Effective Arabic Character Recognition using Support Vector Machines

Mehmmood Abdulla Abd
Ajman University of Science and Technology
Faculty of Computer Science and Engineering
UAE
mahmood_abdullah@hotmail.com

George Paschos
Nth Research
Athens, Greece
gpaschos@netscape.net

Abstract - This paper proposes an Arabic character recognition system. The system focuses on employing Support Vector Machines (SVMs) as a promising pattern recognition tool. In addition to applying SVM classification which is a novel feature in arabic character recognition systems, the problem of dots and holes is solved in a completely different way from the ones previously employed. The proposed system proceeds in several phases. The first phase involves image acquisition and character extraction, the second phase performs image binarization where a character image is converted into white with black background, while the next phase involves smoothing and noise removal. In the fourth phase a thinning algorithm is used to thin the character body. The fifth phase involves feature extraction where statistical features, such as moment invariants, and structural features, such as number and positions of dots and number of holes, are extracted. Finally, the classification phase takes place using SVMs, by applying a one-against-all technique to classify 58 Arabic character shapes. The proposed system has been tested using different sets of characters, achieving a nearly 99% recognition rate.

I. INTRODUCTION

Offline Arabic Character Recognition (OACR) is a challenging problem; systems that address it will have contributed to the improvement of the computerization process [1]. Many scientists have intensively and extensively researched OACR of both printed and handwritten characters. Over the last forty years a great amount of research work in character recognition has been performed for Latin, Hindi and Chinese. The Arabic language serves as a script for several languages in the Middle East, Africa and Asia such as Arabic, Farsi, Urdu, Uygur, Jawi, Pishtu, Ottoman, Kashmiri, Old Hausa, Baluchi, Berber, Dargwa, Ingush, Kazakh, Kirghiz, Sindhi, and others. Moreover, all Muslims can read Arabic script as it is the language of AL-Quran. Despite these facts, research work on Arabic character recognition has not received much attention either because of its difficulties or due to lack of support in terms of funding and other utilities such as Arabic text databases, dictionaries, etc., and of course because of the cursive nature of its writing rules.

The cursive nature of the Arabic script makes the recognition of Arabic distinct from the recognition of Latin or Chinese scripts. In addition, most Arabic characters have from two to four different shapes/forms depending on their position in the word. Arabic writing has different font types. The font styles make Arabic character recognition hard and development of a system that is able to recognize all font styles is difficult. These styles encompass Ruq'a, Nastaliq, Diwani, Royal Diwani, Rayhani, Thuluth, Kufi and Naskh.

Arabic is cursively written from right to left (in both printed and handwritten forms) and the words are separated by spaces. It has 28 characters and each character has two or four different forms/shapes, depending on its position in the word, which increases the number of classes from 28 to 120. However, Arabic has no small or capital letters. The Arabic characters are connected in the word on the base line. Furthermore, some characters in the Arabic language are connectable from right only, these are: ا,د,ذ,ر,ز,و. Some of the right-connectable characters cause an overlapping between subwords, for instance "Waow و ". Overlapping can be addressed with a contour-following algorithm [2]. Therefore, such characters divide the words into two subwords, when they appear in a word.

Some characters have exactly the same shape and some diacritics that make them differ from each other. These diacritics involve a dot, a group of dots, or a zigzag (hamza). The presence or absence of diacritics has a very important effect on Arabic word meaning. For instance, the word "حُب " means, "love " and "حَب " means "grain ",

T. Sobh (ed.), Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, 7–11.

where the meaning completely depends on the diacritics. Diacritics may appear above or below the base line (letter). Some Arabic characters have one to two holes within the character's body. The dot is also another feature that is used to distinguish among similar characters. The maximum number of dots that may appear above the character is three and below the character is two. A thinning algorithm may effectively deal with them.

Arabic character recognition falls into either online or off-line category, each having its own recognition algorithms and hardware. This paper deals with isolated offline Arabic character recognition. The purpose of the proposed work in this paper is to build a high-accuracy Arabic character recognition system using improved feature extraction and optimized Support Vector Machines (SVMs). The objectives of this research are to a) improve the recognition rate in Arabic character recognition, b) improve the performance of SVMs. The proposed methodology is described by the following processing phases:

1. Image acquisition and character extraction
2. Image binarization
3. Smoothing and noise removal
4. Character thinning
5. Feature extraction
6. Classification using multi-class SVMs.

The remaining of the paper is organized as follows. Section II describes preprocessing, section III presents the feature extraction methodology, section IV presents the multi-class SVM classification system followed by the results obtained in section V, while section VI provides a recapitulation and suggestions for future work.

II. PREPROCESSING

The Arabic character features are extracted from gray-level image, which is scanned by a regular flat scanner. The threshold value is chosen based on trial and error. This threshold value is utilized to yield a white character body with a black background. Then, the character's body is isolated from the image background. A binary image is cleaned up and introduced to the feature extraction phase. Mathematical morphology is utilized to remove noise and to smooth the character's body. It is worth mentioning that this technique has not been used in the Arabic character recognition techniques.

Two morphological operations that are mainly used are opening and closing. Closing fills small gaps in an image, which eliminates small holes in the image's contour, and opening opens small gaps or spaces between touched objects in the character's image. This is useful to break narrow isthmuses and eliminate small objects. Both operations employ the same basic morphology operations, which are dilation and erosion, using the same structural elements. Then, a sequential morphological thinning algorithm is used to remove spurious pixels from the edge of the character's body.

III. FEATURE EXTRACTION

The proposed system deals with isolated Arabic characters to recognize an unknown character by deciding to which class it belongs. After extracting all structural and statistical features of the Arabic characters, the feature vector is fed to the SVM classifier. These features consist of the first three moment invariants, the number and position of dots, the number of holes in the character body, and the number of pixels in the dot area, as described below.

A. Moment invariants

The Hu moment invariants are calculated from each character image as described in [3]. These moment invariants are insensitive to image translation, scaling and rotation, thus, they have the desired properties to be used as pattern descriptors.

The first three moment invariants are utilized to decrease the number of features, and consequently speed up training and classification, where the absolute value of the logarithm of each moment invariant is in fact computed instead of the moment invariants themselves. Using the logarithm reduces the dynamic range, and absolute values are taken to avoid dealing with the complex numbers, which may result when computing the negative values of log of moment invariants [3]. The invariance of moments is important and not their signs, therefore absolute values are used.

B. Number of dots and their position

The number of dots and their positions play important roles in Arabic character recognition. Some Arabic characters have the same shape but the number of dots and their positions make them differ from each other. For instance, Ta " ت " and Tha "ث " have the same shape and they differ in the number of dots, and Noon" ن " and Ba " ب "they differ in their dot positions. Consequently,