We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900
Open access books available

185,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Chapter 6

A Deterministic Algorithm for Arabic Character Recognition Based on Letter Properties

Evon Abu-Taieh, Auhood Alfaries, Nabeel Zanoon, Issam H. Al Hadid and Alia M. Abu-Tayeh

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76944

Abstract

Handheld devices are flooding the market, and their use is becoming essential among people. Hence, the need for fast and accurate character recognition methods that ease the data entry process for users arises. There are many methods developed for handwriting character recognition especially for Latin-based languages. On the other hand, character recognition methods for Arabic language are lacking and rare. The Arabic language has many traits that differentiate it from other languages: first, the writing process is from right to left; second, the letter changes shape according to the position in the work; and third, the writing is cursive. Such traits compel to produce a special character recognition method that helps in producing applications for Arabic language. This research proposes a deterministic algorithm that recognizes Arabic alphabet letters. The algorithm is based on four categorizations of Arabic alphabet letters. Then, the research suggested a deterministic algorithm composed of 34 rules that can predict the character based on the use of all of categorizations as attributes assembled in a matrix for this purpose.

Keywords: conditional random field, rule-based mode, word prediction, virtual keyboard, Arabic text entry, enhancement, text entry system, theory of randomized search heuristics, structured prediction, theory of computation

1. Introduction

Arabic language is one of the top five languages spoken in the world. Arabic is used by more than 422 million native and non-native speakers in the world. Also, the letters of the Arabic alphabets are used in other languages like Urdu (65 million natives and 94 million

IntechOpen

© 2018 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. CC BY

non-native) and Persian (110 million) languages. In addition, languages like Baluchi, Brahui, Pashto, Central Kurdish, Sindhi, Kashmiri, Punjabi, and Uyghur are using the Arabic letters. Hence, there is a need to develop an algorithm for character recognition for the Arabic language. Yet, there are major challenges that arise: first, Arabic is a cursive language. Unlike other languages written, Arabic alphabets change shape as written; hence, separate letters in Arabic are usually sub-word rather than stand-alone word. Second, Arabic is written from right to left unlike Latin languages. Third, Arabic has 28 alphabets, with some letters changing shapes based on the location of the letter in the word. Also, some letters are very similar in form yet have secondary marks to differentiate. Furthermore, Arabic is written from right to left cursively. Due to all the mentioned reasons, Arabic character recognition systems are under developed and lacking.

This research is composed of five sections. The first section presents 20 related works. Then, the research explains the letter shapes in Arabic language and the four categories used in the proposed algorithm. The four categorization methods will be employed to develop a deterministic algorithm method of categorization. The first categorization method depends on the number of dots used with each letter. The second categorization method depends on the shape of the letter, with classification to the letters. The third categorization is presented with the shape of the letter as used in the beginning, middle, and end of the word. The fourth categorization method relays on the proportion method, which is a method used in Arabic calligraphy that is based on rhombic dot. Then, the research suggested a deterministic algorithm composed of 34 rules that can predict the character based on the use of all of categorizations as attributes assembled in a matrix for this purpose.

2. Related work

Character recognition is an open-ended problem. A computer cannot recognize character or language alphabets. There is a great progress in character recognition that all can be seen in the different smart application used on smart phones and pad as well as notebooks and PCs. Problems that arise with non-Latin languages are well known, and many researchers have conducted research for their respective languages: Hindi language [1]; Chinese language [2–4]; and Arabic Language [5, 6]. Furthermore, many researchers have conducted research for the Arabic alphabets: Parvez and Mahmoud [7] conducted a survey for text recognition and published their work in a paper titled *Offline Arabic Handwritten Text Recognition: A Survey*. Three researchers [8] conducted a research and published their work in a paper titled *Robust Named Entity Detection Using an Arabic Offline Handwriting Recognition System*; the paper focus was "on extraction of a predefined set of Arabic named entities (NEs) in Arabic handwritten text." Another research by Fouad Slimane, Slim Kanoun, Adel M. Alimi, Jean Hennebert, and Rolf Ingold was conducted in 2010, yet the concentration was on printed character rather than handwritten one. The research conducted by Fard, Moghadam, Bidgoli, and Hussain [9] was very promising and was conducted on Persian language, yet the method used was neural

http://dx.doi.org/10.5772/intechopen.76944

network based which is not deterministic. Another research conducted by Abu-Taieh [10] used an enhanced method of neural network. Another promising research that was studied by Aljarrah et al. [11] is the study concentrated on printed Arabic rather than hand written in order to produce Arabic optical character recognition system. The need for Arabic character recognition is evident according to Ali and Sagheer [12] which addresses the need of smart mobile phones and tablets and hence the need for Arabic character recognition.

Researchers Supriana and Nasution [13] cited nine works including their own research that all are non-deterministic. The research of Sarfraz, Ahmed, and Ghazi [14] developed a license plate recognition system. The research by Izakian, Monadjemi, Ladani, and Zamanifar [15] used chain codes, while Abandah, Khedher, and Mohammed [16] used selected feature extraction techniques. In their research Al-Taani and Al-Haj [17] used structural features, while Kapogiannopoulos and Kalouptsidis [18] used skew angle. The research of Zidouri [19] proposed a general method for Arabic letter segmentation, while Amin [20] used global features and decision tree technique on printed letters not handwritten. Cowell and Hussain [21] used extracting features.

3. Letters in Arabic language

To develop the proposed algorithm, the researchers studied and presented the different categorizations for Arabic letters. Next, each categorization will be explained accordingly. The first categorization method depends on the number of dots used with each letter. The second categorization method depends on the shape of the letter, with classification to the letters. The third categorization is presented with the shape of the letter as used in the beginning, middle, and end of the word. The fourth categorization method relays on the proportion method, which is a method used in Arabic calligraphy that is based on rhombic dot. Each categorization will be explained in Sections 3.1, 3.2, 3.3, and 3.4.

3.1. First categorization: number of dots in the letter

The use of dots to distinguish letters in Latin-based languages is familiar to people. In English, small letters I and J are distinguished by using a dot on top of the letter. In Arabic the use of a dot is used extensively; in fact, only 12 letters out of 28 letters are not doted. Furthermore, some letters use one, two, and three dots. Next, the concept of doted letters will be explained.

The first categorization is according to the number of dots used with each letter. This categorization splits the 28 letters (**Table 1**) into five branches, and from within it breeds two extra letters. The first branch is composed of 12 letters that has no dots whatsoever. The second branch is composed of ten letters: the eight letters have their dot above the body of the letter, and the other two letters have their dot below the letter body. The third branch is composed of four letters: the three letters have their two dots above the body of the letter, and the other one has two dots below its body. The fourth branch has two letters with three dots above the body.

First branch	No dots	12	ا و ه م ل ع ط ص س ر د ح
Second branch	One dot	10	ن ف غ ظ ض ذ ز خ ج ب
Third branch	Two dots	3	ة ق ت ي
Fourth branch	Three dots	2	ش ث
Fifth branch	With hamza	3	ك أ ٱو

Table 1. Arabic alphabets (according to dots).

The fifth branch deals with hamza: there is one basic letter where the hamza is part of the letter "ك," and the other hamza is not part of the letter like the "أ" and "ٱو." The categorization is summarized in **Table 1**.

3.2. Second categorization: letter shape

The second categorization is according to shape of the letter: this categorization splits the 28 letters into 15 branches based on the body of the letter rather than the dots on the letter (see **Table 2**). However, some increase the number of shapes to 18 shapes [22]. The first branch is made of four letters all very similar in shape: two of them are differentiated by one dot (one above the letter and below the letter), and the other two (one has two dots above it and one has three dots above it). The second branch has two letters very similar to each other: one can differentiate between them by the dot above one, while the other one has no dot; furthermore, the third branch and the fourth branch have the same idea similar in shape, yet one dot makes a difference. The same happens with the fifth and sixth branches. The seventh branch has three letters that are similar in shape: one without dot, one with dot above it, and one with dot below it. The eighth branch has two letters that are very similar in shape: one with one dot and the other with two dots. The ninth branch has two letters similar in shape: one with no dots and the other with three dots. The tenth has two letters: one with no dots and other with two dots. The 11th branch has two letters: one with no hamza and the other with hamza shape above the body of the letter. The 12th, 13th, 14, and 15th branches are not similar to each other nor to the rest of the letters.

One may add here a note about the shape of the letters; there are nine letters that have as part of them enclosed space that resembles a circle. These nine letters are (ض ص ظ ط ق ف ه و م). The enclosed circle property is an important aspect of the nine letters that will be used in the algorithm at a later stage.

3.3. Third categorization: letter location in a word

The third categorization of the Arabic alphabets is based on the location of the letter in a word. Generally, shapes of the Arabic alphabets change according to position of the letter in the word itself (beginning, middle, end); some letter can be connected (refers to the letter succeeding or preceding), and others cannot be connected. The shapes of the letters can be generated with ligature or character overlaps [23, 24]. When discussing the letters that start a word, these six letters when falling at the beginning of a word must stand alone; those letters are (ا د ذ ر ز و),

http://dx.doi.org/10.5772/intechopen.76944

Branch	Count	Shape of letter	Proportion
1	4	ن ث ت ب	
2	2	ض ص	
3	2	ظ ط	
4	2	غ ع	
5	2	ذ د	
6	2	ز ر	
7	3	ج خ ح	
8	2	ق ف	
9	2	ش س	
10	2	ة ه	
11	2	ك ل	
12	1	م	
13	1	ا	
14	1	ي	
15	1	و	

Table 2. Arabic alphabets (according to shape).

and the rest of the letters do change form as seen in **Table 3**. Using the same six letters in the middle or end of the word, these letters are only connected and they do not change form. All letters when used at the end of the word have two states: connected and stand-alone.

From the previous one can notice that six letters have special characteristics, namely, (ا, و, د, ذ, ر, ز). These characters when used in the beginning of a word must stand alone, and also when they end a group of characters, they must be followed by independent character. Hence, they only connect to the predecessor not the successor.

Letter	Beginning	Middle	End
أ	ا	ـا	ـا
ب	بـ	ـبـ	ـب
ت	تـ	ـتـ	ـت
ث	ثـ	ـثـ	ـث
ج	جـ	ـجـ	ـج
ح	حـ	ـحـ	ـح
خ	خـ	ـخـ	ـخ
د	د	ـد	ـد
ذ	ذ	ـذ	ـذ
ر	ر	ـر	ـر
ز	ز	ـز	ـز
س	سـ	ـسـ	ـس
ش	شـ	ـشـ	ـش
ص	صـ	ـصـ	ـص
ض	ضـ	ـضـ	ـض
ط	طـ	ـطـ	ـط
ظ	ظـ	ـظـ	ـظ
ع	عـ	ـعـ	ـع
غ	غـ	ـغـ	ـغ
ف	فـ	ـفـ	ـف
ق	قـ	ـقـ	ـق
ك	كـ	ـكـ	ـك
ل	لـ	ـلـ	ـل
م	مـ	ـمـ	ـم
ن	نـ	ـنـ	ـن
ه	هـ	ـهـ	ـه
و	و	ـو	ـو
ي	يـ	ـيـ	ـي

Table 3. Arabic alphabets: stand alone, beginning, middle, and end of a word.

The matrix, seen in **Table 4**, represents the different combination between all 28 letters. The first column in the matrix is the letter coming at the beginning of the order, and the first row is all the letters coming second. Each cell in the matrix shows the two letter shapes and how they change as the order differs. The highlighted letters are the previously mentioned six letters, namely, (ا,و, د, ذ, ر, ز), which if appears at the beginning of the word, then they stand alone. When these letters appear consecutively within a word, they will both be written as stand-alone independent letters.

3.4. Fourth categorization: letter proportion

To keep letters proportional to each other, two ways were used by calligraphers: rhombic dot and circles. Arabic calligraphy was used in mosques and castells as decoration since Islam

http://dx.doi.org/10.5772/intechopen.76944

forbids pictures and statues [22]. Hence, there is a need to decorate with words. Proportion is an essential part of the written word. The circle proportion was suggested by "Ibn Mugla," a well-known calligrapher from the eleventh century [25]. Three elements are the bases of proportion in Arabic calligraphy [26, 27]:

- The height of the *alif*, which is a straight and vertical stroke (3–12) rhombic dots.
- The width of the alif, (the rhombic dot) which is the square impression formed by pressing the tip of the calligrapher's reed pen to paper (see **Figures 1** and **3**).
- An imaginary circle with alif as its diameter, within which all Arabic letters could fit and be written (see **Figure 2**).

The circle is halved vertically and horizontally, with diameter equals the height of the first letter in Arabic alphabets called *alif*. Looking back at **Table 2** and **Figure 4** that represent the shape and form of the letter, the first branch, the shape of the letter, is in the lower half of the circle. The second branch, according to the circle, takes up the first quarter and the third

	ا	ب	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص	ض	ط	ظ	ع	غ	ف	ق	ك	ل	م	ن	ه	و	ي
ا	أا	اب	ات	اث	اج	اح	اخ	اد	اذ	ار	از	اس	اش	اص	اض	اط	اظ	اع	اغ	اف	اق	اك	ال	ام	ان	اه	او	أي
ب	با	بب	بت	بث	بج	بح	بخ	بد	بذ	بر	بز	بس	بش	بص	بض	بط	بظ	بع	بغ	بف	بق	بك	بل	بم	بن	به	بو	بي
ت	تا	تب	تت	تث	تج	تح	تخ	تد	تذ	تر	تز	تس	تش	تص	تض	تط	تظ	تع	تغ	تف	تق	تك	تل	تم	تن	ته	تو	تي
ث	ثا	ثب	ثت	ثث	ثج	ثح	ثخ	ثد	ثذ	ثر	ثز	ثس	ثش	ثص	ثض	ثط	ثظ	ثع	ثغ	ثف	ثق	ثك	ثل	ثم	ثن	ثه	ثو	ثي
ج	جا	جب	جت	جث	جج	جح	جخ	جد	جذ	جر	جز	جس	جش	جص	جض	جط	جظ	جع	جغ	جف	جق	جك	جل	جم	جن	جه	جو	جي
ح	حا	حب	حت	حث	حج	حح	حخ	حد	حذ	حر	حز	حس	حش	حص	حض	حط	حظ	حع	حغ	حف	حق	حك	حل	حم	حن	حه	حو	حي
خ	خا	خب	خت	خث	خج	خح	خخ	خد	خذ	خر	خز	خس	خش	خص	خض	خط	خظ	خع	خغ	خف	خق	خك	خل	خم	خن	خه	خو	خي
د	دا	دب	دت	دث	دج	دح	دخ	دد	دذ	در	دز	دس	دش	دص	دض	دط	دظ	دع	دغ	دف	دق	دك	دل	دم	دن	ده	دو	دي
ذ	ذا	ذب	ذت	ذث	ذج	ذح	ذخ	ذد	ذذ	ذر	ذز	ذس	ذش	ذص	ذض	ذط	ذظ	ذع	ذغ	ذف	ذق	ذك	ذل	ذم	ذن	ذه	ذو	ذي
ر	را	رب	رت	رث	رج	رح	رخ	رد	رذ	رر	رز	رس	رش	رص	رض	رط	رظ	رع	رغ	رف	رق	رك	رل	رم	رن	ره	رو	ري
ز	زا	زب	زت	زث	زج	زح	زخ	زد	زذ	زر	زز	زس	زش	زص	زض	زط	زظ	زع	زغ	زف	زق	زك	زل	زم	زن	زه	زو	زي
س	سا	سب	ست	سث	سج	سح	سخ	سد	سذ	سر	سز	سس	سش	سص	سض	سط	سظ	سع	سغ	سف	سق	سك	سل	سم	سن	سه	سو	سي
ش	شا	شب	شت	شث	شج	شح	شخ	شد	شذ	شر	شز	شس	شش	شص	شض	شط	شظ	شع	شغ	شف	شق	شك	شل	شم	شن	شه	شو	شي
ص	صا	صب	صت	صث	صج	صح	صخ	صد	صذ	صر	صز	صس	صش	صص	صض	صط	صظ	صع	صغ	صف	صق	صك	صل	صم	صن	صه	صو	صي
ض	ضا	ضب	ضت	ضث	ضج	ضح	ضخ	ضد	ضذ	ضر	ضز	ضس	ضش	ضص	ضض	ضط	ضظ	ضع	ضغ	ضف	ضق	ضك	ضل	ضم	ضن	ضه	ضو	ضي
ط	طا	طب	طت	طث	طج	طح	طخ	طد	طذ	طر	طز	طس	طش	طص	طض	طط	طظ	طع	طغ	طف	طق	طك	طل	طم	طن	طه	طو	طي
ظ	ظا	ظب	ظت	ظث	ظج	ظح	ظخ	ظد	ظذ	ظر	ظز	ظس	ظش	ظص	ظض	ظط	ظظ	ظع	ظغ	ظف	ظق	ظك	ظل	ظم	ظن	ظه	ظو	ظي
ع	عا	عب	عت	عث	عج	عح	عخ	عد	عذ	عر	عز	عس	عش	عص	عض	عط	عظ	عع	عغ	عف	عق	عك	عل	عم	عن	عه	عو	عي
غ	غا	غب	غت	غث	غج	غح	غخ	غد	غذ	غر	غز	غس	غش	غص	غض	غط	غظ	غع	غغ	غف	غق	غك	غل	غم	غن	غه	غو	غي
ف	فا	فب	فت	فث	فج	فح	فخ	فد	فذ	فر	فز	فس	فش	فص	فض	فط	فظ	فع	فغ	فف	فق	فك	فل	فم	فن	فه	فو	في
ق	قا	قب	قت	قث	قج	قح	قخ	قد	قذ	قر	قز	قس	قش	قص	قض	قط	قظ	قع	قغ	قف	قق	قك	قل	قم	قن	قه	قو	قي
ك	كا	كب	كت	كث	كج	كح	كخ	كد	كذ	كر	كز	كس	كش	كص	كض	كط	كظ	كع	كغ	كف	كق	كك	كل	كم	كن	كه	كو	كي
ل	لا	لب	لت	لث	لج	لح	لخ	لد	لذ	لر	لز	لس	لش	لص	لض	لط	لظ	لع	لغ	لف	لق	لك	لل	لم	لن	له	لو	لي
م	ما	مب	مت	مث	مج	مح	مخ	مد	مذ	مر	مز	مس	مش	مص	مض	مط	مظ	مع	مغ	مف	مق	مك	مل	مم	من	مه	مو	مي
ن	نا	نب	نت	نث	نج	نح	نخ	ند	نذ	نر	نز	نس	نش	نص	نض	نط	نظ	نع	نغ	نف	نق	نك	نل	نم	نن	نه	نو	ني
ه	ها	هب	هت	هث	هج	هح	هخ	هد	هذ	هر	هز	هس	هش	هص	هض	هط	هظ	هع	هغ	هف	هق	هك	هل	هم	هن	هه	هو	هي
و	وا	وب	وت	وث	وج	وح	وخ	ود	وذ	ور	وز	وس	وش	وص	وض	وط	وظ	وع	وغ	وف	وق	وك	ول	وم	ون	وه	وو	وي
ي	يا	يب	يت	يث	يج	يح	يخ	يد	يذ	ير	يز	يس	يش	يص	يض	يط	يظ	يع	يغ	يف	يق	يك	يل	يم	ين	يه	يو	يي

Table 4. Matrix of the different combinations for all 28 letters.

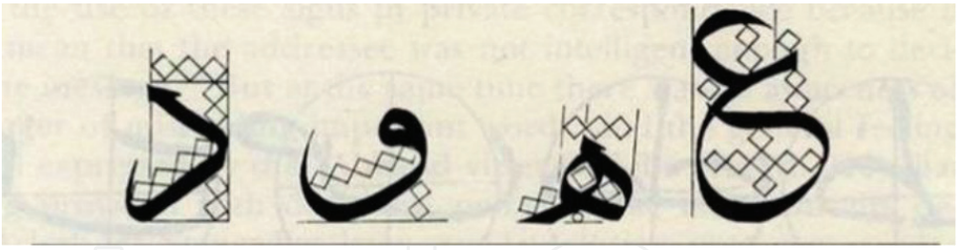


Figure 1. Example of measuring the letter by using rhomboid dots [28].

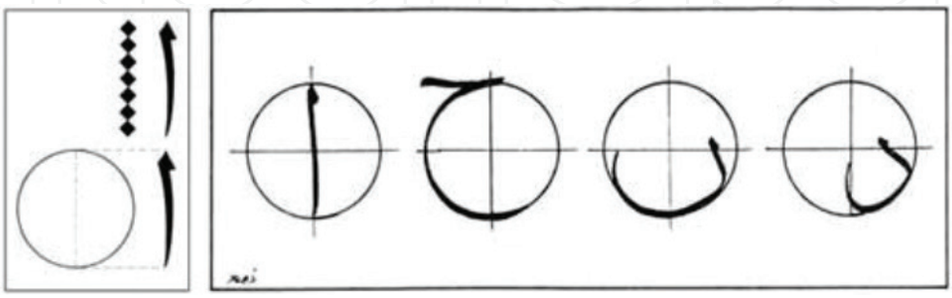


Figure 2. Example of measuring the letter by using circle [28].



Figure 3. The rhombic dot as a guide to proportions [22].

quarter. In the third branch, the letter falls in the first quarter of the circle with the upper half diameter aligned with the half alif of the letter. In the fourth branch, the letter lies on the left half of the circle. In the fifth branch, two letters are located in the fourth quarter of the circle. In the sixth branch, two letters also fall in the fourth quarter of the circle. In the seventh branch, the letter lies on the left half of the circle. In the eighth branch, two letters are both parted in the first and second quarter of the circle with a circular part above the horizontal diameter. In the tenth branch, the letter takes the first quarter of the circle. The eleventh branch takes the second and third quarter of the circle. In the twelfth branch, the letter is at the center of the circle and uses the bottom half the alif. The thirteen branch is the alif itself, which is the diameter of the circle. The fourteenth branch is taking the third quarter of the circle. The fifteenth branch takes the first and fourth quarter of the circle.

http://dx.doi.org/10.5772/intechopen.76944

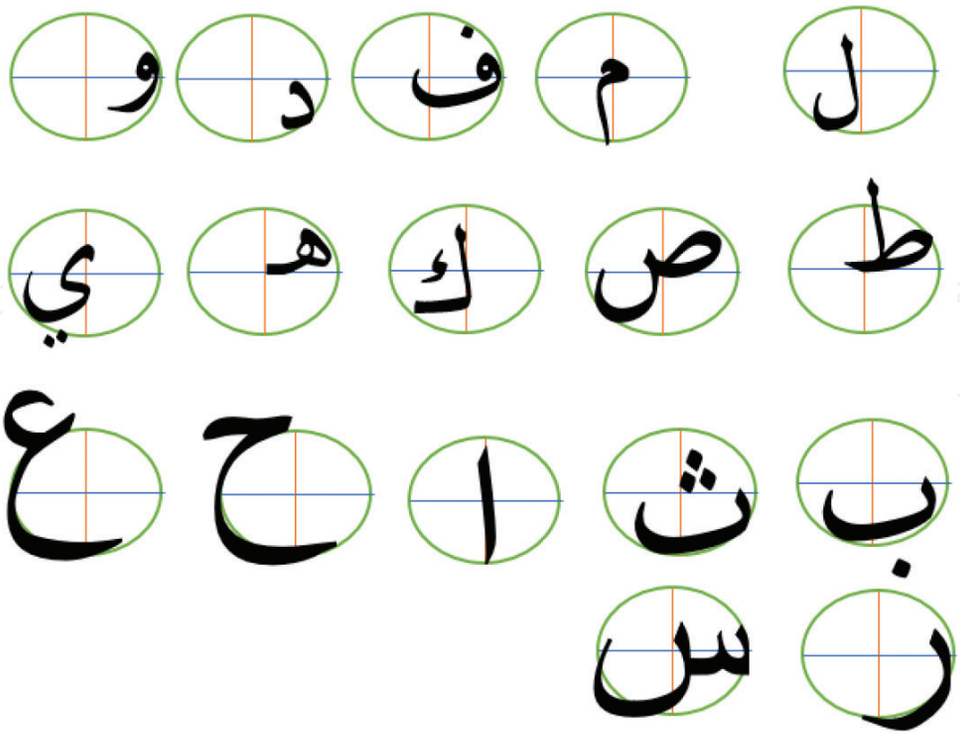


Figure 4. Proportions in Arabic calligraphy.

One can conclude by studying the second categorization and the proportion categorization through the following:

- First, the second branch and ninth branch both (four letters) take same area of the circle.
- Second, the third branch and tenth branch (four letters) both use the first quarter of the circle.
- Third, the fourth branch and seventh branch use the left edge of the circle, yet the differentiation between the two is that one letter is written from right to left and one letter is written from left to right as seen in **Figure 5**.
- Fourth, the fifth and sixth branches (four letters) use the fourth quarter of the circle.

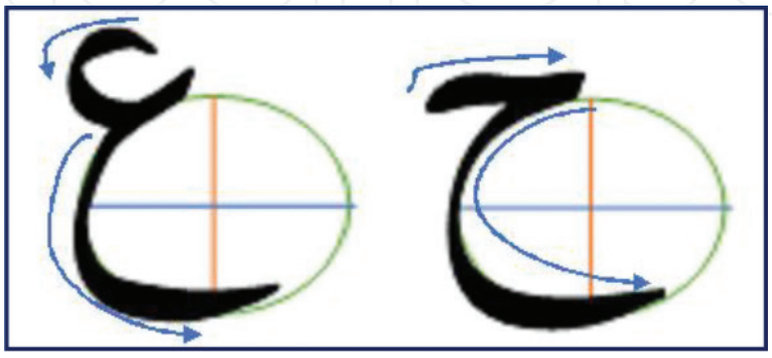


Figure 5. Direction of writing with two circle edge letters.

Hence, give an insight to further classify the letters and manage them into groups. The previous sections explained in details the four categorizations used in the proposed algorithm. Each categorization was an essential in the building blocks and rules of the algorithm.

3.5. Findings of the four categorizations

Based on the four categorizations explained above, a tree of rules can be built as seen in **Figure 6**. The rule tree has five branches: the first branch is for Arabic alphabets that contain *no dots*. The second branch is for letters with *one dot*. The third branch is for letters with *two dots*. The fourth branch includes all letters with *three dots*. The fifth branch is for letters with *hamza*.

For the first branch including 12 letters and in order to distinguish among the letters, the fourth categorization logic was used. Each letter in this branch was located in the quarters of the circle suggested in the fourth categorization. Two letters used the same quarters (ص س); both fall in the first and third quarters of the imaginary circle, which explain the fourth categorization. Still, letter (ص) has an enclosed space, while letter (س) has no enclosed space. Hence, differentiating between the two letters depends on the enclosed space. The enclosed space property is explained previously in the third categorization. The edge of the circle from the fourth categorization was used to differentiate between the ten letters and the letters (ع ح). Furthermore, to differentiate between the two letters, the direction of writing was used. The direction of writing was explained in **Figure 5** previously.

The second branch consisting of all letters with *one dot* included ten letters. The branch spliced further to *dot below* and *dot above* the body of the letter. Again, in this branch the imaginary circle from the fourth category was used. The location of the letters according to the quarters of the imaginary circle was used as seen in **Figure 6**. Also, the distinguishing feature of the letter falling on the edge of the imaginary circle is used, and the property of writing direction seen in **Figure 5** is also used.

The third branch consisting of all letters with *two dots* included three letters. The branch spliced further to *dots below* and *dots above* the body of the letter. There is only one letter in all the alphabets that has two dots below it (ي). And, there are three letters with two dots above the letter body. To distinguish between the three letters, the imaginary circle from the fourth categorization and again none shared the same quarters of the circle.

The fourth branch included all letters with *three dots*. The branch included only two letters both have the dots above their body. Hence, the quarters of the imaginary circle were used to distinguish between them.

The fifth branch is the *hamza* (ء) branch which included three letters, and the distinguishing features were the imaginary circle quarters: letter (ك) is in the second and third quarters of the circle, letter (ؤ) is in the first and fourth quarters, and letter (أ) falls on the diameter of the circle.

The *hamza* (ء) can be seen on top of the letters (أ, ؤ, ك); the hamza sometimes is considered an independent letter when used in some words like (ءالع) and is used as part of the letter in other words. The hamza is a distinguishing character between the two letters (ك, ل). Hence, it is treated as semi-letter and is not listed in the alphabets.

http://dx.doi.org/10.5772/intechopen.76944

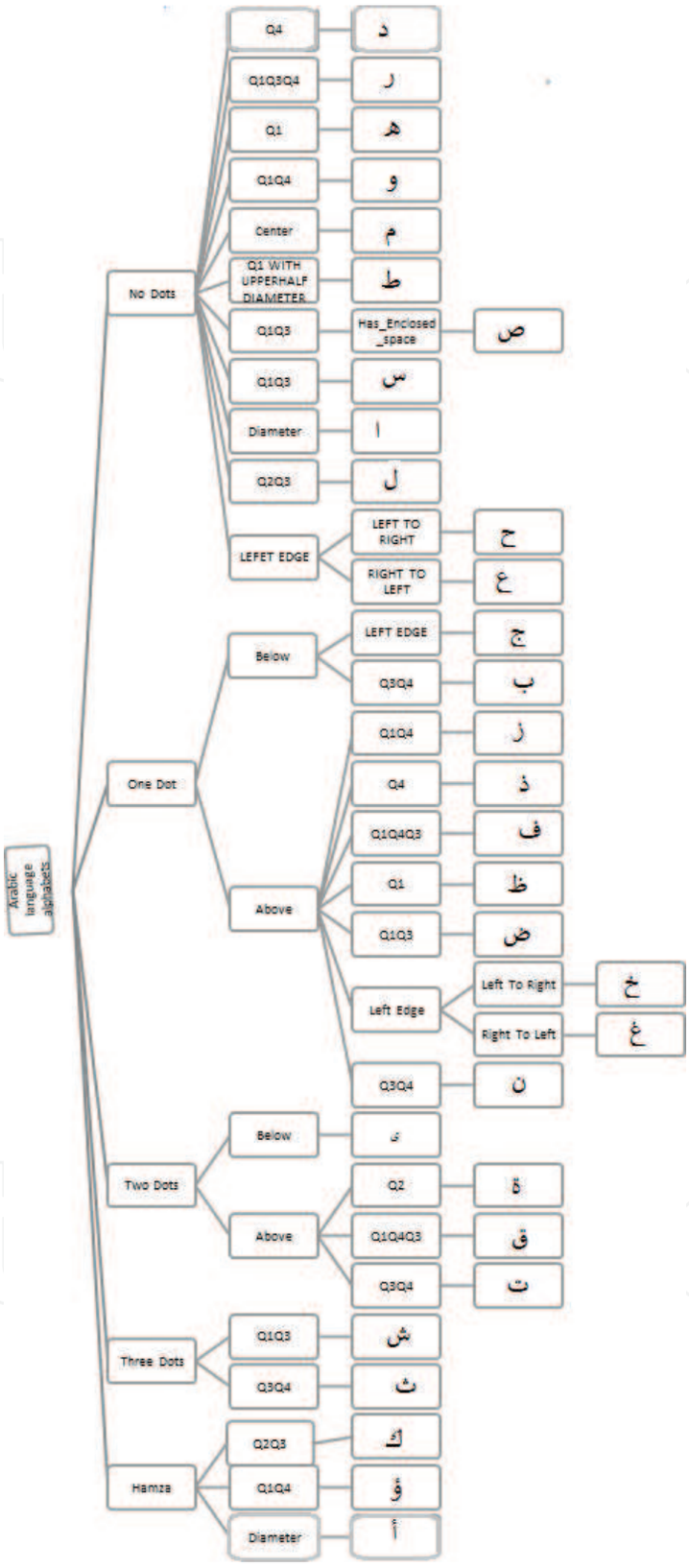


Figure 6. Categorization of the tree drawn based on the four categorizations.

4. Suggested algorithm

After studying all the previously mentioned categorizations, one can reach the conclusion that a deterministic algorithm can predict the character being drawn based on the following matrix in **Figure 7** and along with the matrix is the suggested algorithm in **Figure 8**, hence reducing the determination of a letter to 38 rules.

The suggested algorithm shown in **Figure 7** is composed of five major if-then statements which are based on the first categorization explained above and later summarized in **Figure 7**. The first if-then statement runs from line 1 to 12 in **Figure 8**. The if-then statement really deals with all cases of the letters which have no dots, and their location in the circle is mentioned in proportion categorization. The enclosed space property mentioned earlier was very important to distinguish letter "س" and letter "ص"; both letters fall in the same location in the circle Q1 and Q2, yet the latter has an enclosed space. Also, notice that both "ح" and "ع" have the same properties, yet to differentiate them, the direction of writing is used [9].

The second major if-then statement started at line 13 and dealt with letters with one dot. As shown in **Figure 6**, the dot can be either above or below the body of the letter, i.e., both letters "ن" and "ب" fall in the lower part of the circle quarters 3 and 4. Yet, to make the differentiation, the dot was essential here, the latter had the dot below as seen in lines 23 and 20 in **Figure 8**. Also, line 21, in the same figure, dealt with two letters that are essential falling in the same location, and both had the dot above them, distinguished by the writing direction left to right or right to left. The algorithm can be improved by eliminating line 24, hence reducing the number of rules to 37, since one can use one statement.

The third major if-then statement starts at line 25 and ends at line 31 in **Figure 8**. The if-then statement deals with letters that have two dots according to the first categorization and the matrix seen in **Figure 7**. The nested if-statement deals with the two dots whether above or below the letter. Three letters have two dots above them, yet their location on the circle is very distinguishable; hence, using the attribute "enclosed space" was not necessary. Furthermore, one can eliminate line 31 since this is the only letter in the alphabet that has two dots below it. Still, for the purpose of clarity, line 31 was left in the suggested algorithm. If line 31 was eliminated, the number of rules will be again reduced to 36 rules.

The fourth major if-then statement deals with letters that have three dots; there are only two of them. Both letters can be distinguished based on their respective location according to the proportion categorization. Again, line 34 can be eliminated but was left for the purpose of clarity. If line 34 was eliminated, the number of rules will be again reduced to 35 rules.

The last major if-then statement starts at line 35; the statement deals with case of "hamza." The hamza is an essential part in letter "ك" and is used with other letters like "أ ؤ." The three letters are distinguished by their location within the circle according to proportion categorization. Line 38 can be eliminated but was left to clarify the algorithm, hence reducing the number of rules to 34.

http://dx.doi.org/10.5772/intechopen.76944

		Dot position	Circle quarters	Pen movement	Has enclosed space		Results Letter
First branch	**No dots**					12	
	No dots		Q4				د
	No dots		Q1 Q4 Q3				ر
	No dots		Q1		x		ه
	No dots		Q1Q4		x		و
	No dots		center		x		م
	No dots		Q1 with upper half diameter		x		ط
	No dots		Q1, Q3		x		ص
	No dots		Q1, Q3				س
	No dots		diameter				ا
	No dots		Q2 Q3				ل
	No dots		Left edge	Left to right			ع
	No dots		Left edge	Right to left			ح
Second branch	**One dot**	**above**				**10**	
	One dot	above	Q4 & Q1				ز
	One dot	above	Q4				ذ
	One dot	above	Q1, Q4, Q3		x		ف
	One dot	above	Q1		x		ظ
	One dot	above	Q1, Q3		x		ض
	One dot	above	Left edge	Left to right			خ
	One dot	above	Left edge	Right to left			غ
	One dot	above	Q3 Q4				ن
	One dot	**below**	Left edge				ج
	One dot	below	Q3 Q4				ب
Third branch	**Two dots**					**3**	
	Two dots	**Above**	Q2		x		ة
	Two dots	**Above**	Q1, Q4, Q3		x		ق
	Two dots	**Above**	Q3 Q4				ت
	Two dots	**below**	Q3				ي
Fourth branch	**Three dots**		**Q1, Q3**			2	ش
			Q3 Q4				ث
Fifth branch	**With hamza**					3	
			Q2 Q3				ك
			Q1Q4		x		ؤ
			diameter				أ

Figure 7. The property rules to define each letter in the Arabic alphabets.

```
If Number_of_Dots=0
        If circle_location=Q4  then "د"
        If circle_location= Q1&Q3  then "س"
        If circle_location= Q2&Q3  then "ل"
        If circle_location=Q1&Q3&Q4  then "ر "
        If circle_location=Q1 & Has_Enclosed_space=true then "ه"
        If circle_location= Q1&Q4 & Has_Enclosed_space=true then "و"
        If circle_location= Center & Has_Enclosed_space=true then "م"
        If circle_location=Q1 & with upper half diameter then "ط"
        If circle_location= Q1&Q3 & Has_Enclosed_space=true then "ص"
        If circle_location=diameter then "ا"
        If circle_location= Left_edge then if writing_Direction= left_to_right then "ح"
else "ع"
Else If Number of dots=1
        IF dot_location=Above_letter then
                If circle_location= Q1&Q4 then "ز"
                else If circle_location= Q4 then "ذ"
                else If circle_location= Q1&Q3&Q4 then "ف"
                else If circle_location= Q1 then "ظ"
                else If circle_location= Q1&Q3 then "ض"
                else If circle_location= Q3&Q4 then "ن"
                else If circle_location= Left_edge then if writing_Direction=
left_to_right then "خ" else "غ"
        Elseif dot_location =Below_letter
                If circle_location= Q3&Q4 then "ب"
                else If circle_location= Left_edge then "ج"
Else If Number_of_dots=2
        IF dot_location=Above_letter then
                If circle_location= Q2 then "ة"
                else If circle_location= Q1&Q3&Q4 then "ق"
                else If circle_location= Q3&Q4 then "ت"
        Elseif dot_location =Below_letter
                If circle_location= Q3 then "ي"
Else If Number_of Dots=3
        If circle_location= Q1&Q3 then "ش"
        else If circle_location= Q3&Q4 then "ث"
Else (Hamza case)
        If circle_location= Q2&Q3 then "ك"
        else If circle_location= Q1&Q4 then "ؤ"
        else If circle_location= diameter then "أ"
```

Figure 8. Determine_Character (input:one_character).

5. Conclusion

The proposed algorithm stems from many needs that are more apparent today. First, there is a rise in the use of handheld devices, which use character recognition methods that serves mainly Latin-based languages. Arabic language is one of the top five languages spoken in the world. Arabic is used by more than 422 million native and non-native speakers in the

world. Arabic language is different from other languages: Arabic is a cursive language, written from right to left, and letters change shape according to the position of the word. Hence, there is a dire need to develop an algorithm for character recognition for the Arabic language. However, many algorithms have used artificial intelligent methods to recognize characters that make their algorithms non-deterministic, while the proposed algorithm is deterministic. This research presents four categorization methods that will be employed to develop a deterministic algorithm method of categorization. The first categorization method depends on the number of dots used with each letter. The second categorization method depends on the shape of the letter, with classification to the letters. The third categorization is presented with the shape of the letter as used in the beginning, middle, and end of the word. The fourth categorization method relays on the proportion method, which is a method used in Arabic calligraphy that is based on rhombic dot. Then, the research suggested a deterministic algorithm composed of 34 rules that can predict the character based on the use of all of categorizations as attributes assembled in a matrix for this purpose [29].

The proposed algorithm is only one piece in the whole puzzle. There are many parts that need to be developed. One major part is the input section of the algorithm. Such part needs to exist in order for the puzzle to be complete. The input section needs to parse the word into segments that can detect the shape of the letters, the dots, and the hamza. Furthermore, this research will be a building block for further research and development.

Biography

Evon M. O. Abu-Taieh, PhD, is an associate professor and an author/editor of four scholar books, contributed in more than eight scholar books. She has more than 40 published papers. She was previously the acting dean in the University of Jordan (Aqaba) for 3 years. Dr. Evon is an editorial board member in five renounced journals. She has more than 29 years of experience in education, computers, aviation, transport, AI, ciphering, routing algorithms, compression algorithms, multimedia, and simulation.

Auhood Abdullah Alfaries, PhD, is as assistant professor in the IT Department in King Saud University (KSU). Dr. Auhood received her PhD degree in Semantic Web and Web Services from the School of Computing and Information Systems, Brunel University, UK. She held a number of IT-related academic and administrative positions both in KSU and princess Noura bint Abdulrahman University (PNU). She has experience in quality and program accreditation by serving in a number of quality-related roles since 2011. Auhood is associated with a number of important bodies such as an associate of the UK Higher Education Academy, a member of the Institute of Electrical and Electronics Engineers (IEEE), and a member of the Saudi Computer Society. She is also an ABET program evaluator. She participated as a conference and journal reviewer and a member of a number of national and international workshops and conference program committees. She has served as the vice dean and dean of E-Learning and Distance Learning Deanship in KSU and then in PNU for 2 years and has also served as the assistant general director and then the director for the General Directorate of Information and Communications Technology (ITC) in PNU. Currently, she is the dean of the

College of Computer and Information Sciences. Auhood's research interest includes semantic web, ontology engineering, natural language processing, machine learning, and cloud computing. She is a member of IWAN Research Group.

Dr. Nabeel Mohammed Zanoon received his PhD degree in Computer Systems Engineering, from the South-West State University, Kursk, Russia, in 2011. He is a faculty member with Al-Balqa' Applied University since 2011, where he is currently an assistant professor and the head of the Department of Applied Sciences as well as the director of the ICDL Computer Centre and Cisco Academy Branch of Aqaba University College. He has published several researches in several areas: security of e-banking, algorithm scheduling in grid and cloud, meta-grammar, hardware and architecture, fiber optical, and mobile ad hoc networks.

Issam Hamad Al Hadid is a lecturer at the University of Jordan. He completed his PhD degree at the University of Banking and Financial Sciences (Jordan) in 2010, obtained his MSc degree in Computer Science at Amman Arab University (Jordan) in 2005, and earned his BSc degree in Computer Science at Al-Zaytoonah University (Jordan) in 2002. He has published many research papers in different fields of science in refereed journal and international conference proceedings. His researches focus on self-healing architecture; also, his research interests include AI, knowledge-based systems, security systems, compression techniques and algorithms, and information retrieval.

Alia Abu-Tayeh earned her PhD degree in 1995. She is a lecturer in the University of Jordan (Aqaba) and an ex-lecturer in King Hussein University. She published many scientific articles in renowned journals. Her interest ranges from linguistics to applied mathematics in computer and languages.

Author details

Evon Abu-Taieh[1]*, Auhood Alfaries[2], Nabeel Zanoon[3], Issam H. Al Hadid[1] and Alia M. Abu-Tayeh[1]

*Address all correspondence to: abutaieh@gmail.com

1 The University of Jordan, Aqaba, Jordan

2 King Saud University, Riyadh, Saudi Arabia

3 Al-Balqa' Applied University, Aqaba, Jordan

References

[1] Sharma MK, Samanta D. Word prediction system for text entry in Hindi. 13, 2, article 8 (June 2014). 2014. 29 pages. DOI=http://dx.doi.org/10.1145/2617590

[2] Xue N. Automatic inference of the temporal location of situations in Chinese text. Conference on Empirical Methods in Natural Language Processing (EMNLP '08). Stroudsburg, PA, USA: Association for Computational Linguistics. 2008. pp. 707-714

[3] Xue N, Converse SP. Combining classifiers for Chinese word segmentation. first SIGHAN workshop on Chinese language processing. 18. Stroudsburg, PA, USA: Association for Computational Linguistics. 2002. pp. 1-7. DOI: dx.doi.org/10.3115/1118824.1118839

[4] Xue N, Chen J, Palmer M. Aligning Features with Sense Distinction Dimensions. COLING/ACL on Main conference poster sessions (COLING-ACL '06). Stroudsburg, PA, USA: Association for Computational Linguistics; 2006. pp. 921-928

[5] Green S, Manning CD. Better Arabic parsing: Baselines, evaluations, and analysis. 23rd International Conference on Computational Linguistics. Beijing, China. 2010. pp. 394-402

[6] Marton Y, Habash N, Rambow O. Dependency parsing of modern standard Arabic with lexical and inflectional features. Computational Linguistics. 2013;**39**(1):161-194. DOI: 10.1162/COLI_a_00138

[7] Parvez MT, Mahmoud SA. Offline Arabic handwritten text recognition: A Survey. ACM Computer Survey. 2013;**45**(2):35. DOI: dx.doi.org/10.1145/2431211.2431222

[8] Subramanian K, Prasad R, Natarajan P. Robust named entity detection using an Arabic offline handwriting recognition system. The Third Workshop on Analytics for Noisy Unstructured Text Data (AND '09). New York, NY, USA: ACM. 2009. pp. 63-68. DOI: dx.doi.org/10.1145/1568296.1568308

[9] Fard MM, Moghadam M, Bidgoli BM, Hussain M. Persian on-line handwritten character recognition by RCE spatio-temporal neural network. 5th international conference on Soft computing as transdisciplinary science and technology (CSTST '08). New York, NY, USA: ACM. 2008. pp. 90-94. DOI: dx.doi.org/10.1145/1456223.1456246

[10] Abu-Taieh E. Artificial neural networks: Enhanced back propagation in character recognition. In: Proceedings of Information Technology and Organizations: Trends, Issues, Challenges and Solutions. Volume 1. USA: IGI Global; 2003. pp. 263

[11] Aljarrah I, Al-Khalee O, Mhaidat K, Alrefai M, Alzu'bi A, Rabab'ah M. Automated system for Arabic optical character recognition. In: The 3rd International Conference on Information and Communication Systems (ICICS '12). 5, p. 6. New York, NY, USA: ACM. 2012. DOI: http://dx.doi.org/10.1145/2222444.2222449

[12] Ali A, Sagheer AM. Design and implementation of secure chatting application with end to end encryption. Journal of Engineering and Applied Sciences. 2017;**12**:156-160

[13] Supriana I, Nasution A. Arabic Character Recognition System Development. Juhana Salim MI, editor. Procedia Technology. 2013;**11**:334-341. DOI: 10.1016/j.protcy.2013.12.199

[14] Sarfraz M, Ahmed MJ, Ghazi SA. Saudi Arabian License Plate Recognition System. 2003 International Conference on Geometric Modeling and Graphics (GMAG'03). London,UK: IEEE Computer Society Press. 2003. pp. 36-41. DOI: 10.1109/GMAG.2003.1219663

[15] Izakian H, Monadjemi SA, Ladani BT, Zamanifar K. Multi-font Farsi/Arabic isolated character recognition using chain codes. International Journal of Computer and Information Engineering. 2008;**2**(7):67-70. Retrieved from http://waset.org/publications/301/multi-font-farsi-arabic-isolated-character-recognition-using-chain-codes

[16] Abandah GA, Khedher, Mohammed Z. Analysis of handwritten Arabic letters using selected feature extraction techniques. International Journal of Computer Processing of Languages. 2009;**22**(1):1-25. DOI: doi.org/10.1142/S1793840609001981

[17] Al-Taani AT, Al-Haj S. Recognition of on-line Arabic handwritten characters using structural features. Journal of Pattern Recognition Research. 2010:23-37

[18] Kapogiannopoulos G, Kalouptsidis N. A fast high precision algorithm for the estimation of skew angle using moments. IASTED, International Conference Signal Processing, Pattern Recognition and Application. Crete, Grece: SPPRA. 2002. pp. 275-279

[19] Zidouri A. On multiple typeface Arabic script recognition. Research Journal of Applied Sciences Engineering and Technology. 2010;**2**(5):428-435. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.473.4044&rep=rep1&type=pdf

[20] Amin A. Recognition of printed Arabic text based on global features and decision tree learning techniques. Pattern Recognition. 2000;**33**:1309-1323. DOI: 10.1016/S0031-3203(99)00114-4

[21] Cowell J, Hussain F. Extracting features from Arabic characters. IASTED International Conference on COMPUTER GRAPHICS AND IMAGING (pp. 201-206). Honolulu, Hawai, USA: ACTA Press. 2001

[22] Sood S, Fitzgerald E. Arabic Script and the Art of Calligraphy. Retrieved from The Metropolitan Museum of Art. 2012: www.metmuseum.org/learn/for-educators/publications-for-educators/art-of-the-islamic-world/unit-two/proportional-scripts

[23] Slimane F, Kanoun S, Alim AM, Henneber J, Ingold R. Comparison of global and cascading recognition systems applied to multi-font Arabic text. In Proceedings of the 10th ACM symposium on Document engineering (DocEng '10). ACM, New York, NY, USA, 161-164. 2010a. DOI=http://dx.doi.org/10.1145/1860559.1860591

[24] Slimane F, Kanoun S, Alimi AM, Hennebert J, Ingold R. Comparison of global and cascading recognition systems applied to multi-font Arabic text. In: Of the 10th ACM symposium on Document engineering (DocEng '10). New York, NY, USA: ACM. 2010b. pp. 161-164. DOI: http://dx.doi.org/10.1145/1860559.1860591

[25] Sourdel D. Ibn Muḳla. In: Bearman P, Bianquis Th, Bosworth CE, van Donzel E, Heinrichs WP, editors. Encyclopaedia of Islam. 2nd ed. Edited by: Consulted online on 01 July 2017 http://dx.doi.org/10.1163/1573-3912_islam_SIM_3306. First published online: 2012. First print edition: ISBN: 9789004161214, 1960-2007

[26] Grabar O. The Mediation of Ornament. The A.W. Mellon Lectures in the Fine Arts. Princeton, NJ: Princeton University Press; 1992. p. 38

[27] Osborn J. Narratives of Arabic script: Calligraphic design and modern spaces. The Journal of the Design Studies Forum. 2009;**1**(3). DOI: doi.org/10.1080/17547075.2009.11643292

[28] Schimmel A. (10 14). Styles of Calligraphy, Islamic Art & Architecture. Retrieved from Islamic Art and Architecture. 2011: http://islamic-arts.org/2011/styles-of-calligraphy

[29] Shah NN, Bhatt N, Ganatra A. A unique word prediction system for text entry in Hindi. In: Second International Conference on Information and Communication Technology for Competitive Strategies (p. Article No. 118). ACM. 2016. DOI: 10.1145/2905055.2905334

IntechOpen

IntechOpen

IntechOpen