The 2006 International Arab Conference on Information Technology (ACIT'2006)

ON-LINE RECOGNITION OF ARABIC HANDWRITTEN CHARACTERS

Ahmad T. Al-Taani And Saeed M. Al-Haj
Department of Computer Sciences,
Yarmouk University, Irbid, Jordan.
{ ahmadta, saeedh }@yu.edu.jo

ABSTRACT

In this study, a new approach for the recognition of isolated handwritten Arabic characters is presented. The proposed method places a 5x5 grid on the character to extract the features needed for the recognition step. These features are calculated based on grid calculations. Then these features are feed to the decision tree to classify the character into one of the 28 classes. The classification process depends on the value assigned for each feature which lead to one leaf node in the decision tree that represent the Arabic character to be classified. Experimental results showed the robustness of the proposed approach for the recognition of Arabic handwritten isolated characters of a rate about 80.2%. The test was performed on 1120 different characters written be eight users, 40 examples for each of the 28 Arabic characters.

Keywords: *Arabic Character Recognition, Feature Extraction, Machine Learning, Pattern Recognition.*

1. INTRODUCTION

On-line automatic recognition of handwritten characters has been an on-going research problem for four decades. It has been gaining more interest lately due to the increasing popularity of hand held computers, digital notebooks and advanced cellular phones. A keyboard is very difficult to integrate in small devices and it usually determines the size of the whole apparatus. For example a mouse is insufficient or very slow when used alone in applications in which textual input is also desired.

Because of these problems, new methods for input have been developed; for example systems that recognize speech and handwriting. At first sight handwritten recognition does not appear to be a difficult problem. A recognition system should choose the correct character, usually the character that most resembles the written one from a limited set of characters. Unfortunately, this approach faces a number of difficulties. The most prominent problem in handwriting recognition is the vast variation in personal writing styles. There are also differences in one person's writing style depending on the context. The writing style may also change with time or practice. In addition, the mood of the writer and the writing situation can have an effect on writing style.

Arabic is written and spoken by more than 250 million people, in over 20 different countries. There are 28 characters in the Arabic alphabet. Words are written in horizontal lines from right. Each character has two to four different forms, which depend on its position in the word, see Figure 1.

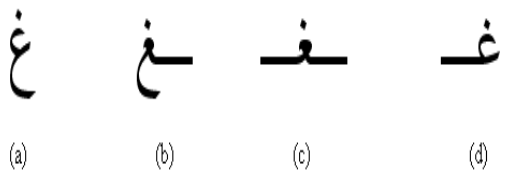


Figure 1: Different forms of "GHYN غ" character. (a) single form, (b) ending form, (c) middle form, (d) beginning form.

There are two main types of written Arabic [6]: classical Arabic; the language of the holly Quran and classical literature and modern standard Arabic; the universal language of the Arabic speaking world, which is understood by all Arabic speakers.

The automatic recognition of Arabic texts is complicated by several properties of the Arabic script:

- Connectivity of symbols,
- Cursive nature of the language,
- Similarity of groups of symbols,
- Highly variable widths, and
- Overlapping between characters.

Arabic character recognition has been one of the major languages to receive attention. This is due, in part to the cursive nature of the task. Two common themes have driven much of the work in Arabic character recognition.

The first is a hierarchical division of the input letter space to simplify the problem. The second theme is heuristically defined rules for classification or feature selection, which tend to be data and writer dependent.

For the past few decades, intensive research has been done to solve the problem of Arabic character recognition. Various approaches have been proposed to deal with problem. To date, challenging problems are being encountered and solutions to these are broadly targeted to improve accuracy and efficiency.

Amin [2] presented a technique for the recognition of Arabic text using the C4.5 machine learning system. The technique can be divided into three major steps. The first step is digitization and pre-processing to create connected component, detect the skew of a document image and correct it. Second, feature extraction, where global features of the input Arabic word is used to extract features such as number of subwords, number of peaks within the subwords, number and position of the complementary character, etc., to avoid the difficulty of segmentation stage. Finally, machine learning C4.5 is used to generate a decision tree for classifying each word.

El-Khaly *et al.* [3] proposed an algorithm for the recognition of machine optically captured Arabic characters and their isolation from the printed text. Recognition algorithms based only on individual characters need to be supplemented with a separation algorithm.

El-Sheikh *et al.* [4] presented an algorithm for the development of a real-time Arabic handwritten character recognition system. The system assumes that characters result from a reliable segmentation stage and the position of the character is known a priori. Four different sets of character shapes have been independently considered (initial, medial, final and isolated). Each set is further divided into four subsets depending on the number of strokes in the character.

El-Wakil *et al.* [5] proposed a method for the recognition of isolated handwritten Arabic characters drawn on a graphic tablet. Proposed features are extracted and used in the recognition process. The proposed system has been implemented and tested on a PDP11/70 mini-computer using BASIC PLUS-2 programming language and real data samples that have been drawn on a Tektronix graphic tablet.

Kharma *et al.* [8] proposed the use of mapping for use in on-line hand-written character recognition. This mapping produces the same output pattern regardless of the orientation, position, and size of the input pattern.

Mezghani et al [10] presented a method for shape representation of Arabic handwritten online character recognition based on a Kohonen associative memory. The features used by the authors are based on tangents and tangent differences at regularly spaced points along the character signal. Experiments are carried out on a database of online Arabic characters produced by a large number of writers. The authors have claimed superior performance of the scheme.

In this study, we present a novel approach for the recognition of Arabic letters. First, features needed for the recognition process are extracted, and then are used by the decision tree to recognize the letter. The architecture of the proposed system is shown in Fig. 1. First, the user writes the letter on a special window, and then the system draws a grid on the letter. Features are extracted based on grid calculations and then used by the decision tree in the recognition process.

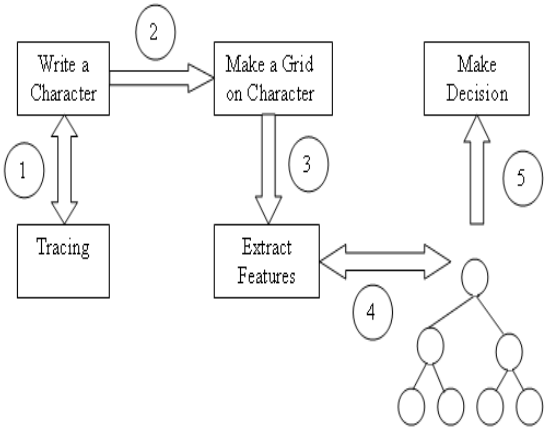


Figure 1: A block diagram of the proposed system.

The rest of the paper is organized as follows. Section 2 introduces some important issues of Arabic letters recognition. Methods and materials are discussed in section 3. Section 4 introduces the experimental results of the proposed method. Finally some conclusions and future work are presented in section 5.

2. ARABIC LETTERS HANDWRITING

Arabic handwriting recognition systems can be subdivided into many different categories. The most important discriminating features of the systems are: the time of recognition, handwriting style variations that includes variations of characters, and alignment of characters [1, 7 and 9]. Recognition can be done either off-line or on-line.

Arabic handwritten letters can vary in both their static and dynamic properties. Static properties are the

underlying, ideal models of the letters and the geometrical properties such as relative positions and sizes of the strokes, corners, retraces, ornamentals, sizes and aspect ratios of the letters, and the general slant of the writing. Dynamic properties are more involved with the generative aspects of the letters. Letters can look similar although their number of strokes, and the drawing order and direction of the strokes may vary considerably.

Fig. 2 shows some variation examples for the letter SHEEN 'ش'. In the Arabic scripts for example, there are four different position-dependent shapes for almost every letter. Also, there are some preference differences between left-handed and right-handed writers for using shapes.

The alignment of letters in Arabic language varies depending on the style of writing and the speed of writing. As shown in Fig. 3, the same word was written in two different ways distinguished by the direction, right-left or up-down.

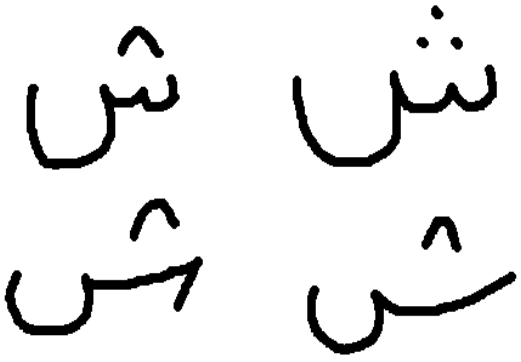
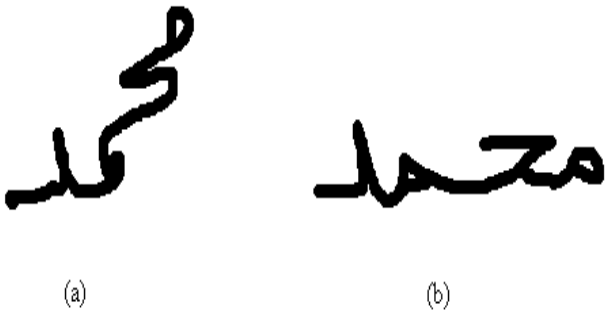


Figure 2: Some variations of writing "SHEEN" letter.

٣. METHODS AND MATERIALS

٣.1 OVERVIEW OF THE PROPOSED METHOD

In this work, we present a novel approach for the recognition of Arabic letters that can be adapted to the demands of hand-held and digital tablet applications. Features needed for the recognition process are extracted, and then are used by the decision tree. These features include: number of segments, left-right density ratio, bottom-up density ratio and others.



(a) (b)

Figure 3: different ways for writing the same word.

The proposed system (Fig. ١) consists of the following phases:

1. A user writes a letter on a special window.
2. The system keeps track of the x-y coordinates of the pixels forming the letter and stores them.
3. The system draws a grid on the letter.
4. Features are extracted from the written letter.
5. Extracted features are used by the decision tree for the recognition process.

These phases are described in more details in the following sections.

3.2 TRACING MODULE

The tracing process must be done in parallel with the writing process (On-Line), so we can keep track of input letter to get the x-y coordinates of every pixel forming the letter.

The outputs of this module are number of segments and a string for each segment that contains the x-y coordinates. Every mouse click is considered as one segment, so the user must keep in mind this rule, for example, the letter SEEN "س" must be written by one mouse click and drag, otherwise it will be considered as two letters (Fig. 4).

The proposed system will not recognize the letter in 4(a) correctly as SEEN, because SEEN letter is classified as one-segment letter. The letter in 4(a) will be stored in two separate lists and to recognize SEEN letter we deal with the first list only.

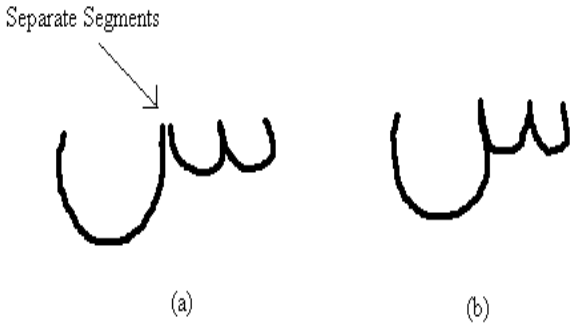


Figure 4: The letter SEEN written two ways (a) Two-segments, (b) Single-Segment.

3.3 DRAWING THE GRID

The main advantage of using the grid is to simplify the feature extraction process that needed to differentiate between letters that have the same shape and number of segments.

For example, Fig. 5 shows two different letters, JEEM and KHAA. They have the same shape but the location of the dot feature is used to differentiate between the two letters. The letter KHAA has a dot above the main shape (Fig. 5-a), while JEEM has a dot location under the main shape.

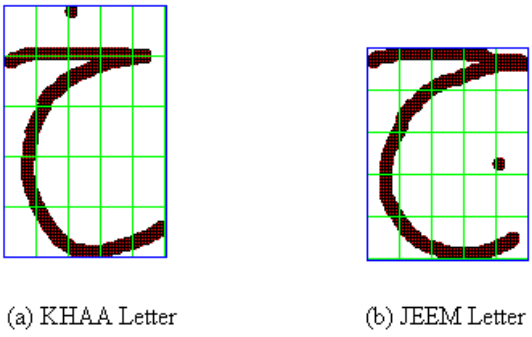


Figure 5: The effect of dot location on letter recognition.

Experiments showed that a grid size of 5x5 is suitable in this study. A 3x3 grid size gave low details about the letters, while using a 7x7 grid size gave more details than we need. Another advantage of using grids is that grid size does not affect the letter size (Fig. 6).

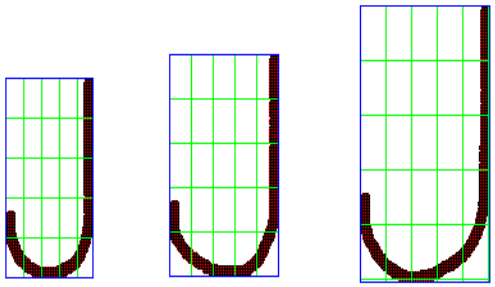


Figure 6: Different sizes of LAAM letter.

3.4 FEATURE EXTRACTION

3.4.1 NUMBER OF SEGMENTS

The most important feature used in this work is number of segments. By segment we mean the separate letter component that must be written without lifting the pen. Fig. 7 shows a character that has three segments.

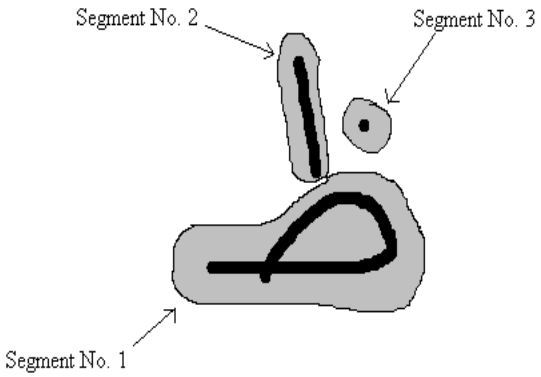


Figure 7: Three-Segment Letter (THAA Letter).

3.4.2 LOOPS

Another feature that is useful is whether the written letter contains a loop or not. Number of Arabic letters that have been considered to contain this feature is nine; such letters that has loop are illustrated in Fig. 8.

To detect a loop in a written letter, we do some processing and calculations on the list that contains the x-y coordinates of the letter.

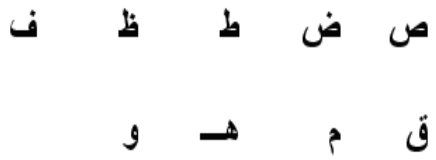


Figure 8: Arabic letters that have loops.

3.4.3 SHARP EDGES

Sharp edge detection is the most difficult feature to be extracted. Sharp edge is similar to angle with property 20-40 degree. Fig. 9 shows some letters that have sharp edges.

There are two types of sharp edges depending on the direction of the edge. In Fig. 9(a) the letter AYN is a y-direction sharp edge type while 9(b) SAAD letter is an x-direction type.

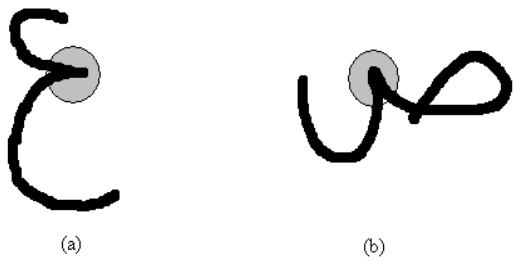


Figure 9: Letters containing sharp edges, (a) y-direction type (b) x-direction type.

3.4.4 SECONDARY SEGMENTS

Any part or component that was written after the primary part is considered as a secondary segment. Secondary segments are stored in lists numbered two to four. There are three types of secondary segments: dot, line, and curve, these types are shown in Fig. 10.

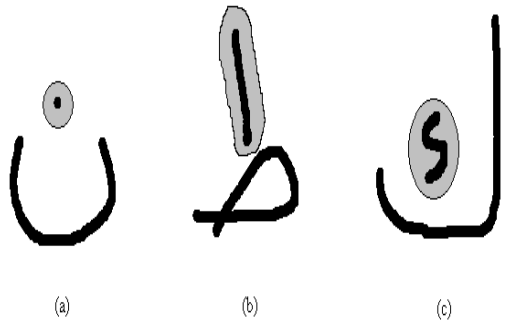


Figure 10: Secondary Segments,
(a) dot, (b) line (c) curve.

4. EXPERIMENTAL RESULTS

We used two different sets of examples, one for training and the other for testing. The method is trained and tested on 1400 different characters written by ten different users. Each user wrote the 28 Arabic characters five times in order to get different writing variations. Experiment results showed the effectiveness of the novel approach for recognizing handwritten online Arabic characters. The proposed method achieved a recognition accuracy of about 90%.

Experimental results showed that the proposed method did not perform well on letters that contains sharp edges, like in the letters "ج، ح، خ، غ، ع". This problem is considered for future work.

Now, let us illustrate the proposed system of the letter "QAAF" in Fig. 11. Features extracted from this letter are: Number of segments: 3, Contains Loop: YES, Contains Sharp Edge: NO, and Type of Secondary Segment: DOT. These features are then used by the decision tree to recognize the letter as QAAf (Fig. 12).

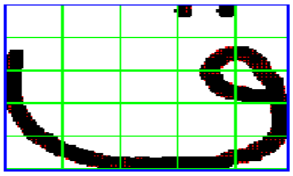


Figure 11: Input letter "QAAF".

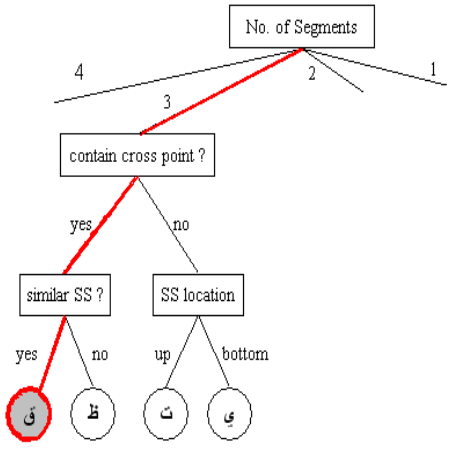


Figure 12: Recognition of the letter QAAF.

5. CONNCLUSIONS AND FUTURE WORK

We have presented a novel approach for the recognition of Arabic letters based on some novel features and the use of decision tree machine learning technique. The proposed method can easily be applied to any application that requires handwritten character recognition.

For future work, some improvements need to be done to enhance the performance of the system for the recognition on letters that contains sharp edges " ع، غ، خ، ح، ج". Another character that has less than average recognition rate is "هـ". Future work, also, will consider the modification of the proposed method to be used for the recognition of an off-line letters.

We are planning to adapt the proposed method to recognize Arabic handwritten digits as well.

Comparing our new proposed method with previous work in the field we found that our proposed method presented new features for the recognition of Arabic handwritten online characters. It also gave promising results; in many cases our proposed method gave better results than previous methods. We came up with this conclusion depending on what the authors presented in their papers, since an empirical analysis of these methods is not from the objectives of this work. This is left for future work as well.

References

[1] Amin A., "Recognition of hand-printed characters based on structural description and inductive logic programming". *Pattern Recognition Journal*, vol. 24, pp 3187–3196, 2003.

[2] Amin A., "Recognition of printed Arabic text based on global features and decision tree learning techniques". *Pattern Recognition Journal*, vol. 33, pp 1309-1323, 2000.

[3] El-Khaly F. and Sid-Ahmed M. A., "Machine recognition of optically captured machine printed Arabic text", *Pattern Recognition Journal*, vol. 23, no. 11, pp 1207-1214, 1990.

[4] El-Sheikh T. S. and El-Taweel S. G. , "Real-time Arabic handwritten character recognition". *Pattern Recognition Journal*, vol. 23, no. 12, pp 1323-1332, 1990.

[5] El-Wakil M. S. and Amin A. Shoukry, "On-line recognition of handwritten isolated Arabic characters". *Pattern Recognition Journal*, vol. 22, no. 2, pp 97-105, 1989.

[6] Hadjar K. and Ingold R., "Arabic Newspaper Page Segmentation". *In proceeding of the seventh international conference on document analysis and recognition* (*ICDAR 2003*), vol. 2, pp 895-899, 2003.

[7] Herrick, E. M., "Letters with alternative basic shapes". *Visible Language*, vol. 9, no. 2, pp 133-142, 1979.

[8] Kharma N. N. and Rabab K. Ward, "A novel invariant mapping applied to hand-written Arabic character recognition". *Pattern Recognition Journal*, vol. 34, pp 2115-2120, 2001.

[9] Kuklinski, T. T., "Components of handprint style variability". *In proceedings of the seventh International Conference on Pattern Recognition*, pages 924-926, 1984.

[10] Mezghani, N., Mitiche, A., Cheriet, M., "A new representation of shape and its use for high performance in online Arabic character recognition by an associative memory". *International Journal on Document Analysis and Recognition*, vol. 7, no. 4, pp 201-210, 2005.