Journal of Mechanical Engineering Research and Developments
ISSN: 1024-1752
Website: http://www.jmerd.org
Vol. 39, No. 2, 2016, pp. 536-543

An Improved Manchu Character Recognition Method

S. Xu[†*], G. Q. Qi[‡], M. Li[§], R. R. Zheng[††], & C. John[ξ]

[†‡]Information Science & Technology, Dalian Maritime University, Dalian Liaoning 116026, China, *Email: xushuangcong@163.com
[†§††]College of Information and Communication Engineering, Dalian Nationalities University, Dalian Liaoning 116605, China
[ξ] School of Engineering, University of St. Thomas, St. Paul, MN 55105-1079, USA

ABSTRACT: To improve the off-line Manchu printed character recognition rate, a method of Manchu recognition based on the letters is presented. Firstly, the preprocessing is performed to segment the Manchu letters aiming at Manchu character image. Secondly, extract the rough grid characteristics and connected domain characteristics of the Manchu letters, then using SVM and BP neural network to recognize the combination features of these ones. Finally, the grid-search method and cross-validation method are used to optimize the SVM kernel function parameters. The result of the experiment shows that the recognition rate of SVM is higher than the BP neural network, and has a better classification results.

KEYWORDS: Off-Line Manchu character; BP neural networks; SVM; Grid-Search; Cross-Validation.

INTRODUCTION

In recent years, the recognition research of printed Chinese characters and numbers in domestic and foreign has achieved a great success. But the study on the minority language is still in the initial stage, especially on Manchu character recognition [1]. Manchu as the national language of Qing Dynasty, a large number of political, cultural, economic, military, diplomatic, astronomy and other aspects of the data are recorded by the Manchu, it has a very high historical value, if it disappeared, these materials also loses its value [2]. More important is that these data scattered throughout the museum and have not been to good use, how to use modern recognition technology to excavate these valuable information so that its value can be fully utilized has become a serious problem, and research on Manchu character recognition system is very important for the protection of cultural heritage in the Qing Dynasty.

DESIGN OF MANCHU CHARACTER RECOGNITION SYSTEM

Manchu language is a branch of Altai phylum, a family of Tungus languages, its written form has punctuate so called "punctuate Manchu". The Manchu script consists of thirty-eight letters, of which there are six vowels and twenty-two consonants, in addition ten specific letters are used for spelling Chinese loanwords specifically. This paper studies on the Manchu letters recognition. First, the preprocessing such as binarization, de-noising is performed on the collected Manchu character, then segment the Manchu letters according to particular segmentation algorithm, extract the feature of rough grid and connected domain characteristics of these letters and use SVM and BP neural network to recognize the combination features of these ones, finally analysis the identification results. The frame of Manchu letters recognition system is shown in Figure 1.

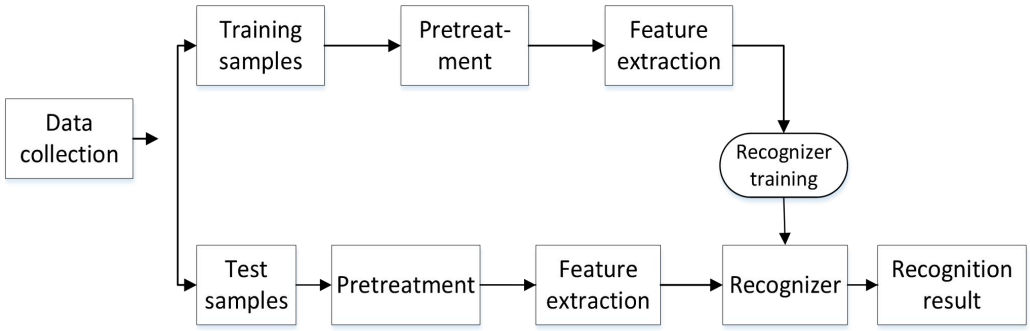


Figure 1. Frame of Manchu letter recognition system.

DOI: 10.7508/jmerd.2016.02.033

PRETREATMENT AND SAMPLE COLLECTION

The first step of Manchu character recognition is to extract the letters from the Manchu character which preprocess is needed. Pretreatment is very important in the recognition system, mainly including binarization, de-noising, tilt correction and other operations. Due to the unique written form of Manchu character, Manchu words can be obtained according to the segmentation on the ranks of Manchu text.

Manchu characters are one of the most complex characters, different letters may be the same variant in the same or different position, one and the same letter may be the same variant in different positions, and even some letters can not appear in certain positions. Therefore, Manchu letters is divided into independent form, first-word, middle-word and final-word according to its position within the Manchu word. The study of Manchu letters recognition method in this paper is based on the initial letters. The segmentation algorithm of letters is as follows:

(1) Confirm the position and width of the spindle. Each Manchu word has a spindle as its trunk, all the letters are conglutinated together through this spindle, moreover the spindle is mostly located in the middle of Manchu word, make of black pixel intensively, therefore doing vertical projection on the Manchu word and differencing the two adjacent rows, find out the maximum and minimum value as the left and right border around the spindle.

(2) Make the pixels in the spindle all white.

(3) Horizontal projection of the Manchu word which spindle is white.

(4) Find the candidate segmentation points. Find the blank space position of Manchu word horizontal projection, calculate the length of blank space, then its midpoint is the candidate segmentation. Finally, merge the over segmentation letters according to the characteristics and rules of word formation of Manchu words .Segmentation of Manchu words as shown in Figure2, part first-word of the Manchu letters as shown in Table 1.

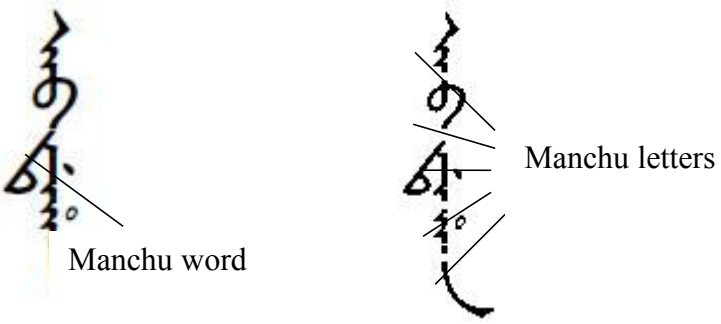
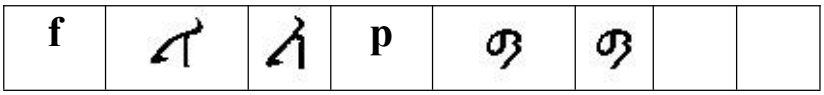


Figure 2 Manchu word segmentation.

Table 1. Part first word of the Manchu letters.

letter	Manchu		letter	Manchu			
a			**h**				
b			**k**				
c			**l**				
d			**m**				
e			**n**				



The 26 letters are collected in 8 different fonts, because the same letters in the same position may be more than a written form, and the letter v do not have first-word form, therefore each font have 39 samples, first-word letters contains 324 samples.

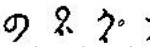
FEATURE EXTRACTIONS

Because the original input image data is very large, in order to achieve the purpose of classify the input document image, characteristics must be extracted, select features with good discrimination and abandon the characteristics of little contribution to the classification. Statistic features and Structure features are commonly used in character recognition, among which structure feature is mainly used to extract some characteristics of character structure and statistic feature is the statistics of whole character [3].The statistical features mainly contains rough grid characteristics, projection transform coefficient characteristics, connected domains, aspect ratio and character template feature etc [4]. This paper mainly extract the rough grid and connected domain features of Manchu letters and combine features of these ones as the recognition features.

Rough Grid Characteristic.

Grid characteristic is one of the features commonly used in character recognition, and it belongs to a kind of local features in statistic features. First divided the image into M×M grids, then count the ratio of black pixels in each grid, finally set the characteristics as the statistical features of recognition [5].In this paper a Manchu letter is divided into 8×8 grids, count the proportion of black pixels within each grid sequentially as one of the recognition features.

Connected Domain Characteristic.

Manchu characters have a very special written structure different from Chinese and English, it is written from top to bottom, from left to right, and all letters are conglutinated together through a Mid-axis. It include connected unites and unconnected unites, such as  .

Combining the features of rough grid and connected domain to get a 65-dimensional feature vector, that is the characteristic of Manchu letter recognition.

CLASSIFIER DESIGN

The Principle of SVM.

Support vector machine (SVM), developed by Vapnik et al., is a new universal learning method based on statistical theory, it is based on the statistical learning theory of VC dimension theory and structural risk minimization principle, it can solve small sample, non-linear, high dimension and local minimum problems, and has become one of the research focus in the field of machine learning. It trains a group of feature subset which are support vectors, making the division of support vector set is equivalent to the division of the entire data set. In the field of Natural Language Processing, SVM is applied to text classification and achieved very good results.

Two Classification Problems.

The SVM method was originally proposed for two classification problems. Its principle is to construct an optimal hyperplane; it can not only separate two class samples correctly but reached the maximum classification intervals. The schematic diagram of SVM is shown in Figure3, squares and circles represent two kinds of samples, H is a classification hyperplane, H1 and H2 are two planes which cover the nearest samples to the classification hyperplane within the two types of samples and parallel to the classification plane. The distance between H1 and H2 is called class interval, the vectors above it is called support vector.

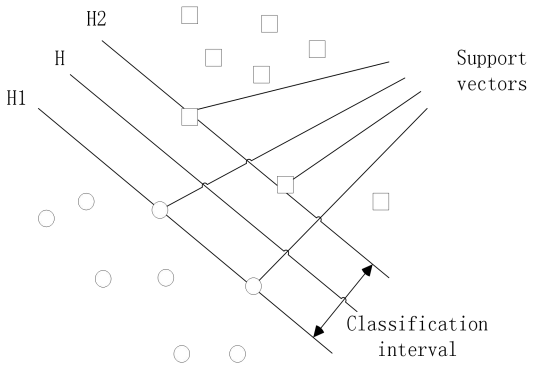


Figure 3. The schematic diagram of SVM.

The specific steps of construct optimal classification plane are as follows:

① Let the sample set $T=\{(x_1,y_1),\cdots,(x_n,y_n)\}\in(X\times Y)^n$, among them $x_i\in X=R^d$, $y_i\in Y=\{-1,1\}$, $(i=1,2,\ldots,n)$; x_i is the feature vector, n is the dimension of sample feature space, y_i is the category which sample x_i belongs to, the general form of a linear discrimination function in D dimensional space is

$$g(x)=\omega^T\cdot x+b \tag{1}$$

The classification plane equation is

$$\omega^T\cdot x+b=0 \tag{2}$$

Normalization the discrimination function so as to all samples of the two classes are satisfied with $|g(x)|\geq 1$, that is to say the samples nearest to the classification satisfied with $|g(x)|=1$,so the classification interval is equal to $2/\|w\|$, thus making the interval maximum equals to minimum $\|w\|$, at the same time the two types of samples can correctly separated, therefore the hyperplane should meet

$$y_i\left[\omega^T\cdot x_i+b\right]-1\geq 0\,,i=1,2,\ldots,n \tag{3}$$

So, the classification plane which satisfied the type above and minimum $\|w\|^2$ is the optimal classification plane.

② The definition of Lagrange function is

$$L(\omega,\mathrm{b},\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{n}\alpha_i y_i\left(w^T x_i+b\right)+\sum_{i=1}^{n}\alpha_i \tag{4}$$

Where $\alpha_i>0$ is the Lagrange coefficient.

③ Problem can be transformed into a simple dual problem based on the original constrains, that is

$$\text{Max}\ \sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\left(x_i{}^T x_j\right) \tag{5}$$

$$s.t.\ \sum_{i=1}^{n}y_i\alpha_i=0\,,\alpha_i\geq 0\ \ i=1,2,\ldots,\mathrm{n} \tag{6}$$

If α_i^* is the maximum value, then $\omega^*=\sum\alpha_i^* y_i x_i$, α_i^* is not zero corresponding samples is the support vectors.

④ Finally, the optimal classification function is

$$f(x)=\mathrm{sgn}\left(\sum_{i=1}^{n}\alpha_i^* y_i\left(x_i\cdot x\right)+b^*\right) \tag{7}$$

SVM is that use the suitable inner product function $K(x_i,x_i)$ instead of the dot product of optimal classification plane, at this point the optimization function becomes

$$Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K\left(x_i,x_j\right) \tag{8}$$

In this case the corresponding discrimination function becomes

$$f(x)=\operatorname{sgn}\left(\sum_{i=1}^{n}\alpha_i^* y_i K\left(x_i,x\right)+b^*\right) \tag{9}$$

Multi-class Classification Problems.

The SVM above can solve two classification problems, but in reality multi-class classification is the main problem. So there are two ways when dealing with multi-classification problems, one method is one against one method, to K categories, construct $K(K-1)$ SVM, if a classifier determines the test sample X belongs to class I, the class I get a ticket, finally the class with most votes is X belongs to. This classification method has a small scale compared with other methods and learned easier. But there are also shortcomings, the more category, the more classifiers is needed, computation will be increase correspondingly; in addition, when decisions are made using the method of voting, there could be more than one category to vote the same circumstances, the indistinguishable category may be exist.

Another method is one against rest classification method. For K classes construct K SVM sub classifiers. For example, labeled the samples belongs to class I as positive class when construct I sub classifier, labeled the samples not belongs to class I as negative class. Calculate the decision function value of each sub classifier when testing the test samples, and then select the category with the most function value as the tested samples category. The speed of this method is relatively faster, however, due to the training of each classifier is trained the entire samples, it requires to solve L quadratic programming problems with n variables, so need longer training time; moreover the situation may occur that the test sample may belongs to multi classes at the same time or do not belong to any class.

LIBSVM is SVM pattern recognition and regression software package designed by Professor Lin Zhiren of National Taiwan University, and it has the advantages of simple, easy to use and effective. A feature of the software is the parameters of the SVM need less regulation, in other words it provides lots of default parameters, and these parameters can solve many problems. The performance of the SVM classifier is affected by many factors, one of the more critical is the error penalty parameter C and the kernel function and its parameters. The penalty parameter is used to adjust machine learning confidence range and empirical risk ratio in a specific space so that the generalization ability of learning machine is the best, its value depend on the specific issues. But for the kernel functions, different kernel function has an effect on the classification performance, the same kernel function have different parameters also affect the accuracy of classification. There are four kinds of commonly used kernel function:

(1) Polynomial kernel function: $K(x,x_i)=[\gamma^*(x\cdot x_i)+coef]^d$,where d is the order of polynomial, $coef$ is the bias coefficient;

(2) RBF kernel function: $K\left(x,x_i\right)=\exp(-\gamma^*\left\|x-x_i\right\|^2)$,where γ is the width of the kernel function;

(3) Sigmoid kernel function: $K\left(x,x_i\right)=\tanh(\gamma\left(x\cdot x_i\right)+coef)$;

(4) Linear kernel function: $K(x,x_i)=x^T x_i$

Parameter Optimization.

The parameter selection process of SVM is actually the process of parameter optimization. Choosing appropriate parameters after the kernel function is determined, the commonly used methods to choose the parameters are grid-search method, cross validation method, genetic algorithm and particle swarm optimization algorithm [6-7], this paper mainly studies the grid-search method and cross validation method.

Grid-search Method.

In the classified using SVM, for example, the values of relevant parameters must be determined when using RBF kernel function, namely the penalty parameter and the kernel function parameters. The grid-search method [8-9] is to give a range of parameter C ,that is $C=2^{-10}\sim 2^{10}$, the search step is -1; the range of γ is $\gamma=2^{-5}\sim 2^{5}$,search step is -1; for each parameters (c,γ) are trained on the original training set, then get the classification accuracy under this

C and G, eventually take the parameters with highest classification accuracy as the model parameters to predict the test set, then get the classification labels and classification accuracy rate of prediction samples.

Cross Validation Method.

Using cross validation (CV) [10] can get the optimal parameters in some sense, effectively prevent the occurrence of over learning and less learning state, and then get the ideal accuracy of the test set. The basic idea is to put the original training data set into two groups, one group will be used as the training set, the other as the test set. Training the classifier by the training set and then use the test set to test the trained models, set the correct classification rate as the performance index to evaluate the classifier, finally select the optimal parameters.

MANCHU LETTER RECOGNITION BASED ON SVM

In this paper, SVM is used to classify the Manchu letter features. One method is one against one method, due to a total of 25 classes, 300 classifiers are needed. The original data is divided into two parts, namely training set and test set; firstly, use the training set to train the classifier, then classify the test set by each classifier, vote for all combinations and finally the category with most votes is the test samples belongs to. The number of samples of training set and test set and the classification accuracy of different kernel functions are shown in Table 2.

Table 2. The accuracy of different kernel functions in one against one method.

Kernel function / Training set and test set	Linear Kernel function	Quadratic Kernel function	Polynomial Kernel function	RBF Kernel function
78 training samples 234 test samples	61.97% (t=6.12 s)	36.75% (t=7.785 s)	42.74% (t=8.673 s)	38.03% (t=4.515 s)
156 training samples 156 test samples	74.36% (t=8.124 s)	53.85% (t=10.198 s)	53.85% (t=11.480 s)	64.1% (t=6.240 s)
234 training samples 78 test samples	80.77% (t=13.009 s)	69.23% (t=12.022 s)	74.36% (t=11.799 s)	78.21% (t=7.307 s)

Another method is one against rest method. The number of training set and test set is the same as above, and the accuracy is obtained by the cross validation method to get the optimal parameters. The data analysis is shown in Table 3.

Table 3. The accuracy of different kernel functions in one against rest method.

Kernel function / Training set and test set	Linear Kernel function	Polynomial Kernel function	RBF Kernel function	Sigmoid Kernel function
78 training samples 234 test samples	69.658% (t=19.025 s)	60.683% (t=17.169 s)	69.658% (t=17.571 s)	55.556% (t=17.995 s)
156 training samples 156 test samples	79.49% (t=65.125 s)	74.359% (t=48.907 s)	83.974% (t=38.121 s)	74.359% (t=39.250 s)
234 training samples 78 test samples	97.436% (t=330.715 s)	93.588% (t=347.411 s)	98.718% (t=318.072 s)	84.615% (t=339.788 s)

As can be seen from the above two tables: when using one against one method to classify the printed Manchu characters, whether using linear kernel function, polynomial kernel function, quadratic kernel function or RBF kernel function, its classification accuracy rate are lower than the one against rest method; in method one the linear kernel function has the highest accuracy rate of 80.77%, while the accuracy rate of RBF kernel function is almost reached

98.718% in method two, an improved nearly 20% than method one, that is because of the RBF kernel function has divisibility and its interpolation ability is strong, and is good at extracting the local properties of samples, so it's have a better classification effect; on the other hand, the number of training samples and classification results have a direct relationship, the more training samples, the higher rate of accuracy. Through the analysis of the classification error categories, find  and , and , and , and always wrongly classified due to its similar writing shape.

In order to verify the accuracy of classification of SVM, classification and recognition based on BP neural network to the same training and test sets, the results are shown in Table 4, compared with the quadratic kernel function SVM, experiment shows that the BP neural network classification accuracy is significantly lower, and the training time is longer. That is because the BP neural network is a kind of optimization method of local search, the network weight is adjusted gradually along the direction of the local improvement, and it will make the algorithm into a local extreme. Coupled with the network's initial weights are randomly selected, the classification accuracy is low and the accuracy of each training time is also different, with instability; meanwhile, because the neural network learning process is repeated in cycles, it need a long time which lead to a slow convergence speed, so the classification accuracy is not satisfactory.

Table 4. SVM and BP neural network classification accuracy.

Training samples and test samples	SVM	BP neural network
78 training samples，234 test samples	36.75% (t=7.785 s)	10.97% (t=143.98 s)
156 training samples， 156 test samples	53.85% (t=10.198 s)	15.64% （t=378.5 s)
234 training samples 78 test samples	69.23% (t=12.022 s)	20.82% (t=529.12 s)

CONCLUSIONS

This paper using SVM and BP neural network to identify the Manchu letters, collect eight kinds of fonts of Manchu head-writing letters and there are 324 original sample data. Set the features of rough grid and connected domain as the Manchu head-writing classification features. BP neural network classifier because of its inherent instability lead to each classification results are different and the classification accuracy rate is relatively low; while the SVM exhibits unique advantages in dealing with small samples, nonlinear and high dimensional pattern recognition, the choice of kernel function is very important, it translate the training samples in the original problem into training samples which is linearly separable in feature space, according to the grid search and cross validation method to optimize parameters, finally use the one against one method and one against rest method to recognize, then compared the results. Experiments shows that the classification accuracy of SVM is obviously higher than that of BP neural network, and the kernel function of SVM is RBF when reached the highest classification accuracy.

ACKNOWLEDGMENT

This works are partly supported by Liaoning Province Natural Science Fund Project (Grant No 2015020084), the Science and Technology Research Project of Liaoning Provincial Department of Education (Grant No L2015127 and L2014548) and the Research Project of State Ethnic Affairs Commission (Grant No 14DLZ007) .

REFERENCE

[1] W. Wei, C. Guo, Off-line Manchu character recognition based on multi-classifier ensemble with combination features, Comput. Eng, 6(2012):2347-2352.

[2] G. Y. Zhang, Study of Offline Handwritten Manchu Character Recognition, Northeastern University, 2006.

[3] Z. H. Li, G. L. Gao, Extraction of Features of Mongolian Printed Character Recognition, Microcomputer, 11(2013):117-119.

[4] J. Z. Jia, The Research of feature selection and classifier design for Printed Offline Uygur character recognition, D. Soochow University, 2008.

[5] Y. L. Wang, Y. Z. Li, R. S. Wang, Application of Rough Grid in Feature Extraction of Printed Tibetan character, Sci. Technol. Eng, 18(2009): 5546-5548.

[6] Y. D. Song, Research of Parameter Selection for Support Vector Machine, Central China Normal University, 2013.

[7] Francesco Camastra, A SVM-based cursive character recognizer, Pattern Recognit, 40 (2007): 3721-3727.

[8] G. H. Feng,. Parameter optimizing for Support Vector Machines classification, Comput. Eng, 3, (2011):123-128.

[9] G. Sun, Research of Multi-class Classification Methods Based on Support Vector Machine, Dalian Maritime University, 2008.

[10] X. L. Wang, Z. B. Li, Identifying the Parameters of the Kernel Function in Support Vector Machines Based on the Grid-Search Method, Journal of Ocean University of Qingdao, 5(2012):859-862.