Chapter 93
High Accuracy Handwritten Chinese Character Recognition Based on Support Vector Machine and Independent Component Analysis

Zhiguo He, Yuquan Zhong and Yudong Cao

Abstract This paper proposed a new method for handwritten Chinese character recognition based on a combination of independent component analysis (ICA) and support vector machine (SVM). First, we extracted independent basis images of handwritten Chinese character image and the projection vector by using fast ICA algorithm, and obtained the feature vector. Then, we used two stage classification methods based on SVM for classification. The scheme took full advantage of good extraction local features capability of ICA and strong classification ability of SVM, thus increasing the system's recognition rate. The experiments show that the feature extraction method based on ICA is superior to that of gradient-based, and the two stage classifiers based on SVM is better than that of modified quadratic discriminant function. On HCL2000, a handwritten Chinese character database, the recognition accuracy of 99.87 % has been achieved.

Keywords Handwritten chinese character recognition · Independent component analysis · Support vector machine · Feature extraction

Z. He (✉) · Y. Zhong · Y. Cao
School of Computer Science, Panzhihua University, Panzhihua, China
e-mail: zhiguohe@126.com

Y. Zhong
e-mail: zhongyuquan@126.com

Y. Cao
e-mail: caoyudong@126.com

W. Du (ed.), *Informatics and Management Science V*, Lecture Notes
in Electrical Engineering 208, DOI: 10.1007/978-1-4471-4796-1_93,

93.1 Introduction

Handwritten Chinese character recognition (HCCR) is an important research topic in pattern recognition, which is widely used in the automatic input of Chinese electronic data processing, in the Chinese text compression, in office automation and computer-aided teaching, etc. It can bring about huge economic and social benefits, but it is also one of the more difficult issues in the field of pattern recognition. Because Chinese characters set [1–3] is very large; the structure of Chinese characters is very complex; Many Chinese characters have high degree of similarity; the writing styles for the same character are many kinds and have large shape variations. These four reasons make HCCR very difficulty. At present, HCCR system has not yet reached satisfactory results, especially for Chinese characters with cursive script. HCCR is divided into four steps: pre-processing, feature extraction, classification and post-processing, among which feature extraction method and classifier is an import factor for recognition performance. To a large extent, the accuracy of an overall recognition system depends on the discriminative capability of features and generalization performance of a designed classifier. Determining how to extract stable and good separable feature for Chinese character is an important research direction. The bottle-neck of feature extraction is the instability of the feature between the different samples of the same Chinese character, so the key for HCCR is to accurately describe the details of the differences for the same Chinese character caused by different writing styles. At present, in HCCR, the method for feature extraction can be mainly divided into two categories: one is based on structural features, which is rarely used because they are difficult to extract and very sensitive to noise; the other is based on statistical features which is widely used in HCCR. Widely used statistical feature extraction methods include gradient features [4] and features based on independent component analysis (ICA) [5]. However, the gradient feature exist deficiencies such as large amount of calculation and high dimension features. Although principal component analysis (PCA) was used to dimensionality reduction, but it is a method based on second-order statistical characteristics and its purpose is to remove the correlation between the components of the image. A large number of studies have shown that the most important information of image is existed in the high-order statistics of image pixels, but the dimensionality reduction method based on PCA did not use the high-order statistical characteristics of images [6]. While ICA is an analysis method based on higher-order statistical characteristics of signal, it is more fully taken into account the statistical independence of the probability density function of the signal. Principal component obtained by PCA is only de-correlation (orthogonal to each other), while ICA not only achieves de-correlation, but also the higher-order statistics obtained are mutually independent. In PCA, the signal to be processed is generally assumed the Gaussian distribution; while in ICA, the signal is assumed non-Gaussian signal which is more in line with the realistic problems. The goal of ICA is separating out the independent component by using linear transform, to remove or minimize the degree of statistical dependence in the image.

At the same time, in HCCR system, the currently widely used classifiers are support vector machine (SVM) classifier [7] and modified quadratic discriminant function (MQDF) [8]. SVM is a new machine learning methods developed in recent years, which uses the principles of structural risk minimization in statistical learning theory, showed strong superiority in dealing with the classification of high-dimensional space and showed strong nonlinear classification ability. While MQDF requires the distribution for each class to be a Gaussian distribution, which often does not match with the reality. Thus, in this paper, SVM is used for classification for Chinese characters. Based on the above analysis, this paper proposes a new method for HCCR. First, we use ICA method for feature extraction for Chinese character image, and then use SVM for classification. Our scheme has reached a higher recognition rate compared to gradient-based feature extraction method and MQDF classifier.

93.2 ICA and Feature Extraction for Handwritten Chinese Character

93.2.1 ICA

The ICA model can be described as:

$$X = AS \tag{93.1}$$

The model describes how the observed data X is obtained by mixing the source S. The source variable S is a hidden variable which can not be directly observed, and the mixing matrix A is also unknown. All the data can be observed is only the random variable X, so it is necessary to estimate the mixing matrix A and the source S. ICA is based on a simple assumption: the source variable S is statistically independent and non-Gaussian distribution. In this basic model, the distribution is unknown.

Extraction image features by ICA is to find a separation matrix W by using linear transform of the observed image, to make the component decomposed by the linear transform as mutually independent as possible and approximation of S. Y is an estimation of S, that is, Y is the extracted feature vector of the image:

$$Y = WX \tag{93.2}$$

where

$$W = A^{-1} \tag{93.3}$$

Through the establishment of the linear model of an image, we can be applied ICA technology to separation the independent component of the observed image, to extract image features, making the separated independent component Y as statistically mutually independent as possible. Y is the estimated mutually independent coefficient, and A is the basis image obtained.

Solving the separation matrix W, we used the FastICA algorithm [9]. The algorithm has the advantage of fast convergence, not requiring a known probability distribution in advance and independent component can be solved one by one, etc., which can reduce the computational cost. FastICA algorithm is to find a direction, namely the unit vector W, through the system's learning to make the projection of W^T X has the largest non-gaussianity. The method of solving the independent variable S is to find the vector W, to make W maximum the non-gaussianity of W^T X. There are a lot of methods to measure the non-gaussianity. We used negentropy, that is:

$$J(W) = [E\{G(W^T X)] - E\{G(v)\}^2 \tag{93.4}$$

where

X is the observation vector,
W is the weight vector,
V is a Gaussian variable with zero mean and unit variance,
G is a nonlinear function.

We took the following functions:

$$G(x) = -\exp\left(-\frac{x^2}{2}\right) \tag{93.5}$$

93.2.2 Feature Extraction for Handwritten Chinese Characters Images

It is necessary to do some preprocessing before Chinese character samples is separated by ICA. Chinese characters are to first normalized, then to whiten and to zero mean value of the input signal. After the processing, the FastICA algorithm was used to solve the separation matrix W. After solving the separation matrix W, we can obtain the basis image Y by using Eq. (93.2). According to the definition of the ICA model, there are: an image X can be obtained by linear combination of the basis image Y, namely:

$$x = \sum_{i=1}^{n} a_i s_i \tag{93.6}$$

where a_i is the feature vector of the Chinese character image x.

Similarly, for any unclassified Chinese characters image, you can use the same method by projection Chinese characters image to the space of basis image Y, then obtaining all projection coefficients and its feature vector.

93.3 Two Stage Classification Based on Support Vector Machine

SVM is a new kind of learning machine based on statistical learning theory. Statistical learning theory is a theory of studying statistical learning for small samples, its main idea is: for linearly inseparable data, first, the input space is mapped into a high-dimensional space by a nonlinear map, to make the data becomes linearly separable or nearly linearly separable in this space, then in this high-dimensional space to obtain the optimal linear classification surface. All operations in the feature space are carried out by the inner product kernel function of the input space. If an inner product kernel function is selected, then it defines a feature space.

SVM is a learning machine following the principle of structural risk minimization. It was first proposed for two-class classification problems, its goal is obtain a linear classification hyper-plane which not only makes two classes separated, but make the separation intervals maximum for the two classes. Assume that the training sample set is: $(x_i, y_i), \quad \mathrm{i} = 1, 2, \ldots, \mathrm{n}$, $\mathrm{x} \in \mathrm{R^d}$, $\mathrm{y} \in \{-1, +1\}$ is category symbol. By solving a condition optimization problem, we can get the optimal classification function:

$$\mathrm{f(x)} = \mathrm{sgn}\left\{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{a_i y_i k(x_i, x)} + \mathrm{b}\right\} \tag{93.7}$$

where

x_i is the support vector;

x is the vector to be classified.

When using SVM for handwritten Chinese character recognition with a large number of categories, it is an important issue to reduce the computational cost and storage cost either in training or recognition stage. If we selected too many samples, the training was too time-consuming, and the number of support vectors is also increased significantly. Based on the above analysis, we adopted two stage classification strategy, the classifier based on Euclidean distance as a pre-classification, the training sample for training SVM using only the candidates obtained by the distance classifier, which made the computational cost and storage cost reduced greatly, and reduced the number of support vectors accordingly. To Train the support vector machine, we used the algorithm proposed by Dong [7].

93.4 Experiments and Results

In this study, we use HCL2000 database, an offline handwritten Chinese character standard database, which is funded by National 863 Program of China, created by Pattern Recognition Laboratory of Beijing University of Posts and Telecommunications. It has become the most influential database for handwritten Chinese character recognition, contains 3,755 frequently used simplified Chinese characters written by 1,000 different persons. It has the characteristics of large sample size, mutual inquiries between sample database of Chinese characters and information database of penman. All the sample of HCL2000 is normalized binary samples, with size 64 (height) by 64 (width). Part of the samples was shown in Fig. 93.1. We chose 700 samples located in xx001–xx500 of HCL2000 as training samples, chose 300 samples located in hh001–hh300 of HCL2000 as test samples. When recognizing Chinese characters, we first used the rough classification (Euclidean distance classifier), and then chose the first 35 candidates obtained by the rough classification as the fine classification (SVM classifier). The candidate characters after fine classification were the final result.

When extraction independent component by using FastICA algorithm, the number of independent components can not be choose too much nor too little. If we chose too many, it may contain a large quantity of noise signals; if too little, it may lose too much feature information of the image. The experimental results show that when the number of principal components is 89, the recognition rate has been reached the highest, shown in Fig. 93.2.

Feature extraction based on ICA was compared with the widely used gradient feature in HCCR. The method for extraction gradient feature please referred to literature [4] and its dimension is 256. In order to facilitate comparison, both

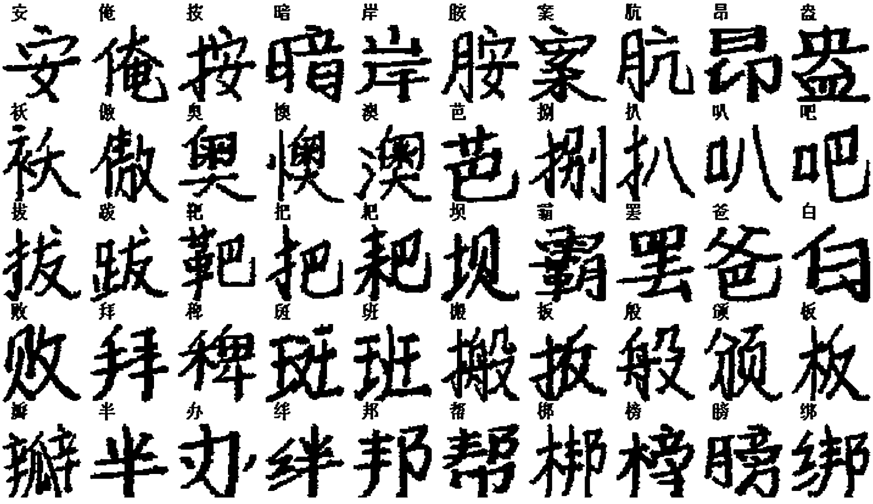
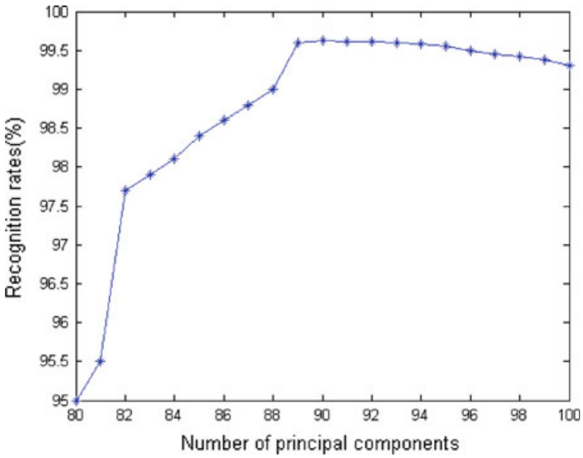

安 俺 按 暗 岸 胺 案 肮 昂 盎
袄 傲 奥 懊 澳 芭 捌 扒 叭 吧
拔 跋 靶 把 耙 坝 霸 罢 爸 白
败 拜 稗 斑 班 搬 扳 般 颁 板
瓣 半 办 绊 邦 帮 梆 榜 膀 绑

Fig. 93.1 Some sample images of HCL2000

Fig. 93.2 The relationship between the number of principal components and its recognition rate



methods used the same MQDF classifier and the results shown in Table 93.1. The experiment was performed on Intel Pentium 4 3.2G with MATLBA 7.0 system equipped with 768 megabytes RAM. From Table 93.1, we known that the recognition rate based on ICA method is higher than that of gradient feature, but the time for feature extraction is quite different. The time for extraction gradient features is about 5 s; while extraction feature based on ICA needs a long time and the time is concerned with the number of principal components and is approximately 2 min when the number of principal components extracted is 89. Recognition time is in a few seconds for each Chinese character with MQDF classifier.

In order to reduce the computational cost and speed up the classification, we use the Euclidean distance classifier as a pre-classifier, and training support vector machine used only the candidate character set obtained by the distance classifier as training samples, which can make the computational cost and storage cost greatly reduced, also make the number of support vectors reduced accordingly. The experiments show that when 35 candidates were chosen, the system has reached very high cumulative recognition rate. For SVM classifier, the one-against-others method was used to construct 35 classifiers. We used polynomial and radial basis function as kernel function, respectively. The parameter selection for kernel function is a rule of thumb. We found by experiments that its recognition rate is

Table 93.1 Recognition performance of HCCR affected by different feature extraction method

The method for feature extraction	The time for feature extraction (The number of principal component)	Recognition rate (%)
Feature based on ICA	90 s (85)	97.57
	127 s (89)	99.45
	158 s (94)	98.72
Feature based on gradient	5 s	97.27

Table 93.2 Recognition rate with different classifiers

Classifiers	Recognition rate (%)
MDQF	99.45
SVM based on two stage classification	99.87

very high by using RBF kernel function. When is 3, it reached the highest recognition rate. The RBF kernel function used is as follows:

$$K(x, x') = \exp\left(-\frac{||x - x'||^2}{2\sigma^2} \right) \tag{93.8}$$

The classification method used in this study was compared with the currently widely used MDQF classifier. When using the same method for feature extraction (based on ICA and the selected principal components is 89), the results was in Table 93.2. Table 93.2 shows that the two stage classifiers used in this study is significantly better than MDQF classifier.

The experimental results show that feature extraction by ICA is superior to the gradient feature. This is because the information between the Chinese character images has certain relevance, is not mutually independent, and this lead to a correlation between the features of different categories, which made the classification accuracy not high. But component obtained by ICA is mutually independent and removes the correlation between the features to a certain extent, thus it may improve the classification accuracy. Meanwhile, the two stage classifiers based on SVM is superior to the widely used MQDF classifiers, because SVM has strong classification capability.

93.5 Conclusion

ICA based on the higher-order statistical correlation between data, extracted internal features of the image, made full use of the statistical characteristic of the input data. ICA as an extension of PCA, it focuses on the higher-order statistical characteristics between data, each of the transformed components is not only unrelated, but also as statistically independent as possible. Therefore, ICA can be more fully reveal the essential features of the input data. But for Chinese character images, a lot of important information is contained in the high-order statistics between the pixels of the image. So the feature extracted by ICA for Chinese character is significantly better than the currently widely used gradient feature. At the same time, the two stage classifier, adopted in this paper, can not only reduce the computational cost for classification with large number of categories by using SVM classifier, but also improve the recognition rate of the entire system. Experiments show that our scheme is superior to the widely used MQDF classifier. But using ICA for feature extraction exist the following deficiencies: the computational cost is very high; iteration of the algorithm depends on the selection of the initial value; the time for feature extraction is very long. All these needs further study.

Acknowledgments The work was partially supported by Key Projects of Education Department of Sichuan Province under Grant No. 10ZA186.

References

1. He ZG, Cao YD (2008) Survey of offline handwritten chinese character recognition. Comput Eng 34(15):201–204
2. Liu CL, Fujisawa H (2008) Classification and learning methods for character recognition: advances and remaining problems. Stud Comput Intell 90:139–161
3. Zhu CH, Shi CY, Wang JP et al. (2011) Study of offline handwritten chinese character recognition based on dynamic pruned FSVMs. In: International conference on electrical and control engineering, vol 123. pp 395–398
4. Liu CL (2007) Normalization-cooperated gradient feature extraction for handwritten character recognition. IEEE Trans Pattern Anal Mach Intell 29(8):1465–1469
5. Rui T, Shen C, Ding J et al (2005) Handwritten digit character recognition by model reconstruction based on independent component analysis. J Comput Aided Des Comput Graph 17(3):455–460
6. Bartlett MS (1998) Face image analysis by unsupervised learning and redundancy reduction, vol 12(31). Dissertation, University of California, San Diego, pp 93–99
7. Dong JX, Krzy_zak A, Suen CY (2005) An improved handwritten chinese character recognition system using support vector machine. Pattern Recogn Lett 26:1849–1856
8. Dai R, Liu CL, Xiao B (2007) Chinese character recognition: history, status and prospects. Frontiers Comput Sci China 1(2):126–136
9. Hyvärinen A, Oja E (2000) Independent component analysis: algorithms and applications. Neural Netw 13(4):411–430