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.

- Calculating a quantity based on measuring step.
- Estimate the fractal dimension based on the slope of the regression line between log (computed quantities) versus log(step sizes).

III. THE PROPOSED FRACTAL DIMENSION

As mentioned before, there are three common stages, for every fractal dimension method:

- The first phase is choosing a measuring step called $S_1, S_2, S_3, \dots, S_N$.
- Second phase is calculating a quantity based on the first step $F(S_1), F(S_2), F(S_3), \dots, F(S_N)$.

$$\begin{aligned} S_1 &\rightarrow F(S_1) \\ S_2 &\rightarrow F(S_2) \\ S_3 &\rightarrow F(S_3) \\ &\vdots \quad\quad \vdots \\ S_N &\rightarrow F(S_N) \end{aligned}$$

- And the third phase is estimating FD using the slope of the regression line between log(computed quantities) versus log (step sizes).

According to the above definitions, we proposed a new fractal dimension algorithm based on Fourier transform coefficients and polar histogram. For this purpose, first, Fourier transform is applied to input images. For every pixel, one coefficient is obtained which has real and imaginary parts. Using these two parts, one radius and one angle is calculated according to (1):

$$a + bj \rightarrow \begin{cases} \theta = \tan^{-1}(\frac{b}{a}) \\ R = \sqrt{a^2 + b^2} \end{cases} \quad (1)$$

Then, angular range of 0 to 2π is quantized into 32 bins, as shown in Fig.1. Each radius which located between two quantized bins $R_n < R < R_{n+1}$ and each obtained angle which located between two quantized angles $\theta_n < \theta < \theta_{n+1}$, is assigned to the corresponding bin and polar histogram is formed.

After calculating described polar histogram, for each R_n there are 32 numbers which can be utilized for fractal dimension calculations as bellow (suppose Ri = 1):

Area	*Scale* $\rightarrow \theta$	$histogram$
$\theta_1 < \theta < \theta_2$	1	h_{11}
$\theta_2 < \theta < \theta_3$	2	h_{12}
$\vdots$	$\vdots$	$\vdots$
$\theta_n < \theta < \theta_{n+1}$	n	h_{1n}

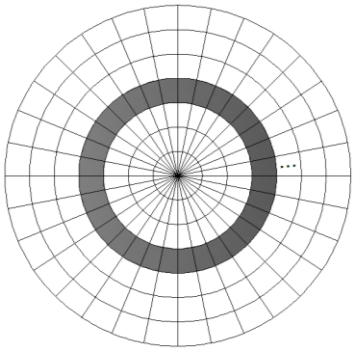


Fig.1:Polar histogram

For every $Ri = r$, the fractal dimension D is estimated through the slop of a linear regression between $log\ (log(h_{rn}))$ and $log(n)$ according to (2). The fractal dimension is usually a fractured number between 1 to 3.

$$D = \frac{log\ (log(h_{rn}))}{log(n)} \quad (2)$$

Experimental result shows that using the function $log(h_{rn})$ converts the proposed FDs to this proper range. In the proposed algorithm, the range and the number of bins for radiuses and angles depend on the specified application. In our experiments, 32 ranges for angles and 6 ranges for radiuses are chosen.

IV. DATA SET CONSTRUCTION

Every day new fonts in books, journals, official letters, and blogs are appeared. But in each language there are some special fonts which are widely used. Lotus, Mitra, Nazanin, Traffic, Yaghut, Zar, Homa, Titr, Tahoma, and Times New Roman are the most popular Farsi fonts [1]. Their differences are shown in Fig.2.

Ben Moussa cited ten popular Arabic fonts as Ahsa, Andalus, Arabic_transparant, Badr, Buryidah, Dammam, Hada, Kharj, Koufi , Naskh [2]. And the most popular fonts in English as reported in [3] are Arial, Bookman, Century, Comic, Courier, Impact, Modern, and Time New Roman. We used the same fonts to deal with Farsi, Arabic and English font recognition.

To evaluate the efficiency of the proposed algorithm, some text images are written by three languages (Farsi, Arabic, and English) and different fonts (as mentioned before). The dataset includes 1400 samples for each language (50 text images for each font).

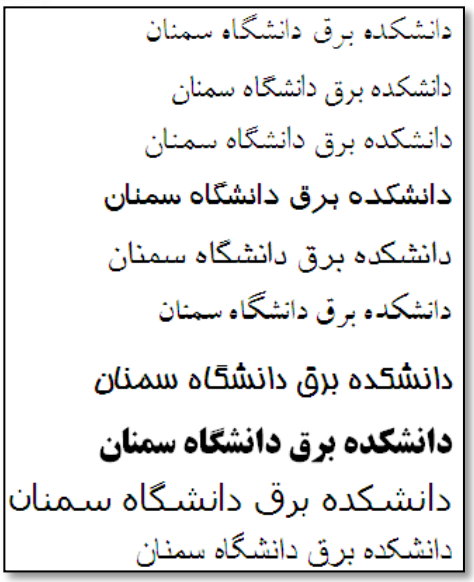


Fig.2: Farsi document written by several fonts: Lotus, Mitra, Nazanin, Traffic, Yaghut, Zar, Homa, Titr, Tahoma, and Times New Roman.

For Arabic and English font recognition, ALPH-REGIM datasets, are utilized [2]. Since there is not standard Farsi dataset for evaluating font recognition algorithms, we provided this dataset ourselves. We used ten popular Farsi fonts as named before. This dataset includes 2000 samples of 10 typefaces with 4 different font sizes (each font presented by 200 samples).

Since ALPH-REGIM dataset is in different sizes, for achieving better recognition rates, we reconstruct 512 × 512 text blocks from primary dataset, according to the texture reconstruction algorithm [1]. For reconstructing, first we find all the lines in the input text image and separate them. Then, all separated lines are aligned in a straight arrangement. Afterwards, these lines are segmented into 512 pixel width. Finally, these broken lines are concatenated vertically to construct 512 × 512 image blocks. Due to lack of enough space, some 128 × 128 text block obtained from Fig.3, Fig.5, Fig.7 are shown in Fig.4, Fig.6, Fig.8.

V. The Advantages of Feature Extraction Method

High-dimensional feature set is one of the problems that many pattern recognition algorithms suffer from it. Such feature vectors make the identification process more complicated. In these situation, using dimensionality reduction algorithms is essential. One of the most important advantages of fractal dimension algorithms as feature extraction method is that these methods lead to low-dimensional feature vectors.

Fig.3: An Arabic text image with several lines.

Fig.4: Three 128×128 text images obtained from Fig.3.

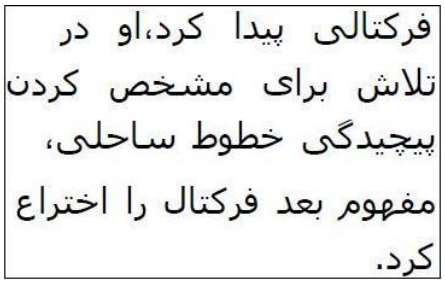


Fig.5: A Farsi text image with several lines.

Fig.6: Three 128×128 Farsi text images obtained from Fig.5.

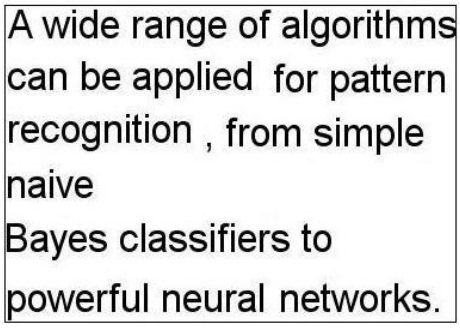


Fig.7: An English text image with several lines.

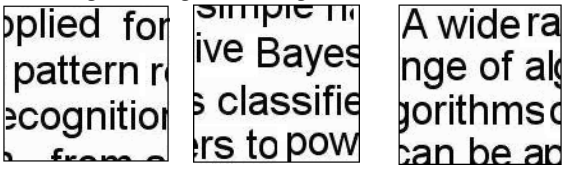


Fig.8: Three 128×128 English text images obtained from Fig.7.

The proposed algorithm produces a 6D feature vector which is very low compared with other feature extraction methods reviewed in the literature.

Document skew is a distortion that every OCR system may encounter. Since many steps used in

these systems only work on aligned images as input, skew correction became a vital part of these systems. One of the most important advantages of the proposed feature extraction algorithm is that it is robust against skew. Rotating an object in the space does not change its dimension. It is true for our fractal dimension measurements too. As shown in Fig.4, the fractal dimension of text images in part (a), (b) ,(c) and (d) are the same, so we won't need skew correction part in our proposed algorithm.

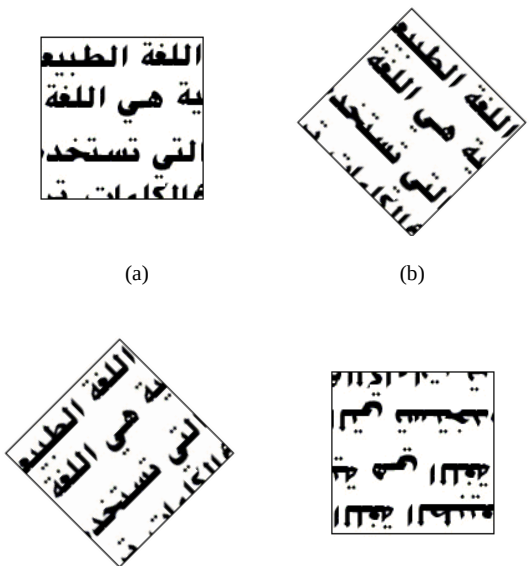
VI. EXPERIMENTS AND RESULTS

After extracting features by our innovative fractal dimension method for both language identification and font recognition, RBF and KNN classifiers are used. The obtained results are presented in section VII and VIII.

VII. LANGUAGE RECOGNITION

As mentioned before, using the proposed algorithms, each text sample is expressed by a 6D feature vector. For language recognition, classifiers deal with 3 classes: Farsi, Arabic and English.

In this case, the utilized dataset contains 1400 samples in which 800 samples used for training and 600 samples for testing. The obtained recognition rates using RBF and KNN classifiers are presented in TABLE.1. Since Farsi and Arabic languages are very similar, some errors are inevitable. Two confusion matrices in TABLE.2 and TABLE.3 show the results.



(a) (b)

(c) (d)

Fig.4: (a) 0° deviation, (b) -45° deviation, (c) 45° deviation, (d) 180° deviation.

TABLE.1: LANGUAGE RECOGNITION RATES (%)

Classifiers	RBF	KNN
Recognition Rates	92.35	90.71

TABLE.2: THE CONFUSION MATRIX USING RBF.

Labs	Farsi	Arabic	English	Total
Farsi	182	19	3	204
Arabic	17	178	4	199
English	1	3	193	197

TABLE.3: THE CONFUSION MATRIX USING KNN.

Labs	Farsi	Arabic	English	Total
Farsi	179	27	2	208
Arabic	19	169	3	191
English	2	4	195	201

VIII. FONT RECOGNITION

In this section some comparative studies between our proposed method and related works are presented. The comparison includes feature dimensions, recognition rates, and their robustness against skew (TABLE 4, TABLE 5, TABLE 6).

One of the major factors of every pattern recognition algorithm is the length of feature vector. The more feature vector length, the more complexity in classifiers. One of the most important advantages of the proposed algorithm is its low dimensionality, especially in comparison with SRF and Gabor algorithms.

According to TABLE 4, TABLE 5 and TABLE 6, it would be clear that the average recognition rates obtained by our algorithm are higher than the maximum recognition rates obtained from other methods.

Skew correction became an inseparable part of every module in OCR systems. Most of the time skew correction algorithms cannot restore primary images correctly so robustness against skew is important. According to TABLE 7, one of the advantages of the proposed method is being robust against skew. Since Gabor and SRF methods are directional, they are very sensitive to skew.

TABLE.4:COMPARATIVE STUDY BETWEEN OUR METHOD AND [1].

Used typefaces	Farsi		
Technique	SRF	OURS	
Number of Features	512	6	
Classifiers	MLP	RBF	KNN
Recognition (%)	94.16	96.91	96.3

TABLE.5:COMPARATIVE STUDY BETWEEN OUR METHOD AND [2].

Used typefaces	Arabic			
Technique	BCD-DCD		OURS	
Number of Features	4		6	
Classifiers	RBF	KNN	RBF	KNN
Recognition (%)	98	96.2	98.41	97.69

TABLE.6: COMPARATIVE STUDY BETWEEN OUR METHOD AND [3]

Used typefaces	English	
Technique	Gabor	OURS

Number of Features	32	6	
Classifiers	WED	RBF	KNN
Recognition (%)	99.1	99.75	99.46

TABLE 7: ROBUSTNESS AGAINST SKEW.

Technique	SRF	BCD-DCD	Gabor	Ours
Skew Robustness	NO	YES	NO	YES

IX. CONCLUSION

In this paper, we proposed a new algorithm for language identification and font recognition in Farsi, Arabic, and English text images. The proposed fractal dimension algorithm is based on the combination of Fourier transform coefficients and polar histogram. It maps each text image into a 6D feature vector. The most important advantages of our proposed method are: low dimensional feature vector, low computational complexity, its high recognition rates and being robust against skew.

REFERENCES

[1] Hossein Khosravi, and Ehsanollah Kabir Farsi font recognition based on Sobel–Roberts features, Elsevier, Pattern Recognition Letters 31 (2010) 75–82.

[2] Sami Ben Moussa, and Abderrazak Zahour, and Abdellatif Benabdelhafid, and Adel M. Alimi ,New features using fractal multi-dimensions for generalized Arabic font recognition,, Elsevier, Pattern Recognition Letters 31 (2010) 361–371.

[3] Yong Zhu, Tieniu Tan, and Yunhong Wang, "Font Recognition Based on Global Texture Analysis", IEEE TRANSACTIONS ON PATTERN NALYSIS AND MACHINE INTELLIGENCE, VOL. 23, NO. 10, OCTOBER (2001).

[4] Yang Z. et al., 2006. An EMD-based recognition method for Chinese fonts and styles. Pattern Recognition Lett. 27, 1692–1701.

[5] Xiaoqing Ding, Member, IEEE, Li Chen, Member, IEEE, and Tao Wu, Student Member, IEEE , "Character Independent Font Recognitionon a Single Chinese Character" , IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 29, NO. 2, FEBRUARY (2007).

[6] R. Lopes, and N. Betrouni, "Fractal and multifractal analysis: A review", Elsevier, Medical Image Analysis 13 (2009) 634–649.