ABSTRACT

Many optical character recognition (OCR) techniques and tools have been developed for plurality of languages. A successful OCR system improves interactivity between humans and computers in many applications such as digitising and recognising written content. With regard to Arabic OCR, the problem of handwriting recognition is challenging because Arabic letters are cursive and shape-changeable depending on their positions. OCR systems have reached nearly perfect acknowledgement of Arabic printed text, yet still in its inception and needs to be greatly improved with handwritten text. Therefore in this study, an approach to recognize Arabic characters based on genetic algorithms (GA) is proposed. The approach requires two separate stages; feature extraction and GA for character recognition development. In the feature extraction stage, six features are detected for each character and denoted as a feature vector of 6 integer numbers. The feature vectors are then utilised in the next stage. Three genetic operators namely selection, crossover and mutation are implemented to search for the similar vectors with the best fitness value to recognise the character. The data used in this study were collected from different resources and stored in a database. It consists of 12,500 printed text words in 50 paragraphs and 15,000 words written by 100 different writers, males and females aged 5 to 60 years. Pre-processing operations are conducted including segmenting paragraphs into lines, segmenting line into words, segmenting words into characters, detecting skeleton, and determining baseline and other horizontal zones. The experimental results have shown that the proposed method has achieved promising accuracy recognition rate with 90.46% for printed text and handwritten characters.

ABSTRAK

Banyak teknik dan alatan Pengecaman Aksara Optikal (PAO) telah dibangunkan bagi mempelbagaikan bahasa. Satu sistem PAO yang berkesan dapat mempertingkatkan interaksi antara manusia dan komputer dalam pelbagai aplikasi seperti pendigitan dan pengecaman kandungan bertulis. Merujuk kepada PAO Bahasa Arab, permasalahan dalam pengecaman tulisan tangan adalah mencabar disebabkan oleh aksara Arab yang berbentuk kursif dan berubah bentuk mengikut kedudukan aksara tersebut. Sistem PAO telah mencapai pengiktirafan hampir sempurna terhadap teks Arab bercetak; namun pembaikan lanjutan terhadap teks tulisan tangan Arab perlu dipergiatkan lagi. Oleh yang demikian, kajian ini mencadangkan satu pendekatan pengecaman aksara Arab berdasarkan Algoritma Genetik (AG). Pendekatan ini memerlukan dua fasa berasingan; penyarian fitur dan pembangunan pengecaman aksara menggunakan AG. Pada fasa penyarian fitur, enam fitur dikesan bagi setiap aksara dan diwakilkan sebagai satu vektor fitur dengan 6 nombor integer. Vektor fitur ini akan digunakan dalam fasa seterusnya. Terdapat tiga operasi genetik iaitu pemilihan, pemindahan dan carian permutasi dilaksanakan terhadap vektor yang serupa dengan nilai muatan terbaik bagi mengecam aksara berkaitan. Kutipan data bagi kajian diperolehi daripada pelbagai sumber dan tersimpan dalam pangkalan data. Ia terdiri daripada 12,500 teks bercetak; 50 perenggan dengan 15,000 perkataan yang ditulis oleh seratus penulis lelaki dan perempuan yang berlainan dengan julat umur di antara 5 hingga 60 tahun. Operasi pra-pemprosesan yang dilaksanakan termasuk segmentasi perenggan kepada garisan, segmentasi garis kepada perkataan, segmentasi perkataan kepada aksara, pengesanan rangka, garis-dasar, dan penentuan zon mendatar. Hasil kajian menunjukkan bahawa kaedah ini berjaya memberikan kadar ketepatan pengecaman yang menyakinkan iaitu sebanyak 90.46% bagi teks bercetak dan aksara tulisan tangan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACR	-	Arabic Character Recognition
BMP	-	Bitmap Picture
CEDR	-	Center of Excellence for Document Analysis and Recognition
CR	-	Character Recognition
GAs	-	Genetic Algorithms
HWR	-	Handwriting Recognition
OCR	-	Optical Character Recognition
OACR	-	Optical Arabic Character Recognition
UR	-	UnRecognized
R	-	Recognized
US	-	UnSegmented

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Problem Background

The character recognition (CR) mechanization is being intensively investigated in the pattern recognition research area. CR automation means translating images of characters into a text; in other words, it represents an attempt to simulate the human reading process. CR is very difficult to accomplish owing to various issues such as the inconsistency of human writing, the segmentation of words into characters, high variability in terms of handwriting styles and shapes, the size of the lexicon, and the writing skew or slant.

There are two main classifications of the problem of handwriting recognition: online recognition and offline recognition; these terms refer to the format of the input handwritings image. Temporal information is available in online recognition, for instance pen tip coordinates as a time function, while in offline recognition, just the handwritings image is obtainable. Several applications require offline handwriting recognition capabilities; these include commercial form reading, bank processing,

document archiving, office automation, mail sorting, etc. Up to the present time, offline handwriting recognition is still an open problem, and has been dealt with by several researchers in this field (Benouaretha *et al.*, 2008;Plamondon and Srihari, 2000; Koerich *et al.*, 2003; Vinciarelli, 2002).

Research into Arabic handwriting recognition has not been conducted as fully as for other scripts, such as Latin, although there has recently been renewed interest in this area (Ben Amara and Bouslama, 2003; Amin, 1998; Lorigo and Govindaraju, 2006; Benouaretha *et al.*, 2008). The developed techniques for Latin handwriting recognition are not suitable for Arabic for the reason that the foundations of Arabic writing, such as its alphabet and grammars, differ from Latin script. In view of the fact that the word is the most innate part of writing, the procedure for recognizing it needs to adopt either an analytic approach that recognizes individual characters in the word or a holistic approach that recognizes the whole word image (Benouaretha *et al.*, 2008). Analytical approaches involve two steps: segmentation and combination (El-Hajj *et al.*, 2005; Kim and Govindaraju, 1997; Benouareth *et al.*, 2006; El-Yacoubi *et al.*, 1999; Koerich *et al.*, 2003). It starts by segmenting the enter image into units that are close to the characters' size. Next, it uses dynamic programming to combine the segmented units to match character forms. In contrast, the holistic approach handles the whole input image. Holistic attributes are typically used to reduce fewer likely selections in the lexicon. Such attributes include word length, translation or rotation invariant quantities, ascenders, descenders, connected components, dots, etc. (Benouaretha *et al.*, 2008). Hence, holistic techniques need to be compared for each word in the lexicon (Souici and Sellami, 2006; Madhvanath and Govindaraju, 2001; Farah *et al.*, 2006), whereas analytical techniques compare for each character.

While many different methods of solving the OCR problem have been explored, the use of a genetic algorithm to recognize characters has been growing in popularity. "Genetic algorithms offer a particularly attractive approach for this kind of problems since they are generally quite effective for rapid global search.

Moreover, genetic algorithms are very effective in solving large-scale problems"(Oliveira *et al.*, 2001).This begs the question of what genetic algorithms (GAs) actually are.

A GAs is an optimization and search technique utilized in computer science to find approximate solutions to problems. It is inspired by processes in biological evolution such as natural selection, inheritance, recombination, and mutation. GAs are generally realized in a computer model, in which a population of runner solutions to an optimization problem progress to better solutions. The evolution starts from a population of completely random individuals and occurs in generations. In each generation, the fitness of the whole population is evaluated, and multiple individuals are selected from the current population based on their fitness. These are modified, mutated, or recombined to create a new population, which becomes current in the next iteration of the algorithm, as shown in Figure 1.1. Usually, the solutions are represented in strings of 0s and 1s, although different encodings are also possible. So, evolutionary algorithms play on populations, instead of coming to one solution.

This research will give an extended illustration of offline Arabic character recognition. The system is based on feature extraction and the genetic algorithms approach.

1.2 Statement of the Problem

One major issue that remains uncharted is how to segment and recognize cursive handwriting, especially Arabic characters. In this regard, our problem statement asks whether genetic algorithms can be effectively adapted for the segmentation and recognition of Arabic characters.

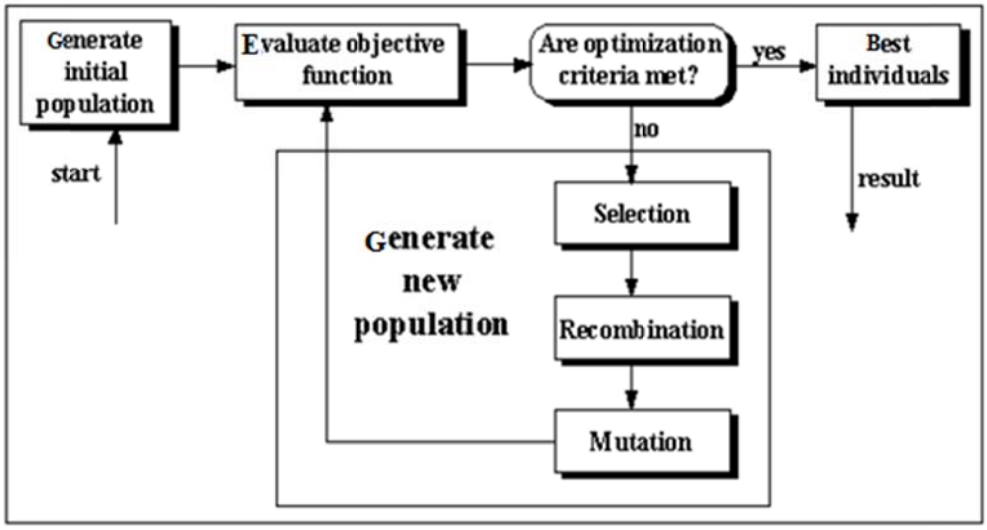

Figure 1.1 Structure of a single population evolutionary algorithm

1.3 Research Aim

This research aims to investigate the use of GAs for the recognition of Arabic handwriting and printed text. In other words, it center of attention on the problem of recognizing Arabic characters to create a successful system.

1.4 Objectives

This research proposes to accomplish the following objectives:

i. To overcome the problems of offline CR.

ii. To prove GAs' ability to recognize offline cursive printed text and handwritten words in Arabic characters.

iii. To develop a feature extraction approach for the shape of Arabic characters.

1.5 Research Scope

This project entails the following scope:

i. The main focus of this work is to recognize and segment cursive Arabic words and printed text in to characters; the sub-word is our region of interest.

ii. The input image may be in one of the following:

- An image containing one and only one sub-word.
- An image containing a word of more than one sub-word, or a text line of several words (see example in Appendix B).
- A document of multiple text lines, such as a paragraph (see example in Appendix C).

iii. GAs techniques will be used and facilities with a population of solutions called *feature vectors* represent the features of the characters.

iv. The algorithm will be programmed using Matlab 2007.

1.6 Significance of the Research

This research will describe an offline handwriting and typed Arabic word recognition and segmentation system based on GAs. As we searched, we found that this is the first work that uses only GAs without any other supplementary method for the problem of recognizing offline Arabic characters.

1.7 Organization of the Report

This report is organized into five chapters. The first introduces the study and gives the problem background, objectives, and scopes. Chapter 2 provides a literature review of the existing techniques to recognize and segment Arabic characters (AC). The methodology of the project is discussed in Chapter 3. Next, the experimental results are given in Chapter 4. The conclusion and suggestions for future work are explained in Chapter 5.