نخستین کنفرانس بین المللی پردازش خط و زبان فارسی
۱۵ و ۱۶ شهریور ۱۳۹۱
دانشگاه سمنان - دانشکده مهندسی برق و کامپیوتر

Farsi Machine-printed Subwords Recognition Using Contour-based Fourier Descriptors

Parnia Bahar
Electrical and Computer Department,
Semnan University, Semnan, Iran
parniabahar@yahoo.com

Saeed Mozaffari
Electrical and Computer Department,
Semnan University, Semnan, Iran
mozaffari@semnan.ac.ir

***Abstract*— This paper presents a fast and simple method for Farsi/Arabic subwords recognition in a large lexicon. By omitting dots and complementary parts of machine-printed characters, a dataset including 9445 Farsi/Arabic subwords written by a single font and single size was obtained. This dataset not only reduces the number of subwords, but makes it suitable for both Farsi/Arabic languages. After normalizing boundary points of each subword, Fourier descriptor features are extracted. Experimental results on 30 plain text shows accuracy of 82.1% on subword level. Considering this large and comprehensive dataset, the obtained results are still promising which can be enhanced in the future by the use of Farsi/Arabic language grammar for connecting subwords.**

Keywords- Farsi/Arabic word recognition, machin-printed documents, Fourier shape descriptors, Large dataset.

I. INTRODUCTION

Spotting words in printed documents in Latin and Chinese has received remarkable attention [1-3] and high precision rates have been reported. Compared to other languages, the work done on Farsi/Arabic words is not sufficient. Special characteristics of Farsi/Arabic alphabet make word recognition more challenging.

Farsi/Arabic optical character recognition and document analysis dates back to a few decades ago. The first attempts returns to printed Farsi text recognition [4] or script recognition [5].

In order to recognize Farsi/Arabic words, some methods have been applied. Spectral features using 2D Fourier transform have been used. Each image is transformed into polar coordinates [6]. The use of hidden Markov model for Arabic word recognition has used in some works such as [7].

Based on the fact that there are a large number of words in Farsi/Arabic, these languages provide large lexicons which include common words. The problem associated with the large lexicons is the time that the input image has to be compared with the words in the lexicon. So, recognition time becomes a critical issue for large lexicons. In this case, a fast approach for eliminating unlikely candidates can be useful to reduce the time needed to compare entities. For more information the reader can refers to [8-9].

In this paper, an easy-to-follow approach to recognize machine-printed Farsi/Arabic subwords is discussed. The aim is to index all subwords in a printed text and identify the appropriate ones from a large lexicon. To this end, a machine-printed subword database including 9445 elements are employed. After removing dots in the text, the subwords are extracted, and contour-based Fourier descriptors are derived as the features.

II. FARSI/ARABIC ALPHABETIC CHARACTERISTICS

Unlike English, Farsi/Arabic scripts are cursively written from write to left. Farsi and Arabic are two similar alphabets having same main characteristics. Farsi/Arabic letters are composed of a main body, in addition to some secondary parts such as dots, zigzag strokes and so on. Some letters have the same main body and differ only in the number and position of their complementary characters, especially dots. An example of different letters with the same body is shown in Fig. 1. But, Farsi letter set includes all 28 Arabic letters plus four additional ones. "پ", "چ", "ژ" and "گ" are these extra letters.

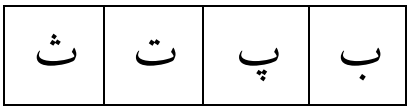

Figure 1. Different letters with a same body

Since Farsi is cursive, most letters in a word are connected to each other. But seven Farsi letters are not allowded to be connected to their adjacent ones. These letters only have isolated form, even if they are used in the middle of a word. Therefore, they make a small gap between characters to form subwords. In general, a word is comprised of one or several subwords. Moreover, this cursive feature causes that the shape of some Farsi/Arabic letters depends on their position in a subword. It can be