Recognition of Arabic characters and fonts

Article *in* International Journal of Engineering Science and Technology · October 2010

Source: DOAJ

6 authors, including:

Ilham Chaker

13 PUBLICATIONS **28** CITATIONS

SEE PROFILE

Qjidaa Hassan

Sidi Mohamed Ben Abdellah University

206 PUBLICATIONS **1,375** CITATIONS

SEE PROFILE

Rachid Benslimane

Sidi Mohamed Ben Abdellah University

70 PUBLICATIONS **378** CITATIONS

SEE PROFILE

RECOGNITION OF ARABIC CHARACTERS AND FONTS

ILHAM CHAKER
LTTI, University Sidi Mohamed Ben Abdellah, Fez, Morocco
Chaker_ilham@yahoo.fr

MOSTAFA HARTI
UFR INTIC, Faculty of Sciences Dhar El Mehraz University Sidi Mohamed Ben Abdellah, Fez, Morocco
mharti@fsdmfes.ac.ma

HASSAN QJIDAH
LESSI, Faculty of Sciences Dhar El Mehraz University Sidi Mohamed Ben Abdellah, Fez, Morocco
qjidah@yahoo.fr

RACHID BENSLIMANE
LTTI, University Sidi Mohamed Ben Abdellah, Fez, Morocco
r.benslimane1@gmail.com

Abstract:

In this paper, we propose a method for recognizing Arabic characters and fonts. This method is based on a retrieval procedure using a dissimilarity measure characterizing the character to be recognized. This dissimilarity measure is calculated on the basis of some polygonal attributes extracting from a polygonal approximation of the character. These attributes are insensitive to the size of the character, its orientation and its translation.

The performance of the proposed method is evaluated by a set of tests made on a database of characters combining 10 classes of fonts of Arabic characters that are mostly used.

Keywords: Arabic Character Recognition, Arabic Fonts, Polygonal approximation, dissimilarity index, polygonal attributes.

1. Introduction:

Nowadays, the recognition of Arabic characters constitutes a major concern by the community of researchers, especially those of Arab countries. Despite the significant progress made over the recent years, the results have not been able to achieve the performance that can match those achieved in the case of other scripts such as Latin or Chinese.

Some researchers have focused on the recognition in real time [1] using the graphic tablets, which simplifies the problem in part by restoring the sense of the outline; others have looked at the printed characters and / or the manuscript in "Off-line" by using a scanner or a camera to capture documents. Since then, several pre-processing tools have been developed for the skeletonization of the image and its smoothness, and for the determination of the edge and the feature extraction. Different approaches of recognition have been developed, these include: the statistical approach [2,3,4,5], the structural approach [6,7,9] and the stochastic approach [8,12]. Generally all these approaches and others, tend to extract each in its way, a class of features and

subsequently evaluate the likelihood between the primitives extracted and those from the prototype forms, already learned by the recognition system.

Generally, Optical character recognition (OCR) systems can be divided into three groups: Mono-font, Multi-font, and Omni-font. Mono-font OCR systems deal with documents written with one specific font; their accuracy is very high but they need a specific module for each font. Omni-font OCR systems allow the recognition of characters of any font, and for this reason their accuracy is typically lower. Finally, Multi-font OCR systems handle a subset of the existing fonts. Their accuracy is related to the number and the similarity of the fonts under consideration.

Character recognition accuracy can be improved using an Optical Font Recognizer (OFR) to detect the font type and subsequently convert the multi-font problem into mono-font character recognition problem [10].

In spite of the importance of the **OFR** in an **OCR** system, it remains a problem often neglected and the studies in this field are few especially, in the case of Arabic [11].

There are three strategies of combining the OFR with the OCR: the a posteriori approach, the a priori approach and the hybrid approach.

The a posteriori approach consists of recognizing the font of a text using the knowledge of characters appearing in it [12]. It helps to correct recognition errors and allows to find the original description of the document.

The a priori approach consists of identifying the text font without any knowledge of the characters that appear in that text. It greatly simplifies the subsequent tasks in an OCR system by optimizing searches by reducing the number of glyphs (the different representations of a character). In addition, it improves system performance while receiving various information about the font, previously available. The hybrid approach can be adapted by combining the a posteriori approach and the a priori one.

In this context, we propose a method for identifying Arabic characters and fonts, based on a description of the character by a dissimilarity index calculated on the polygon representing the character to be recognized. This method is invariant to the size of the character, its orientation and its translation.

The rest of the paper is organized as follows:

Section 2 gives a general description of the proposed recognition method.

Section 3 describes the pre-processing necessary for the recognition operation. This is the edge detection of characters having a width of one pixel.

Section 4 describes the polygonal approximation method of edges.

Section 5 presents the polygonal attributes of characters allowing in a final stage their identification.

Section 6 presents the dissimilarity measure chosen and the last section gives results on test characters.

2. Description of the proposed recognition method

The principle of the suggested recognition method is based on the characterization of each character by a shape index. This index represents a set of parameters, invariant to translation, rotation and scale parameter. In this work, the shape index is not calculated directly on the edge of the studied character, but rather on the polygonal representation of its edge. The polygonal representation of the character edge requires a pre-processing step leading to represent a character by a thin edge with one-pixel width. Finally, the character recognition from models in the database of characters is based on calculating a dissimilarity measure comparing the shape index of the character to recognize with the shape index of the models in the database.

The synoptic diagram of the proposed method is therefore presented as follows:

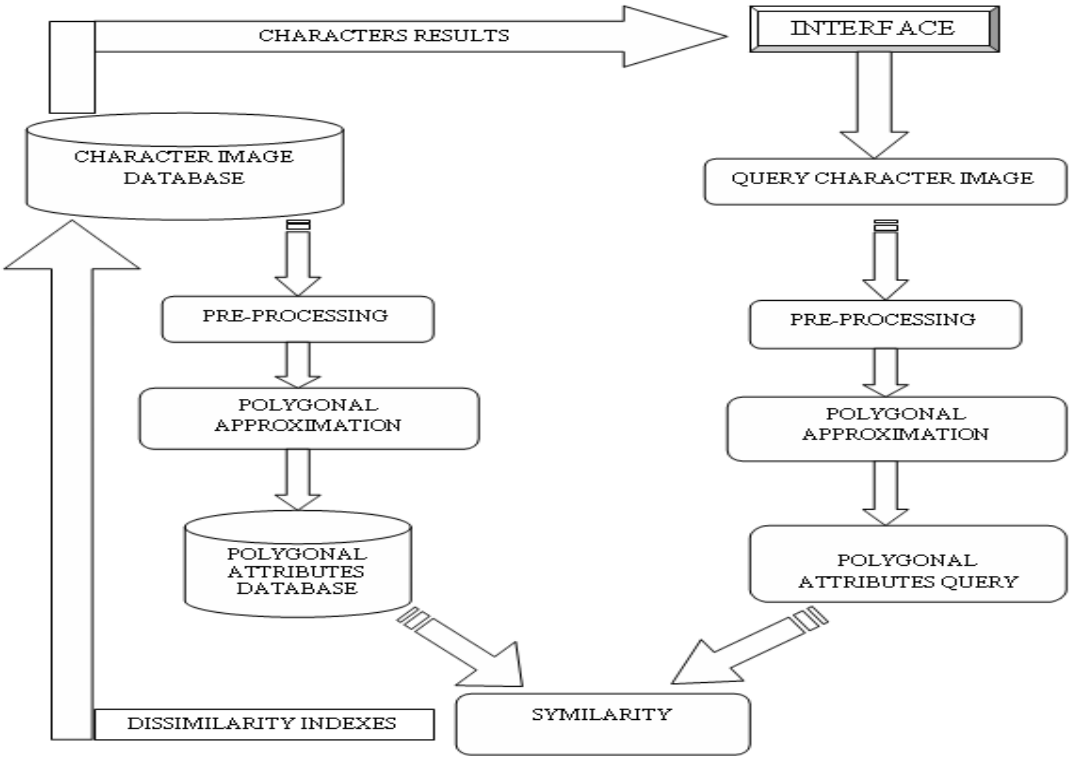


Fig 1. The synoptic diagram of the proposed method

3. Segmentation by edge detection of the characters

According to the image of character to recognize, an edge detection method is applied. In this work, we applied the Canny method [13] followed by a skeletonization procedure in order to obtain a thin edge, with one pixel width. This skeletonization is based on homotopic thinning until idempotence [14].

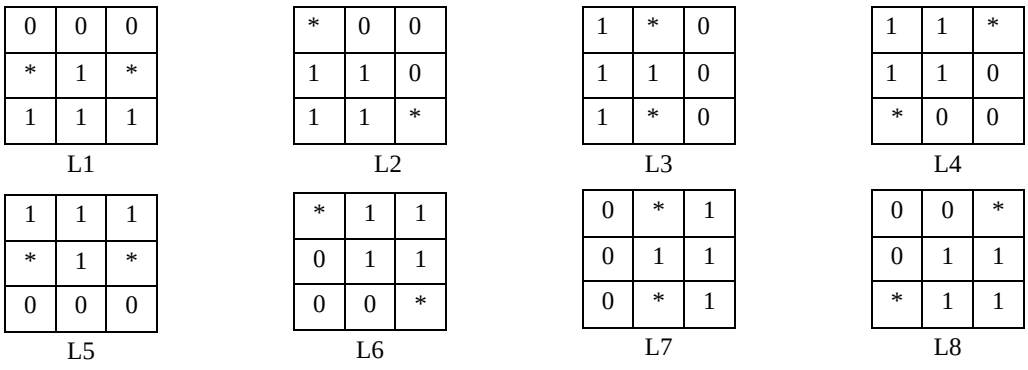


Fig 2. The eight structuring elements of the structuring family L (*: indifferent element: can take the value 1 or 0)

4. Polygonal approximation of the character edge and its normalization

The purpose of a polygonal approximation method of an edge is to extract from a string of edge points, successive segments to minimize a global error criterion or respect a local approximation error. To this end, many methods have been developed [15,16,17].

[18,19,20] examine successively the edge points to determine the longest segment that satisfies a predefined threshold of tolerance. This search process is repeated to find all segments of the polygon approaching the edge in an optimal way.

In spite of the significant number of polygonal approximation methods, there are still major problems of robustness, stability to the geometrical transformations and complexity. Moreover, the algorithms based on the threshold of tolerance of errors which are manually defined without any knowledge on the value of the most relevant threshold. This one may be different from an edge to another. In this work, we have used the method developed by Huang and Wang [21]. This choice is motivated by the simplicity of implementation and the good behaviour with noise.

The algorithm of the Huang and Wang method is given as follows:

1. Find the starting point P_0 which is the farthest point of the edge from the centroid.
2. $P_1 = P_0$ and $P_2 = P_0$.
3. Find the farthest point P_3 from P_2 belonging to the edge.
4. $P_a = P_2$; $P_b = P_3$.
5. If $P_1 = P_3$ stop, if not $P_1 = P_2$; $P_2 = P_3$ and return in 3.

For each segment $[P_i\ P_{i+1}]$ and each part of the curve with the same endpoints P_i and P_{i+1}, we seek the point P_{max} such as the distance (d) from P_{max} to the segment $[P_i\ P_{i+1}]$ is maximal. The P_{max} point is used to build the new polygon which will have a vertex more than the old one. The process is repeated between $P_i\ P_{max}$ and $P_{max}\ P_{i+1}$ until d is less than a tolerance value of the approximation.

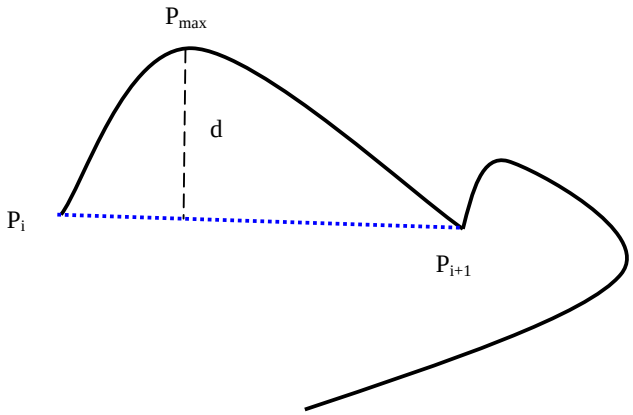


Fig 3. Illustration of the polygonization algorithm.

The recognition method should possess the characteristics of being invariant to the translation, rotation, and scale changes of the character. A normalisation of its polygon is necessary [21]

The polygon can be normalized using the character's centroid, which can be calculated as:

$$(x_c, y_c) = (\frac{1}{N}\sum_i xi, \frac{1}{N}\sum_i yi) = Pc = \text{Centroid}. \quad (1)$$

To this end, we look for the normalization factor η witch is specified as the longest distance from boundary points to the character's centroid.

Then each vertex point *Pi (xi, yi),* on the polygon can be normalized as:

$$x'_i = (xi - xc)/\eta$$
$$y'_i = (yi - yc)/\eta. \quad (2)$$

5. Character characterization by polygonal Attributes

The character recognition requires the characterization of its polygon representative by a set of attributes invariant to the translation, the rotation and the scale parameter.

The following attributes used in this work are those proposed by Huang and Wang [21].

- *Polar distance:* The polar distance *ri* is denoted as the distance between vertex point *P'i* and its respective centroid.
- (ii) *Polar angle:* The polar angle θi is denoted as the slope of the line connecting a vertex point *i* and its centroid.
- (iii) *Vertex angle:* The vertex angle a_i is denoted as the angle between the two line segments $[P'_{i1}, P'_i]$ and $[P'_i, P'_{i+1}]$.
- (iv) *Chord length:* Let l_i denote the i^{th} chord length of the normalized polygon, which is the distance between the two consecutive vertex points, $[P'_i, P'_{i+1}]$.

The above polygonal attributes are illustrated in fig 4:

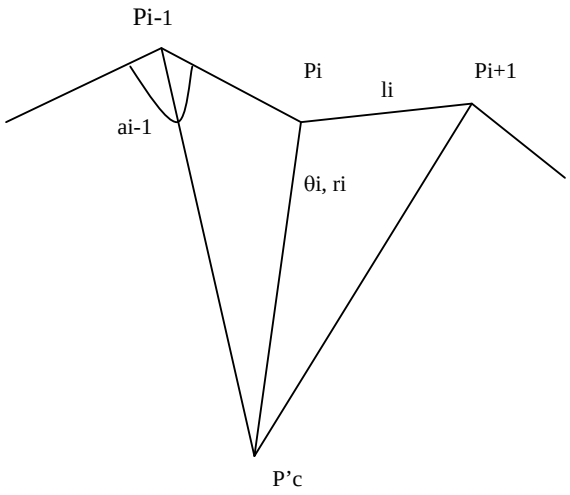


Fig 4. The illustration of the polygonal attributes.

In [21] the authors propose an optimization to avoid comparing the polygon of the character to be recognized to all models of the database. Indeed, two polygons whose edge length is too different from one another are polygons of different objects. We will not use this optimization although it is very simple. It may, however, be useful if the number of models is high.

6. Dissimilarity index

The dissimilarity index gives an idea on the similarity between a given character and a model character. It is based on the comparison of normalized polygons, characterized by their attributes. Let *t* be the polygon of the character to recognize with *M* vertex points and *s* the normalized polygon of the model with *N* vertex point.

Before calculating the dissimilarity measure of two polygons to be compared, it is necessary to obtain some information about their rotation to match the two polygons in any arbitrary orientation.

The algorithm used to retrieve information on the rotation of the polygon is the following [21]:

1. Find the corresponding starting (vertex) points of the two polygons (The farthest point of the edge from the centroid) .
2. Suppose that P^t_1 is the starting point of the test polygon and P^s_1 is the starting point of the model polygon.

3. Calculate the rotation angle: $\Delta\theta = \theta s - \theta t$. where θs and θt are the polar angles of the two starting vertex points.
4. We rotate the polygon object by $\Delta\theta$ so that the vertex angle of the starting point of the polygon *t* corresponds to that of the polygon *s.*

After rotating the test polygon by $\Delta\theta$ we calculate the dissimilarity measure *D* between *s* and *t* which can be denoted as follows:

$$D\ (s,t) = D_m(s,t) + D_m(t,s). \tag{3}$$

$$D_m(s,t) = \frac{1}{M}\sum_{i=1}^{M} d^2\left(p_i^s, E_t\left[p_i^s\right]\right). \tag{4}$$

$$D_m(t,s) = \frac{1}{N}\sum_{i=1}^{N} d^2\left(p_i^t, E_s\left[p_i^t\right]\right). \tag{5}$$

Where *Et[p]* or *Es[p]* is the expected point on polygon *t* or *s* for point *p*, and *d(p, q)* is the (Euclidean) distance between point *p* and point *q*. that is to say between the point and its point estimate on a polygon. The developed calculation of these expected points are given in [21].

This dissimilarity measure is calculated for each vertex point in the polygon to be recognized, and then we get the weakest dissimilarity.

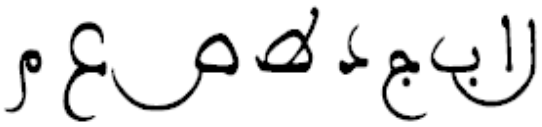
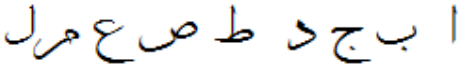
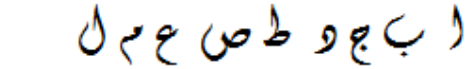
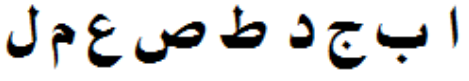
The approximating polygon characterizing the character to be recognized will be compared with all the models by this algorithm and classified as the model having the minimum dissimilarity measure with it.

The font of the recognized character is identified through the font membership of the found model character.

7. Experimental results:

To test the performance of the proposed method, we have developed a database containing Arabic characters written in different fonts. The different characters have been seen in 10 fonts among the most commonly used in applications running under Windows. These fonts are: Tholoth, Diwani, Naskhi, Andalus, Kuffi, Arial, Tahoma, Courier, Arabic Typesetting. And Al_Mabssout font [22] witch is developed in our laboratory (Table 1).

The Arabic Typesetting font and Arial have a very strong morphological similarity; this choice allowed us to evaluate the performance of our method in the case of similar fonts.

	Al_Mabssout [22]
	Tholoth
	Diwani
	Naskhi

ا ب ج د ط ص ع م ل	Andalous
ا ب ج د ط ص ع م ل	Kuffi
ا ب ج د ط ص ع م ل	Arial
ا ب ج د ط ص ع م ل	Courier New
ا ب ج د ط ص ع م ل	Tahoma
ا ب ج د ط ص ع م ل	Arabic Typesetting

Table 1. Representation of the 10 fonts

The performance of this recognition method is experimentally evaluated by calculating recognition and recall rates.

7.1/ Character and font recognition rate:

Three testing sets of characters are used to evaluate the proposed recognition method.

Each test has been applied to a 120 characters randomly chosen and belonging to 10 fonts studied. According to the first character result, all input characters in these tests were correctly recognized.

Test 1: Recognize a character and its font

This test is to identify an input character among the characters of the database. If the character exists in the database the system identifies it, if not, it displays the most similar character.
The Character and font recognition rate is : 100 %

Example:

The character image to be recognized is: Seen of the Arial font

Query	س				
Character & Font	*Seen Arial*	*Seen Arabic Typesetting*	*Seen Tahoma*	*Ain AlMabsout[22]*	*Seen Andalous*
	س	س	س	ع	س
Index	0.0	0.045	0.057	0.081	0.088

Table 2: The first 5 results of test 1

Test 2: Invariance to scale parameter

The input characters, in this test, are chosen in different sizes.

Example:

A character "Hah" of the Tholoth font with a size greater than that of the same character in the database:

Query					
Character & Font	Hah Tholoth	Hah Arial	Hah *Arabic Typesetting*	Hah Courier	Hah Naskhi
					
Index	0.001	0.022	0.034	0.034	0.053

Table 3. The first 5 results of test 2

The Character and font recognition rate is: 100 %

Test 3: Invariance to rotation:

The characters of this testing set have been rotated in different directions

The Character and font recognition rate is : 100 %.

Example:

We put, the character "Tah" of the Andalus font, to a rotation of 90 ° in the opposite sense of the watch needles.

Query					
Character & Font	Tah Andalous	Tah Arial	Tah *Arabic Typesetting*	Reh Kuffi	Tah Tahoma
					
Index	0.01	0.041	0.051	0.059	0.075

Table 4 .The first 5 results of test 3

7.2 / Recall rate:

The recall rate was calculated to quantify the performance of the proposed recognition method. This recall rate is defined as the number of relevant characters retrieved by the recognition method, divided by the total number of existing relevant characters (which should have been retrieved) [23].

$$\text{Recall rate (\%)} = \frac{\text{Number of relevant characters retrieved}}{\text{Number of relevant characters in Database}} * 100 \qquad (6)$$

To calculate this recall rate we used a testing set containing 20 characters randomly chosen (Table 5).

Query	RESULTS										Recall
	1	2	3	4	5	6	7	8	9	10	
AlMabsout Ain	AlMabsout Ain	Arial Sad	Tahoma Seen	Arial Seen	AT* Seen	Tholoth Ain	Diwani Ain	Andalous Seen	Naskhi Sad	Tahoma Ain	**4/10**
AlMabsout Beh	AlMabsout Beh	Courier Dal	AT Beh	Tholoth Lam	AlMabsout Lam	AlMabsout Hah	Tahoma Lam	Arial Lam	Diwani Beh	Courier Beh	**4/10**
Andalous Alif	Andalous Alif	Naskhi Alif	Courier Alif	Kuffi Alif	Tholoth Alif	Tholoth Meem	AlMabsout Alif	Arial Alif	Tahoma Alif	AT Alif	**9/10**
Andalous Dal	Andalous Dal	Kuffi Dal	Tahoma Dal	Naskhi Dal.	Arial Dal	Tholoth Dal	AlMabsout Hah.	AndalousAin	Tahoma Tah	Tahoma Waw	**6/10**
AT Hah	AT Hah	Courier Hah	Tholoth Hah	Arial Hah	Tahoma Hah	Diwani Hah	Naskhi Hah	AT Ain	Tahoma Reh	AT Beh	**7/10**
AT Lam	AT Lam	Arial Lam	Tahoma Lam	Courier Lam	Naskhi Lam	Tholoth Lam	AlMabsoutLam	AT Dal	AlMabsoutHeh	Courier Dal	**7/10**
Arial Heh	Arial Heh	Diwani Heh	Naskhi Heh	Courier Heh	Tahoma Heh	Andalous Heh	AT Heh	Andalous Meem	Kuffi Heh	Kuffi Meem	**8/10**
Arial Sad	Arial Sad	AT Sad	Naskhi Sad	Tahoma Sad	AT Seen	Naskhi Seen	Andalous Sad	Tahoma Seen	Kuffi Lam	Tholoth Sad	**6/10**
Courier Reh	Courier Reh	Arial Reh.	AT Reh	AlMabsout Reh	Diwani Waw	Tahoma Reh	AlMabsout Sad	Diwani Hah	AlMabsoutSeen	Tholoth Reh	**6/10**
Courier Sad	Courier Sad	Diwani Sad	Diwani Seen	Andalous Lam	Naskhi Waw	Courier Seen	Courier Waw	Tholoth Lam	Courier Dal	Courier Lam	**2/10**
Courier Waw	Courier Waw	AlMabsout Waw	Arial Waw	Tahoma Waw	AT Waw	Diwani Waw	AlMabsout Sad	Tholoth Seen	AlMabsoutSeen.	Diwani Sad	**6/10**
Kuffi Seen	Kuffi Seen	Andalous Seen	Arial Ain	Kuffi Meem	Tahoma Seen	Naskhi Ain	Arial Alif	Naskhi Alif	Tahoma Alif	Kuffi Alif	**3/10**
Kuff Tah	Kuffi Tah	Tholoth Heh	Arial Tah	Naskhi Tah	Andalous Heh	Tholoth Tah	Andalous Tah	Courier Tah	Kuffi Reh	Tahoma Heh	**6/10**
Naskhi Reh	Naskhi Reh	Tahoma Reh	Diwani Hah	AlMabsout Reh	AT Meem	Naskhi Lam	Andalous Reh	Courier Reh	Kuffi Lam	AT Reh	**6/10**
Naskhi Ain	Naskhi Ain	Tahoma Ain	Courier Ain	Arial Ain	Andalous Ain	Kuffi Ain	Diwani_ Ain	Tholoth Ain	AT Ain	AlMabsout Ain	**10/10**
Tahoma Alif	Tahoma Alif	AlMabsout Alif	Courier Alif	AT Alif	Tholoth Alif	Naskhi Alif	Arial Alif	Kuffi Alif	Andalous Alif	Tholoth Meem	**9/10**
Tholoth Tah	Tholoth Tah	Arial Tah	Tahoma Tah	AT Tah	Andalous Tah	Courier Tah	Kuffi Tah	AndalousWaw	Kuffi Meem	Kuffi Reh	7/10
Tahoma	Tahoma	Arial	Tahoma	AT	Diwani	Naskhi	Tholoth	Arial	AT	Courier	**5/10**

Dal	Dal	Dal	Waw	Dal	Beh	Dal	Dal	Reh	Waw	Reh	
Tholoth Sad	Tholoth Sad	AT Sad	Kuffi Lam	Tahoma Sad	AlMabsout Sad	Naskhi Sad	Tholoth Seen	AlMabsout Seen	Arial Sad	AT Seen	**6/10**
Tholoth Lam	Tholoth Lam	Courier Lam	AlMabsout Lam	Tahoma Lam	Courier Reh	AT Dal	AT Lam	Arial Lam	Arial Reh	Courier Dal	**6/10**
Average recall rate											**0.615**

Table 5: Average recall rate

* AT= ArabicTypesetting

The average recall rate is 61.5 % for characters randomly chosen and belonging to 10 fonts studied.

Conclusion

In this paper a new method for recognizing printed Arabic characters and fonts is proposed. This method is based on the dissimilarity index computed on the polygonal approximation of the character. This index uses polygonal attributes of the character, which are insensitive to translation rotation and scale parameter. The performance of the proposed method is measured by:

- The speed of recognition;
- The good performance in relation to noise;
- A character and font recognition rate of 100% calculated on the basis of the data used. Of course, this assumes a successful segmentation of characters and enough resolution to safeguard the borders of characters process.

The current system assumes the characters are already isolated and the algorithm can only recognize the isolated characters. For this project to have an additional value, the system should be able to isolate the connected characters from a word. Since there are techniques already developed for isolating Arabic characters, this system could be integrated with one of the existing character segmentation algorithms. So the perspective of this work consists to achieve the segmentation phase to complete this OCR system.

Acknowledgment:

The authors would like to thank the CNRST '*Centre National pour la Recheche Scientifique et Technique'* for the financial support for this research in the context of the pole of competence 2PC and the scholarship program**.**

References :

[1]. El-Badr. B, Mahmoud S.A..(1995) .A Survey and Bibliographie of Arabic optical text recognition. Signal Processing, Vol 41, pp 49-76,
[2]. Nemouchi S, Farah N (1990) ; *Reconnaissance de l'Ecriture Arabe par Systèmes Flous*,
[3]. El-Dabi.S, Ramsis.R and Kamel.A (1990); Arabic Character Recognition System: Statistical Approach for Recognizing Cursive Typewritten Text, Pattern Recognition 23, pp. 485-495.
[4]. Fakir,M ; Hassani. M. M. (1997); Automatic Arabic Characters Recognition by Moment Invariants, Colloque International de Telecommunications, Fes, Morocco, pp. 100-103.
[5]. Margner V, SARAT , (1992)- A System for the Recognition of Arabic Printed Text, Proc. 11th Int. Conf. on Pattern Recognition, pp. 561-564.
[6]. Al-Sadoun H. B.; Amin .A, (1995), A New Structural Technique for Recognizing Printed Arabic Text, Int. J. Pattern Recognition Artificiel Intell. 9,pp. 101-125.
[7]. El-Khaly, F , Sid-Ahmed , M, (1990), Machine Recognition of Optically Captured Machine Printed Arabic Text, Pattern Recognition 23, pp. 1207-1214.
[8]. Zramdini. A and I. rolf,(1993),Optical Font recognition from projection profiles, electroic publishing, Vol6.
[9]. Almuallim. H.; Yamaguch. S. (1987). A Method of Recognition of Arabic Cursive Handwriting, IEEE Trans. Pattern Anal. Machine Intell. PAMI-8(5), pp. 714-722.
[10]. Boukharouba.A , Bennia.A (2005), *Reconnaissance de Caractères Imprimés Omni-fonte*, 3rd International Conference: Sciences of Electronic, Technologies of Information and Telecommunications March 27-31.
[11]. ZAGHDEN.N, ALIMI.A,(2006), Reconnaissance des fontes arabes par l'utilisation des dimensions fractales et des ondelettes, Colloque International Francophone sur l'Ecrit et le Document (CIFED 06),Fribourg(Suisse) Septembre 18-21,pp-277_282
[12]. ANIGBOGU.JJ , (1992): " *Reconnaissance des textes imprimés multifontes à l'aide des modèles stochastiques et métriques ", Thèse de doctorat, Université de Nancy I.*
[13]. Deriche R, (1991) , Fast algorithm for low-level vision . IEEE Transaction on PAMI, Vol. 12, N' 1, p . 78-87.
[14]. Tsao Y. , Fu K. (1981), Parallel Thinning Algorithm for 3-D Pictures, Computer Graphics and Image Processing, 17, 315-331,.
[15]. DAVIS,T. (1999), Fast Decomposition of Digital Curves into Polygons Using the Haar Transform, IEEE Transactions on PAMI, vol. 21, no 8,, pp. 786.790
[16]. ROSIN P. L. (1997), Techniques for Assessing Polygonal Approximations of Curves, IEEE Transactions on PAMI, vol. 19, no 6 , pp. 659.666.
[17]. YIN P., A Tabu , (2000), Search Approach to Polygonal Approximation of Digital Curves, International Journal of Pattern Recognition and Arti_cial Intelligence, vol. 14, no 2, pp. 243-255.
[18]. RAMER U. (1972), An Iterative Procedure for the Polygonal Approximation of Plane Curves, Computer Graphics and Image Processing, vol. 1, pp. 244.256.
[19]. SKLANSKY J., GONZALEZ V. (1980),, Fast Polygonal Approximation of Digitized Curves, PR, vol. 12pp. 327.331.
[20]. Wall K. and Danielsson P.E. (1984),A fast sequential method for polygonal approximation of digitized curves, CVGIP, vol. 28, pp. 220-227,
[21]. HUANG.L.K , WANG J (1996), Efficient Shape Matching Through Model Based Shape Recognition. Pattern Recognition, Vol. 29, No. 2, pp. 207 215, 1996
[22]. CHAKER I, BENSLIMANE R. (2010), '*Création automatique de fontes marocaines'*, Internal report.
[23]. Makhoul J; Kubala F; Weischedel R (1999), Performance measures for information extraction. In: Proceedings of DARPA Broadcast News Workshop, Herndon, VA.