Research Journal of Applied Sciences, Engineering and Technology 6(7): 1283-1287, 2013
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: November 13, 2012 Accepted: January 11, 2013 Published: July 05, 2013

Research on Feature Extraction Method for Handwritten Chinese Character Recognition Based on Kernel Independent Component Analysis

He Zhiguo and Yang Xiaoli
School of Computer Science, Panzhihua University, Panzhihua 617000, China

Abstract: Feature extraction is very difficult for handwritten Chinese character because of large Chinese characters set, complex structure and very large shape variations. The recognition rate by currently used feature extraction methods and classifiers is far from the requirements of the people. For this problem, this study proposes a new feature extraction method for handwritten Chinese character recognition based on Kernel Independent Component Analysis (KICA). Firstly, we extract independent basis images of handwritten Chinese character image and the projection vector by using KICA algorithm and then obtain the feature vector. The scheme takes full advantage of good extraction local features capability of ICA and powerful computational capability of KICA. The experiments show that the feature extraction method based on KICA is superior to that of gradient-based about the recognition rate and outperforms that of ICA about the time for feature extraction.

Keywords: Feature extraction, handwritten Chinese character recognition, independent component analysis, kernel independent component analysis

INTRODUCTION

Handwritten Chinese Character Recognition (HCCR) is an important research topic in pattern recognition, which is widely used in automatic input of Chinese electronic data processing, in 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 subjects in pattern recognition. Because Chinese characters set (He and Cao, 2008; Liu and Hiromichi, 2008; Zhu *et al*., 2011) 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 has not yet reached satisfactory results, especially for Chinese characters with cursive script. HCCR is divided into four steps: preprocessing, 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. This study mainly studies the feature extraction method for HCCR. The feature extraction and selection is the key for pattern recognition. It requires the selected features as much as possible to meet the following three characteristics: with a fast speed for feature extraction; with a good stability for feature extracted, namely, when the same pattern occur small changes, it requires that the feature changed must be small; with strong classification ability for the feature selected. There is always a contradiction between stability and classification capabilities regardless of what features is selected. Determining how to extract stable and good separable feature for Chinese character is an important research direction. The bottleneck 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. For HCCR, we believe that good features should make the differences as small as possible between the different writing samples for the same Chinese character and make the differences as large as possible for different Chinese characters.

At present, in HCCR, the method for feature extraction can be mainly divided into two categories: one is based on structural features, which is based on strokes and obtains a collection of strokes and is often described by the spatial relationship between strokes or radicals. It can embody the essential characteristics of Chinese characters and is not sensitive to shape changes of Chinese characters, but is rarely used because it is difficult to extract and very sensitive to noise. The other is based on statistical feature which is widely used in HCCR. It is based on dot matrix image with binary or grayscale value and is obtained by the dot matrix transformation of the character. It is easy to extract and can tolerate the noise and the distortion of Chinese characters and has good robustness and good anti-

Corresponding Author: He Zhiguo, School of Computer Science, Panzhihua University, Panzhihua 617000, China

interference ability, but has weak ability to distinguish similar Chinese characters. There are many statistical features such as the Fourier transform, Gabor transform (Liu *et al*., 2005) and elastic mesh (Wu and Ma, 2003). At present, the widely used statistical feature includes gradient features (Liu, 2007) and features based on Independent Component Analysis (ICA) (Rui *et al*., 2005). However, the gradient feature exist deficiencies such as high computational cost and too high dimensionality of features. Although Principal Component Analysis (PCA) is 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 do not use the high-order statistical characteristics of images (Bartlett, 1998). While ICA is an analysis method based on higher-order statistical characteristics of signal, which fully taken into account the statistical independence of the probability density function of the signal. In PCA, the signal to be processed is assumed the Gaussian distribution; while in ICA, the signal is assumed non-Gaussian distribution which is more in line with realistic problems. Traditionally, ICA method has the following disadvantages (Rui *et al*., 2005): high computational cost; with a longer time for feature extraction; iteration of the algorithm depends on the selection of the initial value; it bases on linear relationship assumption, thus, the situation of poor separation or even not separation often occur. These all limits its use to some extent. Kernel Independent Component Analysis (KICA) (Bach and Jordan, 2002) draws on the concept of ICA and constructs objective function which measured the component independence in nonlinear function space and uses an entire nonlinear function space instead of a fixed nonlinear function. KICA not only satisfy the nonlinear features existed in realistic signal, but also can apply to a wide variety of source distributions and has better flexibility and robustness than ICA. Based on the above analysis, we propose a new feature extraction method, i.e., based on KICA for HCCR. The experimental results show that feature extraction method based on KICA is better than gradient-based method and the traditional ICA method.

ICA AND KICA

ICA: The ICA (Hyvärinen and Oja, 2000) model can be described as:

$$X = AS \tag{1}$$

The model describes how the observed data X can be obtained by mixing the signal source S. The source variable S is a hidden variable which cannot be direct 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.

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

$$Y = WX \tag{2}$$

where,

$$W = A^{-1} \tag{3}$$

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

KICA: It is important to emphasize that KICA (Bach and Jordan, 2002) is not a "kernelization" of an existing ICA algorithm. Rather, it is a new approach to ICA based on novel kernel-based measures of dependence. KICA based not on a single nonlinear function, but on an entire function space in a Reproducing Kernel Hilbert Space (RKHS). It calculates the objective function in entire kernel Hilbert space Rather than a single nonlinear function by using canonical correlation analysis; it maps data from input space to a feature space by using nonlinear mapping and then in feature space, the mapped data is analyzed and processed. Its significant characteristic is: the nonlinear transformation is realized by using kernel function instead of inner product between two vectors and without the need to consider the specific form of the nonlinear transformation. Namely, it uses nonlinear functions of RKHS as contrast function and the signal is mapped from low-dimensional space to a higher dimensional space.

First, we take two vectors x_1 and x_2 for example to introduce objective function of KICA. Let $x^1, x^2 \varepsilon X = R^N$, K_1 and K_2 is Mercer kernel mapped by Φ_1 and Φ_2 in the feature space F_1 and F_2, F is a set of a vector function from R to R. Then the F-correlation coefficient ρ_F is the maximum value of correlation coefficient between vector function $f_1(x_1)$ and $f_2(x_2)$, that is:

$$\rho_F = \max_{f_1, f_2 \in F} corr(f_1(x_1), f_2(x_2)) = \max_{f_1, f_2 \in F} \frac{\text{cov}(f_1(x_1), f_2(x_2))}{(\text{var } f_1(x_1))^{1/2} (\text{var } f_2(x_2))^{1/2}} \quad (4)$$

From the above formula: If x_1 and x_2 are independent, then $\rho_F = 0$. And, if the set is large enough and $\rho_F = 0$, there are two vectors which are independent each other. This shows that ρ_F can be used as a measure of independence between variables.

Using kernel function in RKHS space, the above equation becomes:

$$\rho_F(K_1, K_2) = \max_{\alpha_1, \alpha_2 \in R^N} \frac{\alpha_1^T K_1 K_2 \alpha_2}{(\alpha_1^T K_1^2 \alpha_1)^{1/2} (\alpha_2^T K_2^2 \alpha_2)^{1/2}} \quad (5)$$

where, K_1 and K_2 are the Gram matrices associated with the data sets $\{x_1^1, \ldots, x_1^N\}$ and $\{x_1^1, \ldots, x_1^N\}$, respectively.

In order to solving F-correlation coefficients well, regularization are introduced, the above formula becomes:

$$\rho_F^k(K_1, K_2) = \max_{\alpha_1, \alpha_2 \in R^N} \frac{\alpha_1^T K_1 K_2 \alpha_2}{(\alpha_1^T (K_1 + \frac{Nk}{2} I)^2 \alpha_1)^{1/2} (\alpha_2^T (K_2 + \frac{Nk}{2} I)^2 \alpha_2)^{1/2}} \quad (6)$$

where k is a small positive constant. The above equation is equivalent to solving the following generalized Eigen value problem:

$$\begin{pmatrix} 0 & K_1 K_2 \\ K_2 K_1 & 0 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \rho \begin{pmatrix} (K_1 + \frac{Nk}{2} I)^2 & 0 \\ 0 & (K_2 + \frac{Nk}{2} I)^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \quad (7)$$

Generalizing to more than two variables, it becomes the following generalized Eigen value problem:

$$\begin{pmatrix} (K_1 + \frac{Nk}{2} I)^2 & K_1 K_2 & \cdots & K_1 K_m \\ K_2 K_1 & (K_2 + \frac{Nk}{2} I)^2 & \cdots & K_2 K_m \\ \vdots & \vdots & \cdots & \\ K_m K_1 & K_m K_2 & \cdots & (K_m + \frac{Nk}{2} I)^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix} = \lambda \begin{pmatrix} (K_1 + \frac{Nk}{2} I)^2 & 0 & \cdots & 0 \\ 0 & (K_2 + \frac{Nk}{2} I)^2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & (K_m + \frac{Nk}{2} I)^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix} \quad (8)$$

Its minimum feature value is represented by $\lambda_f^k(K_1, K_2, \ldots, K_m)$. For ease of calculation, we defined the following contrast function:

$$C(W) = I_{\lambda F}(K_1, K_2, \ldots, K_m) = -\frac{1}{2} \log \lambda_f^k(K_1, K_2, \ldots, K_m) \quad (9)$$

Now, in the following, we gave its specific algorithm:

- Given a set of vector data $y^1, y^2, \ldots, y^N$ and the kernel function *K (x, y)* and the parameter matrix *W*:
- Data were whiten:
- Let $x^i = Wy^i$, for each *i*, we obtained a set of estimated source vector $\{x^1, x^2, \ldots, x^N\}$, then calculated the center Gram matrix $K_1, K_2, \ldots, K_m$, which depends on these vectors.
- Solving the smallest eigenvalue in (8) and according to the formula (9), minimizing the contrast function *C (W)* in the direction of *W*, thus we obtained *W*.

FEATURE EXTRACTION FOR HANDWRITTEN CHINESE CHARACTERS IMAGES

It is necessary to do some preprocessing before Chinese character samples is separated by KICA. First, Chinese characters are normalized, then to whiten and to zero mean value of the input signal. After the processing, the KICA algorithm is used to solve the separation matrix *W*. After solving the separation matrix *W*, we can obtain the basis image *Y* by using (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:

$$X = \sum_{i=1}^{N} a_i S_i \quad (10)$$

where, a_i is the feature vector of the Chinese character image *X*.

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

EXPERIMENTS AND RESULTS

In this study, we used HCL 2000 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

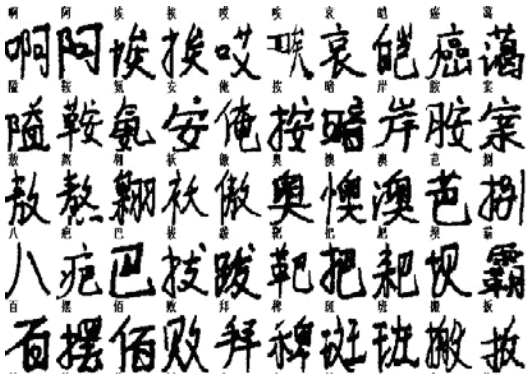

Fig. 1: Some sample images of HCL2000

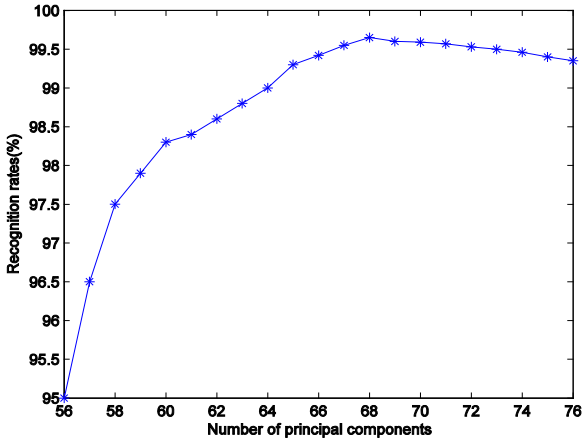

Fig. 2: The relationship between the number of principal components and its recognition rate

information database of writers. All the sample of HCL2000 was normalized binary samples, with size 64 (height) by 64 (width). Part of the samples was shown in Fig. 1. We choose 700 samples located in xx001-xx700 of HCL2000 as training samples, choose 300 samples located in hh001-hh300 of HCL2000 as test samples. When recognizing Chinese characters, for simplicity, we adopted the widely used Modified Quadratic Discriminant Function (MQDF) (Dai *et al*., 2007; Leung and Leung, 2010) as classifier for classification.

When extraction independent component by using KICA 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 67, the recognition rate has reached the highest, shown in Fig. 2.

Feature extraction based on KICA was compared with the widely used gradient feature and ICA method in HCCR. The method for extraction gradient feature please refer to literature (Liu, 2007) and its dimension is 256 and the method for extraction ICA feature please refer to literature (Rui *et al*., 2005). In order to facilitate comparison, three methods used the same MQDF classifier and the results were shown in Table 1. The experiment was performed on Intel Pentium 4 3.2 G with MATLBA 7.0 system equipped with 768 megabytes RAM. From Table 1, we known that the recognition rate based on KICA and 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 seconds; while extraction feature based on ICA needs a long time and the time is concerned with the number of principal components and is approximately two minutes when the number of principal components extracted is 89. The time for extraction KICA feature is also concerned with the number of principal components, but it equals to that of gradient feature and much smaller than that of ICA when the number of principal components is 67. Recognition time is in a few seconds for each Chinese character when using MQDF 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 exist certain relevance, is not independent of each other. This leads to the classification accuracy not high because of the correlation between the features of different categories. 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, KICA has the advantages of ICA and can overcome its deficiencies. The time for extraction KICA features is broadly in line with the gradient method.

Table 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	90s (85)	97.51%
	125s (89)	99.40%
	155s (94)	98.71%
Feature based on KICA	8.5s (57)	97.43%
	9.3s (67)	99.47%
	8.7s (74)	98.52%
Feature based on gradient	5s	97.27%

CONCLUSION

ICA based on the higher-order statistical correlation between data, extracted internal features of the image and 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 characteristics 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. KICA having the advantages of the ICA, while greatly reducing the time required for feature extraction. Therefore, the recognition rate for feature extraction based on KICA is significantly better than that of the current widely used gradient feature for HCCR and the time for feature extraction is roughly the same with gradient feature. This study only researched feature extraction, about classification, which greatly influenced the recognition performance, we will do further research.

ACKNOWLEDGMENT

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

REFERENCES

Bach, F.R. and M.I. Jordan, 2002. Kernel independent component analysis. J. Mach. Learn. Res., 3(12): 1-48.

Bartlett, M.S., 1998. Face image analysis by unsupervised learning and redundancy reduction. Ph.D. Thesis, University of California, San Diego.

Dai, R., C. Liu and B. Xiao, 2007. Chinese character recognition: History, status and prospects. Frontiers Comput. Sci. China, 1(1): 126-136.

He, Z. and Y. Cao, 2008. Survey of offline handwritten Chinese character recognition. Comput. Eng., 34(15): 201-204.

Hyvärinen, A. and E. Oja, 2000. Independent component analysis: Algorithms and applications. Neural Networks, 13(3): 411-430.

Leung, K.C. and C.H. Leung, 2010. Recognition of handwritten Chinese characters by critical region analysis. Pattern Recogn., 43(2): 949-961.

Liu, C., K. Masashi and F. Hiromichi, 2005. Gabor feature extraction for character recognition: Comparison with gradient feature. Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR). Seoul, Korea, Sep. 21-23, pp: 121-125.

Liu, C.L., 2007. Normalization-cooperated gradient feature extraction for handwritten character recognition. IEEE T. Pattern Anal., 29(9): 1465-1469.

Liu, C. and F. Hiromichi, 2008. Classification and learning methods for character recognition: Advances and remaining problems. Stud. Comput. Intell., 90(11): 139-161.

Rui, T., C. Shen and J. Ding, 2005. Handwritten digit character recognition by model reconstruction based on independent component analysis. J. Comput. Aided Design Comput. Graphics, 17(2): 455-460.

Wu, T. and S. Ma, 2003. Feature extraction by hierarchical overlapped elastic meshing for handwritten Chinese character recognition. Proceeding of 7th International Conference on Document Analysis and Recognition (ICDAR'03). Edinburgh, Scotland, Aug. 12-14, 1: 529-533.

Zhu, C., C. Shi and J. Wang, 2011. Study of offline handwritten Chinese character recognition based on dynamic pruned FSVMs. Proceeding of International Conference on Electrical and Control Engineering. Yichang, China, Sep. 16-18, pp: 395-398.