Research on Character Recognition Technology of New Tai Lue Based on Gabor Feature and SVM

Li-xin ZHANG, Peng-fei YU*, Hai-yan LI and Hong-song LI
Information School, Yunnan University,KunMing, China
*Corresponding author

Keywords: Handwritten character recognition, New Tai Lue, Gabor filter, PCA, Support Vector Machine.

Abstract. New Tai Lue script which is also known as Simplified Tai Lue, is an alphabet used to write the Tai Lü language used in Xishuangbanna of Yunnan Province in China. In this paper, we present a handwritten character recognition method of New Tai Lue by Support vector machine (SVM) based classifier on Gabor Features. First, some preprocessing approaches such as superfluous points removing, normalization, smoothing, resampling are applied to preprocess the handwritten character images of New Tai Lue. Second, a filter bank consisting of two-dimensional Gabor filters with various scales and orientations is created to extract features from these images. After that, the result images are processed by PCA to do dimension reduction. Finally, SVM is applied as the classifier to perform recognition task. Experimental results conducted on a database of 16200 characters show that the proposed method are effective.

Introduction

At present, most of the Dai people lived in Xishuangbanna of Yunnan Province in China. After the founding of People’s Republic China, the old Dai characters of xishuangbanna was improved and developed into New Tai Lue. Compared with the studies on Chinese characters[1], Uyghur[2,3] languages, the research of the New Tai Lue character recognition is relatively backward, and there are only a few studies on the recognition of New Tai Lue character. For example, Li et al.[4] presented a printed New Tai Lue character recognition method based on the Back Propagation (BP) Neural Network. A BP neural network is trained with a training set contained 51740 samples, then 4917 samples is applied to test the performance of the proposed approach. The experiment results show that 98.04% recognition rate has been obtained by the proposed method. Yu et al. [5] extracts the character stroke number feature, start and end vector features, additional stroke position features, starting point and end point of the quadrant characteristics, and then apply online random forest algorithm to recognition.

In this paper, we will focus on Gabor-based features and SVM to perform the handwritten character recognition of New Tai Lue. This paper is organized as follows. In Section II describes the sample character collection and preprocessing procedure. Section III presents the Gabor-based feature extraction method and SVM based recognition approach. In Section IV, the experimental results are reported. Conclusions are drawn in Section V.

Sample Character Acquisition and Preprocessing

The Structural Features of the New Tai Lue

The international standard of the New Tai Lue is shown in Figure 1. As can be seen from Figure 1, the structural features of the characters in the New Tai Lue are as follows:

(1) The number of strokes of characters of the New Tai Lue is between 1 and 3, and there are no such strokes as horizontal or vertical strokes.

(2) The characters of the New Tai Lue are composed of left and right parts, upper and lower parts or a whole, and the writing order is relatively simple from left to right, top to bottom.
Some characters of New Tai Lue have strong similarity, which is easy to be confused without attention.

Some complex New Tai Lue characters can be made by adding strokes on simple New Tai Lue characters.

Sample Collection of Handwritten Characters of the New Tai Lue

The collection tool used in this paper is an Android APP, which is developed by our laboratory and used in the collection of handwritten character samples of New Tai Lue. The UI of this App is shown in Figure 2. There are a gray rectangle with red border, which is the input area for the New Tai Lue characters.

Sample Preprocessing

The purpose of character preprocessing is to remove the redundant information in the track points of the character to retain as many useful messages as possible, and to lay a good foundation for the feature extraction of character information.

In this paper, the preprocessing of handwritten of New Tai Lue characters includes the following steps: remove repeating points, normalization, smoothing, resampling, and so on.

Deduplication: in order to eliminate the influence of changes in the number of sample character sampling points caused by handwriting speed, it is important to delete the repeated points in the stroke point sequence.

Normalization processing of characters can be divided into character size normalization and position normalization, in which character size normalization is to convert characters of different sizes into the same size. The normalization of character position is to move the center of mass of the sample character to the desired position. After normalization, the size of each sample is 100 × 100.
In the process of handwriting, characters are often accompanied by various noise, such as quantization noise, random noise, or the writing character is too scribbled to be distinguishable. To solve these problems, smoothing should be done on the characters.

If a character is wrote quickly, a small number of character trajectory points are obtained. As a result, the character recognition rate is affected. To solve this problem, it should be resampled to get new trajectory points along the character tracks. The point sampling interval is as much as possible and the sample character should be resampled.

The resulting image after the preprocessing is shown in Figure 3.

Figure 3. Preprocessed image.

Gabor Feature Extraction of Sample Characters and Support Vector Machine Recognition

In this paper, Gabor filter [6] was used to extract features from handwritten New Tai Lue scripts.

Gabor Filter

The Fourier transform is the integral of the signal in the whole time domain, which reflects the overall characteristic of the signal in the frequency domain, and does not have the local characteristics of a signal. In order to obtain the signal’s local information generated by Fourier transform, the local window function is introduced, and the transformation is called the window Fourier transform. It's only related to partial time signals, and is called the Gabor transform.

Gabor filters are self-similar, and all Gabor filters can be formed by their expansion and rotation.

Two-dimensional Gabor filter is obtained by multiplying sinusoidal carrier by gaussian kernel function. Two-dimensional Gabor function is denoted by:

\[ g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp \left( -\frac{x'^2 + y'^2}{2\sigma^2} \right) \exp \left( i \left( 2\pi \frac{x'}{\lambda} + \psi \right) \right) \]

(1)

\[ \begin{cases} x' = x \cos \theta + y \sin \theta \\ y' = -x \sin \theta + y \cos \theta \end{cases} \]

(2)

In the formula, \( \lambda \) represents the wavelength of sine function; \( \theta \) represents the direction of the Gabor kernel function; \( \psi \) denotes phase offset; \( \sigma \) represents the standard deviation of the gaussian function; \( \gamma \) represents the ratio of width to height in space.

The equation of the Gabor filter used in this paper is:

\[ g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp \left( -\frac{1}{2} \left[ \left( \frac{x'}{sx'} \right)^2 + \left( \frac{y'}{sy'} \right)^2 \right] \right) \sin \left( 2\pi \frac{x'}{\lambda} \right) \]

(3)

where \( sx' \) and \( sy' \) represents respectively Gaussian envelope constants along the x-axis and y-axis.

The output of Gabor filter is divided into two parts: real part and imaginary part. The real part is used in this paper, and the image texture features in different directions are extracted. The scales of the Gabor filter used in this paper are 2, 4, 6, and 8, moreover the directions are 0, \( \pi/4 \), \( \pi/2 \), 3\( \pi/4 \), \( \pi \); the Gabor filter bank has a total of 20 output features, namely a sample image input to the Gabor filter can output 20 texture images.
PCA Dimension Reduction

PCA\[7-9\] is obtained by transformation on the principle of K-L Transform. It can be used to convert a high-dimensional spatial information of the sample to a low-dimensional space. As a result, most of the useful information of the original sample is reserved. In terms of mathematics, this algorithm is the best in statistics. The mean square error before and after transformation is very small, and the low-dimensional space obtained has relatively good discrimination performance. It gets the most expressive feature, not the most discriminative feature.

PCA is a linear transformation. It transforms the data into a new system which retains the features of large variance contribution and removes the features of small variance contribution.

If the number of images in the training sample is M, and the mean u of the input images is:

\[
u = \frac{1}{M} \sum_{i=0}^{M-1} x_i
\]

The scatter matrix is expressed as:

\[
S_x = \frac{1}{M} \sum_{i=0}^{M-1} (x_i - u)(x_i - u)^T
\]

In order to obtain the eigenvalue of the \(S_x\), the following method can be derived to solve the eigenvector obtained by orthogonal normalization.

Let A be a matrix of rank r, Its vector can be expressed as a matrix of \(n \times r\). There are two orthogonal matrixes U, V with rank r and a diagonal matrix with rank r. Its diagonal elements are \(\lambda_0, \lambda_1, ..., \lambda_{r-1}\). They are arranged from big to small. There also exists:

\[
A = U\Lambda V^T
\]

\(\lambda_0, \lambda_1, ..., \lambda_{r-1}\) are the eigenvalues of matrices \(AA^T\) and \(A^TA\), in which the U eigenvectors and V corresponding to the eigenvalues in the matrices are the eigenvectors corresponding to the eigenvalues in the matrices. This is called the singular value decomposition of matrix A, with the opposite singular value \(\sqrt{\lambda_i}\).

From the above:

\[
u_i = \frac{1}{\sqrt{\lambda_i}} X v_i, i = 0,1, ..., M - 1
\]

That's the eigenvector of the image.

The training sample image can be projected into the subspace of \(\nu_i\), and the test sample image can also be projected to it and obtain the coordinate coefficient, so as to prepare for the recognition of sample characters. After feature compression, dimension is reduced from 16,588,800 to 3,709,800, and then input it into a classifier to complete the recognition task of handwritten characters.

Support Vector Machine

The classifier used in this paper is support vector machine (SVM) [10, 11], which was proposed by Corinna Cortes and Vapnik in 1995 and has now become a very important research result in the field of machine learning. Its theoretical basis is obtained from statistical learning. That is, if the provided data follows a certain distribution, in order to truly realize the difference between the ideal output of the machine and the actual output as small as possible, the machine should be replaced by the principle of minimizing structural risks. Other principles are dealt with. This method is not only simple, but also has strong adaptability to new samples and is widely used in machine learning.

When the support vector machine encounters the problem of sample classification, it is usually divided into three types of situations for discussion: linear separability, linear inseparability, and nonlinear separability.

When a linear separable problem is encountered, a space must have a hyperplane that completely separates the sample points. The hyperplane classification equation is:

\[w \cdot x + b = 0\]
And there is an n-dimensional vector, and B is the offset.
The optimal hyperplane is obtained when the plane vector of the separated sample data is farthest from the hyperplane. The optimal hyperplane it creates can usually be transformed into the following minimum cost function:

$$\min \Phi(w) = \frac{1}{2} \|w^2\|$$  \hspace{1cm} (9)

The constraint condition is:

$$y_i[(wx_i) + b] - 1 \geq 0, i = 1, 2, ..., n$$  \hspace{1cm} (10)

According to the above conditions, the Lagrange function is defined as follows:

$$L(w, b, a) = -\frac{1}{2} (w^T w) - \sum_{i=1}^{n} a_i (y_i [(w^T x_i) + b] - 1)$$  \hspace{1cm} (11)

When the number of features is large, the above minimum cost function can be converted to the corresponding dual problem:

$$\max W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i, x_j)$$  \hspace{1cm} (12)

The constraint condition is:

$$\sum_{i=1}^{n} y_i \alpha_i = 0, \quad \alpha_i \geq 0, \quad i = 1, 2, ..., n$$  \hspace{1cm} (13)

When solving and analyzing the above optimization problems, the solution of (13) is satisfied under KKT(Karush-Kuhn-Tucker) condition:

$$\alpha_i^* [y_i (w \cdot x + b) - 1] = 0$$  \hspace{1cm} (14)

In order to solve the above problem, the obtained weight coefficient and classification label must also be used. The weight coefficient expression is as follows:

$$w^* = \sum_{i=1}^{n} \alpha_i y_i x_i$$  \hspace{1cm} (15)

The above Eq. 13, only then did the samples be classified, which is called the support vector.

According to the above conditions, the final classification optimization function is:

$$f(x) = \text{sgn}\{\sum_{i=1}^{n} \alpha_i^* y_i (x_i^T x) + b^*\}$$  \hspace{1cm} (16)

The solution of Eq.14 can be obtained.

When linear is not separable, you need to add a relaxation term $$\xi_i \geq 0 (i = 1, 2, ..., n)$$ and penalty factors C, then Eq. 10 becomes:

$$y_i[(w^T x_i) + b] - 1 + \xi_i \geq 0$$  \hspace{1cm} (17)

Similarly, the optimization problem is changed from Eq.9 to

$$\Phi(w, \xi) = \frac{1}{2} \|w^2\| + C \sum_{i=1}^{n} \xi_i$$  \hspace{1cm} (18)

In the presence of linear inseparable, the main method of SVM is shown in the input vector is mapped to a high dimensional linear separable feature vector space, and construct the optimal classification plane in the space.

In the case of non-linear separability, the sample cannot be separated by a linear model. At this point, the kernel function model is introduced. At this point, the nonlinear transformation corresponds the input space to the feature space one by one, and the original hypersurface model is transformed into the hyperplane model in the corresponding feature space. And what you do is to take the inner product and replace with the kernel $$k(x_i \cdot x_j) = \phi(x_i) \cdot \phi(x_j)$$, the Eq.12 becomes

$$\max Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_j, x_j)$$  \hspace{1cm} (19)

The constraint condition is changed from Eq.13
\[ \sum_{i=1}^{n} y_i \alpha_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, 2, \ldots, n \]  

(20)

**Experimental Results**

In the Gabor filter, the orientation and the scale are taken as $3\pi/4$ and 7, respectively. Then 16200, 12150 and 8100 training samples and 2430 test samples are taken respectively. The results are shown in Table 1.

| Training Samples | 16200 | 12150 | 8100 |
|------------------|-------|-------|------|
| SVMs             | 11005 | 8828  | 6386 |
| Recognition Rate (%) | 83.09 | 82.80 | 80.70 |
| Run Time (s)     | 1245  | 579   | 406  |

Because the Gabor filter can be used to extract different features in different scales and directions, the Gabor filter can be applied to obtain sample image output with different features. Some attempts have been made in this paper to superimpose the obtained feature images to build a feature image. The Gabor filter has scales of 2, 4, 6, and 8 with directions of 0, $\pi/4$, $\pi/2$, $3\pi/4$, and $\pi$. The image obtained by the filter bank with different scales and directions are accumulated and sent to a support vector machine. The results with different numbers of training samples are shown in Table 2.

| Training Samples | 16200 | 12150 | 8100 |
|------------------|-------|-------|------|
| SVMs             | 10389 | 8280  | 6033 |
| Recognition Rate (%) | 81.73 | 80.12 | 79.79 |
| Run Time (s)     | 1337  | 506   | 317  |

For the Gabor filter, the scales are 2, 4, 6, and 8 and the directions are 0, $\pi/4$, $\pi/2$, $3\pi/4$, $\pi$, and the images obtained by the filter bank are spliced and sent to a support vector machine. The results of testing the same batch of test samples and training samples are shown in Table 3.

| Training Samples | 16200 | 12150 | 8100 |
|------------------|-------|-------|------|
| SVMs             | 13110 | 10199 | 7319 |
| Recognition Rate (%) | 87.08 | 85.47 | 82.59 |
| Run Time (s)     | 17360 | 8975  | 4241 |

It can be seen from the above tables that, as the number of training samples increases, the recognition rate of test samples increases, and the highest recognition rate is 87.08\%. Under the same training samples, the more support vector machines are obtained, the higher the recognition rate of the test sample is. The third method used in this paper is the best, but it takes longer than the first two methods.

This paper makes a simple analysis of the first method to identify the wrong samples: It is found that the recognition errors of similar samples are relatively large, such as "$\theta$", "$\delta$", "$\emptyset$", and "$e$". At the collection stage of the sample character, as much as possible, the sample character of the person who is familiar with the writing specification and better familiarity with the New Tai Lue is collected.

**Summary**

At present, there are relatively few researchers on character recognition of New Tai Lue scripts. In this paper, the recognition of New Tai Lue characters is carried out based on Gabor features and...
SVM. As a result, the recognition rate of the New Tai Lue character the proposed method in handwritten script reached 87.08%.

In the future, it should be pay more attention to the research of New Tai Lue character recognition. It is hoped that the recognition of handwritten New Tai Lue characters can be put into practice as soon as possible, and it will contribute to the local economic and cultural development for Dai people.

Acknowledgement

The authors would like to thank the financial supports by National Natural Science Foundation of China (Grant No.61462094, 61561050) and Natural Science Foundation of Yunnan Province of China (Grant No.2015FB116).

References

[1] X. Y. Zhang, Y. Bengio, C. L. Liu. “Online and offline handwritten chinese character recognition: A comprehensive study and new benchmark,” Pattern Recognition, 2017, vol. 61, pp. 348-360.
[2] C. Yan, H. Xie, S. Liu, J. Yin, Y. Zhang, Q. Dai. “Effective Uyghur language text detection in complex background images for traffic prompt identification,” IEEE transactions on intelligent transportation systems, 2018, vol. 19(1), pp. 220-229.
[3] W. Simayi, M. Ibrayim, D. Tursun and A. Hamdulla. “Research on on-line Uyghur character recognition technology based on center distance feature,” 2013 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), IEEE, 2013, pp. 293-298.
[4] D. F. Li, P. F. Yu, H. Y. Li and G. Peng. “Printed New Tai Lue character recognition based on BP neural network,” IEEE International Conference on Signal and Image Processing (ICSIP), IEEE, 2016: 339-342.
[5] Y. Yu, P. F. Yu, H. Y. Li and H. S. Li. “Online Handwritten New Tai Lue Character Recognition Using SVM,” Advances in Intelligent Systems and Computing, 2018, vol. 686, pp. 426-434.
[6] M. Yang, L. Zhang. “Gabor feature based sparse representation for face recognition with gabor occlusion dictionary,” European Conference on Computer Vision, 2010, vol. 6316, pp. 448-461.
[7] A. Subasi, M. Gursoy. “EEG signal classification using PCA, ICA, LDA and support vector machines,” Expert Systems with Applications, 2010, vol. 37, 8659-8666.
[8] K. Delac, M. Grgic, S. Grgic. “Independent comparative study of PCA, ICA, and LDA on the FERET data set,” International Journal of Imaging Systems & Technology, 2005, vol.15, 252-260.
[9] L. I. Smith. “A Tutorial on Principal Components Analysis,” 2002.
[10] B. E. Boser, I. M. Guyon and V. N. Vapnik. “A Training Algorithm for Optimal Margin Classifiers,” Proceedings of the Fifth Annual Workshop on Computational Learning Theory 5 144-152, Pittsburgh, 1992.
[11] C. Cortes and V. Vapnik. “Support-Vector Networks. Machine Learning,” 20(3), 273-297, 1995.