Optimized Segmentation Techniques for Handwritten Arabic Word and Numbers Character Recognition

M. Mansour, M. Benkhadda & A. Benyettou
Department of Electronics, Electrical & Electronics Engineering Faculty
University of Sciences and Technology of Oran – B.P. 1505 El-M'Naouer Oran, Algeria
E-mail : ik_mansour@yahoo.fr

Abstract

The segmentation of the Arabic words for recognition still remains a problem to rise for the variability of the writing styles. The adopted technique rests on the hybridization of several segmentation methods based on the various reference levels. The segmentation process led to the decomposition of the writing or words in elementary entities of morphology relatively simple. The choice of primitives of various types is essential in order to ensure an understanding description.

The work presented in this paper reveals the approach of the technique used and proposes an algorithm which has been tested with a set of Arabic words by different writers ranging from poor to acceptable quality. A particular attention was fixed to segmentation techniques that gave suitable results. The initial experimental results are very encouraging and promising.

1. Introduction

The recognition, in general, is strongly based on a suitable procedure of segmentation. The latter process is considering being difficult to achieve in a formal approach. Many works, worldwide, uses a pre-segmentation instead of direct segmentation. However, this technique requires a complete data base of words. A dictionary of all the radical words must be available. Most of researchers rely on the procedures of segmentation, since this one represents a dominant step in the sequence of words recognition.

It should be noted that the process of segmentation encounters several problems with the variety of the characteristics of the Arabic handwriting. The Arabic language presents 28 basic characters which vary according to their position in the word (Beginning, Middle, End and Isolate). Sixteen between these characters have points like secondaries, above or below the principal line. There are the characters which having the same body but differing in the number or the dots position, either above or below the character body (e.g. bah "ب ", tha " ث "). In this situation, the method applies a segmentation technique that produces the character as two separate parts (the primary part (i.e. body of the character) and secondary part (i.e. dots and hamza "ع")).

The most of characters have the secondaries which are a point ".", two points "..", three points "..." and zigzag "ع". It can be significant during the classification (reduction of the class number).

The Arabic words are composed of one or several under words. The pseudo word can contain one or more characters (example: "لر " pseudo word).

The Arabic language has a whole of orthographical accessories or diacritic which can change the meaning and the pronunciation of a word:

a) The short vowels which are placed above or below a letter, assign a sound to it (Dhamma, "ُ", Fatha "ـَـ ", Kasra "ـِـ ");

b) Soukoun "ْ" which is placed above the consonant to mark the vowel absence;

c) Tanwine for example "نورٌ" read "Nouroon" which is a redoubling of short vowel (characters underlined);

d) Shedda " ّ " which is placed above a letter to indicate a redoubling;

In addition to the 28 principal letters, Arabic characters have additional such as TA_MARBOUTA " ة " , LAM_ALIF (لا). Moreover, some characters contain the zigzags like fill characters.

In this article, we propose some methods of segmentation studied and exploited in our research tasks.

2. Segmentation Techniques

The objective was focused on the behaviour of the various techniques of segmentation applied and exploited according to the nature of the disposition

of the numbers and the characters in the image. They are based on some characterization approaches.

2.1. Structural methods with explicit segmentation

The structural methods are based upon a process of three phases:

- The representation,
- The extraction of primitives,
- Description and classification.

When we talk about an analytical approach with explicit segmentation or dissection of the layout, it is acted in fact of cutting out the word in basic unit (segments, lines, graphemes) that identified by a method of classification (compared to prototypes). The words are represented, either in the form of tree, or in the form of graph. This method generally uses assumptions based on the models of Markov.

2.2. Structural methods with implicit segmentation

In order to reduce the combinative aspect of the analytical approach to explicit segmentation, some
methods guide the segmentation according to the part of the word already recognized, we talk then about implicit segmentation.

We can include in this strategy, the affixed method, which consists in separating the radical part (the verbal origin of the word) of the other components such as the prefixes, the suffixes, after their
recognition, and the radical is then segmented with whole share (see figure below).

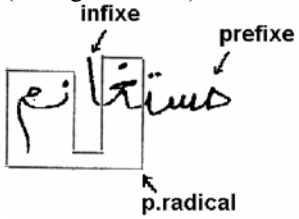


Figure 1: Affixed Method

The disadvantage of this method is that the algorithms of recognition are made much more brittle and difficult to set up.

2.3. Segmentation by emission assumptions

In order to look at the problem of the interdependence of the segmentation operations and recognition, the strategy classically used, consists in emitting a whole of assumptions of segmentation which are then checked during the phase of recognition. Figure 2 illustrates a method of segmentation based on the vertical optimums in "y" detected in the layout. The elementary segments considered are portions of layout separated by two successive optimums in "y".

The main difficulty encountered by this approach is corresponding to obtain the good assumptions of segmentation by robust approach (corresponding to the cutting of the letters or analyzed traced figures) between the whole of the assumptions put forth. While minimizing the number of these assumptions, they likely generate a combinative explosion of the various solutions to be considered.

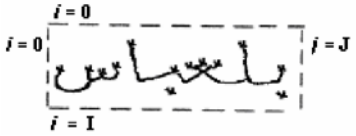


Figure 2: Assumptions of segmentation starting from the vertical optima

The determination of the vertical optimum can be established according to our own observation by the algorithm proposed in figure 3.
The advantage of this algorithm with set assumptions, generally, converges to a breaking desired point even in the case of an overlapping. Figure 4a and figure 4b illustrate clearly the situation. The disadvantages at this allow Algorithm appears on the level of the checking method in order to avoid an over-segmentation.

```
For all element T [i, j] of image MAKE:
    IF ( T [ i, j ] =1 ) AND ( T[ i-1, j-1 ] + T [ i+1, j+1 ] =1 )
       OR
      ( T[ i, j ] =1 and ( T[ i-1, j+1]+T[ i+1, j-1]=1 ))
      jtest =j;
      itest =i;
                                  i=itest
            CALCULATE: s1= Σ T [i, jtest] ;
                                  i=I
                                  i=0
            CALCULATE: s2= Σ T[i,jtest] ;
                                  i=itest+1
    IF ((s1=0 ) OR (s2 = 0)) : jopt=jtest ;
       iopt=itest ;
    IF NOT see the element T[i][ j] according;
    END IF
END.
```

Figure 3: Algorithm Structure

i_{opt}, j_{opt} corresponding, respectively, to the line and the column of the vertical optimum to record and check.

Figure 4a: Projection of the segmentation assumptions in the event of overlapping
▙: Points defined in the vertical optimums

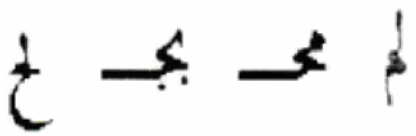


Figure 4b: Overlapping Arabic characters

2.4. Method of Almuallim & Yamaguchi

The algorithm of segmentation using this method is based on the following phases:

- Segmentation of the word,
- Detection of the limits of the word to be recognized by the determination of following parameters:

(1) Determination of the line with the maximum of black pixels, known as base line (I_{mid}).

(2) The first line containing black pixels, by sweeping the image from the top to the bottom, knows as the line (I_{min}).

(3) Determination of the last line containing the black pixels (I_{max}).

(4) The first column containing the black pixels, by sweeping the image from the left to the right (J_{min}).

(5) The last column with black pixels (J_{max}).

(6) Scanning the image from the top to the bottom, referring to (I_{mid}), and right to left until detection of the first black pixel (the pixels on I_{mid} are not considered).

(7) Scanning of the character and designation of its end by (Je_1).

(8) Detection of the beginning of the next character (Je_2).

(9) Separation of the two characters at the average between the two limits as shown in figure 5.

$J_b = (Je_1 + Je_2) / 2$.

(10) Reset J_{e1} and J_{e2} and the operation continues until the end of the word.

The convergence of this method is as good as filtering and fine segmentation.

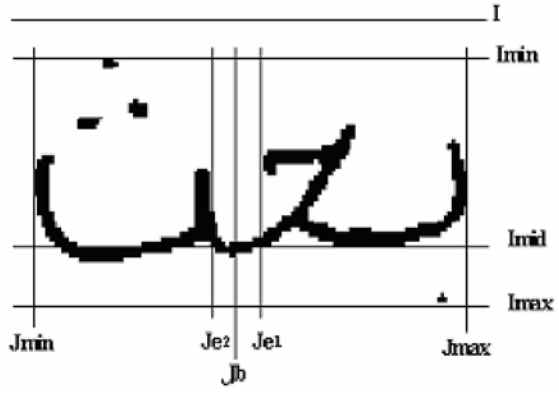


Figure 5: Limitation of the word

The principal disadvantages are located in the following constraints:

* This technique relies on the base line by supposing that it carries the maximum of black pixels. In general it's not always easy to determine the base line because there are varieties in the styles of writing like it is shown in the figure 6.

* The segmentation can be false for the below mentioned case (figure 7) where the J_b average is not well defined. In this situation we have what we call an over-segmentation.

Figure 6: Difficulties to determine the base line

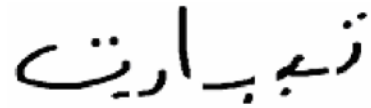


Figure 7: Error made by on-segmentation by using the algorithm of Almuallim &Yamaguchi

2.5. Average Method

This method is divided into tasks that are:

a) Segmentation of the word in parts related

b) Segmentation of the related parts in isolated characters.

2.5.1. Segmentation of a related parts word

To allow a localization of the constituent parts, we suppose that there is at least a white column which separates two related parts. Their insulation is carried out by a simple vertical projection as shown in figure8.

Segmented image ***Segmented image***

Figure 8: Segmentation related parts word.

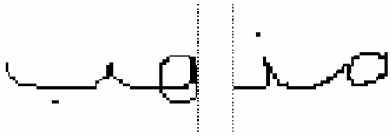


Figure 9: Another case of segmentation of the related parts word.

2.5.2. Segmentation of a related parts word with isolated characters

Since we are interested in the ascending approach; the problem is reduced to the isolated

recognition characters constituting the layout. The segmentation in characters is thus necessary and crucial for the reproduction of the original image.

The point of connection is presented by the smallest sum of the average value in that:

$$Moy=(\frac{1}{N})\sum_{i=1}^{N}X_i$$

Where N is Number of columns,

X_i: the number of the black pixels of column i.

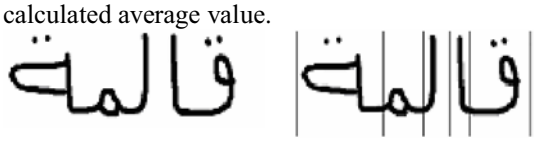
Consequently, every part verifying a sum lower to Moy must be segmented in different characters.

At the end of the word or a related part, the following rule must be applied:

$$L_i+1 > 1.5\ L_i$$

Where L_i is maximum value of **i** in the histogram.

In the example of the figure 10, we show that the segmentation must occur in relation to the calculated average value.



Original image ***Image result***

Figure 10: Segmentation of an Arabic word

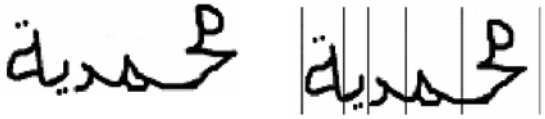
3. Comparison synthesis

The problem appears in certain characters which overlap and which we call bindings shown in figures 11, 12 and 13.

In general, the main disadvantages which reduce the effectiveness of such method or other segmentation are:

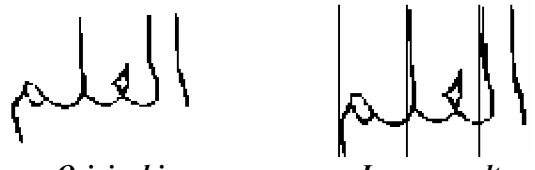
Over segmentation which causes the cutting of a single nature in several no significant segments,

The overlapping which allows, on the other hand, generate more than one character with the same segment which gives a wrong recognition.



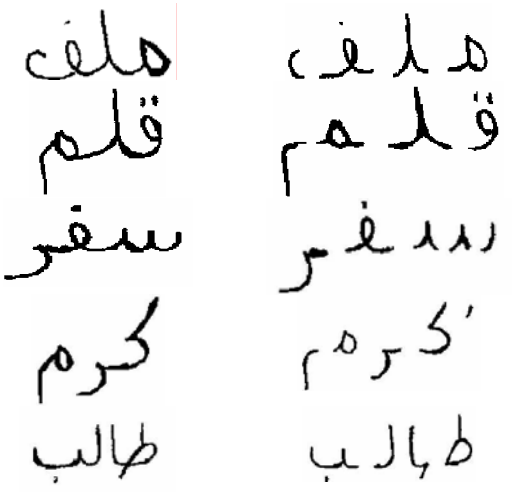
Original image ***Image result***

Figure 11: Case of false segmentation



Original image ***Image result***

Figure 12: Another case of false segmentation



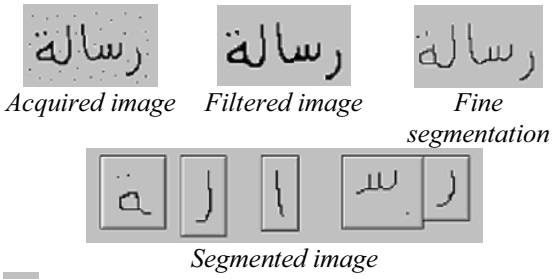
Original image ***Segmented image***

Figure 13: Examples of critical case for the segmentation of an Arabic word.

4. Experimental results and comparisons

In this part, we expose some results which were obtained by the various techniques of segmentation.

Various illustrations putting in evidence: the segmentation, the filtering and the fine segmentation of the handwritten Arabic words and numbers that are presented in figure 14, figure 15, figure 16 and figure 17.

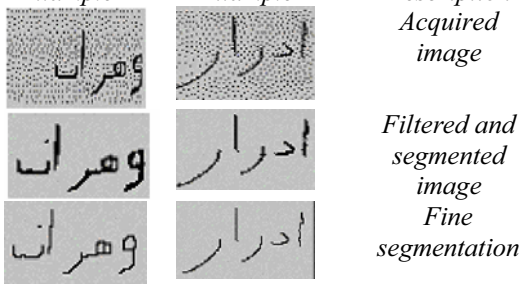


Acquired image *Filtered image* *Fine segmentation*

Segmented image

: Background of the OCR window program.

Figure 14: Segmentation of characters



Example 1 *Example* 2 *Description*

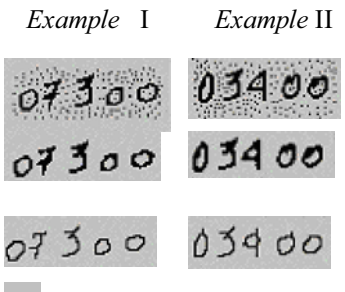
Acquired image

Filtered and segmented image

Fine segmentation

: Background of the OCR window program.

Figure 15: Segmentation of words

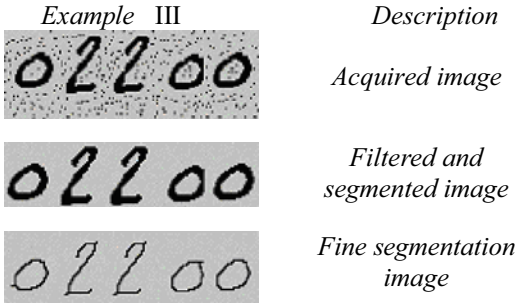


: Background of the OCR window program.

Figure 16: Segmentation of numbers

The rate of segmentation per word varies from 40% to 90% according to the method exploited on a set of 300 written words by several script writers and with different styles.

As concerning the numbers, the method allows, on a set exceeding the 300 numbers to segment correctly (3/5) with (5/5) numbers.



: Background of the OCR window program.

Figure 17: Segmentation of numbers

It is noticeable that it is always difficult to reach of exactitude great rate; since some types of library documents or historic articles present the blurred characters for example or unfamiliar polices.

Facing this great morphological variability of the handwritten characters, it would be desirable that the morphological variation met between the various possible layouts of the same character is less significant than the morphological variation present between layouts of characters belonging to different classes.

There are also many forms of characters which intrinsically ambiguous and thus are very muddled as shows it the following examples of figure 18.

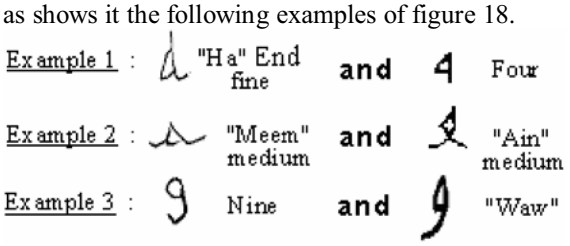


Figure 18: Ambiguity of certain forms of characters

The obtained results presented in this paper (deduced from the various segmentations), is reinforced by a multitude of tests related to the fixed thresholds according to the tested set of the handwritten Arabic words and numbers.

5. Conclusion and future work

This work presents part of a system of recognition of the handwritten Arabic words and numbers based on system OCR.

By looking of the results obtained, we can conclude that the greatest disadvantage of our system lies in the segmentation which is very sensitive to the variation between the styles of writing and the size of the resulting characters. To improve this procedure, we propose the following solutions:

Considering that the direct methods of segmentation of words in characters often provide errors, we could attenuate these latter, by combining the methods with syntactic and semantic methods aiming to correct the errors by successive passages of segmentation until obtaining a good recognition. Among these methods, we can indicate:

- Direct method of comparison with a dictionary of pre-established words. This method gives good results of segmentation (therefore recognition), however it requires the introduction of a data base of at least hundred thousands of words which constitutes a tiresome work and requires a very space memory capacity.
- The segmentation method in split words or affixed method, which aims at separating the parts of word constituting the prefixes, the suffixes, and the infixes of the radical part.

This method presents as the disadvantage of the need for the introduction of a dictionary of Arabic radical words (with a number smaller as the previous method), but it makes possible the reduction of the problem of the segmentation by partition of the words to be recognized.

The tests which we made on several styles of writing increased each time the ambiguities especially relating to the phase of segmentation before the recognition. Therefore, it should be noted that to make an acceptable classification, it is always necessary to be ensured to have used a good segmentation technique.

6. References

[1] A. Zahour, B. TACONET, P. MERCY, S. RAMDANE, “Arabic Handwritten Text Line Extraction”, Actes de ICDAR’2001, Seattle, 8-11 septembre 2001,p. 281-285.

[2] U. Pal, S. Datta, “Segmentation of Bangla Unconstrained Handwritten Text”, Actes de ICDAR’2003, Edinburgh, 3-6 Août 2003, Scotland.

[3] K.Hadjar, R. Ingold, “Arabic Newspaper Page Segmentation”, Actes de ICDAR’2003, Edinburgh, 3-6 Août 2003, Scotland, p. 895-899.

[4] F. Venturelli, "A Successful Technique for Unconstrained Hand-Written Line Segmentation" Progress in Handwriting Recognition", p 563-568.

[5] Wu. Xiaoying, Graham Leedham, "Separating Lines and Words in Unconstrained Handwriting", IGS’97 Proceeding Eighth Biennial conference of the International graphomocs society, Porto Antico di Genova, Italy August 24-28, 1997, pp117-118

[6] H. Oliver, H. Miled, K. Romeo, Y. Lecourtier "Segmentation and Coding of Arabic Handwritten Words", Laboratoire d’informatique Industrielle Image. Université de Rouen. IEEE 1996.

[7] C .K. Lee, S. P. Wong, ″A mathematical morphological approach for segmenting heavily Noise corrupted images″, Pattern Recognition, Vol.29,n .8, app.1347-1538,1996.

[8] C .K. Lee, S. P. Wong, "A mathematical morphological approach for segmenting heavily Noise corrupted images", Pattern Recognition, Vol.29,n .8, app.1347-1538,1996.

[9] J.R. Parker, "Practical Computer Vision Using C", John Wiley & Sons, Inc, Toronto, 1994.

[10] A. Amin, "Off-line Arabic characters Recognition", The State Of the Art, Pattern Recognition, vol. 31, n°5, 1998, p. 517-530.

[11] B. Al-Badr, S. Mahmoud, ″Survey and bibliography of Arabic Optical text recognition″, Signal Processing, 41, pp. 49-77, 1995.

[12] M. Ha Thien, Matthias Zimmermann, Horst Bunke, "Off-line handwritten numeral string recognition by combining segmentation-based and segmentation-free methods", Pattern Recognition, Vol.31, Issue 3, March 1998, p.257-272.

[13] Zhizhen Liang, Pengfei Shi, "A Metasynthetic Approach for Segmenting Handwritten Chinese character strings", Pattern Recognition Letters, Vol. 26, Issue 10, 15 July 2005, Pages 1498-1511.

[14] Ashraf Elnagar, Reda Alhajj, "Segmentation of connected handwritten numeral strings", Pattern Recognition, Vol. 36, Issue 3, March 2003, Pages 625-634.

[15] U. Pal, A. Belaïd, Ch. Choisy, “Touching numeral segmentation using water reservoir concept”, Pattern Recognition Letters, Vol.24, Issues 1-3, January 2003, Pages 261-272.

[16] Kye Kyung Kim, Jin Ho Kim, Ching Y. Suen, “Segmentation-based recognition of handwritten touching pairs of digits using structural features”, Pattern Recognition Letters, Vol.23, Issues 1-3, January 2002, Pages 13-24.

[17] Karim M. Hussein, Arun Agarwal, Amar Gupta, Patrick S. P. Wang, “A knowledge-based segmentation algorithm for enhanced recognition of handwritten courtesy amounts”, Pattern Recognition, Vol.32, Issue 2, February 1999, Pages 305-316.

[18] N.W.Strathy, C.Y.Suen, A.Krzya, “Segmentation of Handwritten Digits Using Contour Features”, Proc. Int'l Conf. Document Analysis and Recognition, pp. 577-580, 1993.

[19] Z.Chi, M.Suters, H.Yan, “Separation of Single- and Double Touching Handwritten Numeral Strings”, Optical Eng., vol. 34, pp. 1,159-1,165, 1995.

[20] Z.Shi, S.N.Shrihari, Y.-C.Shin, V.Ramanaprasad, “A System for Segmentation and Recognition of Totally Unconstrained Handwritten Numeral Strings”, Proc. Int'l Conf. Document Analysis and Recognition, vol. 2, pp. 455-458, 1997.

[21] J.Hu, D.Yu , H.Yan, “Algorithms for Partitioning Path Construction of Handwritten Numeral Strings”, Proc. Int'l Conf. Pattern Recognition, vol. 1, pp. 372-374, 1998.

[22] M.Cheriet, Y.S.Huang , C.Y.Suen, "Background Region-Based Algorithm for the Segmentation of Connected Digits”, Proc. 11th Int'l Conf. Pattern Recognition, vol. 2, p. 619, Sept. 1992.

[23] Z.Lu, Z.Chi , P.Shi, “A Background-Thinning-Based Approach for Separating and Recognizing Connected Handwritten Digit Strings”, Pattern Recognition, vol. 32, no. 6, pp. 921-933, 1999.

[24] B.Zhao, H.Su, S.Xia, “A New Method for Segmenting Unconstrained Handwritten Numeral String”, Proc. Int'l Conf. Document Analysis and Recognition, vol. 2, pp. 524-527, 1997.

[25] N.Arica, F.T.Yarman-Vural, “A New Scheme for Offline Handwritten Connected Digit Recognition”, Proc. Int'l Conf. Pattern Recognition, vol. 2, pp. 1,127-1,129, 1998.

[26] S.W.Lee, S.Y.Kim, “Integrated Segmentation and Recognition of Handwritten Numerals with Cascade Neural Network”, IEEE Trans. System, Man, and Cybernetics, vol. 29, no. 2, pp 285-290, 1999.