PAWs Segmentation by intrinsic/discriminant dual modeling based on a hierarchical organization of Arab grapheme and the curves Bezier

Aissa KERKOUR ElMIAD, Azzedine MAZZROUI
Laboratoire de Recherche en Informatique
Faculty of Sciences and Technology of Al Hoceima, Faculty of Sciences Oujda
University Mohammed 1er, Oujda University
{mid.kerkour,azze.mazroui}@gmail.com
MAROCCO

Abstract: - We propose in this work an approach to segmentation of Arabic texts printed offline open vocabulary. The method is based on human perception using their natural power for segmentation. The novelty of our work is in the analysis of this type of approach on complex fonts and calligraphy has strong ligatures; we propose an artificial segmentation feature extraction based on the use of structural primitives for a robust description of the different morphological variability font considered. The proposed algorithm yielded results quite satisfactory with an extension for the resolution of a particular case of the sub-segmentation. This model has enabled us to increase significantly the recognition performance and develop a recognition system open vocabulary.

Key-Words: - OCR Arabic, skeletonization, diacritics, ligatures, segmentation-recognition, multi-font

1 Introduction

The use or not of phase segmentation in character or in grapheme makes the distinction between two possible strategies for word recognition be they printed or handwritten: overall approaches that consider the word as a w hole without trying to identify each of the letters that compose, which opposes the analytical approaches that seek in the first instance to cut the word into letters and then seek to recognize them.

In the case of cursive writing, the problem is even more complex. In the community of the printed writing recognition, he was admitted that it is impossible to directly segment a word in cursive letters. To segment in characters, they must have previously been identified; and for recognize a character, it must be properly segmented beforehand. This is the paradox of K.M. Sayre [1]. In order to resolve this dilemma, it is necessary to cut the word into subparts letters. For more information on the segmentation of cursive writing, the synthetic work of X. Dupré [2]. Generally two approaches are used: the segmentation in grapheme(explicit segmentation) [3] which consists a segments the word into subparts that are almost all letters. And analysis by the sliding windows fenêtres glissantes (implicit segmentation) where the word is divided into vertical stripes.

Arabic printed is known for its rich fonts and styles: there are over 450 different styles and fonts. On the one font to another, the morphological characteristics of Arabic character change considerably. In [4,5], we find a fairly detailed study of the morphological characteristics of Arabic. Thus, the structure of the Arabic word printed is in part responsible for the quality of segmentation and the degree of complexity of recognition. Several solutions have been proposed for the segmentation of the Arabic script. Thus, overall and analytical methods were tested. Lists of work already carried out are described in [6].

In this paper, we will discuss first the general principle of our strategy. The approach of the proposed problem is inspired by the standard methodology for segmentation by a human. We then provide a brief overview of these techniques emphasizing commonalities. We w ill discuss modeling Bezier curves in image processing and algorithms traverses PAWs. Finally, we explain in detail the operating principle of the proposed segmentation algorithm, and is the heart of our system.

2 Problematic

Because of certain characteristics of the Arabic script, the solution for the segmentation is not always trivial. Indeed, the morphological study of the Arabic script shows that it is difficult to operate the segmentation at the level the character. All these reasons have led us to ask the question: how humans can identify the characters even though they are ill-posed in the word. This question was already asked by the scientific community and researchers. But the answer was to say that humans at such a developed brain at such that it can not artificialised. That is why we have applied our approach to segmentation based on research in attachment points of the characters in writing. The search for these attachment points present an easy way to separate grapheme (separation into connected components, contour extraction ..)

In view of these problems, we have chosen a global modeling at the level PAW. This choice is penalized by the large number of points that can be selected to achieve the segmentation of PAW in specific locations. However, this solution is very interesting in the case of cursive writing.

A PAW is a connected component of black pixels grouping one or more letters. An Arabic word is a sequence of related entities entirely separate from those PAWs (or PAWs: Pieces of Arabic Words). A word can be composed of one or more PAWs (see Fig 1). Each PWS is a sequence of related letters, giving the appearance of recursiveness at this writing.

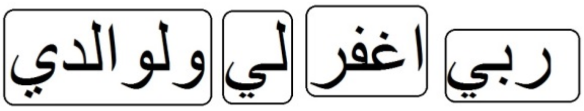


Fig.1 : An Arabic word may consist of several connected components (PAWS): from right to left,Two Paws, Two Paws, a PAW PAWS and five per word

Taking account of the notion of PAW, inherent in the Arabic script, is advantageous, because the interpretation of PAWs rather than words, duct, particular for this vocabulary, a reduction in the size and complexity of the problem. As a result, it seems very useful to process the selected vocabulary recognition based on analytical modeling nickname that also offers the advantage of avoiding the delicate problem of segmentation into characters related to the analytical approach.

3 General description of the proposed system

Figure 2 shows our system of segmentation and recognition of Arabic script printed multifonte off-line.

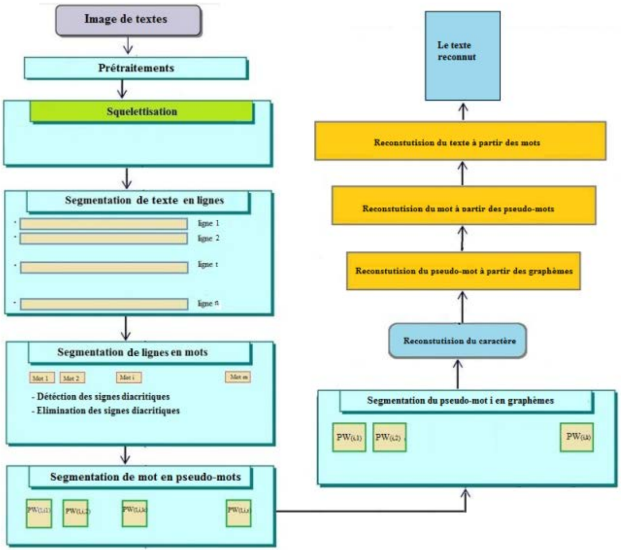


Image de textes
Prétraitements
Squelettisation
Segmentation de texte en lignes
ligne 1
ligne 2
ligne t
ligne n
Segmentation de lignes en mots
Mot 1
Mot 2
Mot i
Mot m
- Détéction des signes diacritiques
- Elimination des signes diacritiques
Segmentation de mot en pseudo-mots
PW(1,i,1)
PW(1,i,2)
PW(1,i,k)
PW(1,i,r)
Segmentation du pseudo-mot i en graphèmes
PW(i,1)
PW(i,2)
PW(i,k)
Reconstutision du caractère
Reconstutision du pseudo-mot à partir des graphèmes
Reconstutision du mot à partir des pseudo-mots
Reconstutision du texte à partir des mots
Le texte reconnut

Fig.1 : General System Diagram

In the first place, image text passes through two stages of pretreatment which consist in squeletiser the text from the image then eliminate diacritics after identifying and memorized their positioning. Then, the segmentation phase of the text in lines is triggered. This segmentation is based on the method of horizontal projection. Then, to word segmentation we use the segmentation method by the vertical projection. . The word segmentation is carried in paws by identifying connected components, then use the technique [7] the isolated characters are detected. Segmentation PAWs of words Arabic in grapheme is then applied and is based on a study of topological characteristics of the Arabic script. Indeed, given a pseudo Arabic word is generally obtained by connection of several letters, the approach that we have developed consists in detecting the positions of these connections.

At the level of recognition, the approach adopted is to use a similar approach for isolated characters and that is developed in the [7]. In this case the vocabulary of isolated characters is not limited to 28 characters, but it is extended to all forms of writing characters according to their positions in the word (starting, middle or end of the word).

To implement this approach of pseudo-analytical recognition, we propose a new segmentation concept vocabulary and a new database image printed of Paws generate from three different Arabic fonts complexities. Besides images of pseudo-words the new database also contains files with the line position of the page image in the base of each PAW.This information is exploited evaluated in the step of feature extraction to consolidate the overall robustness of the feature vectors.

Given an image I text, the phase of recognition of the image I therefore goes through the following steps:

- Squeletiser the text in the image I,
- Detect diacritics then eliminated while retaining information on their positions,
- Segment the image into lines and I associate an identifier with each line specifying its position in the image,
- Segment lines into words and associate with each word an identifier specifying its line of belonging and its position in the line,
- Segment words into Paws and associate with each PAW an identifier specifying the word which it forms part and its position in the word,
- Segment Paws in grapheme associate with each grapheme an identifier specifying the associated PAW and its position in the PAW
- Identification the characters and the ligatures of the PAW
- Recognition of each character and of each ligation
- Reconstruct the text by proceeding by linking them grapheme recognized of a sam e PAW, then link in the PAWs of the same word, Then arrange these words in line and finally put the lines one below of the other. To achieve this work, we use the identifiers of previous steps.

The recognition process is evaluated by realizing tests on several levels:

- Relative test performance the step of segmenting the image into rows I
- Relative test performance the step of segmenting words into lines,
- Relative test performance the segmentation step in words Paws
- Assessment of the rate of recognition of graphemes
- Evaluation of overall recognition rate.

3.1 Preprocessing

These preprocessing are identical to those made on isolated characters: binarization, separation of diacritical dots, noise suppression and turnaround forms. We use the same processes of preprocessing therefore presented in [7]

3.2 Skeletonization

This step directly influences our approach to segmentation of a PAWs in grapheme, because the algorithms of route of trace of PAW can be misdirected by wrong segmentation of a grapheme. These errors to segmentation occur due of the trace noises that are generated a wrong skeletonization (see Fig 3). Treat these noise requires a preliminary phase of treatment. The segmented page undergoes the same algorithm Zhang-Wang modified skeletonization carried out on isolated characters [7]. The results of their application on a text sample are shown in Fig 3.

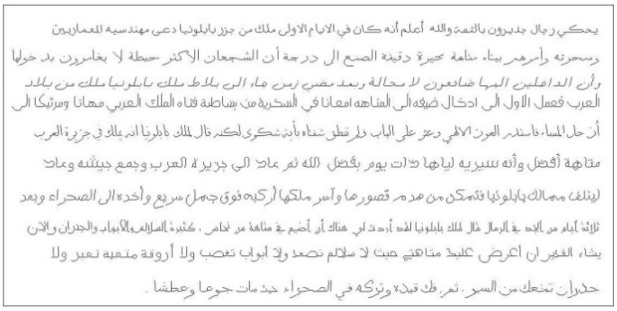


Fig.3: The Skeletonization algorithm modified of the text Zhang-Wang

3.3 Detection of diacritics

Diacritics must be removed before extraction PAWs and their positions are stored. This information of the presence or not of diacritics will be useful during the recognition step of the graphemes. This elimination step we prevents such disturbances during extraction of PAWs

The algorithm used for the extraction of diacritics is a modified version of the one proposed in [8]. Cet algorithme se base sur l'aire (qui englobe le mot), height and superposition vertical of connected components (see Fig 4). The thresholds are fixed empirically by Menasri et al [8] and adjusted made by our tests.

After step skeletonization the number of pixels of big connected components that do no t match the diacritical marks is very large compared to the connected components that represent diacritics number. The second stage carried keeps small letters such as د رand containing few pixels and may resemble diacritical marks. The third test is based on the relative position of the connected component and its neighbors: If a connected component C1 of area reduced is located above another connected component C2, then C1 is a diacritical mark. However, the algorithm may fail when the diacritical mark treated is of important large size in relation to word size.

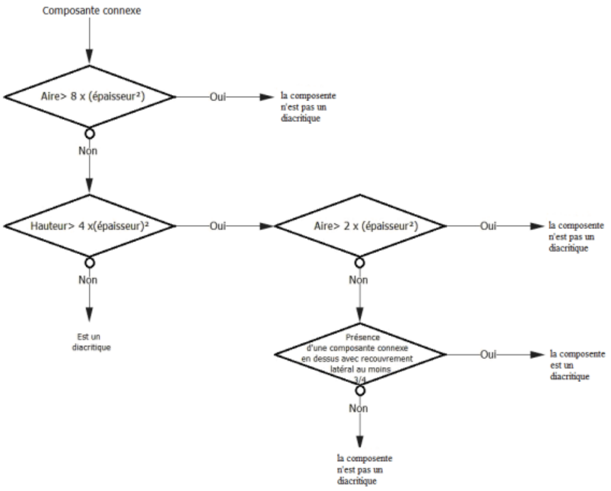

Composante connexe
Aire> 8 x (épaisseur²)
Oui
la composente n'est pas un diacritique
Non
Hauteur> 4 x(épaisseur)²
Oui
Aire> 2 x (épaisseur²)
Oui
la composente n'est pas un diacritique
Non
Est un diacritique
Non
Présence d'une composante connexe en dessus avec recouvrement latéral au moins 3/4
Oui
la composente est un diacritique
Non
la composente n'est pas un diacritique

Fig.4: Mensari algorithm [8] after renewal for the detection of diacritics mark printed writing.

3.4 Segmentation in line

Is based on the histogram of horizontal projection, This method assumes that the majority of pixels are arranged on the base line. For details on these methods, the art developed by AM Al-Shatnawi [9] and Kh. Omar [10] is rich enough. We tested this segmentation approach on a text written in 14 fonts selected to represent the different font families. The total number of lines is equals 934 The test results are given segmentation a rate of 99.17%.

3.5 Segmentation in word

By a horizontal projection we consider that all spaces between two successive clusters are candidates for separation between words. considering that the spaces between words change the along with font change (see Fig 5), the size of the space used in our approach to identify words is calculated from the frequency of appearance of these sizes in the total text.

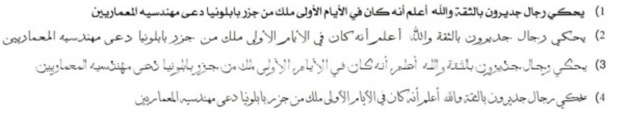

1) يحكي رجال جديرون بالثقة والله أعلم أنه كان في الأيام الأولى ملك من جزر بابلونيا دعى مهندسيه المعماريين
2) يحكي رجال جديرون بالثقة والله أعلم أنه كان في الأيام الأولى ملك من جزر بابلونيا دعى مهندسيه المعماريين
3) يحكي رجال جديرون بالثقة والله أعلم أنه كان في الأيام الأولى ملك من جزر بابلونيا دعى مهندسيه المعماريين
4) يحكي رجال جديرون بالثقة والله أعلم أنه كان في الأيام الأولى ملك من جزر بابلونيا دعى مهندسيه المعماريين

Fig.5: Space change between words in the texts for the following fonts

We tested the approach of segmentation the lines into words upon lines obtained in the previous step, with text image contains 1000 words. The wrong words segmented can be grouped into two classes:

- The first class contains the words which the system of segmentation judges as P AWs and not as whole words their number is 6 words
- The second class contains compound words of at least two PAWs and for which the system judges one of their PAWs as whole word, their numbers is 2

As it appears on the examples of release, the system considers some words as being PAWs because the distance between these words and the words which succeeding is very small. The rate of segmentation of the text to words for 14 font is 99 %.

3.5 Localization of the PAWs

Before to proceed to localization of the PAWs, We begin by eliminating the diacritic points by proceeding according to the approach developed in the section 3.3. The related components remaining correspond to the main components PAWs and to the PAWs trained by a only letter.

We tested this approach of segmentation in PAWs on the same text of test used in the evaluation of the stages of segmentation in lines and in words and which made up o f 1000 words. The rate of localization of the PAWs is 99.75 %. The number of PAWs badly segmented for these 14 cast irons is 7 PAWs, As it appears on us examples, the main reason is due to the presence of a connection between two pixels belonging to two different PAWs. that prevents the system from detecting them as being two related components.

4 Segmentation of a P AW into graphemes

The segmentation approach Paws into grapheme that we adopted is based on the human perception of segmenting words into characters. It is based on two principles:

- the writing of Arabic words is done from right to left and from top to the bottom,
- the concatenation points between characters of the same word are always of the nodes

Therefore, the points which will be candidates to be points for clipping are thus necessarily of the nodes. We shall explain in the following how to identify these points among the set of nodes in a PAW. In addition, the search for these cutoff points will follow a route travels from the PAW favoring, each time it is possible, the right-left direction and top to the bottom.

To explain our approach to segmentation Paws into grapheme, we will need the following definitions.

Definition: an extremity point of a skeleton PAW is a pixel which admits one neighboring pixel belonging the skeleton (see Fig 6).
Definition: a multiple point or node of skeleton of a PAW is a pixel that has at least three neighboring pixels belonging to the Hskeleton. Les trois pixels voisins qui sont appelés nœuds simples et celles qui ont quatre voisins sont appelés multi-nœud (voir la Fig 6).

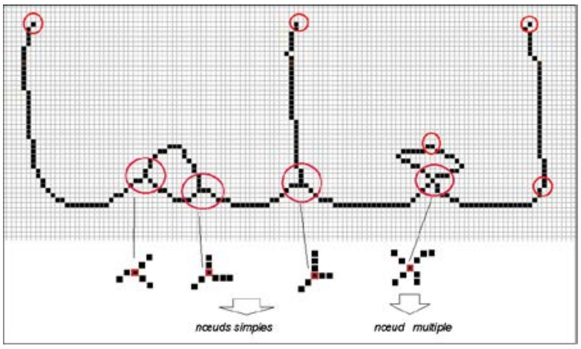


Fig.6: extremities points of the PAW لعلها; Single and multiple nodes PAW لعلها; Simple points of PAW

Our approach the segmentation into graphemes paws goes through the following steps:

4.1 Starting point of the route the trace of PWS

The starting point of the route (course) is the point extremity most to the right and the highest of the skeleton (when it exists). Thus, to identify this point we distinguish the following three cases:

(a) The skeleton of the PAW does not admit of end points: in this case the characters of the PAW admit all loops. Ainsi, on effectue une coupe de la patte en deux parties au milieu de la plus grande partie formée par les points ayant un seul pixel(For this we carry a horizontal projection of the PAW) Each of the parties thus obtained admits an extremity point that will be used subsequently as a starting point of the segmentation process of the part in grapheme (see Fig 8).

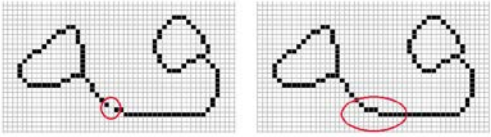


Fig.8: Example of PAW without a extremity point

(b) The skeleton of the PAW admits one extremity point: in this case the starting point will be this extremity (see Fig 9).

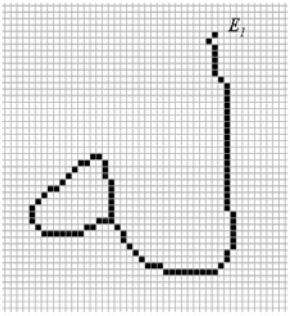
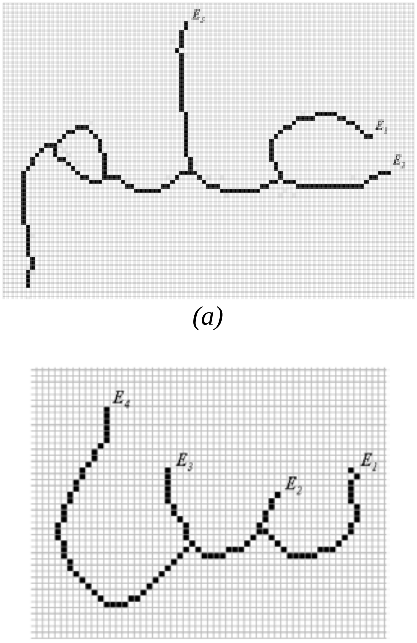


Fig.9 : PAW with a single extremity point E_1

(c) This case we retain both ends points lying furthest to the right of the skeleton. If the distance between these two points is below a threshold (that we have calculated follows an analysis of the different characters of the Arabic language), then we judge that these two point extremity belong to the same grapheme, and thus the starting point will be the highest among these two points extremites. In the opposite case(ie the distance between these two point extremity is greater than threshold), well the two points extremites do no t belong to in the same and in this case the starting point will be the rightmost of these two points (see Fig 10).



(a)

(b)

Fig.10: (a) Both ends E_1 and E_2 belong to the same character "ع" of the word "علم"

(b) Both ends E_1 and E_2 of the word "بين" does not belong to the same character

4.2 Search of the point first of cut

Once the starting point E_1 identifed, we browse from this point the trace of PAWs until attain the extremity point or a node

1) if this point is a extremty point E_2, then the trace E_1E_2 is a grapheme(see Fig 11)(because all this pixles by located on the trace EE are simples)

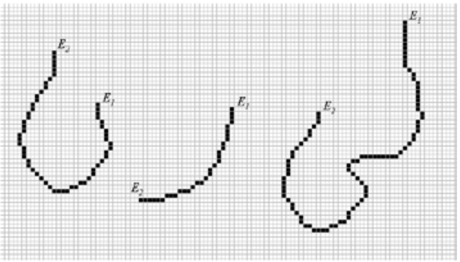
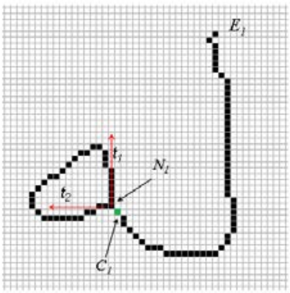


Fig.11 : Examples of grapheme knotless

2) Alternatively, ie if the first point is a node N_1, then we consider the two t_1 and t_2 directions other than that between the points N_1 and E_1, which are classified according to the priority order defined in Fig 7 (b). Thus, if t_1 and $t_2 = d_i = d_j$ then necessarily $i < j$. Cet ordre permet de classer les chemins partants de N_1 selon un ordre qui donne la priorité au chemin allant du haut vers le bas et de la gauche vers la droite(thus our approach systematically seeks to identify the highest cut-off point and the rightmost)

Following direction t_1, we browse the trace from the node N_1 and we stop once we meet extremity point or a node. We distinguish the following cases:

a. The point encountered is no other than the point N_1: in this case the grapheme after the nodeN1 following the direction t_1 contains a loop and consequently the point cut-off C_1 will before last pixel the trace E1N1(the pixel just lying before leaving E_1 N_1 as shown in Fig 12)



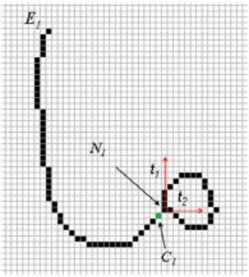


Fig.12 : Example where the N_1 node belongs to a loop

b. The point is encountered point extremity s noted E_2: in this case we distinguish two cases depending on w hether the two points extremities E_1 and E_2 belongs or not the same grapheme.

- If the distance between these two extremities points is less than the threshold calculated in paragraph previous research relating to the starting point, then these two extremities points belong to the same grapheme and the cutoff point C_1 will be the first pixel after N_1 in the direction t_2(see Fig 14)

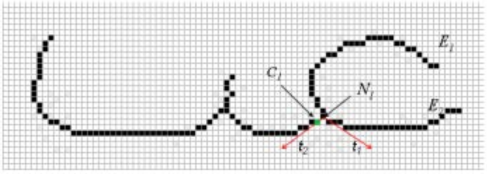


Fig.13: Example where the node N_1 does not belong to a loop

- Otherwise, we conclude that the two endpoints(remplace par), E_1 and E_2 does not belong to the same grapheme and subsequently the cutoff point C_1 will be before last the pixel of trace E_1N_1 (the pixel be located just prior N_1 starting E_1)

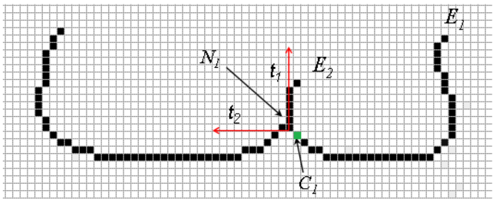


Fig.14: Example where the node N_1 does not belong to a loop

- The point is a encountered node noted N_2: in this case we browse the trace from the node N_1 and in following the direction t_2 until reaching an endpoint or an node. We thus distinguish the following cases:

- The point encountered not other than the node N_2: In this case both nodes N_1 and N_2 belong to the same loop and therefore cutoff C_1 will be pixel the penultimate of trace E_1N_1(the pixel be located just before N_1 starting of E_1)

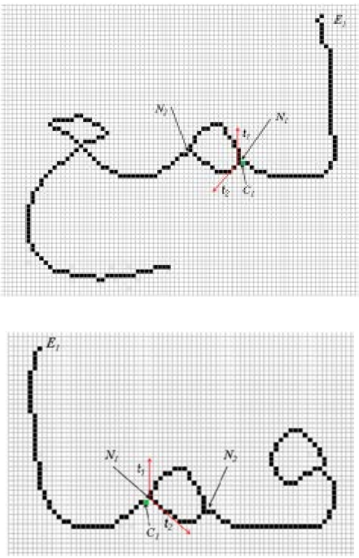


Fig.15: Example where the nodes N1 and N2 are of the same loop

- The met point M is a node N_3 other than N_2 (see Fig 17) or a extremity point E_3(see Fig 16): the point cut-off C_1 will be in the case the pixel be located just will in this case the pixel located just right the node N_1 (see Fig 7.24)

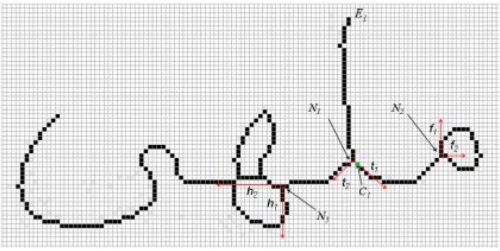


Fig.16 : Exemple où le nœud N_1 n'appartient pas à une boucle, le nœud N_2 appartient à une boucle et le point M est un nœud N_3

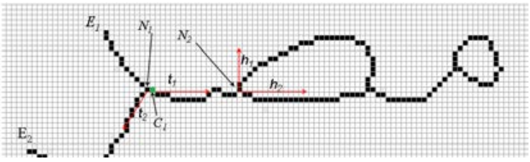


Fig.17: Example where the node N_1 does not belong to a loop, the node N_2 belongs to the loop and the point M is an extremity point E_2

4.3 Paws Segmentation algorithm in grapheme

Segmentation of graphemes PAW C goes through the following steps:

a. Identify the starting point.
b. to Search cutoff point.
c. Proceed with a segmentation the PAW in this cutoff point.
d. Apply the stages a), b) and c) in each of both segments of the PAW obtained.
e. Repeat this process until the obtained segments be are constituted or only simple dots with two points extremities (graphemes without loops), Either of simple loops or two attached loops(It is the case of the characterه)

The example in Fig 18 below shows the different stages of segmentation PAW "مستمع" into grapheme.

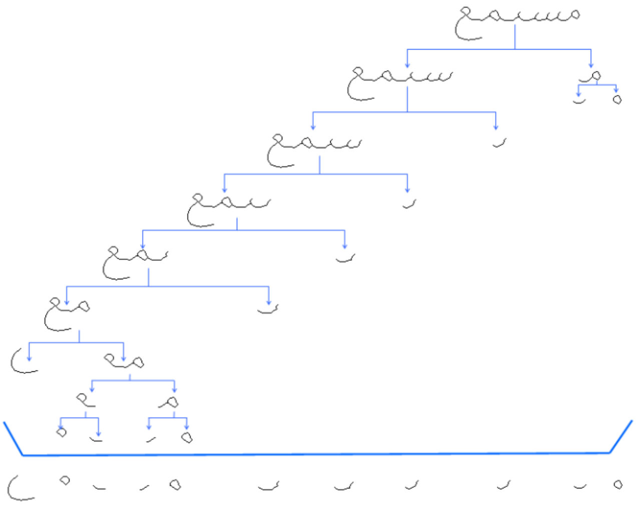


Fig.18: The different stages of the segmentation of the word "مستمع" in grapheme

5 Identification of characters of the pseudo-word

Those parts of pseudo-word obtained as a result this first phase belonged to three classes, For this, we recover the diacritical points we have eliminated in Section 3.2, and calculate for each grapheme G_i obtained its characteristic matrix using the approach adopted in the chapter of the first part to the recognition of isolated characters. Thus at step i and according to the nature of the grapheme G_i and those of its successors we distinguish the following cases:

a) If graphemes G_i and G_{i+1} are vertically aligned (The one over the other one), then these two graphemes will be concatenated to form a single character and the algorithm will pass then in the grapheme G_{i+2}.

b) If the grapheme G_i is a loop then it will be concatenated with the grapheme G_{i+1} to form a single character and the algorithm will pass then in the grapheme G_{i+2}.

c) Otherwise we distinguish two cases:

I. If the grapheme G_i has points (below or above) then in this case the grapheme G_i is a character and the algorithm will pass then to grapheme G_{i+1}.

II. Otherwise, we consider the following grapheme G_{i+1} and we distinguish the two following cases:

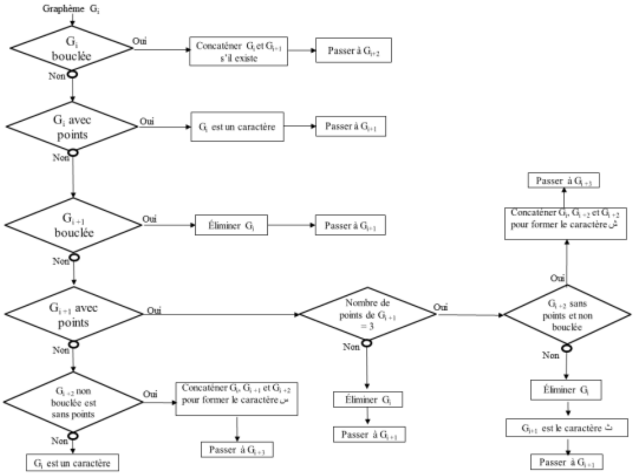
α. If the grapheme G_{i+1} is a loop then the grapheme G_i belongs to the third class and therefore it will be eliminated.

β. Otherwise, (ie the grapheme G_{i+1} is not a loop) then we conclude with the following cases:

1. The grapheme G_{i+1} does not have any point and the grapheme G_{i+2} is not a loop and does not have points, and in this case the graphemes G_i, G_{i+1} and G_{i+2} will be concatenated to form the "س" character. The algorithm then pass to grapheme G_{i+3}.
2. The the grapheme G_{i+1} does not have any points, but the grapheme G_{i+2} is a loop or he has the points, then the grapheme G_i is a character and the algorithm will pass ensuite to grapheme G_{i+1}
3. The grapheme G_{i+1} possesses one or two points, then the grapheme Gi belongs to the third class and therefore it will be eliminated. The algorithm then will pass grapheme G_{i+1}.
4. The grapheme G_{i+1} possesses three points, then we are facing two cases:

a. The grapheme G_{i+2} is not a loop and does not have any dots, in this case the graphemes G_i, G_{i+1} and G_{i+2} will be concatenated to form the character "ش". The algorithm then will pass grapheme Gi+3.

b. Otherwise (ie grapheme G_{i+2} is a loop or has spots) then the grapheme Gi belongs to the third class and therefore it will be eliminated, and the grapheme G_{i+1} is a character(is character"ث"), The algorithm then will pass the grapheme G_{i+2}.



Algorithm identification characters of pseudo-word

By applying this algorithm to the example of pseudo-word"مستمع" of Fig 19 above, we get the following characters:

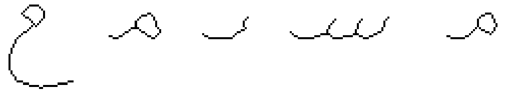


Figure -19 : Exemple de segmentation du pseudo-mot "مستمع" en caractères

6. Segmentation of ligatures

For some fonts, we note the existence of characters concatenated vertically and this makes segmentation them a very difficult task (see Fig 20). The recognition of such forms as much as individual characters is task a very difficult [11]. Since our segmentation system to return the pieces of a pseudo-word are chained upright as a single grapheme (Part 4.3)it would therefore be judicious to seek to recognize these characters concatenated vertically as being a single the grapheme in instead of the segmented and seek to recognize them separately.

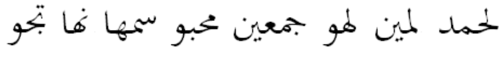
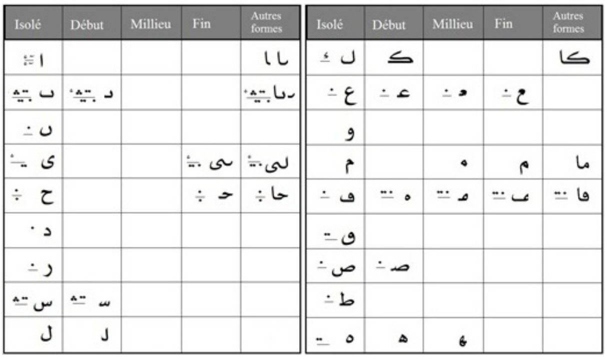


Fig.20: Examples of vertical ligatures

To take into account these ligatures, we have enriched the grapheme database formed by the various forms writing of Arabic characters (as a function of the character position in the word) with the forms different the ligation existing in texts which recognizes.

7. Learning

To evaluate our system, we proposed a new alphabet which operates a number of specificities of Arabic writing, especially the myriad forms that may take one letter a function to its position in the word, and redundancy the letter forms which only differ than by the presence, the position or the number of points. This alphabet also contains all forms of concatenation of two characters than our segmentation system can generate as outputs and the various ligatures observed in a an alysis of the writings studied in this phase of the test (see Fig 21 for all 28 of these forms).

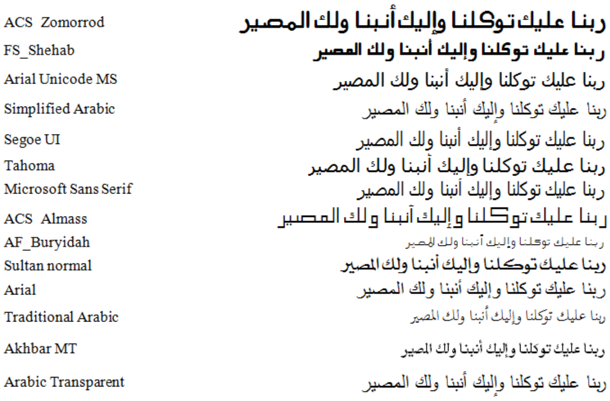


Isolé	Début	Millieu	Fin	Autres formes	Isolé	Début	Millieu	Fin	Autres formes
ا				ٮا ا	ك	كـ			كا
ٮ	ٮـ			ٮا	ع	عـ	ـعـ	ـع	
ں					و				
ى			ـى	لى	م		ـمـ	ـم	ما
ح			ـح	حا	ڡ	ڡـ	ـڡـ	ـڡ	ڡا
د					ٯ				
ر					ص	صـ			
س	سـ				ط				
ل	لـ				ه	هـ	ـهـ		

Fig.2 : Alphabet graphemes

7.1 The bass of test

Saw that there no basis of standard images of Arabic texts, we built our own database of multifontes printed Arabic texts. We selected 14 Arab fonts among those most used: *FS_Shehab, Arial Unicode MS, Microsoft Sans Serif, Segoe UI, Tahoma, Simplified Arabic, AF_Buryidah, Sultan normal, Traditional Arabic, Akhbar MT, and Arabic Transparent.* These images were recorded with the free format "PNG" and with a resolution of 300 ppi (pixels per inch). The images taken into considered are grayscale. Chaque page contient une image du texte arabe a 1000 mots and 2026 PAWs written with size 18 (see sample in Fig 22 above) written each time with 11 fonts. Ces polices ont la particularité de la morphologie relativement complexe, they have several ligatures, overlap and collisions between characters and diacritical marks of a single word, in more horizontals lengthenings in the case of Sultan normal. Then we are interested in this work also, the Akhbar MT font the most used in Arab newspapers. Fig 1 presents an example of text generated by 14 different fonts.



ACS Zomorrod
ربنا عليك توكلنا وإليك أنبنا ولك المصير
FS_Shehab
ربنا عليك توكلنا وإليك أنبنا ولك المصير
Arial Unicode MS
ربنا عليك توكلنا وإليك أنبنا ولك المصير
Simplified Arabic
ربنا عليك توكلنا وإليك أنبنا ولك المصير
Segoe UI
ربنا عليك توكلنا وإليك أنبنا ولك المصير
Tahoma
ربنا عليك توكلنا وإليك أنبنا ولك المصير
Microsoft Sans Serif
ربنا عليك توكلنا وإليك أنبنا ولك المصير
ACS Almass
ربنا عليك توكلنا وإليك أنبنا ولك المصير
AF_Buryidah
ربنا عليك توكلنا وإليك أنبنا ولك المصير
Sultan normal
ربنا عليك توكلنا وإليك أنبنا ولك المصير
Arial
ربنا عليك توكلنا وإليك أنبنا ولك المصير
Traditional Arabic
ربنا عليك توكلنا وإليك أنبنا ولك المصير
Akhbar MT
ربنا عليك توكلنا وإليك أنبنا ولك المصير
Arabic Transparent
ربنا عليك توكلنا وإليك أنبنا ولك المصير

Fig.3 : Arabic text on every typeface used in test data

The test pages sudden the following treatments:

- Skeletonization;
- Image segmentation of text into lines;
- Segmentation of lines into words;
- Segmenting words into pseudo-words;
- Segmentation of pseudo-words into characters and ligatures;
- Recognition of characters and ligatures

7.2 Learning

we note a relationship which brings together us fonts by three class C_1, C_2 et C_I. For the classes C_1 and C_2 are natures various morphologically and C_I is an intermediate class of both. This intuition is found for the OFR (Optical Font Recognition) [12]. In our strategy it allows to improve the performances our system and reduce the complexity of the task by bringing back an problematic to multifonte a the problematic monofontes: It considerably simplifies later tasks by optimizing of search by reducing the number of various characteristics at considered.

For fonts used in the learning phase, we realized multiple tests and those are the fonts ACS Zomorrod, ACS Almass and Arial take each of the classes C_1, C_2 et C_I respectively. These fonts have it possible to obtain the best results in the test phase. in the test phase. Choosing these fonts as was dictated by the fact that during the phases of segmenting text in lines, then lines into words and finally the words into pseudo-words we obtained a segmentation rate equal to 100% for the fonts, and on the other by the fact that the two fonts ACS

Zomorrod and A CS Almass g iving the best recognition rate during the learning phase in the case of character recognition isolated[7]. And we improved by a third font Arial which includes some characteristic which we resembles good correct for the learning phase for both phases of segmentation and characters classification.

C_1	C_I	C_2
ACS Zomorrod *FS_Shehab* *Arial Unicode MS* *Microsoft Sans Sarif* *Segoe UI* *Tahoma* *Simplified Arabic*	*ACS Almass* *AF_Buryidah* *Sultan normal*	*Arial (Corps CS)* *Traditional Arabic* *Akhbar MT* *Arabic Transparent*

Table 1. Classes of fonts

8. Results and Discussion

8.1 Segmentation

The segmentation process into PAWs proposed does not allow the resolution of the problems of over-segmentation and sub-segmentation perfectly. However, the detection of the presence of these two problems is possible with a comparison of the number extracts of PAWs by this method and the number of paws of the annotation of the image. The images present that these types of problems number will be comprehensive in our images. These results are close to those obtained by Abdulkader A. [13]. A thorough analysis of these images indicates to us that:

- Most often, over-segmentation in PAWs is mainly due to the phenomenon of ligature'سمع' for us our approach can not complete to make segmentation by against the system gives only just 'س ع '. and what types of problems they found Traditional Arabic.

- The sub-segmentation in Paws is due to PAWS which contains the final character 'ص' as t he extremity point in all the fonts it gives through our approach Segmentation two graphemes: 'ں'et' صـ' whereas they should not (a only character 'ص' end). This phenomenon sub-segmentation undesirable for these paws is due to the small expressed see Figure 23. Another kind of sub-segmentation for this character 'ه' see Fig 23, Our system gives as result two characters 'م' and 'ح' this problem we find for fonts : FS_Shehab, Microsoft Sans Serif, Segoe UI, Arial Unicode MS. In our future work, we will present a direction for reflection for a solution to provide a solution the problem considered.

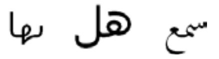


Fig.4 : Examples of some segmentation error ligature

8. 2 Recognition

If we evaluate the male characters recognized in analyzing their shapes after the step of segmenting by our approach, is observed the percentage of error comes from the recognition of bad character 'ع' in their middle and final position. Our classification process which is based on Bezier [7] approach he classified as the character 'ح. Other recognition errors are due to the fact of the great similarity which exists between these characters, vision even by the human eye made these mistakes for these fonts. And finally us giving the picture below which shows the percentage of segmentation correct of Paws and word for each fonts (reps. percentages of correct recognition of Paws and fonts for each word).

		Segmentation Correct %		Reconnaissance Correct %	
	Font	PAWs	Mots	PAWs	Mots
	ACS Zomorrod	99,61	99,20	99,36	98,70
	FS_Shehab	96,00	91,90	91,76	83,30
C_1	*Arial Unicode MS*	98,22	96,40	98,37	96,70
	Simplified Arabic	99,61	99,20	95,36	90,60
	Segoe UI	98,22	96,40	98,57	97,10
	Tahoma	96,00	91,90	96,40	92,70
	Microsoft Sans Serif	98,22	96,40	98,22	96,40
	ACS Almass	98,22	96,40	98,62	97,20

C_1	*AF_Buryidah*	99,36	98,70	96,10	92,10
	Sultan normal	99,61	99,20	95,36	90,60
C_2	*Arial*	99,61	99,20	95,36	90,60
	Traditional Arabic	99,31	98,60	94,13	88,10
	Akhbar MT	99,61	99,20	95,36	90,60
	Arabic Transparent	99,61	99,20	95,36	90,60
	Average	98,66	97,28	96,31	92,52

Table 2. Recognition results

The results obtained showed that the system has succeeded in recognize all of the characters and ligatures(see Table2). This performance is explained by the robustness of the approach and regularity of font being tested. We plan to test this approach less regular fonts in order to see their limitations and overcome these potential limitations.

In Table 3 presents the recognition rates of some character recognition systems developed recently by several research teams.

Systems	**Recognition rate of Paws**
Iping Supriana, Albadr Nasution, [14]	74.33%
STARS Company, [15] Arabic OCR	95.00%
Andrey Stolyarenko et Nachum Dershowitz, [16]	71.00%
Our system	96,31%

Table 3. Comparison of recognition rate for different Arabic recognition systems

9. Conclusion

Convinced of the superiority of the concept of PAW in Arabic script, the analytical approach pseudo is became our first choice. In this work, we presented a system for offline segmentation of Arabic script calligraphy printed multifonte complex and exhibiting strong ligatures. It is based on well-chosen points and their relationships which exist for the seal of character in Arab PAWS. Through a efficient selection of appropriate features for each of the three fonts and detailed management of errors the output we have obtained interesting results reflecting a robust and efficient recognition; This is an interesting contribution to the field of the recognition of complex multifonte Arabic script with solid ligatures.

References

[1] K.M. Sayre, Machine recognition of handwritten words : A project report. Pattern Recognition, vol. 5, no . 3, pp. 213-228, septembre 1973.

[2] X. Dupre. Contributions à la reconnaissance de l'écriture cursive à l'aide de modèles de Markov cachés. PhD thesis, Univérsité de Rene Descartes - Paris V, 2003.

[3] E. Lecolinet. Planar markov modeling for arabic writing recognition : Advancement state. In Proceedings of the first IEEE Int. Conf. on Document Analysis and Recognition (ICDAR), pp. 740-748, Saint Malo, France, Sept. 1991.

[4] N. Essoukri Ben Amara, F. Bouslama. Classification of Arabic Script Using Multiple Sources of Information: State of the Art and Perspectives. Int. Journal on Document Analysis and Recognition, vol. 5, no. 4, pp. 195-212, 2003.

[5] S. Gazzah, N. Essoukri, Ben Amara. Utilisation des Ondelettes en Caractérisation des Fontes Arabes. Int. Conf. On Image and Signal Processing, ICISP'2003, Maroc, June 2003.

[6] R. G. Casey, E. Lecolinet, A Survey of Methods and Strategies in Character Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, pp. 690-706, 1996.

[7] A. Mazroui, A. Kerkour Elmiad, Recognition of multifont Isolated Arabic characters by Bézier Curves, International Journal of Computer Applications, vol. 100 - No. 6, 2014

[8] F. Menasri. Contributions à la reconnaissance de l'écriture Arabe manuscrite. PhD thesis, Université de Paris Descartes, 2008

[9] A.M. AL-Shatnawi, S. AL-Salaimeh, F. H. AL-Zawaideh, and O. Khairuddin, Offline Arabic Text Recognition-AnOverview, World of Computer Science and Information Technology Journal (WCSIT), vol. 1, no. 5, pp. 184-192, 2011.

[10] M. S. Khorsheed, Off-Line Arabic Character Recognition - A Review. Pattern

Analysis & Applications vol. 5, no. 1, pp. 31-45, 2002.

[11] U.-V. Marti, H. Bunke, Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition systems, World Scientific Publishing Co., Inc., pp. 65-90, 2002.

[12] C.A.CRUZ, R.R.KUOPA, M.R.AYALA, A.A.GONZALEZ, R.E.PEREZ, "High order statistical texture analysis-font recognition applied", *Pattern Recognition Letters, vol 26, N° 2*, 2005, pp 135-145.

[13] AdbulKader A., Two-tier approach for Arabic offline handwriting recognition. In The Tenth International Workshop on F rontiers in Handwriting Recognition (IWFHR 10). (2006).

[14] Iping Supriana, , Albadr Nasution, Arabic Character Recognition System Development, Vol 11, June 2013, p: 334–341, 4th International Conference on E lectrical Engineering and Informatics, ICEEI 2013

[15] STARS Company, Arabic OCR, 2010

[16] Andrey Stolyarenko & Nachum Dershowitz. OCR for Arabic using SIFT Descriptors with Online Failure, , O CR for Arabic using SIFT Descriptors with Online Failure, חמישי, כ"ט אייר התשע"אISCOL 2011, Thursday, June 2, 2011.