Feature Extraction Phase for Offline Arabic Handwritten Character Recognition

Dr. Rawia I. O. Ahmed
Department of Computer Science
College of Community Female
University of Ha'il
Saudi Arabia

Dr. Mohamed E. M. Musa
College of Computer Science and Information Technology Sudan
University of Science and Technology Khartoum, Sudan

Abstract: —In this paper we reviewed the importance issues of the optical character recognition, gives more emphases for OCR and its phases. We discuss the main characteristics of Arabic language, furthermore it focused on Feature Extraction phase of the character recognition system. We described and implemented the algorithm of Feature Extraction based on Freeman Chain codes. The algorithms are tested using 47,988 isolated character sample taken from SUST/ ALT dataset and achieved better results. The Feature Extraction phase developed by using MATLAB software.

Keywords: optical character recognition; offline recognition; online recognition; handwritten, preprocessing; Feature Extraction: Freeman Chain codes.

1. INTRODUCTION

Over the past three decades, many studies have been concerned with the recognition of Arabic words. Offline handwritten Arabic characters' recognition have received more attention in these studies, because of the need to Arabic document digitalization.

In this paper, Feature Extraction system for an isolated Arabic handwritten are design and tested by using SUST/ ALT dataset. it's a new dataset developed and published by SUST/ALT (Sudan University of Science and Technology-Arabic Language Technology group) group. It contains numerals datasets, isolated Arabic character datasets and Arabic names datasets[1]. 40 common Arabic (especially in Sudan) males and females' name[2]. Each form written by one writer resulting 40,000 sample. it used for researching purpose.

The rest of the paper is organized as follows: Then the main characteristics of Arabic language are discussed in Section 2. Then the concepts and phases of OCR system are described in Section 3. The proposed Feature Extraction phase based on Freeman Chain codes discussed in Section 4. conclusion and future work are presented in Section 5.

2. The MAIN CHARACTRERISTICS OF ARABIC LANGUAGE

Many studies have been conducted on recognition of Chinese, Japanese and Latin languages, but few were done on Arabic handwritten recognition[3]. One of the main reasons for this is that characteristics of Arabic language do not allow direct implementation of many algorithms used in other languages. The characteristics of Arabic language can be summarized as follows:

- Arabic language is represented in 28 characters and appears in different four shapes isolated, initial, medium or final.
- Arabic language is written from right to left, rather than from left to right this is useful for human reader rather than for the computer. As seen in Figure.1

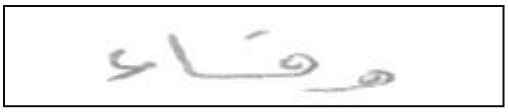

Figure.1 Arabic Word Written from Right to Left

- Arabic characters of a word are connected a long baseline, and character position above and below the baseline.
- Some Arabic character have the same shape and differ in the number of dots by which it will be identified, for example characters ث ,ت ,ب have the same shape but differ in number of dots, one dot in character Baa, two dots in character Taa, and three dots in character Thaa.
- Some Arabic character have the same shape and differ in the position of dots by which it will be identified, for example characters ن ,ب the two characters have the same shape and identify with one dot, but they differ in position of dot one is above the baseline (character Noon), and other under the base line (character Baa), this differentiation can change the meaning of a word.
- The width and high of Arabic characters are differ from one character to another.
- The shape of Arabic character varies per writer.
- Arabic writing is cursive, most of Arabic characters are connected from two sides; right and left, only six characters are connected from right side only, as shown in Figure.2.

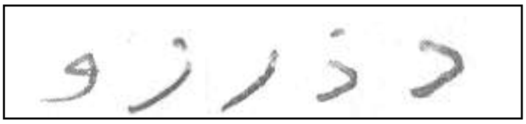

Figure.2 Arabic characters which can be connected from right to left

- Moreover, Arabic language has some diacritics called Tashkeel. The names of these Tashkeel: Fatha, Dhamma, Kasra, Sukun, Shadda, Fathatain, Kasratain, Dhammatain also combination of them are possible. These diacritics may change the meaning of specific word, for example: when we put Fatha diacritic on the word “حر” it became “حَر” which meaning “hot weather”, when we put dhamma diacritics on the same word, it became “حُر” which meaning “free”.

- Some Arabic words consists of more than one sub-words. A sub-word is the basic standalone pictorial block of the Arabic writing [4]. A brief details of Arabic handwritten characteristic were reviewed by Lorigo [5].

3. THE OPTICAL CHARACTER RECOGNITION SYSTEM

The Optical Character Recognition (OCR) is one of important tasks in computer area. It has many definitions, OCR defined as a process that attempts to turn a paper document into a fully editable form, which can be used in word processing and other applications as if it had been typed through the keyboard[3]. Also OCR was defined by Srihari et al. as the task of transforming text represented in the special form of graphical marks into its symbolic representation[6] .

The recognition of handwritten can be applied in many areas such as names of persons, companies, organizations, newspapers, letters, archiving and retrieving texts, proteins and genes in the molecular biology context, journals, books, bank chequs, personal signatures and digital recognition, etc.[7]. A recognition system can be either online or offline[8]. It is online if the data being captured during the writing process. It always captured by special pen on an electronic interface. Online recognition has several interesting characteristics: firstly, recognition is performed on one dimensional rather than two dimensional images, secondly, the writing line is represented by a sequence of dots which its location is a function of time[3].A recognition system is offline if its data scanned by scanner after writing process is over, such as any images scanned in by a scanner. In this case, only the image of the handwriting is available.

When we compared online handwriting recognition systems with offline systems, we found that offline systems are considered more difficult than online systems. This difficulty due to several reasons, out of which online handwriting recognition depends on temporal information, which facilitate the recognition system, but the temporal information is lacked in offline handwriting, it depends on passive images stored in files. This lemma makes offline systems less accurate than online systems. Furthermore, offline systems are more complex than online systems, because they depend on human writing which had more feature and characteristic specially for Arabic language.

There is no great variation exists between phases of the online and offline handwriting recognition systems.

the general phases of OCR systems are: data capture, preprocessing, segmentation, feature extraction, classification and, post processing as shown in Figure .3

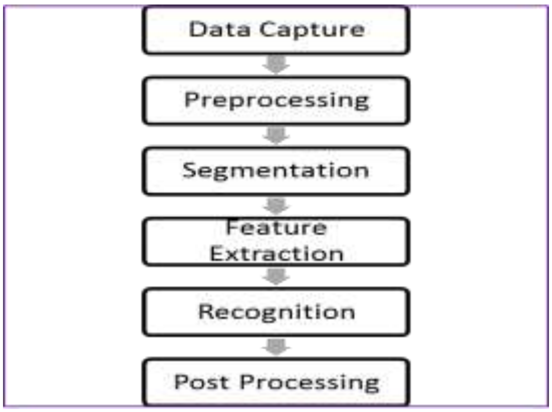

Figure.3 General recognition system phases

4. THE PROPOSED FEATURES EXTRACTION PHASE

Prior to the features extraction phase preprocessing phase must be done. Pre-processing of the handwritten character image is an important factor, to simplify the task of recognition. Usually several operations can be performed in this phase. Since in SUST isolated characters dataset, some preprocessing method are done during the development stage[1],minimal number of preprocessing processes are used in this work. An image file of isolated handwritten character will first be introduced to the system as gray scale bmp image. Then obtained images are binarized to be in digital form. When the study focus on characters' body only, dots is removed from some characters. Thinning is very important process in OCR, therefore we applied it the binary images. The next sub sections give a brief detail of these phases were discussed in our previous work [9].

The next step is to extract the useful features that will have used in classification phase. Many researchers agree that feature extraction phase play an important role in a handwriting recognition systems.

A human being can differentiate between various objects by observing their colors, shapes and attributes. To simulate this intelligence idea into a computer system, we need to implement geometrical and topological representation methods to help the system in recognizing the shapes of objects.

After studying several features methods, we found that geometrical and topological representation methods are reverent methods to recognize the isolated handwritten character.

To represent each image in a feature vector form, a mathematical model with a finite number of parameters is required. But unfortunately, they are no reasonable mathematical model currently exists.

In this section, we describe an efficient method for extracting features from handwritten Arabic character body using freeman chain code

4.1 Freeman Chain Code Algorithm

FCC can be 4-connectivity or 8-connectivity. It traced the boundary of an image in a clockwise or anticlockwise directions. The main weakness of 4-connectivity is that we be unable to find the transverse points[10]. These points are very valuable in image recognition. So, to overcome the weakness of 4-connectivity we use 8-connectivity FCC. In

8-connectivity each code can be considered as the angular directions (8 directions).

In our experimentation, we labeled the direction from 0 to 7 and decide to trace the image in a clockwise direction, it is commonly in many research and suitable for Arabic characters, because it written from right to left [11-13].

The big challenge of FCC is how to find the starting pixel and the directions of image traverses. We have produced different chain code for the same character, if we start with different starting points or traverse on different directions. Therefore, consistency plays an important factor to overcome the variations of the chain code for the same character and preserve the success of the algorithm.

As mention before, we obtained the boundary of the character image by traverse the FCC in a clockwise direction. We assigned numbers 0 to 7 to each direction.

The character images are binarized to 1's and 0's pixels in preprocessing phase. Instead of storing the absolute location of each 1 pixel, we stored it is direction from its previously coded neighbor. A neighbor is any of the adjacent pixels in the 3x 3 pixels' neighborhood surround that current pixel as shown in Figure .4

The literature of Freeman chain code was introduced in [14]. To define the starting point, we move from the top of character body image to the bottom raw by raw to find first nonzero pixel. Furthermore, we assume that this first pixel has one neighbor. When we find the first nonzero pixel we defined it as starting point of the chain code and stored it in chain code list. In some cases, the first nonzero pixel has two neighbors. This case holds when the character body written as loop, for example, see the "Haa" character in Figure in Figure .4, If this case arises, we assume the starting chain code is zero. After finding the start point of the chain code we traverse to the next neighbor pixel in the image of character body.

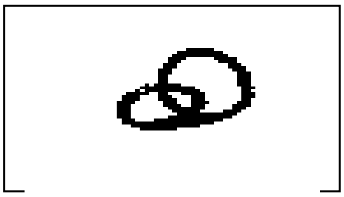

Figure.4 "Haa" Character Written as Loop

In fact, there must be at least one neighbor boundary pixel at one of the eight locations surrounding the current starting boundary pixel (note each location has one value from 0 to 7 marked per the chain code direction). Again, the starting point can have more than one neighbor. In this case the chain code direction plays and important factor, to determine which neighbor will be chosen. In our experimentation, we chose a clockwise direction. When the neighbor found, we stored it in chain code list. Then finding the next neighbor.

The process of finding the next neighbor continuous until we reached to the starting point. This algorithm followed to find the FCC are summarized in Figure 5.

```
    Input: Thinning binary character body image.
    Output: 8-connective Freeman chain code
    While there are still images to be traversed
    Begin
    move from the top of character body image to the bottom, raw
by raw to find first nonzero pixel
If one nonzero pixel then
Begin
Assume that it is a starting pixel
    Stored it is direction in chain code list
    End
    Else
    Stored in chain code list 0 value as starting pixel
    End
      From starting pixel to end pixel do
        Assign 0-7 values to the eight directions
        Travels the neighbor pixel in clockwise direction
         Find and store the direction code of the neighbor pixel in
chain code list
        Move to next position
      End
    End
```

Figure.5 The Algorithm for Generating

8- Connective Freeman Chain Code

4.2 Normalized Freeman Chain Code

In this section, the feature vector for each character body image is obtained by applying 8-connective chain code. Then the two-dimensional (2D) matrix is converting to one-dimensional (1D) contain 10 digits by applying the proposed normalized chain code algorithm14]. Samples of results obtained from implementation these algorithms on character "ف" samples are shown in Table 1.

Table .1 The Results Obtained from Implementation of

Normalized Chain Code on Character "ف"

No	Original Image	Image After Preprocessing Operations	Normalized Chain Code
1			5 5 6 6 7 0 0 1 2 3
2			5 5 6 7 7 0 0 1 2 3
3			5 5 6 7 1 1 0 2 3 4
4			5 6 7 0 0 1 2 3 4 4

The proposed algorithm worked well and gives 99.23% result, this makes the algorithm more effective only about 193 of the samples their chain code be zero, this due to unconnected properties of these samples.

In making comparisons with the existing work, it is difficult to compare with work in[17] since those authors used other dataset and they have chosen 200 images so the proposed algorithm cannot be implemented on the same data.

5. CONCLUSION AND FUTURE WORK

In this paper, we present a review about optical character recognition system and its phases. Also, we list the characteristics of Arabic language, and focused in one of important phases in recognition systems which is feature extraction. Moreover, we described and implemented feature extraction algorithms based on free man chain code on isolated Arabian characters. In the future, we will have implemented this algorithm to Arabic word handwritten.

6. REFERENCES

[1] Musa, M.E. Arabic handwritten datasets for pattern recognition and machine learning. in 2011 5th International Conference on Application of Information and Communication Technologies (AICT).

[2] Wahby, T.M., I.M. Osman, and M.E. Musa, On Finding the Best Number of States for a HMM-Based Offline Arabic Word Recognition System. 2011.

[3] Mori, S., H. Nishida, and H. Yamada, Optical character recognition. 1999: John Wiley & Sons, Inc.

[4] Abed, M.A., Freeman chain code contour processing for handwritten isolated Arabic characters recognition. Alyrmook University Magazine,Baghdad, 2012.

[5] Lorigo, L.M. and V. Govindaraju, Offline Arabic handwriting recognition: a survey. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2006.

[6] Srihari, S.N., A. Shekhawat, and S.W. Lam, *Optical character recognition (OCR).* 2003.

[7] Amin, A., Off-line Arabic character recognition: the state of the art. Pattern recognition, 1998.

[8] Khorsheed, M.S., Off-line Arabic character recognition– a review. Pattern analysis & applications, 2002.

[9] Rawia, Ahmed; Mohammed, Musa, Preprocessing Phase for Offline Arabic Handwritten Character Recognition, International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064 Volume 5 Issue 11 ,2016.

[10] Yang, M., K. Kpalma, and J. Ronsin, A survey of shape feature extraction techniques. Pattern recognition, 2008.

[11] Sampath, A., C. Tripti, and V. Govindaru, Freeman code based online handwritten character recognition for Malayalam using backpropagation neural networks. International journal on Advanced computing, 2012.

[12] Omer, M.A.H. and S.L. Ma. Online Arabic handwriting character recognition using matching algorithm. in Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on. 2010. IEEE.

[13] Izakian, H., et al., Multi-font Farsi/Arabic isolated character recognition using chain codes. World Academy of Science, Engineering and Technology, 2008.

[14] Ahmed, Rawia Ibrahim Omer. Offline Recognition System for Isolated Arabic Handwritten Characters using Hidden Markov Models. Diss. Sudan University of Science and Technology, 2017.

[15] Abed, M.A., Freeman chain code contour processing for handwritten isolated Arabic characters recognition. Alyrmook University Magazine,Baghdad, 2012.