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

Language Discrimination and Font Recognition in Machine Printed Documents Using a New Fractal Dimension

Akram Alsadat Hajian Nezhad
Electrical and Computer Engineering
Department, Semnan University, Semnan, Iran
a_hajiannezhad@sun.semnan.ac.ir

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

***Abstract*—This paper focuses on language separation and font** recognition in multilingual **and multi-font texts. The purpose of this task is to improve performance of general OCR systems, dealing with omni-fonts and different languages. The proposed method is based on an innovative fractal dimension measurement. The extracted features with this method are independent of document contents and considers language and font recognition problem as texture identification task. Experimental results on three different languages namely, Farsi, Arabic and English with their most popular fonts show that the proposed method not only separates these languages but recognizes their font types accurately.**

Keywords-Optical Character Recognition (OCR), Optical Font Recognition (OFR), Language Discrimination, Fractal Dimension (FD).

I. INTRODUCTION

Nowadays, OCR systems are utilized by many individuals to convert scanned text images into machine readable form. Every OCR system is made of several modules such as image acquisition, preprocess, layout analysis, character recognition and document regeneration [1]. To increase the accuracy of these systems, some new modules are added every day. Language identification and font recognition are two pre-processing stages recently emerged in many OCR systems. Multilingual OCR systems must deal with variety of languages and lack of such ability decreases their recognition rates. Moreover, the operation of those OCR systems handling multi-font document images is more difficult than those deals with single-font document.

There are different language identification and font recognition systems based on SVM, Wavelet transform, Gabor filter, Sobel-Robert gradient, and Fractal dimension for Latin documents. However, due to the complexities of Farsi and Arabic languages, number of papers in these fields are limited.

The utilized technique for font identification problem in [1], is based on combination of directional gradients, Sobel and Roberts for identifying ten popular Farsi fonts. In [2], Sami Ben Moussa used two fractal dimension methods called BCD and DCD for the purpose of ten Arabic font recognition.

In [3], a multi-channel Gabor filtering technique is proposed for English font recognition. In [4], a font recognition method based on empirical mode decomposition (EMD) is proposed. Five basic strokes was used to characterize the stroke attributes of six Chinese fonts. Ding et al employed a 3-level wavelet transform for font identification of seven Chinese fonts [5].

II. FRACTAL GEOMETRY AND DIMENSION

In 1983, Mandelbrot established fractal geometry to describe every complex phenomenon that Euclidean geometry fails. Fractal geometry contains different areas and one of the most important one is fractal dimension (FD).

Fractal geometry, unlike Euclidean geometry, deals with fractional objects. In terms of fractal geometry, fractal objects have these three properties:

1) Being self similar.

Self similarity categorize to three categories:

- Perfect self similar objects such as Broccoli cabbage.
- Imperfect self similar objects such as mountains.
- Statistical self similar objects such as text document images.

2) Being complicated in tiny scales.
3) Having fractured dimensions.

Researches show that a huge number of environs objects are located in statistical self similar objects category, including text images [2].

So we decided to use fractal dimension for font recognition. In this paper we proposed an innovative fractal dimension method and then use it for the purpose of language and font recognition.

According to [6], all the fractal dimensions algorithms obey these three stages:

- Choosing a measuring step.