International Journal of Computer Applications Technology and Research
Volume 5 –Issue 12, 760-763, 2016, ISSN:-2319–8656

Preprocessing Phase for Offline Arabic Handwritten Character Recognition

Rawia I. O. Ahmed
College of Computer Science and Information Technology Sudan University of Science and Technology Khartoum,
Sudan

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 the pre-processing phase of the character recognition system. We described and implemented the algorithms of binarization, dots removing and thinning which will be used for feature extraction phase. The algorithms are tested using 47,988 isolated character sample taken from SUST/ ALT dataset and achieved better results. The pre-processing phase developed by using MATLAB software.

Keywords: optical character recognition; offline recognition; online recognition; handwritten, preprocessing.

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, preprocessing 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: the concepts of OCR approaches are described in Section 2. Then the main characteristics of Arabic language are discussed in Section 3. Then phases involved in OCR system are discussed as general in Section 4, These phases are: preprocessing, segmentation, feature extraction and classification. The proposed Preprocessing phase discussed in Section 5. conclusion and future work are presented in Section 6.

2. THE OPTICAL CHARACTER RECOGNITION APPROACHES

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[4] .

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.[5]. A recognition system can be either online or offline[6]. 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. Table.1 summarizes the differences between online and offline recognition systems.

Error! Reference source not found. **The differences between online and offline recognition systems**

Criteria of recognition	on-line recognition system	off-line recognition system
Data Capture	during the writing process	scanned in by a scanner or camera.
Data Type	Temporal information	Not temporal information
Accuracy	More accurate	Less accurate
Complicity	Less complex	More complex

3. 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[7]. 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.
- Arabic characters of a word are connected a long baseline, and character position above and below the baseline. As seen in Figure.1 which illustrates the word "samah"; the character "Seen" appear above the baseline, while character "Meem" appears below the baseline.

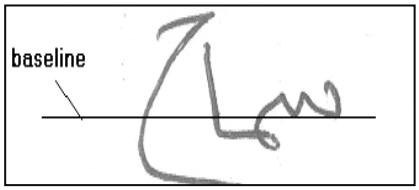

Figure.1 Arabic characters of a word are connected a long baseline

- Some Arabic character have the same shape and differ in the number of dots by which it will be identified, for example characters ب, ت, ث 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 ب, ن 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 [8]. A brief details of Arabic handwritten characteristic were reviewed by Lorigo [9].

4. RECOGNITION SYSTEM PHASES

OCR systems either can be online or offline. There is no variation between phases of both systems. It depends on lexicon nature, and the recognition approach. The lexicon is a key point to the success of any OCR system. As the size of lexicon grows, the recognition efforts and the complexity are increased. So, the general phases of OCR can be described by six phases[10]. First, the data can be captured by several ways depending on the system (online or offline). Then the scanned text image may need to be passed through several preprocessing steps. After the preprocessing process, the text image may need to be segmented into lines, words, pieces of words, characters or pieces of character. To facilitate the recognition phase, useful features are extracted from the text image, then the valuable classifier methods were used to build the model. Finally, to improve the recognition rate, some post-processing operations may be applied on the model. But post processing phase can scarcely apply and limited to few systems as in [11, 12]. So, the general phases of OCR systems are: data capture, preprocessing, segmentation, feature extraction, classification and, post processing as shown in Figure .3

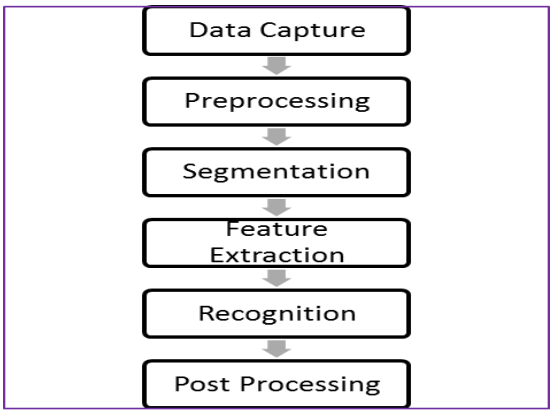

Figure.3 General recognition system phases

5. THE PROPOSED PREPROCESSING 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 cans 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 operations.

5.1 Binarization

Binarization operation attempts to converted the gray scale image into a binary image based on threshold. So, the bitmap images are threshold and converted into 1s and 0s forms. Two types of thresholding are existing. These types are global and local thresholding. In global thresholding, threshold selection leads to a single threshold value for the entire image[13].This value is often based on an estimation of the background intensity level of the using an intensity histogram. In local thresholding different values are used for each pixel according to the local area information[14].

Since the proposed system implemented on simple isolated handwritten character images, where the characters can be to distinguish into background and foreground pixels, the global thresholding methods are sufficient for this type of images. Therefore we use Otsu's method[15]. It applied on dataset to converted the image into 1s (background) and 0s(foreground).

5.2 Dots Removing

When the goal of the system is to design an offline handwritten recognition system to deal with isolated Arabic handwritten character body written by multiple writers, before recognition all dots need to be removed. Some of Arabic characters, may have three, two or one dots such as " ب, ت, ث " characters or may be without any dot such as "ح , د , و" characters. We removed dots from the (Baa, Faa, Noon and Yaa) characters images to extract the character's bodies only. When applying this operation some of data are loss during this process (about 601 samples), Table.2 illustrated the accuracy rate of this process.

Table .2 The accuracy rate of dots removing

character	Accuracy rate
Baa	90.28%
Fah	94.33%
Noon	95.39%
Yaa	77.38%

Samples are loss due to two reasons. The first reason, is the writing style of the writers, for examples some writers connected dots with the main body of characters "Baa, Fah and Yaa", or writing dot inside character "Noon" as shown in Figure.4a & Figure.4b.

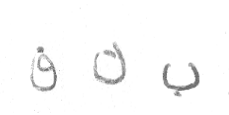

Figure .4a

The original images

Figure .4b

Images after dots removing process

The second reason, due to unclear samples from the original dataset, Figure.5a displays some unclear samples from the original dataset, and Figure.5b displays the same samples after removing dots.

Figure .5a

Unclear samples from the original dataset

Figure .5b

The same samples after dots removing.

5.3 Thinning

Finally, the character body image is thinned by T.Y. Zhang and C. Y. Suen algorithm [16] to maintained the connectivity of skeleton and extracted the edges which is be the input to the feature extraction phase. Figure.6a displays an image for character "Baa" body before thinning, and Figure.6b displays the same image after thinning.

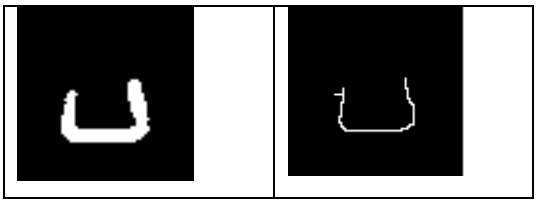

Figure .6a

Image Before Thinning

Figure .6b

Image After Thinning

6. CONCLUSION AND FUTURE WORK

In this paper, we present a review about optical character recognition and its importance, and its main approaches and techniques. Also, we list the characteristics of Arabic language, and focused in one of important phases in recognition systems which is preprocessing. Moreover, we described and implemented preprocessing algorithms to binarized, dots removing and thinning for Arabian characters. In the future, we will use the result from this phase to extract features and design recognition system.

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

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

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

[7] Al-Emami, S. and M. Usher, On-line recognition of handwritten Arabic characters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990.

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

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

[10] O'Gorman, L. and R. Kasturi, Document image analysis. Vol. 39. 1995: IEEE Computer Society Press Los Alamitos, CA.

[11] Amin, A., G. Masini, and J. Haton, Recognition of handwritten Arabic words and sentences. ICPR84, 1984.

[12] Amin, A. and J.F. Mari, Machine recognition and correction of printed Arabic text. IEEE Transactions on systems, man, and cybernetics, 1989.

[13] Gatos, B., I. Pratikakis, and S.J. Perantonis, Adaptive degraded document image binarization. Pattern recognition, 2005.

[14] Suliman, A., et al., Chain Coding and Pre-Processing Stages of Handwritten Character Image File. electronic Journal of Computer Science and Information Technology, 2011.

[15] Otsu, N., A threshold selection method from gray-level histograms. Automatica, 1975.

[16] Zhang, T. and C.Y. Suen, A fast parallel algorithm for thinning digital patterns. Communications of the ACM, 1984