Machine Printed Arabic Character Recognition Using S-GCM

Liying Zheng[1] Xianglong Tang[2]

[1] *School of Computer Science and Technology, Harbin Engineering University*
[2] *School of Computer Science and Technology, Harbin Institute of Technology*

Harbin, Heilongjiang, 150001, *P.R.China*

Email: tiankai_@mail.hrb.hl.cninfo.net

Abstract

Arabic characters are widely used in Arabic countries. However, there is a little work has been done on recognition of Arabic characters. This paper proposed a new method for recognition machine printed Arabic characters. The proposed method employs Ishii et al's chaotic neural network model, which is called globally coupled map using the symmetric map (S-GCM), for recognizing Arabic characters. The proposed method is tested on two fonts, Simplified Arabic and Arabic Transparent, and 9 sizes, 8, 9, 10, 11, 12, 14, 16, 18, 20. The recognition rate is greater than 97%.

Keywords: Arabic character recognition, chaotic neural networks, associative memory, covariance matrix, minimum distance classifier.

1. Introduction

The Arabic language has a rich vocabulary. More than 200 million people speak this language as their native speaking, and over 1 billion people use its character set, such as Persian and Urdu. However, compared with other language, such as Chinese, Japanese and English, the work done on Arabic character sets is little [1][2]. The main reasons are as the following

- Research in the area of Arabic character recognition is relatively recent
- The characteristics of Arabic language do not allow direct application of the techniques for classifying characters in other language
- There is no adequate support in term of funding and other utilities such as Arabic text database

The research on Arabic character recognition is began in 1980s, since then there are various recognition methods have been proposed, such as the methods based on image density[3]-[7], the methods based on moment invariants and back-propagation neural networks (BPNN)[8][9], and the methods based on primitive features and decision tree [10][11].

Recent developments in the field of artificial neural networks (ANNs) provide an alternative approach for Arabic character recognition. One of the developments in this field is chaotic neural networks (CNNs). By far, different types of CNNs have been proposed, such as Wang-Smith's chaotic simulated annealing (CSA) model [12] and its improvement version [13], Chen-Aihara's transiently chaotic neural network (TCNN) [14], Kaneko's model called globally coupled map (GCM) model [15] and Ishii et al's model called globally coupled map using the symmetric map (S-GCM) [16]. Among these CNNs, S-GCM is an improvement version of Kaneko's GCM model, and it can be used for associative memory. Ishii et al's studies show that both the memory capacity and the basin volume for each memory in S-GCM model are larger than those in the Hopfield model employing the same learning rule [16].

In this paper, an Arabic character recognition method using Ishii et al's S-GCM model is proposed. First, some simple image processing methods are used to process the original character image. Next, using the 24*24 pixel matrix for the input of S-GCM model and computing the output of the S-GCM model. Finally, a minimum distance classifier is used to recognize the input character. The proposed method is tested on two fonts, Simplified Arabic and Arabic Transparent, and 9 sizes, 8, 9, 10, 11, 12, 14, 16, 18, 20. The recognition rate is greater than 97%.

The characteristics of Arabic characters and the S-GCM model are introduced briefly in section2 and section3, respectively. The Arabic character recognition method using S-GCM is clarified in section4. The Experimental results and some analysis are given in section5. Some conclusions are given in section6.

2. The Characteristics of Arabic Characters

Some of the characteristics of Arabic characters are the following [8], [11], [17], [18].

1) In Arabic alphabet, there are 28 isolated characters, each of which has two to four representations, depending on the position (beginning, middle, end or isolated) in the word, which increases the number of classes from 28 to 100.
2) Generally speaking, different Arabic characters have different sizes.
3) Characters may have a dot, two dots, or three dots, which are called secondary part of the character, associated with its main part, and can be above or below or even inside the character.
4) When the character Alif is written immediately after the character Lam, a new character, Lam-Alif, is created.
5) There are pairs and groups of characters, which differ only by the secondary parts.

Fig. 1 introduces Arabic characters in all forms and their characteristics.

Fig. 1: Arabic characters in all forms and their characteristics. IF: isolated form, BF: beginning form, MF: middle form, EF: end form. 1: different characters with different sizes, 2: different characters with different numbers of dots, 3: the character Lam-Alif, 4: different characters with the same main parts, 5: some characters have only two forms.

3. The S-GCM Model

This section provides a brief introduction of S-GCM. The system is given by [16]

$$x_i(t+1) = (1-e)\, f_i(x_i(t)) + \frac{e}{N}\sum_{j=1}^{N} f_j(x_j(t)) \quad (1)$$

$$f_j(x) = a_j x^3 - a_j x + x \quad x \in [-1,1] \quad (2)$$

where $x_i(t)$ denotes the i^{th} unit value at time t, N is the number of units, and t is the discrete-time, ε is a constant parameter. The cubic function $f_j(\ )$, which has a symmetric function shape, can produce chaos with a specific value of its bifurcation parameter α_j.

The characteristics of S-GCM are determined mainly by the values of its parameters, ε and α. The parameter α indicates the strength of each unit chaotic, and the parameter ε indicates the strength of the coupling. Therefore, as α becomes large the S-GCM becomes chaotic, and as ε becomes large the S-GCM becomes coherent.

The above S-GCM model can be used for associative memory, and Ishii et al's studies show that both the memory capacity and the basin volume for each memory in S-GCM model are larger than those in the Hopfield model employing the same learning rule.

By using S-GCM for associative memory, one of the most crucial works is to convert the stored patterns into the parameters of S-GCM. There are two methods to perform this work: adjusting parameter α_i and adjusting parameter ε_i. The principles of these two methods are similar. Both of them employ the covariance learning rule, which is broadly used in associative memory neural networks. So, here we only introduce the method for adjusting parameter α_i (note: the value of parameter ε in Eq.(1) is a constant with this method).

Let $\{X^1, X^2, \ldots, X^M \mid X^k \in \{-1,+1\}^N\}$ be a set of N-dimensional binary patterns to be stored. x^k_i denotes the i^{th} unit value in the k^{th} binary pattern, and M is the number of stored patterns. The evolution of parameter α_i in Eq.(2) is given by

$$a_i' = \{a_i + (a_i - a_{\min})\tanh(bE_i)\}^{\#} \quad (3)$$

$$\{x\}^{\#} = \min\{\max\{x, a_{\min}\}, a_{\max}\} \quad (4)$$

$$E_i = -x_i \sum_{j=1}^{N} w_{ij} x_j \quad (5)$$

$$w_{ij} = \frac{1}{M}\sum_{k=1}^{M} x_i^{\,k} x_j^{\,k} \quad (6)$$

where α_{max}, α_{min} and β are constant parameters. x_i and E_i are system state and partial energy, respectively. Matrix $[w_{ij}]$ is called covariance matrix. Generally speaking, the evolution of α_i is done every other steps, so Eq.(3) does not using t.

4. The Recognition Method Using S-GCM

Generally speaking, the form of each Arabic character can be determined during the process of the character segmentation stage. So, we can use 4 classifiers to recognize 4 different forms of Arabic character, respectively. Since the method for recognizing each form is the same, we only introduce the method for

recognizing isolated form of machine printed Arabic characters. The main steps are the following.

1) Using the image pre-processing method to process the original character image. This step include
   i. Using the median filter to smooth the image.
   ii. Using the horizontal and the vertical projection of the character image to find the bounding rectangle of each character.
   iii. Scaling this bounding rectangle to a 24*24 pixel matrix.
2) Using the above 24*24 pixel matrix for the input of the S-GCM model, and computing the output of the model according to Eq.(1)-Eq.(5). This step incude
   i. Let t=0
   ii. If t is the times of 3, then using Eq.(3)-Eq.(5) to compute the new value of α_i
   iii. Using Eq.(1) and Eq.(2) to compute the value of $x_i(t+1)$, and let $t=t+1$
   iv. Using the following formula to compute the energy of the S-GCM

$$E = \sum_i E_i \qquad (7)$$

   Where E_i is computed by Eq.(5).
   v. If $|E|<0.01$ or $t>=500$, then go to step vi, otherwise go to step ii
   vi. Saving the output of the S-GCM.
3) Using the minimum distance classifier to recognize the output of the S-GCM, and saving this classifier's result as the final recognition result.

Note that

- Here, the parameter α_i is adjusted every other 3 times
- Before using the above method to recognize the input Arabic character, the covariance matrix [wij] must be computed. First, selecting some character images as standard patterns (for isolated form of Arabic character recognition, 29 standard patterns were selected). Next, using the above step 1), to process these standard patterns. Finally, using Eq.(6) to compute the covariance matrix [wij]

5. Experimental Results and Analysis

The recognition method, which is discussed deeply in section 4, has been implemented using Visual C++ program and tested with more than 1000 samples of Arabic character. These samples, which are written in two fonts, Simplified Arabic and Arabic Transparent, and 9 sizes, 8, 9, 10, 11, 12, 14, 16, 18, 20, are captured using a scanner with a resolution of 300 DPI. Figure 2 shows some of these samples.

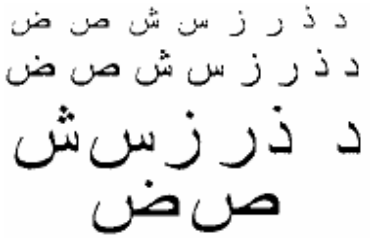

Fig.2: tested samples

We use four similar classifiers, C1, C2, C3 and C4, to recognize isolated, beginning, middle and end form of Arabic character, respectively. The initial value of parameter α_i is a random number between α_{min} and α_{max}, and the values of other parameters, which are contained in Eq.(1)-Eq.(6), are shown in Table 1. The recognition rate of each form of Arabic character is shown in Table 2.

Table1: The parameter values

	N	ε	α_{max}	α_{min}	β	M
C1	24*24	0.001	2.8	3.7	0.99	29
C2	24*24	0.004	2.8	3.7	0.98	22
C3	24*24	0.004	2.8	3.7	1.0	22
C4	24*24	0.001	2.8	3.7	0.99	29

Table2: Recognition rates (REC: recognition rate, IF: isolated form, BF: beginning form, MF: middle form, EF: end form)

Font	Arabic Transparent			
Form	IF	BF	MF	EF
REC	98.6	97.2	98.4	97.1
Font	Simplified Arabic			
Form	IF	BF	MF	EF
REC	98.5	97.1	98.5	97.1

From table 2 we know that

- The recognition rates for these two fonts are similar. This is because that these two fonts are alike
- The recognition rate is greater than 97%, which clearly is a high rate
- The recognition rates for middle and end form are lower than those of isolated and beginning form. This is due to that the middle (or end) forms of some Arabic characters, such as the character noon and baa, are very similar, and sometimes the proposed method can not classify these characters correctly

6. Conclusions

Character recognition is one of the most important stages for any character recognition system. By far, the field of Arabic character recognition is still an open field. Therefore, a new machine printed Arabic

character recognition method, which is based on Ishii et al's S-GCM model, is proposed. The new method has been tested with a lot of samples, and the results show that it has a very high recognition rate.

7. References

[1] A. Amin, "Off-line Arabic character recognition-a survey," Proc. of the 4th International conference on Document Analysis and Recognition, pp. 596-599, 1997.

[2] M.S. Khorsheed, "Offline Arabic character recognition-a review," Pattern Analysis & Applications, pp. 31-45, 2002, 5.

[3] H. M. M. Hosseini and A. Bouzerdalm, "A system for Arabic character recognition," Proc. of the 2th Australian and New Zealand conference on Intellegent Information Systems, pp. 120-124, 1994.

[4] J. Alherbish and R. Ammar, "High Performance Arabic Character Recognition," The Journal of Systems and Software, pp. 53-71, 1998, 44.

[5] N.M.Wanas, M.R.El-Sakka and M.S.Kamel, "Multiple classifier hierarchical architecture for handwritten Arabic character recognition," International Joint Conference on Neural Networks, pp.2834-2838, 1999, 4.

[6] F.Bouslama and H.Kishibe, "Fuzzy Logic in the recognition of machine printed Arabic characters ", The 6th International Conference on Neural Information Processing, pp. 16-20, 1999.

[7] H.Al-Yousefi and S.Udpa, " Recognition of Arabic characters," IEEE Trans. Pattern Analysis Machine Intell., pp. 853-857, 1992, 14(8).

[8] M. M. Altuwaijri and M. A.Bayoumi, "Arabic text recognition using neural networks," IEEE International Symposium on Circuits and Systems, pp.415-418, 1994.

[9] H.Y.Y. Sanossian and M.Al-karak, "An Arabic character recognition system using neural network," Proc. of IEEE Signal Processing Society Workshop, pp. 340-348, 1996.

[10] A.Amin, "Prototyping structural description using decision tree learning techniques," The 16th International Conference on Pattern Recognition, pp. 76-79, 2002.

[11] B.M.F.Bushofa and M.Spann, "Segmentation and recognition of Arabic characters by structural classification," Image and Vision Computing, pp. 167-179, 1997, 15.

[12] L.Wang and K.Smith, "On Chaotic Simulated Annealing," IEEE Transactions on Neural Networks, pp. 716-718, 1998,9(4).

[13] L.Zheng, K.Wang and K.Tian, " An approach to improve Wang-Smith chaotic simulated annealing", International Journal of Neural Systems, pp. 363-368, 2002, 12(5).

[14] L.Chen and K.Aihara, "Chaotic Simulated Annealing by a Neural Network Model with Transient Chaos," Neural Networks, pp. 915-930, 1995, 8(6).

[15] K.Kaneko, "Clustering, coding, switching, hierarchical, ordering, and control in a network of chaotic elements," Physica D, pp.37-172, 1990, 41,

[16] S.Ishii, K.Fukumizu and S.Watanabe, "A net work of chaotic elements for information processing," Neural Networks, pp. 25-40, 1996, 9(1),

[17] B. Al-Badr and S. A. Mahmoud, "Survey and bibliography of Arabic optical text recognition," Signal Processing, pp. 49-77, 1995, 41.

[18] L. Zheng, Abbas H. Hassin and X.Tang, "A new algorithm for machine printed Arabic character segmentation," Pattern Recognition Letters, pp. 1723-1729, 2004, 25.