SCIENCE CHINA

December 2011 Vol. 53 No. 1: 1–10
doi:

Handwritten Chinese Character Recognition Using Eigen Space Decomposition

LONG Hui[1]*, ZHANG Xiaochen[1] & Kuruoglu E Ercan[1,2]

[1]*School of Electrical and Computer Engineering, Georgia Institute of Technology, Shanghai* 200240, *China*
[2] *Institute of Science and Technology of Information (ISTI), "A. Faedo", Italian National Council of Research (CNR), Pisa, Italy*

Abstract In this paper, we mainly describe a new approach of Handwritten Chinese Character Recognition (HCCR), which is based on eigen-character extraction. The procedure of the eigen-character extraction method is explained including initialization, eigen character extraction (or eigen spaces generation) and character recognition. Two different methods are presented to do eigen character recognition respectively. Besides, k Nearest Neighbor (kNN) is implemented to improve the recognition rate of the new approach. In the end, a comparison is made between the eigen-character extraction approach and other existing approaches through simulation based experiments. The results show that our approach has a satisfying rate and could be further improved if combined with some other methods such as elastic matching and wavelet methods.

Keywords handwritten Chinese character, character recognition, eigen character, eigen space decomposition, k Nearest Neighbor

1 Introduction

The Chinese characters play a major role in conveying ideas in our daily life. Handwritten Chinese Character Recognition has a fundamental importance, being extremely valuable to the hand-written material restoration, thus could be used for long-term storage of useful information, as well as handwriting input for portable computers and smart phones. As a result, hand-written Chinese character recognition motivates great research efforts.

However, Tang et al.[1] describes off-line handwritten Chinese character recognition as one of the most challenging topics in pattern recognition, since it involves a large number of characters with complex structure, serious interconnection among the components, and considerable pattern variation. Given these difficulties, we believe that a theoretically well-founded approach is necessary.

A Handwritten Chinese Character Recognition (HCCR) system is a computer application for automatically identifying or verifying a Chinese character from a digital image. Although HCCR is difficult, the human ability to recognize handwritten Chinese characters is remarkable. It has been studied intensively since 1990s, and many effective methods have been proposed. Some important techniques, including directional feature extraction, nonlinear normalization, and modifications of quadratic classifiers, have contributed to today's high accuracies on hand-printed character recognition.

To extract features holistically from the character image or decompose characters structurally into component parts-usually strokes, by applying wavelet transformation[2] and elastic grid[3] are two popular approaches for HCCR. Here a novel approach is followed to do HCCR, using Eigen-Characters.

*Corresponding author (email: hlong6@gatech.edu,xzhang322@gatech.edu,ercan.kuruoglu@isti.cnr.it)

2 Eigen characters and Eigen faces

The first part of the article will place more emphasis on the introduction of some definitions of basic concepts being used in our paper, including eigen faces Fig.1 and eigen character recognition of Latin letters.

Figure 1: Eigen faces [4]

A state of the art of technique in face recognition is eigenfaces which is based on the methodology of eigenspaces. An eigenspace[5] of a square matrix is the set of all eigenvectors with the same eigenvalue together with the zero vector. Eigen faces are constructed from the eigenvectors calculated via eigenvalue analysis on a collection of face images. With these eigen faces, the input facial images can easily be classified based on Euclidian distance by applying principal component analysis (PCA). [6] Similarly, a state of the art character recognition method for Latin letters also utilizes eigen space decomposition[7]. Eigen characters have found various applications, the most notable being eigen digits which aim recognizing hand written numbers [8]. (See Fig.2) Various extensions of this method also have been suggested such as the eigen deformation method for more elastic matching of Latin letters recognition[9].

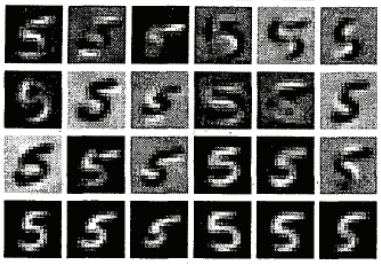

Figure 2: Eigen digits

Eigen space-based approaches approximate the character vectors (character images) with lower dimensional feature vectors. The main supposition behind this procedure is that the character space (given by the feature vectors) has a lower dimension than the image space (given by the number of pixels in the image), and that the recognition of the characters can be performed in this reduced space. This approach requires a training process, in which the eigen character database is created. The dimension reduction is achieved by using the projection matrix, which is obtained from the training database.

3 Eigen Character Recognition Method

The basic steps involved in Character Recognition using Eigen characters approach are as follows:
Step 1 Initialization

- Acquire initial set of character images known as Training Set, as shown in Fig.3.

- Transfer these characters from image format to Matrix format and then resize them to vectors. Assume that each character is acquired as a $M \times N$ matrix. Then resize the matrix into a $MN \times 1$ vector $T_{MN\times 1}$.
- Center the character in every image. Find the left-side, right-side, upper-side and lower-side blank spaces in order. Then crop the images to the edges.

Figure 3: Data set as the training set

Step 2 Eigen character extraction

Two methods of eigen character extraction are in this section.

Method 1

- Assume that the training database consists of k different characters and the total sample size is n. Hence, let $[T_1, T_2, \cdots, T_n]$ represents the database;
- Then compute the mean value of all T_i as $T_{i,mean}$, and replace T_i with $T_i' = T_i - T_{i,mean}$. Let $A_{MN\times n} = [T_1', T_2', \cdots, T_n']$;
- Next compute eigenvectors v_i using the power method[21]. Let $V = [v_1, v_2, \cdots]$, where v_i is an eigenvector;
- Then a threshold λ-th is set in order to sort and eliminate eigenvalues and calculate the eigenvectors of covariance matrix C, which are called eigen characters. And $V = [v_1, v_2, \cdots]$ are the so-called eigen characters.

In this method, different Chinese characters are not differentiated. In contrast, we just deal them as the same. Thus the eigen character corresponding to this method contains the contribution of all different characters in the training database, thus cannot represent any intuitive characteristic of any Chinese character. See Fig.4.

Method 2

Figure 4: An eigen character from all training data

1. To begin with, each different Chinese character is grouped into a class. Let the space $C = [c_1, c_2, \cdots, c_k]$ stand for the entire Chinese character set. Where, class c_i, $i = 1, 2, \cdots, k$, stands for a certain character, and k is the total number of Chinese characters that could be recognized in the system. Let $B_{c_i,n} = [b_{c_i,j}|j = 1, 2, \cdots n_i])$ stand for the set of training database for class c_i. And n_i stands for the sample size of the class c_i.

2. For each class, PCA[10] is performed. To begin with, let $[b_{c_i,1}, b_{c_i,2}, \cdots, b_{c_i,n_i}]$ be a $MN \times n_i$ matrix, which consists of all information of class c_i training data.

3. Then, compute the mean value of all $b_{c_i,j}$ as b_{c_i}, mean, and replace $b_{c_i,j}$ with $b'_{c_i,j} = b_{c_i,j}, -b_{c_i,mean}$. Let $A_{MN\times n_i} = [b'_{c_i,1}, b'_{c_i,2}, \cdots, b'_{c_i,n_i}]$.

4. Next, compute eigenvectors of class c_i using power method. Let $V_{c_i} = [v_{c_i,1}, v_{c_i,2}, \cdots]$, where $v_{c_i,j}$ is an eigenvector of class c_i.

5. In the end, sort and eliminate those whose eigenvalue is less than 1, while calculating the eigenvectors of covariance matrix C, which is called eigen-characters. And $V_{c_i} = [v_{c_i,1}, v_{c_i,2}, \cdots]$ is the so-called eigen-character space of class c_i.

In this method, eigen-spaces are created for each different character, which is shown in Fig.5.

Figure 5: Eigen characters formed from training database (four different characters database and four corresponding eigen character space)

For example, if we have 4 different characters such as “ 狗”, “ 鸡”, “ 猪” and “ 鼠”, each character with a database of 20 training samples. By method 2 we could generate 4 corresponding eigen-spaces. The input test sample is projected to these 4 eigen spaces respectively. Then, calculate the distances between the projected array and the array of the original input sample. The one with minimum distance is chosen. In the example, 4 distances are calculated.

The main differences between the two methods is that method 1 does not sort the training database with respect to different characters and creates just one eigen-space for the entire training database, while method 2 sorts the training database into different classes with respect to different characters and creates an eigen-space for each different character.

Step 3 Perform character recognition Assume one handwritten character is chosen as a test character. The two slightly different methods of eigen extraction require two approaches to perform character recognition.

For method 1, the character recognition is performed by projecting the test character and all training characters into the same eigen-character space generated by the whole training samples, see Fig.6. Then calculate the Euclidian distances between each the test sample and every training sample as d_i. Let $d_i = min(d_1, d_2, \cdots)$, and the algorithm returns the character that class c_i represents as the final character recognition result.

For method 2, the character recognition is performed by projecting the test character into several eigen-spaces generated by different classes $C = [c_1, c_2, \cdots, c_k]$ and comparing the measured Euclidean distance d_{c_i} among them, as Figure 6. The algorithm returns $d_{c_i} = min(d_{c_1}, d_{c_2}, \cdots)$ and the character which class c_i represents is returned as the final character recognition result.

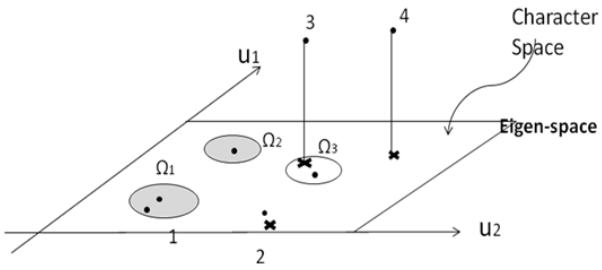

Figure 6: Eigen-space projection

To summarize the whole process of character recognition using eigen, we will see the diagram shown below as the logical process Fig.7 and the system process Fig.8.

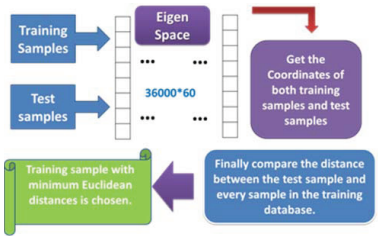

Figure 7: The logical process

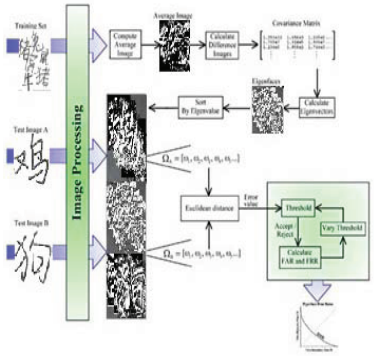

Figure 8: The system-level process

The training set of images is given as input to find Eigen space. Using these images, the average character image is computed. The difference of these images is represented by covariance matrix. This is used to calculate Eigenvectors and Eigen values. These are the Eigen characters which represent various character features. Sort the Eigen values, and consider larger ones of them since they represent maximum character features. Eigen space spanned by the Eigen characters has lower dimension than original images.

The Euclidean distance is calculated, and minimum distance is chosen as the final matching. In this way Character Recognition is carried out using Eigen character Approach.

4 KNN method for character recognition

In order to improve the performance, we introduced kNN (k Nearest Neighbor) method [11]. kNN is a method for classifying objects by choosing the closest training examples in the feature space. It is also a

type of instance-based learning, or lazy learning where the function is approximated only locally and all computation is deferred until classification. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of its nearest neighbor.

In our experiment, $k = 3$, shown in Fig.9, where the characters circled out are the initial matching results. Since the index of " 狗" and " 鸡" are respectively tagged twice and once, we identify the input character to be " 狗". When matching the sample character with the training database, kNN method could eliminate the ones with too large differences caused by handwritten inaccuracy. Thus, by using kNN, the performance of our proposed approach gains perceivable improvement in the accuracy rate of recognition.

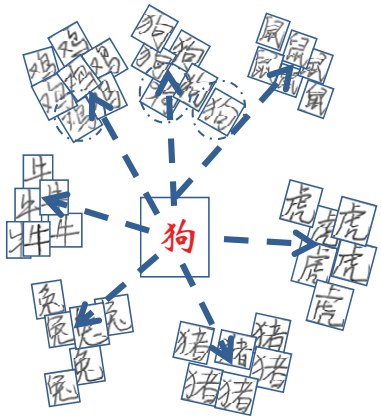

Figure 9: Recognition result using kNN

5 Results and Performance Evaluation

In our experiment, we applied our proposed method using a publicly available database CASIA-OLHWDB1[12] as the training set. This database contains unconstrained handwritten characters of 4,037 categories (3,866 Chinese characters included) produced by 420 persons, and 1,694,741 samples in total. A sample set of regular writing is shown in Fig.10.

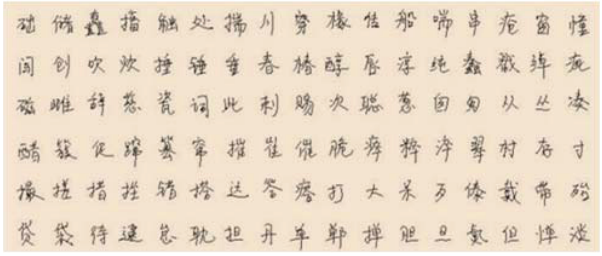

Figure 10: A sample set of regular writing from CASIA

After preliminary experiments, we give out the matching results of seven animal characters. Initial results show that our proposed algorithm performs well and all test samples have been correctly recognized in our system. The test and equivalent characters are show in Fig.11.

In our experiment, we also do a comparison between our proposed method and the existed and currently popular methods, such as Discriminant Analysis[13], C4.5[14] and FNN[15]. We find that although a little weaker on the error rates, on character learning-and-training stage, our proposed method is running at a

Figure 11: Recognition result

higher speed and at the same time storage-economical. We also find that the recognition performance is superior for a small number of different Chinese characters.

Characters	**DA**	**C4.5**	**FNN**	**Eigen**
100	86.7	90.1	97.4	**98.2**
200	85.0	87.8	98.5	**94.0**
500	84.5	85.6	93.9	**90.4**

Table 1: Comparison of Recognition Rates (%) with DA, C4.5, FNN and Eigen Character Method

6 Conclusion

This paper mainly represents a new approach to handle Chinese character recognition. The new method is based on eigen character decomposition.

Two different ways of character recognition after eigen character extraction are studied. Through the experiment, the two methods are suited to different situations. If the character number is relatively small and each character has a relatively low sample size, method 1 is preferred. If the character number is very large and each character has a large enough sample size, method 2 would no doubly generate a better result.

The new method is also compared with other Chinese handwriting character recognition methods as well. From the experiment, it is clear that although the recognition rate is acceptable. And the recognition rate is much higher when it comes to smaller database. Besides the high speed and storage advantages of the eigen character extraction method, the new method does not necessarily to be implemented alone. In fact the new method could also combined with other recognition methods such as elastic matching and wavelet method, which would further improve the recognition rate especially in large database cases.

7 Acknowledgements

E.E. Kuruoglu gratefully acknowledges partial support from Chinese People's Republic 111 Programme "Bringing Foreign Experts of Talent" at Shanghai Jiao Tong University on Video Science and Technology.

References

1 Tang Y, Tu T, Liu J, S W Lee, W W Lin, I S Shyu, Off-line Recognition of Chinese Handwriting by Multifeature and Multilevel Classification. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 5, pp 556-561, 1998.
2 Zhong G, Jin L, A New Elastic Circle Meshing Feature Extracting Method for Handwriting Chinese Character Recognition, In: Computer Engineering, vol. 28, no. 11, pp 61-62, 2002.
3 Uchida S , Sakoe H, Eigen-deformations for Elastic Matching Based Handwritten Character Recognition. In: Pattern Recognition, 2003.
4 Kshirsagar V P, Baviskar M R, Gaikwad M E, Face Recognition Using Eigenfaces, 2011 3rd International Conference on Computer Research and Development (ICCRD), On page(s):302-306, Volume: 2, Issue: 11-13, March 2011.
5 Turk M, A Random Walk Through Eigenspace, IEICE Trans. Inf. and Syst., vol. E84-D, no. 12, pp. 1586-1695, December 2001.
6 Gomathi E, Baskaran K, Recognition of Faces Using Improved Principal Component Analysis, ICMIC, pp.198-201, 2010 Second International Conference on Machine Learning and Computing, 2010.
7 Mitoma H, Uchida S, Sakoe H, Online Character Recognition Using Eigen-Deformations, IWFHR, pp.3-8, Ninth International Workshop on Frontiers in Handwriting Recognition (IWFHR'04), 2004.
8 Bishop C M, Winn J M, Non-linear Bayesian image modelling, 6th European Conference on Computer Vision, Dublin, Ireland, Jun 26-Jul 01, 2000 ECCV 2000, PT I, Lecture Notes in computer Science, Vol: 1842 pp: 3-17, 2000.
9 Tappert C, Cursive Script Recognition by Elastic Matching. IBM, Journal of Research Development, 26(6):765-771, 1982.
10 Smith L I(2002), A Tutorial on Principal Component Analysis, Journal of Measurement, vol 51, 2005.
11 Shang-Hua Teng , Frances F. Yao, k-Nearest-Neighbor Clustering and Percolation Theory, Algorithmica, v.49 n.3, p.192-211, October 2007.
12 Wang D H, Liu C L, Yu J L , Zhou X D, CASIA-OLHWDB1: A Database of Online Handwritten Chinese Characters, ICDAR09(1206-1210). IEEE DOI Link 0907.
13 Etemad K andChellappa R, Discriminant Analysis for Recognition of Human Face Images. JOSA A, Vol. 14, Issue 8, pp. 1724-1733 (1997).
14 Amin A and Singh S, Recognition of Hand-printed Chinese Characters using Decision Trees/Machine Learning C4.5 System, Pattern Analysis and Applications, Vol. 1, no. 2, pp.130-141, 1998.
15 Tan T X, Ng G S, Quek C, C. L.Stephen, Ovarian cancer prognosis by hemostasis and complementary learning, Koh. Pages:145-154.