Recognition of Individual Handwritten Letters of the Farsi Language using a Decision Tree

Atefe Matin Niya
Department of Computer Engineering
Dezful Branch
Islamic Azad University, Dezful, Iran

Hedieh Sajed
Department ofComputer Engineering
Amirkabir University of Technology,
Tehran, Iran

ABSTRACT

In this study, in order to recognize Farsi handwritten letters, firstthe pre-processing operation is done on the letters' images including normalization, thinning, reduction, noise reduction, etc.,and then the feature vector of the letters is extracted using the first to the forth momentums fromthe second level of wavelet transform and contourlettransform.A combination of decision-tree methods is used for the final recognition of letters.

The database used in this study is the "Hoda" handwritten letter collection. The mean recognition rate in this combinational method is 97.89%.

Keywords

recognition of Farsi letters, handwritten recognition, pattern application, decision tree, wavelet transform, contourlet transfor.

1. INTRODUCTION

Optical character recognition is animportant, applied and very active branch inpattern recognition.The main problem in this branch of computer sciences is to recognize characters, subwords and words whose images are available. Solving this problem can have a tremendous effect on the effective relationship between man and machine and can also be a great help in the automation of written document processing. Generally, a document recognition system can be on-line or off-line. If time information on the writing sequence is not available, the system will be off-line and its data willbe usually obtained by scanning the images of previously written texts. If the points' time sequence at the time of writing is available, on the other hand, the system will deal with on-line data.

This study aims to present a new method for the off-line recognition of Farsi individual letters.

To date, relatively fewer studies have been conductedonthe off-line recognition of Farsi letters compared to Latin languages. A concise review of the most important studies in this field will be presented in section 2. Section 3 will examine the proposed method and its differentsubsections. Section 4 will present the experimental results of theproposed method and section 5 will expressthe conclusion and future solutions.

2. LITERATURE REVIEW

Recently, different approaches for writeridentification have been proposed. A scientificvalidation of individuality of handwriting is performedby Srihari et al. [1]. In this study handwriting samplesof 1500 individuals, representative of the U.S.population with respect to gender, age, ethnic groups,etc., were obtained. The writer can be identified basedon Macro features and Micro features that areextracted from handwritten documents. Said et al. [2]proposed a global approach based on multi-channelGabor filtering, where each writer's handwriting isregarded as a different texture. Bensefia et al. [3] usedlocal features based on graphemes extracted fromsegmentation of cursive handwriting. Then writeridentification is performed by a textual basedinformation retrieval model. Schomaker et al. [4]presented a new approach, using connected-componentcontours codebook and its probability-densityfunction. Also combining connected-componentcontours with an independent edge-based orientationand curvature PDF yields very high correctidentification rates. Schlapbach et al. [5] propose aHMM based approach for writer identification andverification. Bulacu [6] evaluated the performance ofedge-based directional probability distributions asfeatures in comparison to a number of non-angularfeatures. Marti et al. [7] extracted a set of features fromhandwritten lines of text. The features extractedcorrespond to visible characteristics of the writing, forexample, width, slant and height of the three mainwriting zones. In ois et al. [8] a new feature vectorisemployed by means of morphologically processingthehorizontal profiles of the words.Because of the Lack of the standard database for writer identification, the comparison of the previous

3. THE PROPOSED METHOD FOR FARSI LETTER RECOGNITION

2.1The pre-processing phase

In the pre-processing phase, images are first transformed into equal dimensions of 24×24 pixels for normalization. Since the image lengths and heights may not change by the same ratio (i.e. they may stretch in one direction), images must bethinned. Figure (1) shows an example of thinning.

Fig 1: The thinning process

The noise caused by optical scanners leads to points such as stains, discrete line segments, connection between lines, filling of the existing holes in someletters' images, etc.

Also, different distortions should be considered including local changes, the curvature of letters' angles, letters' deformation or corrosion. Prior to the recognition stage, these defects should be fixed. The most important reason for noise reduction is to reduce error in the recognition phase. Also, noise reduction will reduce the image file size which in turn will reduce the time needed for future processing and storing. In this method, by determining a proper threshold level, we shall reduce noise and then transform grey images to black & white ones. This operation reduces the data size considerably.

In the next stage, images should be binarized so that an operation such asthinning is possible.

2.2 The representation phase (feature extraction)

After the pre-processing phase, properties must be extracted from images. The representation phase is a very important phase in OCR systems because the results obtained from this phase have direct impacts on the recognition phase quality. In the representation phase, each input pattern is assigned a feature vector or code which represents that pattern in the feature space and makes that pattern different from other patterns.

In this article, the feature extraction phase is divided into two parts:

The first partisfeature extraction from wavelet transform coefficients: first, we will apply two levels of wavelet transform (DWT) on images usingthe Haar scaling function. The wavelet transform will cause images to crosshigh-pass and low-pass filter sets continuously so that data rate remains unchanged. Then, among the four obtained subbands (cA, cH, cD, cV), we will select the most effective one by means of the trial and error method. The selected subbands in this experiment are 3 subbands of cA, cH, cVthat have the most similarities to the existing images. In the 2D wavelet transform, a 2D scaling function is required like $\varphi(x,y)$ in which each one is the product of two 1D functions. Except for multiplications that produce 1D results like $\varphi(x)\Psi(y)$, the four remaining multiplications produce aseparable scaling function:

$$\varphi(x,y)= \varphi(x)\ \varphi(y) \quad (2\text{-}1)$$

andseparable "direction sensitive" wavelets:

$$\psi^{H}(x,y)=\psi(x)\varphi(y) \quad (2\text{-}2)$$

$$\psi^{V}(x,y)=\varphi(x)\psi(y) \quad (2\text{-}3)$$

$$\psi^{D}(x,y)=\psi(x)\psi(y) \quad (2\text{-}4)$$

These wavelets measure function changes (severity changes for images) in different directions:

ψH: measures changes along columns (such as horizontal edges)

ψV: respond to changes along rows (such as horizontal edges)

ψD: corresponds to changes along diagonals.

2.2.1The Haar transform

One of the operations related to imaging in multi-precisionanalyses is the Haartransform whose basic functions are the simplest and oldest of orthogonal wavelets. The Haartransform is expressed as matrix as follows:

$$T=HFHT \quad (2\text{-}5)$$

where F is a N×N matrix, and H is the N×N Haartransform. The matrix transpose is required because H is not symmetric. For Haartransform, H includes Haar basic functions, i.e. hk(z). They were defined on the continuous range ofz ϵ[0, 1] for k= 0, 1, 2, … N-1, where N=2n.

The second phase of feature extraction from the contourlet transform: A contourlet transform is a non-separable directional 2D transform that is used to describe subtledetails and curves. Contourlet transforms efficiently describe smooth contourlets that are main and important components in natural images. Unlike other transforms which are createdfirst in a continuous area and are then discretized for data sampling, a contourlet transformstarts from a discretized area with the help of bank filters; then, it converges to a continuous area through an analytical multi-resolution framework. Like the first phase of images, we will obtain the contourlet transform coefficients in two levels and select the best subbands. To select the best subbands, we will use the trial and error method. In this phase, the best results were obtained from the union of four subbands.

Once the coefficients are obtained from the above transforms, image properties are extracted by the first to the fourth momentums, i.e. mean, var, skewness, and kurtosis; subsequently, they are stored in a file.

Using the feature extraction method and a 2D wavelet transform, 36 properties are obtained, whereas using a contourlet transform, 40 properties are obtained; finally, a vector with 76 properties isformed. We obtain this vector for all of the images in the database and store them in a file. The group number is added to the end of the feature vector. This numbering is done based on the following table:

Table 1.numbering is done based

Lable	Persian Spell	Lable	Persian Spell
0	ا	18	ط
1	ب	19	ظ
2	پ	20	ع
3	ت	21	غ
4	ث	22	ف
5	ج	23	ق
6	چ	24	ک
7	ح	25	گ
8	خ	26	ل
9	د	27	م
10	ذ	28	ن
11	ر	29	و
12	ز	30	ه
13	ژ	31	ی
14	س	32	ئ

15	ش	33	أ
16	ص	34	هـ OR ـهـ
17	ض	35	ـه

2.3 The recognition phase

In this phase, we will present a classification model for letter recognition based on the obtained feature vectors and Table 1.

2.3.1 The decision tree

Trees in artificial intelligence are used to represent different concepts such as sentence structures, equations, game modes, and so on.

Decision tree learning is a method for approximation of objective functions with discrete values. This method is resistant to data noise and can learn the seasonal composition of conjunctive statements. It is a popular inductive learning algorithm that has been used in several applications successfully.

A decision tree is a tree in which samples are classified in a way that they grow down from the root and finally reach leaf nodes.

In this article, the C4.5 decision tree is used for letter recognition.

2.3.2 The C4.5 decision tree

C4.5 builds decision trees from a set of training data in the same way as ID3 using the concept of information entropy. The training data is a set $S = s_1, s_2, \cdots$of already classified samples. Each sample $s_i = x_1, x_2, \cdots$is a vector where $x_1, x_2, \cdots$represent attributes or features of the sample. The training data is augmented with a vector $C = c_1, c_2, \cdots$where $c_1, c_2, \cdots$represent the class to which each sample belongs.
At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurses on the smaller sublists.

This algorithm has a few base cases.

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.

- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.

- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

4. IMPLEMENTATION AND EXPERIMENTAL RESULTS

MATLABsoftware has been used for implementation.

The "Hoda" data set is used as atraining and experimental data set. This set includes 88351 samples, of which 70645 samples are used for training and 17706 are used for experimentation.

In the proposed method, the mean recognition rate of Farsi letters was obtained as 97.89%.

Table 2 shows the implementation results in the confusion matrix and table 3 shows the final results for each classes indata base.

5. SUMMARY, DISCUSSION AND CONCLUSION

This article presents a simple method for the recognition of Farsi letters based on recognition with the C4.5 recognition method. Thus, first the feature vector was extracted by the second level wavelet transform and the second degree contourlet transform; then, the image vectors were classified using theses vectors. The pre-processing phase plays an important role in reaching desired results.

Based on the experimental results, it seems that the recognition of the letters' main bodies has proper results with theC4.5 method, but these tiny movements (including points, hats, etc.) eventually specify the letters' final shapes. Of course, this problem is more frequent in the groups where the differences between members lie in points, such as " ب، پ، " ت، ث and "ج، چ، ح، خ"groups. Despite problems in the recognition of tiny movements, it is probably better to use a more complicated decision tree in addition to the initial recognition.

Table2.confusion matrix

ق	ف	غ	ع	ظ	ط	ض	ص	ش	س	ژ	ز	ر	ذ	د	خ	ح	چ	ج	ث	ت	پ	ب	ا	
0	7	4	0	3	8	2	0	0	0	10	13	6	5	7	0	0	0	0	2	3	4	9	2027	ا
4	11	1	0	2	1	3	0	0	0	0	2	0	1	5	1	0	0	1	6	16	18	1996	0	ب
5	1	3	5	3	3	2	0	6	0	6	0	0	3	5	3	3	3	6	7	8	1994	7	9	پ
15	8	5	1	1	1	1	0	1	1	1	0	0	4	5	0	0	0	1	25	1966	3	7	0	ت
10	11	4	3	1	1	2	0	3	1	1	1	0	8	0	2	5	3	6	1721	12	0	4	1	ث
6	0	9	4	3	4	0	0	3	0	2	0	0	0	0	27	34	36	1975	3	1	1	0	0	چ
16	0	1	4	12	5	0	1	10	0	2	0	0	1	0	14	18	1894	19	2	0	1	0	0	ه
8	2	12	2	1	2	2	0	0	0	3	0	0	2	0	33	1979	5	26	1	1	3	0	0	ح
3	0	10	3	2	0	0	0	1	0	10	1	0	7	1	1934	3	2	9	4	0	3	0	5	خ
3	2	1	2	0	0	0	0	1	0	9	6	2	15	2022	1	1	0	0	6	3	1	4	4	د
0	6	4	1	0	0	0	0	0	0	17	19	0	1765	10	2	0	0	0	13	7	7	0	2	ذ
0	0	0	0	0	0	0	0	0	0	0	9	2050	0	0	0	0	0	0	0	1	0	4	11	ر
0	13	0	0	0	1	0	0	0	0	22	2004	3	1	1	0	0	0	0	1	0	2	5	4	ز
0	3	9	1	0	0	2	0	4	0	1944	10	0	4	3	6	2	0	0	4	4	1	2	3	ژ
0	7	0	0	3	1	18	40	6	2013	0	0	0	0	0	0	0	0	0	0	6	0	0	0	س
8	7	2	1	6	8	20	6	2001	4	0	0	0	0	0	1	3	6	6	2	0	3	3	0	ش
1	5	0	2	2	2	38	2000	7	25	0	0	0	0	0	0	0	3	1	3	2	1	0	0	ص
5	5	2	4	6	6	1958	12	5	4	0	0	0	0	0	0	2	5	3	3	2	2	0	1	ض
3	1	4	4	27	1991	1	0	7	0	1	0	0	0	0	1	1	0	0	0	1	4	0	0	ط
4	4	3	3	1832	6	2	1	3	3	0	0	0	0	0	2	0	0	0	0	0	2	0	0	ظ
11	0	29	1961	3	2	2	0	0	0	3	0	0	1	1	4	12	4	4	4	5	3	0	3	ع
2	0	1916	4	0	3	1	0	3	0	7	0	0	4	0	18	3	3	4	1	0	2	1	2	غ
6	1955	2	2	2	1	1	1	3	1	1	1	2	1	0	2	0	1	1	4	5	3	6	0	ف
1926	1	0	8	3	3	1	0	0	0	0	0	0	0	1	2	3	8	1	3	8	6	0	0	ق
2	1	5	0	0	1	7	2	2	1	5	1	2	1	1	2	1	0	1	2	0	1	2	2	ک

Table.3 The final results for each classes in data base

Results	character	Class #	Results	character	Class #	Results	character	Class #
97.66%	ک	24	98.76%	ز	12	95.93%	ا	0
96.59%	گ	25	97.60%	ژ	13	97.46%	ب	1
97.98%	ل	26	97.64%	س	14	97.35%	پ	2
96.56%	م	27	96.55%	ش	15	96.75%	ت	3
97.70%	ن	28	98.56%	ص	16	96.43%	ث	4
96.49%	و	29	98.04%	ض	17	95.39%	ج	5
96.29%	ه	30	98.32%	ط	18	94.89%	چ	6
96.95%	ی	31	99.12%	ظ	19	97.31%	ح	7
98.11%	ئـــ	32	98.51%	ع	20	95.58%	خ	8
99.23%	آ	33	97.51%	غ	21	98.42%	د	9
99.49%	هـــ ـهـــ	34	97.31%	ف	22	98.18%	ذ	10
99.71%	ـــه	35	97.97%	ق	23	97.06%	ر	11

6. REFRENCES

[1] F. Solimanpour, J. Sadri, C.Y. Suen, "**Standard databases forrecognition of handwritten digits, numerical strings, legalamounts, letters and dates in Farsi language**", in: Proceedings of the10th International Workshop on Frontiers of Handwriting Recognition, La Baule, France, 2006, pp. 3–7

[2] M. Hanmandlu, O.V. Ramana Murthy, Vamsi KrishnaMadasu,"**Fuzzy Model based recognition of handwrittenHindi characters**", Digital Image Computing Techniquesand Applications, 0-7695-3067-2/07 © 200 9IEEE.

[3] Abulhaiba S. H., Mahmood S. A., Green R. J., "**Recognition of handwritten cursiveArabic characters**", IEEE Trans. On PAMI, Vol. 16, No. 6, pp. 664-671, 1994.

[4] Altuwaijir M. and Bayoumi M., "**Arabic text recognition using neural networks**",Proc. Int. Symp. On Circuit and Systems (ISCAS'94), pp. 415-418, 1994.

[5] R.Plamondon; S. Srihari; "**On-line andOff-line Handwritten Recognition: AComprehensive Survey**"; *IEEE Trans. On Pattern Analaysis and MachineIntelligence*; Vol. 22; No. 1; January 2000;pp. 63-83.

[6] N. Arica; F.T. Yarman-Vural; "**Anoverview of caracter recognition focusedon off-line handwriting**"; *IEEE Trans. OnSystems; Man and Cybernetics-Part C:Application and Reviews*; Vol. 31; No. 2;May 2009; pp. 216-233.

[7] M. Ziaratban, K. Faez, F. Faradji, "**Language-based featureextraction using template-matching in Farsi/Arabic handwrittennumeral recognition**", in: Proceedings of the 9th InternationalConference on Document Analysis and Recognition, vol. 1,Curitiba, Brazil, 2007, pp. 297–301.

[8] Teltscher, H. O. "**Handwriting-Revelation ofSelf**", Hawthorn Books Inc. Publishers, New York,1971.

[9] J. C. Simon., "**Off-line cursive word recognition. Proc**". IEEE, 80(7):1150–1160,1992.

[10] Tal Steinherz, Nathan Intrator, and Ehud Rivlin. "**Skew detection via principalcomponents analysis. In Proceedings of the 5th International Conferenceon Document Analysis and Recognition**", pages 153–156. IEEE Computer Society

[11] Ebrahimi.A. Kabir.E.,"**A pictorial dictionary for printed Farsi sub words**", Pattern Recognition Letters 29 (2008) 656–663.

[12] R. Jacobs, "**Methods for combining expert's probability assessments**", Neural Computation, vol. 7, pp. 867-888, 1995.