Chapter 38
New Mechanisms to Enhance the Performances of Arabic Text Recognition System: Feature Selection

Marwa Amara
SOIE Laboratory, Tunisia

Kamel Zidi
University of Tabouk, Saudi Arabia

ABSTRACT

The recognition of a character begins with analyzing its form and extracting the features that will be exploited for the identification. Primitives can be described as a tool to distinguish an object of one class from another object of another class. It is necessary to define the significant primitives during the development of an optical character recognition system. Primitives are defined by experience or by intuition. Several primitives can be extracted while some are irrelevant or redundant. The size of vector primitives can be large if a large number of primitives are extracted including redundant and irrelevant features. As a result, the performance of the recognition system becomes poor, and as the number of features increases, so does the computing time. Feature selection, therefore, is required to ensure the selection of a subset of features that gives accurate recognition and has low computational overhead. We use feature selection techniques to improve the discrimination capacity of the Multilayer Perceptron Neural Networks (MLPNNs).

INTRODUCTION

Man-machine communication is marked by its tendency to constrain human intervention. This can be achieved if machines that are able to listen to and recognize word, read the documents and correctly handle characters that form them are employed. Optical characters recognition (OCR) was the subject matter of multiple researches. Its purpose is to convert the scanned images of a printed or handwritten

DOI: 10.4018/978-1-5225-2229-4.ch038

document into a computerized file (a machine-encoded text) which can be manipulated by word processing software. Reading a printed and even a handwritten document can be of great benefit in various domains. It would be a breakthrough, for instance, if the computer could read fluently, sort mail automatically, treat invoices and checks and access all written information whose very existence begins with a mere sheet of paper. In recent years, the recognition of Arabic scripts has received increasing attention. Many approaches for the recognition of Arabic characters have been proposed. However, no high recognition rate has been achieved from existing recognition systems (Alaei et al., 2012, Al-Zoubaidy, 2006, Amara et al, 2016 a, Tharwat et al 2015 a). The main reason for getting low accuracy is accounted for by the particularity of the Arabic script. Unlike other languages, the Arabic script has morphological characteristics that are the cause of the failure of treatment. Writing recognition is part of pattern recognition which is concerned with the shapes of characters. Researchers have realized intensive work that led to the publication of several articles bearing on character recognition. Historical overviews about recognition methods can be found at (Gaikwad et al, 2008; Lee et al, 1996; Khorsheed, 2002). Recognition of the Arabic script can be traced back to the 80s. However, most of the already published work have focused on Latin characters and then applied them on for the recognition of Arabic script. For an overview in the field of Arabic handwriting recognition, we include articles (Amara et al, 2014; Al-bader et al, 1995; Ahmad et al, 2012; Parvez et al, 2013; Amin et al, 1997). As found in (Nasien et al, 2014)a presentation of lines recognition. In addition, other studies describe methods handwritten (Impedovo et al, 1991) and printed (Suen et al, 1980; Amara et al, 2014; Amara et al, 2015; Amara et al, 2016a; Amara et al, 2016b) can be consulted. There is no universal system of OCR that can handle all cases of writing but rather different approaches depending on the type of data processed and the intended application.

In this research, we concentrate on improving the feature extraction stage by selecting efficient features to extract. We use genetic algorithm (GA) as a feature selection technique to select best feature subsets. We analyze the recognition accuracy as a function of the feature subset size using a perceptron multilayer (PML) classifier.

Our chapter is organized as follows: In Section *2,* we provide an overview of letters recognition. In Section *3*, we present the characteristics of the Arabic script. Section *4* thoroughly exposes the details of the proposed system. Section *5* will be devoted to the experimentation and evaluation. To conclude, we discuss the results in Section *6*.

BACKGROUND

Writing recognition is part of pattern recognition which is concerned with the shapes of characters. Researchers have realized intensive work that led to the publication of several articles bearing on character recognition. There is no universal system of OCR that can handle all cases of writing but rather different approaches depending on the type of data processed and the intended application. The figure 1 shows the general structure of an OCR system.

As stated earlier, in order to recognize a given character, every character needs to be analyzed in terms of its form, and its features need to be extracted for characters identification purposes. Since the selection of the best discriminant features is considered as a crucial step in the recognition system, many researches have stressed that the inclusion of additional features leads to a worse rather than better performance. The preliminary works on feature selection started in the 1960s (Al-Zoubaidy, 2006). Figure 2 summarizes the methodology used to select features. The feature selection method is commonly used to