatively small interline or having annotation between lines as in BSB and KSU documents.

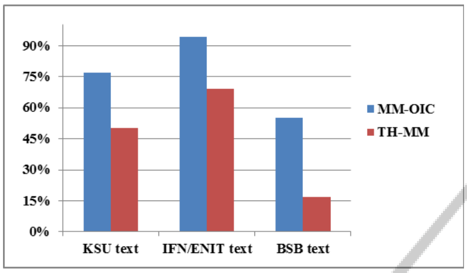


Figure 12: Lines extraction rate for the two methods.

The results of the two methods for the case of the line segmentation into PAWs are illustrated in Figure 13. We notice that the best extraction rate is given by the MM-OIC method also. This is explained by the fact that the proposed method can find the OICs of all components in the treated document. The extraction rate for the IFN/ENIT lines is approximately the same for the two methods. However, for the case of the KSU and BSB lines, the MM-OIC gives a greater extraction rate than of the HT-MM ones. The very low extraction rates given by the HT-MM method are due also to the utilization of the structural element by this method in order to resolve the problem of the discontinuity in the Arabic words.

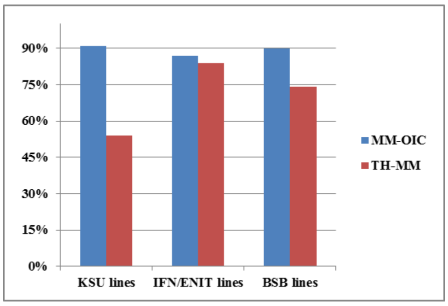
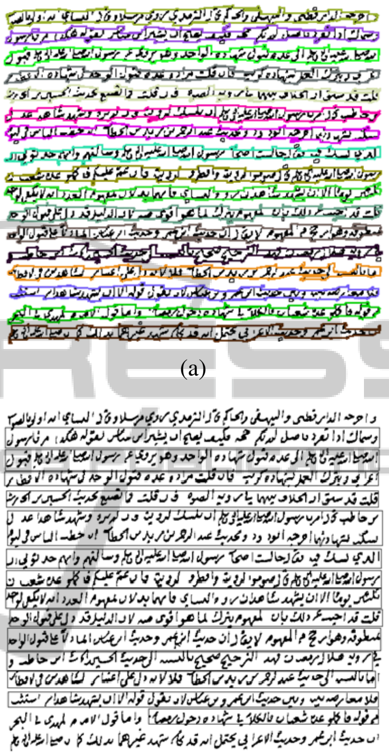


Figure 13: PAWs extraction rate for the two methods.

4.2 Comparison of the Two Methods

The chosen evaluation method favors the second method because it considers the extracted blocks as geometric segments. However, we can deduce that the MM-OIC segmentation method is better than HT-MM. But this method has a bad rate of text line detection by the use of the KSU and BSB database. In Figure 14, we illustrate the result of the two methods applied on the same document for the text line segmentation. Figure 14(a) clearly shows that the MM-OIC method has given a better result than the HT-MM method. (Figure 14(b)). For the first method, each colored polygon represents a text line and for the second method, each extracted line is delimited by a rectangle.

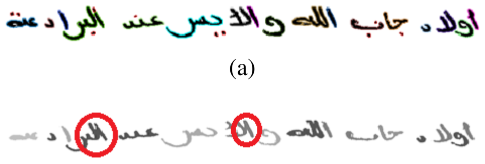


(a)

(b)

Figure 14: Text lines segmentation. (a) MM-OIC method, (b) HT-MM method

In the same, we illustrate in the Figure 15 two examples of segmentation of text line into PAWs with the two method. Figure 15(a) present a good PAWs segmentation by the MM-OIC method. On the other side, the HT-MM method has given an under segmentation result for the same segmented line (Figure 15(b)). The under segmentation is circled in red color.

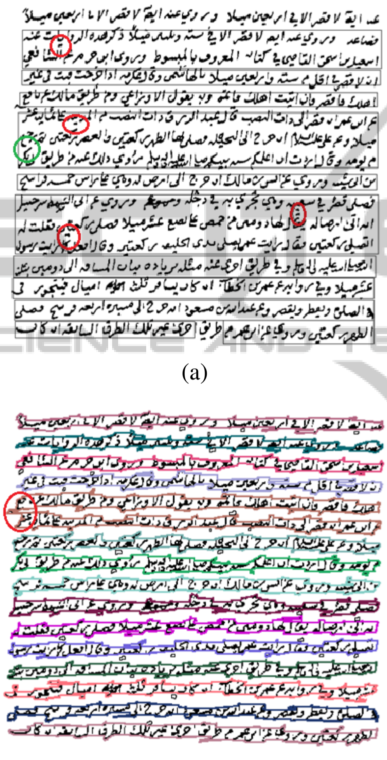


(a)

(b)

Figure 15: PAWs segmentation.(a) MM-OIC method, (b) HT-MM method (under segmentatio)

Another advantage which classifies the method MM-OIC compared to the HT-MM method is that the first method, in some cases, can overcome the problems of lines overlap and diacriticals symbols which generally generate false lines. In Figure 16, we present an example of this case. Figure 16(a) shows clearly that the HT-MM method gives a bad text lines segmentation caused by overlapped lines (circle in green color) and diacritical symbols (circle in red color) respectively. On the contrary, in Figure 16(b), the MM-OIC method gave better results of segmentation by exceeding these problems.


(a)

(b)

Figure 16: Text lines segmentation with presence of the problems of overlapped lines and diacritical symbols. (a) Result of the HT-MM method, (b) Result of the MM-OIC method

The main reason of this difference is due to the fact that the shape of extracted component for second method is rectangular (Figure 9(d)). This shape can increase overlapping between lines or PAWs. However, for the first method the shape tries to approximate the shape of the script (Figure 7(c)).

In the other hand, MM is used for the two methods to connect desired extracted components. For the MM-OIC the extraction step is done by the OIC algorithm. However, the HT in the second method is not used for the extraction but to improve the connection step. The extraction is then done by a labeling step.

4.3 Comparison with Other Methods

Several approaches have been developed for the segmentation of the handwritten Arabic script, especially for text lines segmention. The obtained segmentation results depend of each method and of the databases used for the test. To classify the two proposed method, we present in Table 1 the results of some approaches with each used database. This table gives an idea about the obtained segmentation rates but cannot be used for comparison only if all these methods are evaluated on the same databases.

Table 1: Results of some existing methods of segmentation of handwritten Arabic documents in lines.

Approaches	Test Database	Extraction rate
(Bennasri et al., 1999)	100 documents	98.6%
(Zahour et al., 2007)	100 documents	96%
(Zahour et al., 2008)	160 documents	96.6%
(Li et al., 2008)	2691 Arabic text lines	84.6%
(Ouwayed et al., 2012)	36 documents	98.35% (samples) 65.07% (muti-size)
Our approaches	63 documents	MM-OIC: 75% HT-MM: 45%

5 CONCLUSIONS

Two segmentation methods of handwritten Arabic documents are presented. The first one is the MM-OIC method and the second one is the HT-MM method. Both methods allow the extraction of text lines from handwritten Arabic document and the segmentation of lines in PAWs. An evaluation method is proposed in order to assess and to compare the efficiency of these two segmentation methods. The evaluation criterion used is the limit positions of zones (lines or PAWs). Evaluation is done on a set of documents selected from three sources: IFN/ENIT-database, the two online libraries of BSB and KSU. The average rate line segmentation of MM-OIC is 75%, and of HT-MM is 45%. But the average rate of PAW segmentation of MM-OIC is 89%, and of the HT-MM is 70%.

We should focus in the future works on the improvement of the MM stage of both segmentation methods in order to improve the segmentation rate.

Furthermore, we should test these two segmentation methods on a larger and a complex database. And, we think to evaluate the proposed methods on other criteria such as the number of detecting segments or the recognition rate for PAWs extraction. A comparison to other developed method is also one of our perspectives on the same databases.

REFERENCES

Abdulkader, A. Two-tier approach for arabic offline handwriting recognition. In *IWFHR06*, pages 65000T.1–65000T.11.

Bennasri, A., Zahour, A., and Taconet, B. (1999). Extraction des lignes dun texte manuscrit arabe. In *Vision Interface99*, pages 42–48.

Biswas, A., Bhowmick, P., and Bhattacharya, B. (2010). Construction of isothetic covers of a digital object: A combinatorial approach. *Journal of Visual Communication and Image Representation*, 21:295–310.

Bouafif, S. F., Snoussi, S. M., and Ellouze, N. (2006). Détection des lignes pré-imprimées de chèques bancaires tunisiens par la transformation de hough en vue de l'extraction de l'écriture manuscrite. In *Séminaire Automatique Industrie (SAI)*, pages 45–52.

Bouafif, S. F., Snoussi, S. M., and Ellouze, N. (2012). A hybrid method for three segmentation level of handwritten arabic script. *IAJIT'12*, 9(2):117–123.

Bukhari, S. S., F., S., and Breuel, T. M. (2009). Use of the hough transformation to detect lines and curves in pictures. In *Communication of the ACM*, pages 446–450, Barcelona, Spain.

Duda, R. O. and Hart, P. E. (1972). Script-independent handwritten textlines segmentation using active contours. In *ICDAR'72*, pages 11–15.

Li, Y., Zheng, Y., Doermann, D., and Jaeger, S. (2008). Script-independent text line segmentation in freestyle handwritten documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1313–1329.

Malleron, V., Eglin, V., Emptoz, H., Dord-Crouslé, S., and Régnier, P. (2009). Text lines and snippets extraction for 19th century handwriting documents layout analysis. In *ICDAR'09*, pages 1001–1005.

Miled, H., Cheriet, M., and Olivier, C. (1998). Multi-level arabic handwritten words recognition. In *The International Workshops on Advances in Pattern Recognition (IAPR)*, pages 944–951, Sydney, Australia.

Nicolaou, A. and Gatos, B. (2009). Handwritten text line segmentation by shredding text into its lines. In *ICDAR'09*, pages 626–630, Barcelona, Spain.

Ouwayed, N., and Belaid, A. (2012). A general approach for multi-oriented text line extraction of handwritten documents. *IJDAR12*, 15(4):297–314.

Sakar, A., Biswas, A., Bhowmick, P., and Bhattacharya, B. (2010). Word segmentation and baseline detection in handwritten documents using isothetic covers. In *ICFHR10*, pages 445–450.

Sarfaz, M., Nawaz, S. N., and Al-Khuraidly, A. (2003). Offline arabic recognition system. In *International conference on geometric modeling and graphics*, pages 30–35.

Snoussi, S. M. (2003). *Modèle prespectif neuronal à vision globale-locale pour la reconnaissance de mots arabe omni-scripteurs*. PhD thesis, ENIT.

Snoussi, S. M., ElAbed, H., Bouafif, F. S., Bouriel, K., and Ellouze, N. (2008). Baseline extraction : Comparison of six methods on ifn/enit database. In *ICFHR'08*, page 1170.

Zahour, A., Likforman-Sulem, L., Boussalaa, W., and Taconet, B. (2007). Text line segmentation of historical arabic documents. In *ICDAR'07*, pages 138–142, Curitiba, Paran, Brazil.

Zahour, A., Taconett, B., Likforman-Sulem, L., and Boussellaa, W. (2008). Overlapping and multitouching text-line segmentation by block covering analysis. *Pattern Analysis and Applications*, 12(4):335–351.