

Freeman Chain Code Contour Processing For Handwritten Isolated Arabic Characters Recognition

Assistance Prof. Majida Ali Abed

College of Computers Sciences & Mathematics – University of Tikrit – Iraq

Abstract:

The Arabic characters are being used in various languages. Recognition of handwritten characters of Arabic alphabet set is an important area of research. The work on recognition of Handwritten Isolated Arabic Characters (HIAC) is still an open research problem and it has numerous applications. Machine reading of optically scanned text is usually called optical Character recognition (OCR). In this paper we present an OCR for Handwritten Isolated Arabic Characters. Basic Characters are recognized by Freeman chain code. The proposed method obtains the contour of the character and then generates a Freeman Chain Code for contoured character. The obtained Freeman chain code is unique for handwritten isolated characters. The present method is based on template matching to recognize handwritten printed isolated Arabic characters. It was trained and validated on 200 images (consisting of 7400 Arabic characters written by 18 writers). Our experimentation observed the overall recognition rate is 95%. It is carried out in order to improve the rate and the performance of an Arabic handwritten word recognition system. The proposed system has been implemented and tested on Matlab R2008b environment

Keywords: Arabic characters, Handwritten Isolated Arabic Characters (HIAC), Freeman Chain Code (FCC), Contoured character.

سلسلة فريمان لمعالجة الأحرف العربية المعزولة المكتوبة يدويا وتمييزها
الأستاذ المساعد ماجدة علي عبد
كلية علوم الحاسوب والرياضيات- جامعة تكريت

المخلص

اللغة العربية هي مصدر اللغات السامية ، وحروفها تستخدم في مختلف اللغات . التمييز للحروف المكتوبة يدويا من مجموعة حروف اللغة العربية هو من أهم مجالات البحث. العمل على تمييز الحروف العربية المعزولة و المكتوبة يدويا (HIAC) لا يزال قيد البحث وله العديد من التطبيقات. في هذا البحث، نستعرض عملية التمييز للحروف العربية المعزولة والمكتوبة بخط اليد. الهدف من البحث هو تمييز الحروف العربية المعزولة المكتوبة يدويا باستخدام سلسلة فريمان. الطريقة المقترحة تتضمن الحصول على مخططات الحروف ومن ثم توليد سلسلة فريمان الخاصة بها. النتيجة الحاصلة من سلسلة فريمان تكون وحيدة للحرف العربي المطلوب تمييزه. تستند الطريقة المقترحة على مطابقة القوالب للتمييز. تم التدريب والتحقق من صحة الطريقة المقترحة على 200 صورة (التي تتكون من 7400 الحروف العربية والتي كتبت بواسطة 18 شخص). أظهرت النتائج العملية ان معدل التمييز هو 95 ٪ . يتم تنفيذه من أجل تحسين معدل وأداء أنظمة التمييز للحروف العربية المكتوبة يدويا. تم تطبيق هذا النظام المقترح واختباره على البيئة الماتلاب R2008b

الكلمات المفتاحية: الحروف العربية ، الحروف العربية المعزولة المكتوبة يدويا ، سلسلة فريمان ، مخطط الحرف

1. Introduction

The goal of Character Recognition (CR) is to transform the input data, such that handwritten character ,text written document ,text typed on document online and offline writing into a digital format. In pattern recognition field, language recognition is considered as one of the most complicated problem in artificial intelligence field. Character Recognition (CR) automation occupies a big and intensive research zone of pattern recognition research area. CR automation means translating images of characters into an editable text, in other words , it represents an attempt to simulate the human reading process [5][7]. The CR can be an on-line or off-line type. If the CR system has the ability to trace the points generated by moving a special pen on a special screen, the system belongs to the on-line type, while the system belongs to the off-line type when it accepts only the pre-scanned text images to perform the recognition function. The off-line type deals with printed and handwritten texts, while the on-line type deals with handwritten texts only .The researches on the recognition of the handwritten writing tend progressively toward the conception of efficient systems of recognition [9]. The handwritten recognition is generally considered as a difficult task because of the differences of handwritings and of the irregularity of the writing of the same writer. The complex nature of the shape of its characters and their resemblance give back irregular the tracing of a written Arabic word. The Arabic handwriting is of semi-cursive nature, indeed, an Arabic word is a sequence of disconnected entities called pseudo-words. Several works were dedicated to the handwritten Arabic characters recognition. Some classifiers were based on Artificial Neural Networks (ANN) [1], on Hidden Markov Models (HMM) [2], on Freeman chain code [3][13] or on the k -nearest neighbors [4]. One of the ways to represent an image efficiently is by using Freeman chain code because it is one of simple image representing ways and represent an image based on its boundary .This paper is organized in 6 sections, Section 2 is Arabic writing characteristics include overview of Arabic characters and the Handwritten Isolated Arabic Characters (HIAC) . Section 3 describes the Freeman chain code and its applications .Section 4 describes the proposed system used in this research. Section 5 describes the experimental results. Finally, Section 6 states the main conclusions

2. Arabic Writing Characteristics

The Arabic language has a very rich vocabulary. More than 300 million people in the world speak and write in Arabic, and over one billion people use the Arabic language in several religion-related activities. Arabic alphabet contains basically 28 characters are written from right to left. Fifteen of them have dots and 13 are without dots. Dots above and below the characters Ten of them have one dot (Baa,Jeem,Khaa,Thal,Zai,Dhad,Dha,Gahin,Faa,Noon) three have two dots (Ta,Qaf) and two have three dots (Tha,Sheen)Each character can take from two to five different shapes, according to its position (beginning, middle, end or isolated)[6] .This is one reason makes Arabic recognition complex. The other reason is the similarities among the different characters and the differences among the same character, for example character Baa, Taa, Thaa they have similar body shape but differ in number and position of dots also Jeem, Haa, Khaa also Faa and Qaf. The Arabic handwriting is also rich in diacritics that allow to differentiate the notion of character and the notion of character. Several Arabic characters are based on the same character and differ only by the number and the position of their diacritics. Following the position that they take care in the word : isolated, in the beginning, in the middle or at the end, some Arabic characters change of shapes. For example, the letter [ain] has the four following shapes respectively: ع , ا , ا , ا . Arabic text is written from right to left and is always cursive. Table (1) shows the Arabic characters set in the four different shapes. Some characters that differ only by the number or location of dots. Arabic characters do not have fixed width or fixed size, even in printed form .Arabic writing is known to be cursive even in printed form. However, it differs from cursive handwriting of English in that some characters can be connected from one side only. Out of the 28 basic Arabic characters, six

can be connected from the right side only while the other 22 can be connected from both sides. These six characters are: dal (د), raa (ر), waw (و), alef (ا), thal (ذ), and zay (ز). These six characters have only two forms, the stand-alone form and the final form. Whereas the rest of the characters can appear in any of four forms: the beginning, the middle, the final, and the stand-alone form. Consequently, an Arabic word may consist of one or more sub-words. A sub-word can be defined as the basic stand-alone pictorial block of the Arabic writing. Researchers found many other characteristics of Arabic characters. These include:[8][10] :

- Arabic characters are written from right to left.
- Some of Arabic characters are not connectable with the succeeding Character . Those characters are (ا ذ ر ز و), which are fundamental for Defining the notion of sub-word in Arabic writing, because the user is Constrained to cut the word into several entities each time he Encounters one of these characters .
- Arabic characters contain many fonts and shapes (up to four different shapes) depending on there relative position in the sub-word, Table (1)
- Characters of the same font have different sizes (i.e. characters may have different width even though they have the same font and point size). Hence, adding to the complexity of the segmentation algorithm.

Table (1): The Different Forms of Arabic Characters

Character	Isolated character	Initial	Middle	Final
Alef	ا			ا
Ba	ب	ب	ب	ب
Ta	ت	ت	ت	ت
Tha	ث	ث	ث	ث
Jeem	ج	ج	ج	ج
H'a	ح	ح	ح	ح
Kha	خ	خ	خ	خ
Dal	د			د
Thal	ذ			ذ
Ra	ر			ر
Zai	ز			ز
Seen	س	س	س	س
Sheen	ش	ش	ش	ش
Sad	ص	ص	ص	ص
Dhad	ض	ض	ض	ض
Tta	ط	ط	ط	ط
Dha	ظ	ظ	ظ	ظ
Ain	ع	ع	ع	ع
Gahin	غ	غ	غ	غ
Fa	ف	ف	ف	ف
Qaf	ق	ق	ق	ق
Kaf	ك	ك	ك	ك
Lam	ل	ل	ل	ل
Meem	م	م	م	م
Noon	ن	ن	ن	ن
Ha	ه	ه	ه	ه
Waw	و			و
Ya	ي	ي	ي	ي

3. Freeman Chain Codes

Freeman Chain codes FCC are one of the shape representations which are used to represent a boundary by a connected sequence of straight line segments of specified length and direction the boundary of an active region is represented by a freeman chain code which starting from any boundary pixel lists the direction required to move from one pixel to the next anti-clockwise round the boundary until the starting pixel is reached. The direction to the next boundary pixel is represented by an ASCII codes for the 1 to 8 .The first approaches for representing digital curves using chain code was introduced by Freeman in 1961 and it is known as Freeman chain code (FCC).The representation of the chain is based on 4-connectivity or 8-connectivity of the segments [11].The direction of each segment is coded by using a numbering scheme as shown in Figure (1)

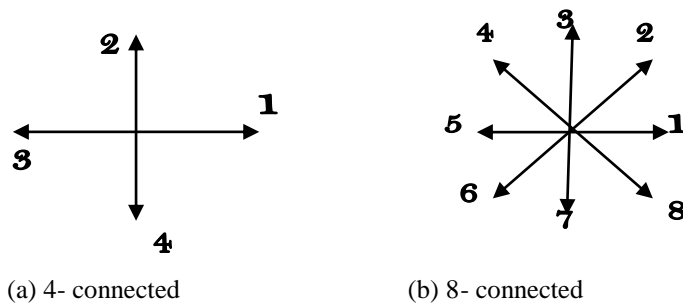


Figure (1): Boundary Freeman Chain Codes

A coding scheme for line structure must satisfy three objectives [12]:

1. It must faithfully preserve the information of interest;
2. It must permit compact storage and convenient for display
3. It must facilitate any required processing.

A Freeman chain code can be generated by a boundary of an object in a clockwise direction and assigning a direction to the segments connecting. There must be an adjoining boundary pixel at one of the eight locations surrounding the current boundary pixel. By looking at each of the eight adjoining pixels, we will find at least one that is also a boundary pixel. Depending on which one it is, we assign a numeric code of between 1 and 8 as already shown in Figure (1). If the pixel found is located at the right of the current location, a code "1" is assigned, the pixel in the upper right direction, a code "2" is assigned, the pixel in the lower right direction, a code "8" is assigned, and the pixel in the upper left direction, a code "4" is assigned. The process of locating the next boundary pixel and assigning a code is repeated until we came back to our first location or boundary pixel. Freeman Chain codes have been claimed as one of the techniques that are able to recognize characters and digits successfully . This is because of several advantages listed by :

- 1- Representation of a binary object is that the chain codes are a compact representation of a binary object.
- 2- Translation invariant representation of a binary object. Due to that, it is easier to compare objects using this technique.
- 3- Freeman Chain codes is a complete representation of an object or curve. This means that we can compute any shape feature from the Freeman chain codes. The normalized Freeman chain code is obtained by transforming it to a two dimensions matrix. The first row of this matrix contains the value of the chain code, and the second row contains the frequency of occurrence of that value. For example, if the Freeman chain code of two given characters are:

66882266667788812224, 66882666678881111223 then it can be converted into the following form of a 2×9 matrix:

6	8	2	7	1	4
6	5	5	2	1	1

6	8	2	7	1	3
6	5	3	1	4	1

we remove all values whose frequencies are 1 For instance, in the above example, the chain code will be reduced to:

6	8	2	7
6	5	5	2

6	8	2	1
6	5	3	4

The process of removing the less-frequent digits can be continued. For instance in our test, the frequencies less than or equal to five were deleted. Again in the resulted chain code the frequency of each remained digit is summed. Then to transform the chain code matrix to a normalized chain code with length of 10, the relative frequency of each digit is computed using .

$$F_i^n = \frac{F_i}{\sum F_i} \times 10$$

Where F_i^n is the normalized frequency and F_i is the frequency of each digit in the chain code respectively. In the above example we will obtain:

6	8	2	7
3.33	2.77	2.77	1.11
6	8	2	1
3.33	2.77	1.66	2.22

Then the normalized frequency would be rounded of to nearest decimal which in turn would be concatenated to generate the length 10 chain code:

6668882227

6668882211

4. Proposed System

The general steps of the proposed system which used in this paper Figure(2) show the block diagram of this system . Characters tested in our proposed system have been written 26 times by 18 writers to obtain a database of about 7400 samples.

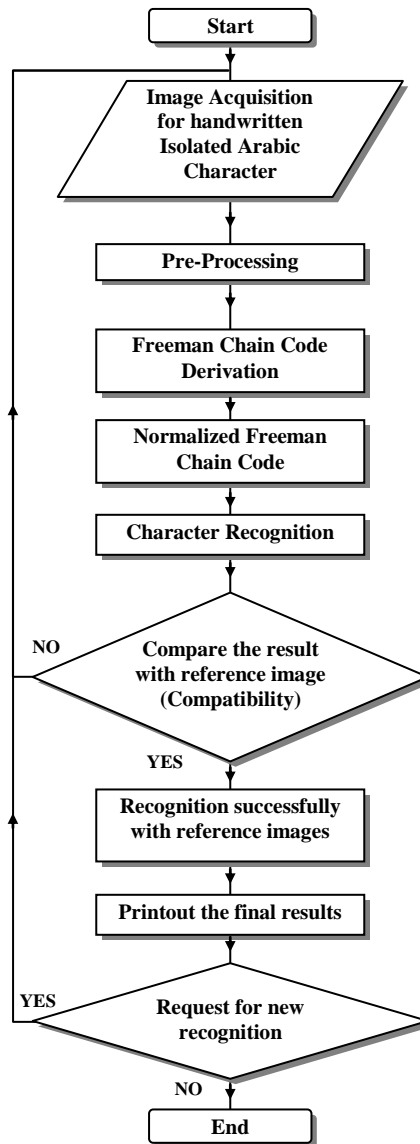


Figure (2): Block diagram of proposed system

5. Experimental Results

In Arabic language, there are ten different characters as shown in Figure (3) that only differ in number of and the position of dots between their and the rest characters because this similarity in shape and to simplify our experimental we neglect these dots and take main body for characters like (ب , ت , ث) we take " ب " without dot , (ح , خ , ج) we take " ح " without dots and so as for other characters [(د , ذ) , (ر , ز) , (س , ش) , (ص , ض) , (ط , ظ) , (ع , غ)] characters

without the dots The eighteen Arabic characters as shown in the Figure (4), was written by 18 writers from different educational backgrounds by hands. Input image consists of the isolated handwritten Arabic character to proposed system , this image of the handwritten Arabic character which different from one to another. After obtaining the image of the character, the coordinates of the Boundary pixels are obtained which is sent to obtain the Freeman chain code. The obtained Freeman chain code is then normalized as shown in the Figure (5) . This normalized Freeman chain code is compared with the existing database. If match occurs then respective character is identified. Else there is the message shown print error. Matlab R2008b environment was used to implement the proposed system and database of all characters along with normalized Freeman chain code is used. After getting the normalized Freeman chain code for handwritten isolated Arabic characters, the codes were compared with existing database to recognize the character which gives the 95% accur

ع	ظ	ض	ش	ز	ذ	ح	ج	ث	ت
ع	ط	ص	س	ر	د	ح	ب		

Figure (3) : Arabic characters with different dots and similar shapes

Seen	Ra	Dal	Ha	Ba	Alif
Kaf	Qaf	Fa	Ain	Tah	Sad
Ya	Waw	Haa	Naun	Meem	Lam

Figure (4) : Examples of (18) characters of handwritten isolated Arabic used for examination

Isolated Arabic characters	Handwritten isolated image	Image	Normalized Freeman chain code
ا			1133555777
ب			7776665544
ج			1166633888
د			1188877755
ر			8887776655
س			5566777733
ص			2277766655
ط			7772288555
ع			6668882211
ف			5554433777
ق			5555447722
ك			6668885555
ل			7777665544
م			2288777666
ن			8877553322
هـ			4422666655
و			5544888866
ي			6668885544

Figure (5) : Results Obtained From Implementation of the isolated handwritten Arabic characters

For example we take the isolated handwritten Arabic character " ع " written by several writers as shown in the Figure (6) After obtaining the image of the character " ع " , the coordinates of the Boundary pixels are obtained which is sent to obtain the Freeman chain code. The obtained Freeman chain code is then normalized as shown in the Figure (7) .

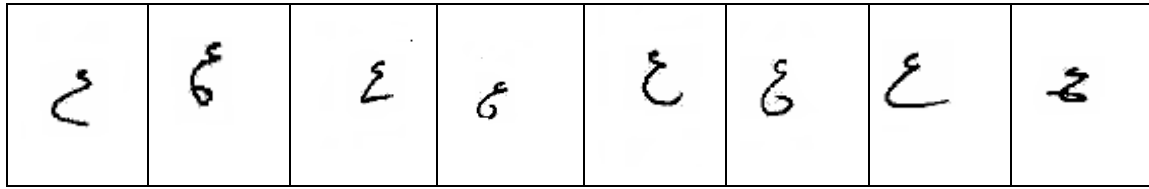


Figure (6):The " Ain " character written by several writers



















Isolated Arabic character "ع"	Handwritten isolated image	Image	Normalized Freeman chain code
ع			6826628811
ع			6826668822
ع			6826881155
ع			6826668222
ع			6826688822
ع			6826666822
ع			6826882255
ع			6826885525
ع			6826688882

Figure (7) : Results Obtained From Implementation Of The character "ع"

6. Conclusion

In this paper, an attempt has been made to develop a simple and effective method for the recognition of isolated Arabic characters . The present method is based on template matching to recognize isolated Arabic characters uses Freeman chain code based on contour tracing method. The proposed method obtains the contour of the character and then generates a freeman chain code for contoured character. The obtained Freeman chain code is unique for isolated Arabic characters . Experimental results using 18 characters written by several writers show that the accuracy of the proposed system is 95%.The relatively high accuracy of the method when it is tested on isolated Arabic characters.

7. References

1. G. Looney 1997 "Pattern Recognition using Neural Networks " New York Oxford Carl University Press,
2. G. Abandah .and Khedher,2004 , "Printed and Handwritten Arabic Optical Character Recognition" Initial Study, A report on research supported by the Higher council of Science and Technology ,Jordan.
3. Gonzales, R. C and Woods, R. E. 2002. "Digital ImageProcessing".2nd Ed .Upper Saddle River, N. J.: Prentice-Hall,
4. H. AL-Yousefi and S. S. Udpa,1988 "Recognition of handwritten Arabic Characters," in Proc. SPIE 32nd Ann. Int. Tech. Symp. Opt. Optoelectric Applied Sci. Eng .Aug.
5. H. El-Abed and V. Märgner,2007"Comparison of different preprocessing and feature extraction methods for offline recognition of handwritten Arabic words", in Proc. Int'l Conf. Document Analysis and Recognition (ICDAR'07), pp. 974–978.
6. H. Izakian, S. A. Monadjemi, B. Tork Ladani, and K. Zamanifar. 2008, " Multi-Font Farsi/Arabic Isolated Character Recognition Using Chain Codes". Department of computer engineering, faculty of engineering, University of Isfahan, Iran.
7. I. Siddiqi and N. Vincent,2008 "Combining Global and Local Features for Writer Identification", In Proc of 11th Int'l Conference on Frontiers in Handwriting Recognition (ICFHR), Canada
8. Khorsheed, M. S.,2002 "Off-Line Arabic characters recognition " Pattern Analysis and Applications.
- 9.M. Bulacu and L. Schomaker:2007" Text-Independent Writer Identification and Verification Using Textural and Allographic Features", IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 701-717.
- 10 .Nor Amizam Jusoh and Jasni Mohamad Zain. 2009 "Application of Freeman Chain Codes: An Alternative Recognition Technique for Malaysian Car Plates". University Malaysia Pahang.
- 11.S. Madhvanath, G. Kim and V. Govindaraju. 1999 " Chain Code Contour Processing for Handwritten Word Recognition". IEE Transactions on Pattern Analysis and Machine Intelligence
- 12.Yang Mingqiang, Kpalma Kidiyo and Ronsin Joseph. 2008." A Survey of Shape Feature Extraction Techniques ". Shandong University, Jinan
- 13.Yong Kui Liu and Borut Zalik. 2004." An efficient chain code with Huffman coding." Dalian Nationalities University.

