Font Recognition Using Variogram Fractal Dimension

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

Abstract: *This paper is dealing with font recognition problem in Farsi, Arabic, and English documents. It considers font recognition as texture identification task and the extracted features are independent of document content. The proposed method is based on one of the fractal dimension techniques which is called Variogram Analysis. The average recognition rates using RBF, and KNN classifiers are respectively %95.5, %96 for Farsi fonts, and % 96.9, %98.84 for Arabic fonts, and % 98.21, %99.6 for English fonts. The most important advantages of our algorithm are low feature dimensions, low computational complexity, and high speed compared with the previous efforts.*

Key words: Optical Character Recognition (OCR), Optical Font Recognition (OFR), Fractal Dimension (FD), Variogram Analysis.

1. Introduction

Nowadays optical character recognition (OCR) systems are utilized by many people to convert scanned text images into machine-encoded forms. Previous OCR systems were made of several modules such as image acquisition, preprocess, layout analysis, character recognition and document regeneration. To increase their accuracy, Optical Font Recognition (OFR) module recently added to many OCR systems.

The idea of adding OFR module to an OCR system is the fact that, the average discriminating rates in an OCR system with font recognition is significantly higher than those without OFR part.

Since, font recognition problem is still a challenging issue in many OCR systems, we focused on this problem in this paper. The utilized dataset for evaluating of our algorithm includes Farsi, Arabic and English text images.

2. Related Works

Font recognition is becoming a fundamental issue in document analysis and recognition, in this section we present a short review of some previous related works. All the previous OFR algorithms are based on one of these two realms: typographical or textural features.

Typographical features include character skews, character weights, space width, projection in upper, centre and lower zones of the line and etc, as shown in Fig.1. Although, typographical based algorithms usually perform font recognition well, they are sensitive to noise

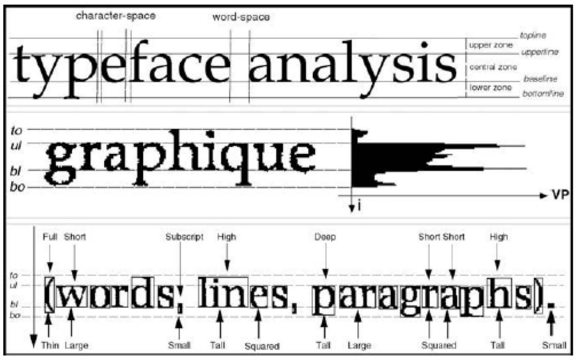

Fig. 1: Typographical features [2].

and require high resolution scanned images. Some previous related works using these kinds of features are [5], [6], [7].

Previous efforts demonstrated that textural features are more applicable than typographical features. Some of textural based algorithms for OFR are Wavelet algorithm, Gabor filter, Sobel-Robert gradient, Fractal dimension and etc [1], [2], [3], and [4].

Now, we shortly review some textural algorithms which will be used for comparison purposes in the succeeding sections.

As recent researches show, fractal dimensions are very useful to quantify the complexity of images. Whereas Farsi and Arabic scripts are complex patterns, these methods sound proper for font recognition. Sami Ben Moussa and et.al, proposed some fractal dimension methods for the purpose Arabic font recognition [1]. The two utilized fractal methods in this paper are Box Counting Dimension (BCD), and Dilation Counting Dimension (DCD).

In [2], Hossein Khosravi and et.al used a gradient method for Farsi font recognition problem. This method which called SRF is based on combining two different kinds of directional gradients, Sobel and Roberts.

And finally in [3], Yong Zhu and et.al used directional multi-channel Gabor filter for English fonts recognition.

3. Feature Extraction Using Fractal Geometry

Feature extraction is a crucial step in every pattern recognition system. Recently, fractal geometry was used

978-1-4673-1148-9/12/$31.00©2012IEEE

for feature extraction task in different applications. Mandelbrot established fractal geometry to describe the complexity of natural phenomenon in 1983. After that, a lot of analytical methods for fractal feature extraction have been proposed and numbers of fractal geometric methods are increasing every day.

Fractal geometry contains different areas and one of the most important is fractal dimension. Fractal dimension (FD) proceed to describe the dimension of objects which Euclidean geometry fails to describe. In fact, Euclidean geometry only deals with objects with integer dimensions but fractal geometry deal with fractional objects.

In terms of fractal geometry, fractal objects have three properties:

- Being self similar.
- Being complicated in tiny scales.
- Having fractured dimension.

Self similar objects can be categorized into three categories: perfect self similar objects such as Broccoli cabbage, imperfect self similar objects such as mountains, and statistical self similar objects. Researches show that a huge number of environs objects are located in third category, and according to [1] text images are not exception.

Sami Ben Moussa and et.al in [1] applied combination of two fractal dimensions BCD and DCD for Arabic font recognition problem. Furthermore some previous efforts such as [2] and [3] show that directional features are suitable for font recognition. So, we decided to use a directional fractal dimension method.

There are a lot of methods for calculating fractal dimension such as Box counting, Differential box counting, Extended counting, Triangular prisms, Covering blanket, Power spectrum, Isarithm, Multifractal Renyi's dimensions, Discrete wavelet transform, Variogram and etc, we chose Variogram method which is a directional fractal dimension method.

The Variogram method, is a FD method which describes differences in value between pairs of sample with a given relative orientation. The mathematical definition of the Variogram algorithm is as follows [8], [9], [10]:

$$\gamma(h) = \frac{1}{2n(h)} \sum_{i=1}^{n(h)} [Z_i - Z_{i+h}]^2 \qquad (1)$$

$$FD = 2 - \frac{1}{2}\left(\frac{log\ (\gamma(h))}{log\ (h)}\right) \qquad (2)$$

In the "Equation (1) and (2)":

- h is the lag step (the utilized h in this paper are $h = 1, 2, \ldots, 6$).
- Z_i, Z_{i+h} are the image pixel values.

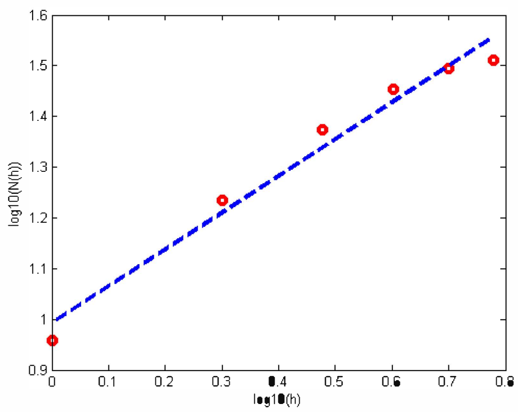

Fig.2: graphical figure based on log(h) and log(γ(h)).

- n(h) is the number of pairs or differences for each lag or ***h***.

To estimate FD through these equations:

1) Choose some directions for feature extraction (we used vertical, horizontal, and diagonal direction in this paper).
2) For computing vertical direction $Z_i \sim im(x,y)$ and $Z_{i+h} \sim im(x+h,y)$.
3) For computing horizontal direction $Z_i \sim im(x,y)$ and $Z_{i+h} \sim im(x,y+h)$.
4) For computing diagonal direction $Z_i \sim im(x,y)$ and $Z_{i+h} \sim im(x+h,y+h)$.
5) Now through "Equation (1)", the Variogram dimension can be estimated using h = $1, 2, \ldots, 6$.
6) For every h one γ(h) is obtained, for estimating the Variogram dimension we need a linear regression between log[h] and log[γ(h)] . A graphical figure based on these two parameters is shows in Fig.2.
7) Finally for computing fractal dimension through "Equation (2)", replace $log\ (\gamma(h))/log\ (h)$ with obtained slope.
8) Moreover we used the intercept of the linear regression as second feature, whereas 3 directional FD were computed our feature vector is 6D.

4. Data Set and Data Reconstruction

Every day new fonts in books, journals, official letters, and blogs are appeared. In every language there are some special fonts which widely used. Sami Ben Moussa and et.al in their paper, introduced ten popular Arabic fonts: Ahsa, Andalus, Arabic_transparant, Badr, Buryidah, Dammam, Hada, Kharj, Koufi , Naskh [1].

According to [2], Lotus, Mitra, Nazanin, Traffic, Yaghut, Zar, Homa, Titr, Tahoma, and Times New Roman are the most popular Farsi fonts.

Zhu explored that the most popular fonts in English language are [3]: Arial, Bookman, Century, Comic, Courier, Impact, Modern, and Time New Roman. We used the same fonts mentioned above in our work for Farsi, Arabic and English font recognition.

To evaluate the efficiency of the proposed OFR, some text images with different kinds of fonts are utilized. The Arabic and English text images used in this paper are ALPH-REGIM datasets, provided by Sami Ben Moussa [1].

Unlike Arabic and English, there is not standard Farsi dataset for evaluating OFR algorithms. So, we provide such dataset ourselves to test the efficiency of our algorithm. We used ten popular Farsi fonts as introduced in [2]. This dataset includes 2000 samples of 10 typefaces with 4 different font sizes, each font presented by 200 samples. We used 100 samples for training and 100 samples for testing.

Since our proposed algorithm works in block level and ALPH-REGIM dataset are in different sizes, for achieving better recognition rate, we reconstruct 512×512 text blocks from primary dataset according to the texture reconstruction algorithm described in [2], three text block shown in Fig.3, Fig.5, Fig.7. For reconstructing 512*512 text block 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 the lack of enough space for showing, 128*128 text block obtained from Fig.3, Fig.5, Fig.7 are shown in Fig.4, Fig.6, Fig.8.

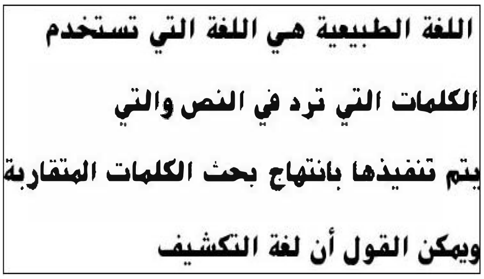
اللغة الطبيعية هي اللغة التي تستخدم
الكلمات التي ترد في النص والتي
يتم تنفيذها بانتهاج بحث الكلمات المتقاربة
ويمكن القول أن لغة التكشيف

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

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

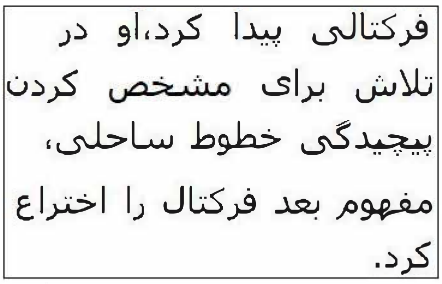
فرکتالی پیدا کرد،او در
تلاش برای مشخص کردن
پیچیدگی خطوط ساحلی،
مفهوم بعد فرکتال را اختراع
کرد.

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

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

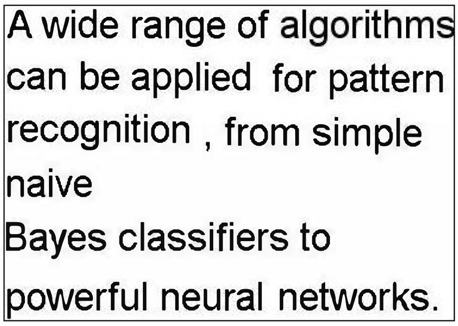
A wide range of algorithms
can be applied for pattern
recognition , from simple
naive
Bayes classifiers to
powerful neural networks.

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

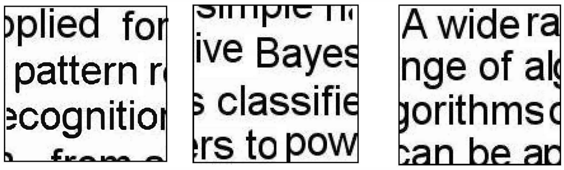

Fig.8: Three 128*128 English text images obtained from Fig.7.

5. Experimental Results

After extracting features by Variogram method, RBF and KNN classifiers are used to classify the font datasets. In this algorithm, each sample is expressed by a 6D feature vector. The utilized features are obtained using vertical, horizontal, and diagonal directions.

Two figures for showing the capability of extracted features to separate different fonts were plotted in Fig.9 and Fig.10. These figures show that how these 6 features can separate different fonts. These figures are based on the average extracted features of 10 Farsi fonts with the font size 16.

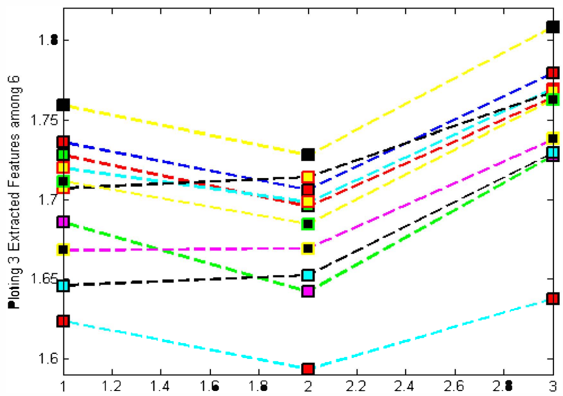

Fig.9: Plotting the average of 3 extracted features among 6 (feature 1, 2, and 3 which contains the slopes of linear regression for the 10 Farsi fonts with font size 16).

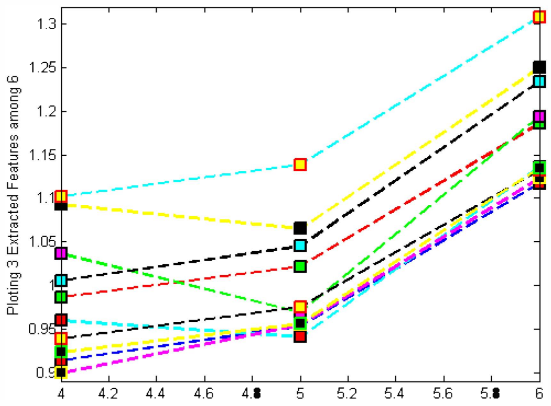

Fig.10.Plotting the average of 3 extracted features among 6 (features 4, 5, and 6 which contains the intercepts of linear regression for the 10 Farsi fonts with font size 16).

Two confusion matrixes using RBF and KNN classifiers are presented in TABLE 1, TABLE 2. According to these tables, good discrimination rates are achieved and few errors occur among some similar prototypes, these two confusion matrixes can obviously show the capability of font identification through the proposed algorithm.

Some error estimation graphs using KNN and RBF classifiers were plotted in Fig.11, Fig.12, and Fig.13. As mentioned before occurring these errors are due to the existence of some similar fonts in our utilized dataset.

According to TABLE 3 and Fig.11, Fig.12, and Fig.13 the best recognition rates and minimum errors are related to English fonts. It shows that the complexity of English text images with discrete characters are comparatively less than Farsi and Arabic text images whit cursive scripts.

TABLE 1: The confusion matrix of 10 Farsi fonts with 4 font size (the test dataset) using RBF classifier.

True Labs Estimated Labs	1	2	3	4	5	6	7	8	9	10	Total
1	91	0	4	0	0	7	0	0	0	0	102
2	0	100	0	0	0	0	0	0	0	0	100
3	4	0	95	0	0	4	0	0	0	0	103
4	0	0	0	100	0	0	0	0	0	0	100
5	0	0	0	0	94	0	0	0	0	8	102
6	5	0	1	0	0	89	0	0	0	0	95
7	0	0	0	0	0	0	100	0	0	0	100
8	0	0	0	0	0	0	0	100	0	0	100
9	0	0	0	0	0	0	0	0	100	0	100
10	0	0	0	0	6	0	0	0	0	92	98
Total	100	100	100	100	100	100	100	100	100	100	1000

TABLE 2: The confusion matrix of 10 Farsi fonts with 4 font size (the test dataset) using KNN classifier.

True Labs Estimated Labs	1	2	3	4	5	6	7	8	9	10	Total
1	94	0	6	0	0	6	0	0	0	0	106
2	0	100	0	0	0	0	0	0	0	0	100
3	2	0	93	0	0	4	0	0	0	0	99
4	0	0	0	100	0	0	0	0	0	0	100
5	0	0	0	0	92	0	0	0	0	5	97
6	4	0	1	0	0	90	0	0	0	0	95
7	0	0	0	0	0	0	100	0	0	0	100
8	0	0	0	0	0	0	0	100	0	0	100
9	0	0	0	0	0	0	0	0	100	0	100
10	0	0	0	0	8	0	0	0	0	95	103
Total	100	100	100	100	100	100	100	100	100	100	1000

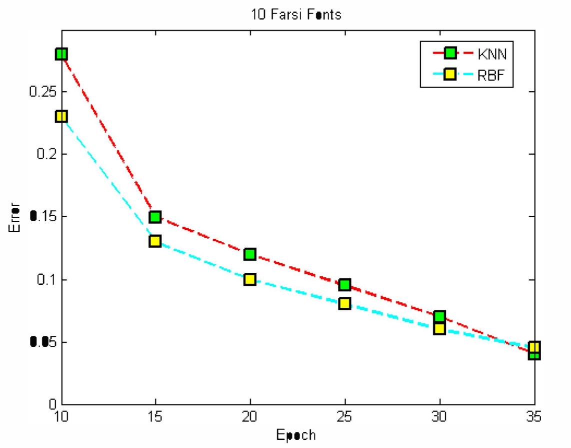

Fig.11: Estimating the classification error of Farsi fonts using KNN and RBF classifiers.

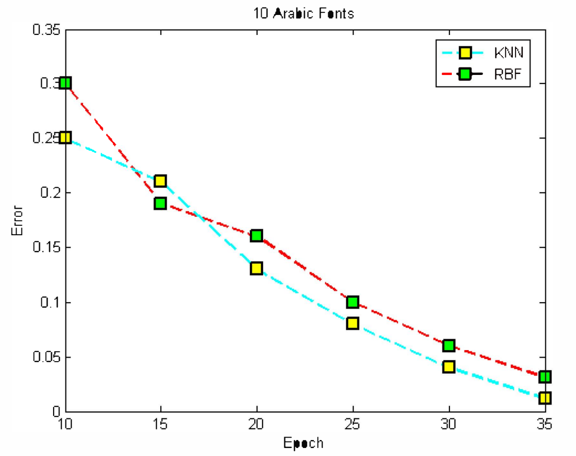

Fig.12: Estimating the classification error of Arabic fonts using KNN and RBF classifiers.

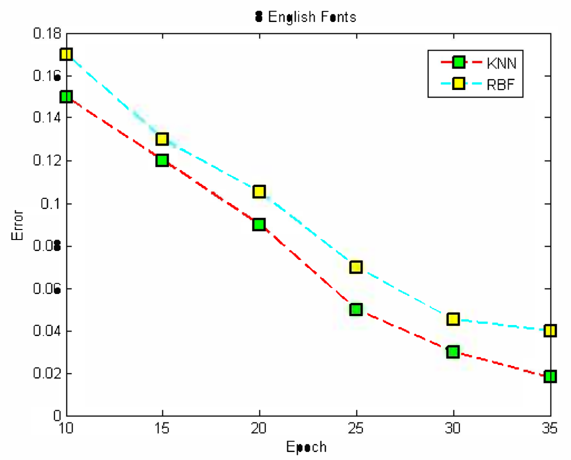

Fig.13: Estimating the classification error of English fonts using KNN and RBF classifiers.

TABLE 3: The average discrimination rates of our work (%) .

Font / Classifiers	Farsi	Arabic	English
RBF	95.5	96.9	98.21
KNN	96	98.84	99.6

Some comparison of our work with some related works which mentioned in section 2, are presented in TABLE 4, TABLE 5 and TABLE 6. According to TABLE 4, for Farsi font recognition our algorithm has some advantages over Sobel- Robert algorithm. Not only the average recognition rates of the proposed method are higher but also our fractal based features have lower dimensionality. By comparing the results in TABLE 5 and TABLE 6, it would be clear that recognition rates obtained in our algorithm is higher than the maximum recognition rates obtained in BCD-DCD and Gabor algorithms using the same datasets.

Moreover our experimental results showed that the proposed method is a high speed algorithm, it is about 254 times faster than BCD-DCD, and 18 times faster than Gabor method.

TABLE 4: Comparative study between SRF and our algorithm

Technique	SRF	Ours	Ours
Feature numbers	512	6	6
Used typefaces	Farsi	Farsi	Farsi
Classifiers	MLP	RBF	KNN
Time (s)	0.060	0.159	0.159
Recognition (%)	94.16	95.5	96

TABLE 5: Comparative study between BCD-DCD and our algorithm

Technique	BCD-DCD	Ours	BCD-DCD	Ours
Feature numbers	4	6	4	6
Used	Arabic	Arabic	Arabic	Arabic
Classifier	RBF	RBF	KNN	KNN
Time (s)	40.470	0.159	40.470	0.159
Recognition	98	96.9	96.2	98.84

TABLE 6: Comparative study between Gabor and our algorithm

Technique	Gabor	Ours	Ours
Feature numbers	32	6	6
Used typefaces	English	English	English
Classifiers	WED	RBF	KNN
Time (s)	2.848	0.159	0.159
Recognition (%)	99.1	98.21	99.6

6. Conclusion

In this paper, we proposed a new algorithm for font recognition problem, in Farsi, Arabic, and English text images. Our proposed algorithm is based on a directional fractal dimension. According to previous works, directional methods and fractal dimensions perform well for feature extraction in OFR system, so we chose a combination of these methods, which is called Variogram fractal dimension. With these directional fractal feature extraction method, each document image is mapped into a 6D feature vector. The most important advantages of the proposed method over previous approaches are: low dimensional feature vector, low computational complexity, it 's high speed and high recognition rates.

References

[1] 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.

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

[3] Yong Zhu, Tieniu Tan, and Yunhong Wang, 2001. Font recognition based on global texture analysis, IEEE Trans. Pattern Anal. Machine Intell. 23 (10), 1192–1200.

[4] Xiaoqing Ding, Li Chen, and Tao Wu, Character Independent Font Recognitionon a Single Chinese Character, IEEE Transactions on Pattern Analysis and Machine Intelligent, Vol. 29, No. 2, February 2007.

[5] Chaudhuri, B.B., and Garain, U., Automatic detection of italic, bold and all-capital words in document images. In: 14th Internat. Conf. on Pattern Recognition(1998) pp. 610–612.

[6] Zramdini, A., Ingold, R., Optical font recognition using typographical features. Pattern Anal. Machine Intell (1998), 20 (8), 877–882.

[7] Jeong, C.B. et al. Identification of font styles and typefaces in printed Korean documents (2003). Lect. Notes Comput. Sci. 2911, 666–669.

[8] R.D. Valdez-Cepedaa, E. Olivares-SaÂenz, Fractal analysis of Mexico's annual mean yields of maize, bean, wheat and rice, Elsevier , Field Crops Research 59 (1998) 53±62

[9] T. Babadagli, and K. Develi, Fractal characteristics of rocks fractured under tension, Elsevier, Theoretical and Applied Fracture Mechanics 39 (2003) 73–88.

[10] Ricardo David Valdez - Cepeda, and Daniel Hern _ Andez-Ram_ Irez, Fractality of Monthly Extreme Minimum Temperature, Fractals, Vol. 11, No. 2 (2003) 137{144}.

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