Online System for recognizing Arabic Numerals Based on Matching Alignment Algorithm

Mustafa Ali Abuzaraida[1], Akram M. Zeki[2], Ahmed M. Zeki[3]
[1]*Misurata University, Libya*
[2]*International Islamic University Malaysia, Malaysia*
[3] *College of Information Technology University of Bahrain*

Abstract

Online text recognition systems have been continually given due importance these days globally because of the rapidly developing touch screen gadgets. However, it has been difficult to utilize keyboards and external mouse-like inputs in significantly tinier devices which consequently paved the way to current researched based scientists to look for some newer techniques which could design such type of online systems which could further deal with different kinds of texts for example, digits, symbols and alphabets. In the present paper, an online system for recognizing manually written Arabic numerals is being given. This paper will show digit acquisition, preprocessing, feature extraction and recognition phases in detail. The set of the data was gathered from 100 writers using a touch screen PC with 100 samples of every digit. The average accuracy rate of the outcome of the test of this proposed system was 98%, which is a significant accuracy rate.

1. Introduction

The field of text recognition is considered as one of the major fields of the pattern recognition area which has been the subject of various researches for more than thirty years [1].

Generally, offline text recognition approaches are designed to convert scanned scripts into a text documents. In contrast, online approaches capture the text by writing on touch screen devices or recording the movements of a stylus and convert the action into a text format.

Online recognition field has been gaining more interest lately due to the increasing of pen computing applications like tablet devices, digital notebooks, and advanced cellular phones [2].

These days, such types of devices are high in use throughout the world which persuaded various companies to improve their previously manufactured devices to cope with poly-lingual functions. Although, these devices can deal with a variety of languages for instance, Chinese, Latin, Indian, Arabic, Korean, and many others in use by millions and millions of speakers all over the world with textual or speech form [3] [4].

From the literature in the text recognition field, it is noticeable that most of the research works were dedicated to offline approaches for Latin characters and other languages such Chinese. On the other hand, a few researches and studies have been published to develop online approaches using new methods and algorithms in this area for texts in general and digits in particular [4].

A series of researches have been in publication in this field in the recent decades. However, majority of the researches were covered of solving segmentation problem and recognizing detached letters while identifying numbers and mathematical symbols could not get any spectacular concentrations.

Arabic digits are commonly used by billions over the world. The shape of the digits (0,1,2,3,4,5,6,7,8,9) were originally designed by Arabic Mathematician scholars and upgraded by the Muslim scholar Al-Khwarizmi who invents the zero in the ninth century [5]. These styles were used in the western part of Arab world which located in North Africa and Alandalus "Spain" in the 10 century [6].

In this research paper a presentation of identifying manually or handwritten Arabic numerals is showed. The system has four main phases which are text acquisition, preprocessing, feature extraction, and recognition phases.

The rest of this paper is organized as follows: Section 2 summarizes the architecture of the proposed system and each step of the system is explained. Section 3 presents the results of testing the system while the conclusion and the summary of the paper is presented in section 4.

2. Architecture of the proposed system

This anticipated system has four main stages: text acquisition, preprocessing, feature extraction and recognition phases which implemented distinctive pattern recognition system architecture [1] as shown in Figure 1. Although, each and every manually written numeral has been processed separately as one single block, segmentation has been excluded from this system. Accuracy rate of recognition can be improved and the time process can be made limited by the segmentation free strategy [4]. However, each stage of the proposed system has a single or more

aims for reaching the system's objectives and moreover, to improve the overall recognition percentage. The stages or the phases of this system have been elaborated in the following manners:

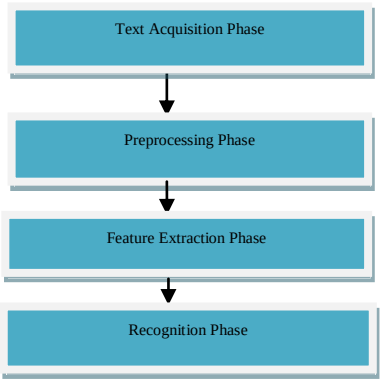

Figure 1. Phases of online text recognition System

2.1 Data Collection Stage

Data collection stage is the initial step of any pattern recognition system and aims to get raw data which is used later for training and testing manner. In this stage, every handwritten digit is captured by writing it on an interface device that records the handwritten digit in time stamped coordinates of pen trajectory (x, y) [3].

Here, for the purpose of collecting the training and testing datasets, 1.5 GHz core i3 Acer Tablet has been used to collect the dataset "same used in [7]". This touch screen computer can easily be used to acquire the handwritten Arabic digits with a simple way of normal writing on the touch screen using a special pen. The way of writing on this Tablet can minimize the noise and errors while recording on the Tablet surface.

A platform was developed using Matlab with graphic user interface for collecting manually written Arabic numerals. In this system, first, the writer can begin writing the digit or any letter in the field given immediately after entering their identification number. Second, the writer writes the digit that is visible in the given image above the acquiring data platform. Then, having written the number in the given writing area, the writer should click on the next button to write the next digit. Finally, in case, the writer wants to rewrite the current number prior to beginning writing the next one, they should click on the reset button for rewriting the number another time.

Data collection by using this natural way of writing can provide closely resembles, smoothed, and filtered data collected from the computer Tablet. Figure 2 shows the data collection platform.

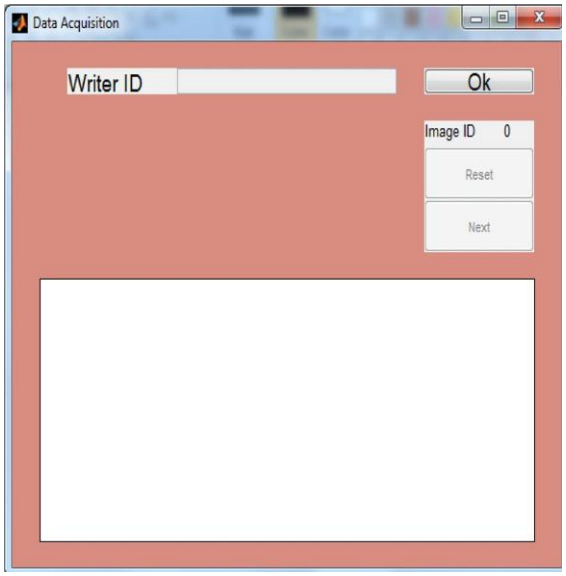

Figure 2. Data acquisition platform

2.2 Preprocessing Phase

In both types of handwriting recognition systems, preprocessing phase structure is required. [8]. But because of the text acquisition way, the stages of preprocessing may differ because of a lot of reasons for example, the scanner quality, quality of the paper, and skew of the text. Furthermore, noise can also occur in online systems due to some other reasons like the form of sharp edges, non-centered text, uneven sizes of text and missing points in text trajectories due to high handwriting speed. [3] [9]

To minimize the noise which might occur in the handwritten text as discussed before, the preprocessing phase in online handwriting recognition is done. In preprocessing phase, numerous multi-steps are incorporated and every step performs a particular task to sieve the set of data. In addition, this can perform better overall recognition rate and therefore, it is considered as one of the crucial phases of online handwriting recognition system and for which it has been the matter discussions among the researchers [10].

Nevertheless, performing and many preprocessing steps in this phase may cause some problems in online handwritten recognition systems. For instance, delay may take place to overall time processing [4]. Also, it may affect and reduce the recognition accuracy rate by complicating the processing which can lead to omission of some important parts and features of the text [3].

Generally, data collection in online handwriting recognition stores the stylus movements on the writing surface. These movements are distributed at various positions on writing area of the acquisition platform and then joined from the first position (x_1, y_1) to the last (x_n, y_n) to present the appearance of drawn text. Although, the stylus

movements consist of three actions which are: Pen Down, Pen Move and Pen Up actions. The serial of points is collected when the writer presses, moves, lifts the stylus up consecutively. Pen Move function records the movements of the stylus on writing tablet from the writing start point (x_1, y_1) until the last point (x_n, y_n) where 'n' is the total number of points in the writing movements' list [11].

There are four steps which are included in preprocessing phase in the proposed system as follow:

2.2.1 Digit Smoothing. In the proposed system, a smoothing technique is used to smoothen the handwritten curves called Loess filter. This filter is based on conducting the local regression of the curves points using weighted linear least squares and a second degree polynomial model.

In this technique, each smoothed value is determined locally by neighboring data points defined within the writing curve. The process is weighted and a regression weight function is defined for each data point contained within the writing curve [12].

2.2.2 Digit Simplification. Data point's simplification is the process of reducing the number of data points acquired by a digital device through removing the redundant points which could be inappropriate for pattern classification. This processing directly affects and enhances the recognizer performance. However, Douglas Peucker's algorithm [13] was adopted to simplify the acquired handwritten digit point sequence.

2.2.3 Digit Size Normalization. The size of the acquired handwritten digits depends on how the writer moves the stylus on writing area. The handwritten digits are generally written in different sizes when the pen is moved along the border of writing area that may cause some ambiguity in the next phases. Size normalization is a necessary step that should be performed in order to recognize any type of text. This can be achieved by converting the acquired handwritten digit with assumed fixed size format [14].

2.2.4 Centering of the Digit. After resizing the acquired handwritten digit, the current coordinates are needed to be shifted to the centering axis (x_0,y_0) to make sure that all handwritten digit points are in the equal formatting and all data are translated to the same spot relative to the origin.

2.3 Features Extraction Phase

Getting an appropriate set of features and an efficient extraction method are considered as the most vital factors for achieving a high recognition of performance for pattern recognition research field in general and text recognition in particular [15].

In the feature extraction phase of the proposed system, each handwritten Arabic digit is described using a set of features that distinguishes it from other handwritten digit in the dataset. The features of each handwritten digit represent in a ordered format called features vector. The features vector is then used in the next phase by the classifier to match the closest class using a classification criterion. In addition, the purpose of performing feature extraction phase is to realize that just a part of data points are equally important to the pattern recognition task.

In online recognition systems, the information about how the character has been written is found. Although, complex preprocessing steps cannot be performed in practical online systems such as Tablets and (personal digital assistant) PDAs since data is collected as the text is being written. Hence, taking advantage of the dynamic characteristics of the data is crucial such as the speed, angular velocity, and other features of this kind. These features remain available for processing as the character is written on the Tablet [15].

Choosing a proper type of features depends on the nature of the text, the type of the system processing which may be online or offline, and the texts types that can be handwritten or printed. However, feature types of recognizing any text can be categorized into three main types: structural features, statistical features, and global transformation [15].

Here, in this system, structural features are used to extract the handwritten Arabic digits features. While the system is not performing the segmentation part, the proposed system uses light amount of features which can help to avoid the complexity during the system excision.

Structural features are used to be the main features of the proposed system. They take the pen trajectory directions as the main feature representing handwriting movements. Freeman Chain code is used to create the direction matrix for each handwritten digit. Freeman Chain code [16] represents the pen movements directions by a numeric code consisting of 8 digits. These directions are listed from 1-8 to represent the eight main writing directions as illustrated in Figure 3.

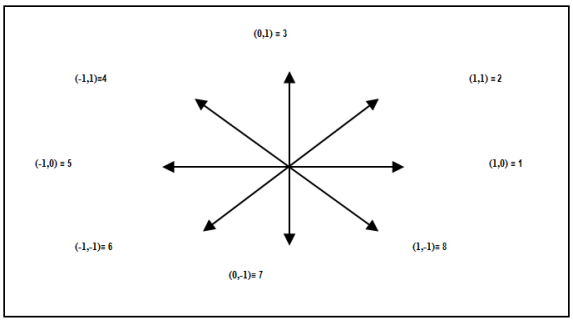

Figure 3. Freeman chain code

The action of extracting the features begins from the first point of writing and continues till the last point. The procedure of algorithm has been explained in steps as under.

- Read the points sequence S of the handwritten digit.
- Find the *d(x)* and *d(y)* values for the $S(x_2,y_2)$ point indexed by $S(x_1,y_1)$.
- Find the Freeman Chain Code for this pair from Table 1.

Table 1. Finding the freeman chain code

D(x)	D(y)	Code
0	+1	3
0	-1	7
-1	+1	4
-1	-1	6
+1	+1	2
+1	-1	8
-1	0	5
+1	0	1

- Make $S(x_2,y_2)$ as the first point and eliminate the previous point.
- Repeat Steps 2-4 for all the points.
- Record the Freeman Chain sequence to represent the handwritten digit movements' directions.

After completing the algorithm steps, a code for the directions of pen movements are stored. Each sequence is presenting a symbol of Arabic digit formatted in Freeman Chain Code. These sequences will be used in the next phase in order to distinguish and compare the handwritten digits by the classifier.

2.4. Recognition phase

The Global Alignment Algorithm (GAA) has been used in this research study as a recognition engine to identify and distinguish the Arabic numerals. Next to this phase, the proper digit from the set of data can be classified by the system [17].

Sequences Alignment or sequences comparison conceder is the heart of the bioinformatics field. It describes the way to arrange the DNA, RNA, or protein sequences by identifying the regions of similarity among them. Furthermore, it is used to conclude structural, functional, and evolutionary relationship between the matched sequences. Also, alignment algorithm finds the similarity level between query sequence and different database sequences. The algorithm is designed based on dynamic programming approach which divides the problem into smaller independent sub problems. It finds the alignment more quantitatively by assigning the matching scores [18].

In fact, the most well-known and widely used methods for sequences alignments are: Local and Global Alignment Algorithms. Local Alignment Algorithm compares the sequences which are suspected to have similarity or even dissimilar sequences length to find the local regions with high level of similarity. On the other hand, it is very much appropriate to use Global Alignment Algorithm for comparing the closely related sequences which are of same length. Here, the alignment is carried out from the beginning until the end of the matched sequence to find out the best possible alignment [17]. However, Global Alignment (Needleman-Wunsch algorithm) Algorithm is used in the proposed system as a classifying tool.

GAA was developed by Saul B. Needleman and Christian D. Wunsch in 1970 [17], which is basically a dynamic programming algorithm for sequence alignment. The dynamic programming can solve the original problem by dividing it into smaller independent sub-problems. The algorithm explains global sequence alignment for aligning nucleotide or protein sequences in general. However, these alignment techniques could be used in many different aspects of computer science approaches.

Basically, dynamic programming is used to find the optimal alignment of two sequences. It finds the alignment in a quantitative way by giving score values for matches and mismatches. The alignment is accurately obtained by searching the highest scores in the matrix [18].

For matching any two amino acid sequences, the algorithm is designed to find the highest score value of the sequences by building a two- dimensional matrix. Basically, the algorithm procedure is defined with three following steps.

- Assuming an initialization score matrix with the possible scores.
- Filling the matrix with maximum scores.
- For appropriate alignment, tracing back the previous maximum scores.

3. Results of testing the system

50 writers were requested to write the Arabic numerals to test this proposed system and every writer was asked to repeat writing every number in their own writing style to get 150 cases for all the 10 numbers in various writing manners.

On applying 80% for training and 20% of the numerals dataset for testing, the system recognized most of the digits successfully with an exception of digits such as 9 and 4 which were recognized with some irregularities due to similarities of the shapes of the digits.

4. Conclusion and Summary

Many countries have initialed using smart tablets in the classes to help the students to write on smart tablets in lieu of writing on papers in the past few

years which has made the learning and teaching much easier. These devices have helped in recognizing text software of any kind of manuscripts and for this viewpoint, Online recognition system is required to recognize the Arabic texts, numbers, and symbols.

This paper highlights a brief description of the design of a recognition system for the Arabic numerals. The principal goal of this paper is to open the research gate to such kind of researches. Further, this system can help in manufacturing education software to recognize the Arabic mathematical actions. In addition, this software can provide with more significance in learning mathematical subjects at high school or at university.

5. References

[1] M. A. Abuzaraida, A. M. Zeki and A. M. Zeki, "Recognition Techniques for Online Arabic Handwriting Recognition Systems," In Proceeding of the International Conference on Advanced Computer Science Applications and Technologies (ACSAT2012), Kuala Lumpur, Malaysia, 2012.

[2] Mustafa Ali Abuzaraida, Akram M Zeki and Ahmed M Zeki, "Online Recognition System for Handwritten Hindi Digits Based on Matching Alignment Algorithm," In Proceeding of the Third International Conference on Advanced Computer Science Applications and Technologies (ACSAT2014), Amman, Jordan, 2014.

[3] Mustafa Ali Abuzaraida, Akram M. Zeki and Ahmed M. Zeki, "Problems of writing on digital surfaces in online handwriting recognition systems," In Proceeding of the Information and Communication Technology for the Muslim World (ICT4M), 2013 5th International Conference on, 2013, pp. 1-5.

[4] M. A. Abuzaraida, A. M. Zeki and A. M. Zeki, "Segmentation Techniques for Online Arabic Handwriting Recognition: A survey," In Proceeding of the 3rd International Conference on Information and Communication Technology for the Moslem World: ICT Connecting Cultures, ICT4M 2010, Jakarta, Indonesia, 2010, pp. D37-D40.

[5] R. Kaplan and E. Kaplan, The Nothing that Is: A Natural History of Zero: Oxford University Press, 1999.

[6] Solomon Gandz, "The Origin of the Ghubār Numerals, or the Arabian Abacus and the Articuli." vol. 16, T. U. o. C. Press, Ed., ed: The University of Chicago Press, pp. 393-424, 1931.

[7] Mustafa Ali Abuzaraida, Akram M Zeki and Ahmed M Zeki, "Online Database of Quranic Handwritten Words," Journal of Theoretical & Applied Information Technology, vol. 62, 2014.

[8] Mustafa Ali Abuzaraida, Akram M Zeki, Ahmed M Zeki and Nor Farahidah Za'bah, "Online Recognition System for Handwritten Arabic Chemical Symbols," In Proceeding of the Computer and Communication Engineering (ICCCE), 2014 International Conference on, 2014, pp. 138-141.

[9] M. A. Abuzaraida, A. M. Zeki and A. M. Zeki, "Difficulties and Challenges of Recognizing Arabic Text," in Computer Applications: Theories and Applications, ed Kuala Lumpur: IIUM Press Malaysia, 2011.

[10] N. Tagougui, M. Kherallah and A.M. Alimi, "Online Arabic handwriting recognition: a survey," International Journal on Document Analysis and Recognition, pp. 1-18, 2012.

[11] Mai Al-Ammar, Reham Al-Majed and Hatim Aboalsamh, "Online Handwriting Recognition for the Arabic Letter Set," Recent Researches in Communications and IT, 2011.

[12] Loader Clive, Local Regression and Likelihood vol. 47: springer New York, 1999.

[13] Douglas David and Peucker Thomas, "Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature," Cartographica: The International Journal for Geographic Information and Geovisualization, vol. 10, pp. 112-122, 1973.

[14] Mustafa Ali Abuzaraida, Salem Meftah Jebriel, "The detection of the suitable reduction value of Douglas-Peucker algorithm in online handwritten recognition systems". IEEE International Conference on Service Operations And Logistics, And Informatics (SOLI), Hammamet, Tunisia, 2015. pp 82-87.

[15] M. A. Abuzaraida, Akram M Zeki and Ahmed M Zeki, "Feature Extraction Techniques of Online Handwriting Arabic Text Recognition," In Proceeding of the 5th International Conference on Information and Communication Technology for the Muslim World (ICT4M), 2013, pp. 1-7.

[16] Freeman Herbert, "Computer Processing of Line-Drawing Images," ACM Comput. Surv., vol. 6, pp. 57-97, 1974.

[17] R Durbin, S Wddy, A Korgh and G Mitchison, Biological sequence analysis: probabilistic models of proteins and nucleic acids: Cambridge University Press, 1998.

[18] Neil C. Jones and Pavel A. Pevzner, An Introduction to Bioinformatics Algorithms, illustrated ed. Cambridge, Massachusetts London, England: Massachusetts Institute of Technology Press, 2004.

