A Novel Blocked Winner Sequence Feature Extraction Method Based SOM For Large Set Chinese Character Recognition

Lei Wang*,Wenjuan Zhang*,Lianming Wang*†

**Institute of Computational Intelligence,Northeast Normal University, Changchun, Jilin, 130024.*
†corresponding author:e-mail:wanglm703@nenu.edu.cn.

Keywords:feature extraction; blocked winner sequence; SOM neural networks; Chinese characters recognition; large character set.

Abstract

A novel feature extraction method named blocked winner sequence feature extraction (BWS) method based on SOM is proposed for recognizing large set of Chinese characters. Firstly, a Chinese character sample is blocked to make sub block set according to certain principles; Secondly, every sub-block is passed into a SOM network orderly for learning, and the location number of the winner for each sub-block is stored; Finally, location numbers of all winner neurons are combined into a sequence as the feature of the character. Compared with traditional SOM neural networks, the method using orderly combination of multiple winner neurons instead of single winner to represent feature can reduce the network size and computation as well as improve network capacity effectively. Using block can improve the anti-interference ability and the recognition accuracy of the network. The method is used for the extract feature of 3500 Chinese characters in GB2312 Chinese character library with Euclidean distance recognition, and the recognition accuracy reaches 99.897% and 93.274% respectively in the cases of adding 10% and 20% random noise.

1 Introduction

Self-organizing Map (SOM)，invented by Professor Teuvo Kohonen, is a data visualization technique which can reducethe dimensions of data with the use of self-organizing neural networks[1]. SOM neural networkshas been widely applied tospeech recognition, image processing, data dimension reduction, data mining and other fields[2,3] as one of the hot research topics in the field of artificial neural networks.

Similar to self-organizing characteristic of the human brain, SOM neural networks have a simple structure，which gets wide attention from many computer vision scientists. When the traditional SOM neural network is used for image recognition, each neuron in the input layer need correspond to a pixel of the image, and each sample image need correspond to a neuron in the output layer at least, which causes large network size and computational complexity for large samples.

To reduce the size and increase processing speed，SOM neural networks has been improved by many researchers in different applications. The neuron of output layer is improves for spike neuron applied in the classification of cancer dataset and improves the processing speed of SOM in document[4]. Genetic algorithm is adopted to choose network weights in document[5]. The number of neurons in the competitive layer is decided by formula in document[6] and initial weight vectors confirmed by sample center vector are applied to for remote sensing image classification to improve the classification efficiency and accuracy. The Supervised network, whose weight is adjusted according to different input samples to predict classification and choose different formulas by means of adding the output layer to the input layer and the competitive layer, is proposed for data set KDD CUP99 to finish invaded test experiment and improve invaded detection rate in document[7]. Semi-supervised pattern classification using tree topology SOM and a few neurons is realized in document[8] which is more reasonable than other classification schemes only using marked instances in training stage. Associated SOM is set up based on C-V model for image segmentation in document[9], and different SOM networks is used to process foreground and background to improve the identify of network. Some methods of combining contour models and SOM network in image segmentation are summarized in document[10] to extract image edge information better.

The recognition difficulty of Chinese characters is far greater than other language because of the multiple quantity and changeable structure. At present, some progress has been made in traditional text recognition technology and document analysis technology. At present, traditional text recognition technology and document analysis technology have made some progress, and the recognition rate of optical character recognition (OCR) can currently reach 99.99%, but high quality must be required for the picture and it is by scanning software and scanning process. Existing Chinese character recognition methods are mostly targeted at Chinese character recognition of small character set, and recognition rate is affected by font and character set[11].

Prepossessing and invariant feature extraction are combined on the basis of the stroke tangent and the invariance of the bending in document[12] based on HMM method. With the use of point and stroke oriented, the recognition accuracy of small character set, medium character set and large character set are

98.4%, 96.5% and 91% respectively. Recognition experiments for Chinese characters of large character set based on HMM show that to improve the recognition rate of large character set is still a problem to be solved by HMM. Enhanced Bias classifier is designed for handwritten numeral recognition in document[13], and recognition rate of Bayesian network is 99.29% which is much higher than traditional Euclidean distance, but it needs more samples and complex analysis and calculation. Recognition system based on multi features and parallel neural networks computing is constructed in document[14]. Odate, R et al.[15] formalized the candidate reduction technique for the Nearest Neighbor (NN) problems,and proposed an improved method that works fine with Chinese character sets with a faster and more accurate. Based on recurrent neural networks, Dapeng Tao et al.[16] proposed an effectively method by integrating a principal component convolution layer with the 2-D long short-term memory to recognize Chinese character. Xiaoqing Ding et al. [17] extracted wavelet features from the image transformed by wavelet transform, and classified font class with Modified Quadric Distance Function classifier, which achieved a recognition rate of 90.28% on a single unknown character and 99.01% for five characters. The subsystem composed of parallel model solves problems with cluster computer because of the large amount of computation. Deep convolution neural network is used for the recognitionof similar handwritten Chinese characters with higher recognition rate and larger network size in document[18]. Therefore, The Chinese character recognition of larger characters set is studied less at present. In this paper, feature extraction method of Blocked Winner Sequence (BWS) feature extraction is proposed to simplify the network structure, reduce the amount of computation and improve the recognition accuracy targeting at The Chinese character recognition of larger characters set.

2 SOM neural networks

2.1 Topology structure of SOM neural networks

The human brain is composed of a large number of cells. Functions of different parts of the human brain are different, and sensitive degrees from an aspect or a particular stimulus signal for different regions of the cells are also different. According to these characteristics, the theory of self-organizing map was proposed by Professor Kohonen, a neural networks expert in Finland in 1981[19].SOM is one of the most widely used architectures for unsupervised neural networks, and the core idea is that neural network is divided into different corresponding regions automatically which respond differently to input mode when it accepts external input[20]. SOM neural networks architecture shown in Figure 1 is a feed forward network with full connection and double layers containing an input layer and an output layer. The input layer is used to transmit external information into network. The output layer which is a low dimensional space is also known as the competition layer. The neurons in competition layer are connected with other neurons around laterally to simulate the inhibition function of neurons in the human brain. Arbitrary high dimensional inputs are mapped to low dimensional space by network, and some similar properties of the input data are reflected to adjacent feature maps geometrically that are two-dimensional discrete graphics with the same topology structure for the output layer.

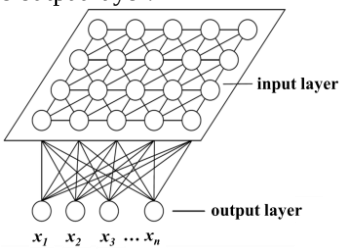


Figure 1:Topology structure of SOM neural networks.

2.2 Learning algorithm of SOM neural networks

Supposing that the input vector is an *n*-dimensional data set, two dimensional SOM neural networks with *m* output neurons is founded. W_{ij} are the connection weights between the ith input vector and the jth neuron in the output layer. The training procedures of SOM neural networks are as follows:

(1) When iterations *t* is equal zero, the initial value of the weight w_{ij} are random numbers between 0 and 1 or from the input vector randomly. Here $N_c(0)$ is called the initial neighborhood function of the winning neuron, and α_0 and *T* are called learning rate and total number of iterations respectively.

(2) If the iteration number *t* is less than total iteration number *T* in the initialization, the learning steps from 3 to 7 are repeated.

(3) Input samples are normalized, and input vector *x* is extracted randomly.

(4) The distance between input vector and all neurons in output layer is calculated when iteration number is *t*. All Euclidean distances are calculated according to Equation (1) to find the minimum, and the corresponding neuron is the winner neuron.

$$i(x) = \arg\min_j \left\| x(n) - w_j \right\|, j = 1,2,...,m . \quad (1)$$

(5) Weight vectors of winner neurons and all neurons in neighborhood are updated.

$$w_j(t+1) = w_j(t) + \alpha(t)(x_i - w_j(t)), j \in N_{j^*}(t), \quad (2)$$

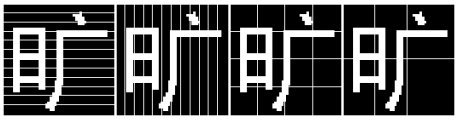
here $\alpha(t)$ is called learning rate function, which will decrease with the increase of the iterations to ensure the convergence of the training process.

(6) Learning rate function and neighborhood function are adjusted in step 6.

(7) Supposed that $t=t+1$, it will be returned to step 3 to resume if $t<T$, otherwise, the cycle will be stopped and the algorithm will be over.

3 Image Block

The image block was first proposed by foreign researchers. According to low face recognition accuracy affected by illumination and posture. Pentland et al. introduced the idea of block for face recognition and proposed a method of block feature space[21]. The more important parts on the face, such as eyes, nose and so on, were chosen and then features extracted from these parts were gathered for recognition. The basic principle of image block is to divide the image into a number of blocks, and to extract features from each sub block image respectively for later use[22]. There are many ways to block the image, row block, column block, equal block and unequal block. Several image block modes are shown in Figure 2.



(a) (b) (c) (d)

Figure 2: Several image block modes, (a) Row block, (b) column block, (c) equal block, (d)unequal block.

The research has shown that block method can not only reduce the dimension of the network input samples, but also extract the local features effectively and make more use of image information. When a piece of image is changed, the image as a whole is not affected. Therefore, the method has good robustness. In the practical application, the different ways of blocking are selected according to the characteristics of the image. The row block is used for simplifying process according to the characteristics of more horizontal and vertical strokes of Chinese characters in this paper.

4 BWS feature extraction and recognition

The schematic diagram of BWS feature extraction is shown in Figure 3.

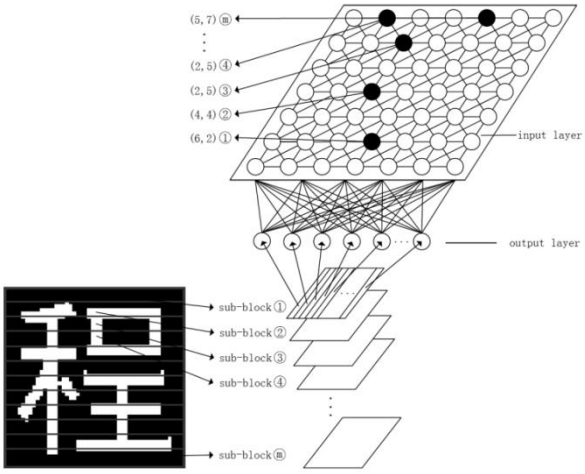


Figure 3: Schematic diagram of feature extraction based on BWS feature extraction.

To take the Chinese character "cheng" for instance, it is divided into m blocks according to a certain principle of block. Each sub-block is arranged in order to build sub-block set, and then input networks orderly. Each sub-block is mapped to a neuron of output layer through the input layer, and the location of neuron is noted down as the sub-block feature. Winner neurons sequence made up of features of the entire sample characters is produced after importing all sub-blocks to the network. Each Chinese character has a corresponding feature sequence, consisting of standard feature library.

BWS feature extraction method adopted in this paper is realized with arrangement feature of winner neurons orderly. The Equation (3) can calculate network capacity *C* based on the above method.

$$C = P_n^r = \frac{n!}{(n-r)!}, \tag{3}$$

here *C* is the classifiable ample capacity of network. P is the permutable number. n is the number of neurons in the output layer of the network. r is the number of winner neurons in the feature sequence of Chinese characters samples, and also the block numbers of original image. According to the method proposed in this paper, the capacity of network depends on the number of neurons in output layer and the block numbers of samples. It is easy to improve the capacity of network through increasing the number of neurons in the output layer and changing the sample blocks.

The flow of BWS feature extraction is shown in Figure 4.

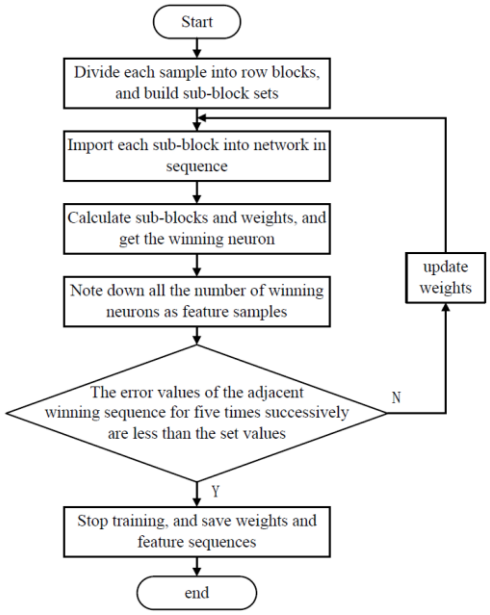


Figure 4: Flow chart of feature extraction based on BWS.

When the error values of the adjacent winner sequence for five times successively are less than the set values in network training process, the output of the network at this time is saved as the features of standard sample and stored in property database for the continued recognition. Error of adjacent winning sequence *e* is calculated as follows:

$$e = \sum_{k=1}^{l} \sqrt{(x_k(i) - x_k(i-1))^2 + (y_k(i) - y_k(i-1))^2}, \tag{4}$$

here $(x_k(i), y_k(i))$ is the number of neuron *k* while the epoch of network training is *i*.

The flow chart of recognition is shown in Figure 5. Firstly, the BWS feature extraction method is used to extract features of sample to form the feature sequence; Secondly, the matching operation is carried out with the feature database of standard sample and the feature sequence obtained above, and the Euclidean distance method is used to calculate the distance. Finally, the sample with minimum distance is taken as the recognition result.

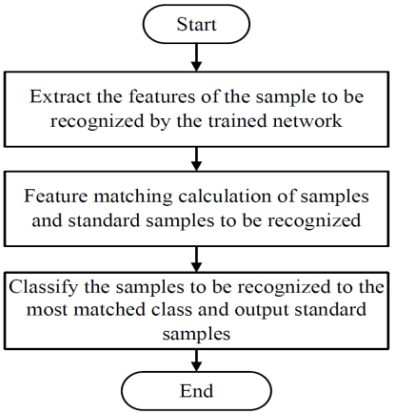


Figure 5:Flow chart of recognition.

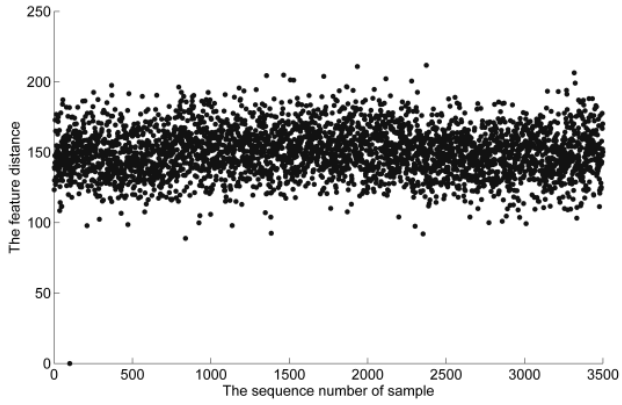
Euclidean distance algorithm is shown as Equation (5), where d_{ij} is the distance between sample i which is to be recognized and standard sample j, (x_k^i, y_k^i) is the number of winner neuron of sub-block k of sample i and (X_k^j, Y_k^j) is the number of winner neuron of sub-block k of sample j.

$$d_{ij} = \sum_{k=1}^{m} \sqrt{(x_k^i - X_k^j)^2 + (y_k^i - Y_k^j)^2} . \quad (5)$$

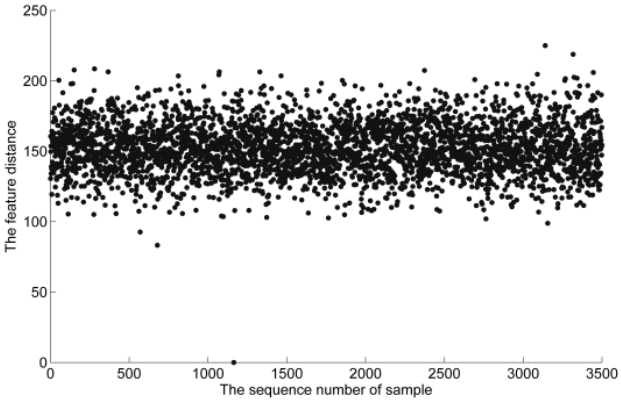
5 The process of Chinese character recognition and results analysis

In order to verify the effectiveness of the method, 3500 commonly used Chinese characters in GB2312 Chinese character library are taken as the sample set, and each character is expressed with a 48*48 pixel binary image.

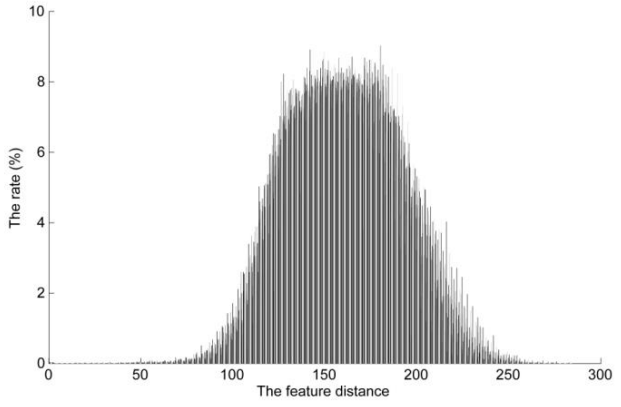
In this paper, all the Chinese characters are blocked by the principle of equal row that each row is a sub-block. Fully connected SOM neural networks with two layers is used, in which the number of neurons in the input layer is equal to the number of a row pixel of samples that the value is 48 and the number of neurons in the output layer is 49 with a two-dimensional structure of 7*7.While one sub-block is sent to the network, it will produce one winning neuron whose sequence number (X_m^n, Y_m^n) will be written down, where m is the sequence number of sub-block and n is the sequence number of Chinese character samples. The sequence number of winner neurons $\{(X_1^n, Y_1^n), (X_2^n, Y_2^n), ..., (X_{48}^n, Y_{48}^n)\}$ corresponding to the 48 sub-blocks is the feature sequence of the nth Chinese character samples. Compared to traditional methods with single winner neuron, the way of multi competitive combination can greatly reduce the scale of the network.



(a)



(b)



(c)

Figure 6: Feature distance results, (a) Distance between 100th and other samples, (b) Distance between 1164th and other samples, (c) Feature distance statistics.

The distance between each sample feature and other samples feature is calculated in order to verify the method of feature extraction with separability. Feature distances of 100th and 1164th samples to other samples selected randomly are shown in Figure 6(a) and Figure 6(b) respectively. X-axis is the sample number, and Y-axis is the feature distance between selected sample and other samples. It can be seen from the Figure 6 that the distance between each sample and other samples is not 0, that is the features of them are different and more than 100 of the distance exceed 90%. Figure 6(c) is a

histogram in which each color represents the distance from a sample to the other sample. X-axis is the feature distance which is divided into 100 intervals, and Y-axis is the number of sample for each distance interval. Likewise, it is shown that more than 100 of the distance exceed 90%. Therefore, there is better discrimination for feature.

The effectiveness of the proposed method in this paper is verified with three experiments. The input of network is the standard sample in the first experiment. Samples added 10% and 20% of the random noise to the standard samples are used for the second experiment and the third experiment respectively.

Parameter and results comparison of three experiments are shown in Table 1. It can be seen that the recognition accuracy is 100% for standard samples, 99.897% for samples with 10% random noise and 93.274% for samples with 20% random noise.

Experiment	Noise of samples	Samples recognized	Recognition accuracy（%）
First	0	3500	100
Secondly	10%	17500	99.897
Third	20%	17500	93.274

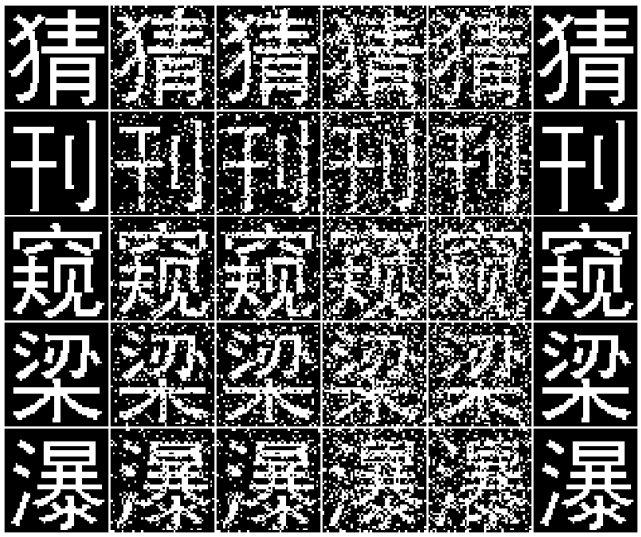
Table 1:Parameter and results comparison of three experiments.

When features of 3500 Chinese characters are extracted and recognized, parameter comparison between proposed method and others networks is shown in Table 2. It can be seen that the proposed method can reduce the size of the network and computation infinitely while dealing with the same problems.

methods	Number of neurons in input layer	Number of hidden layer	Number of neurons in output layer	The scale of weight matrix
BP	2304	No theoretical derivation	13	No theoretical derivation
Hopfield	2304	No	No	2304×2304
Tradition-al SOM	2304	No	≥3500	≥2304×3500
proposed	48	No	49	48×49

Table 2: Parameter comparison between proposed method and others networks.

Parts of the training samples and recognition results are shown in Figure 7.The column on the left are training samples. The middle columns are to be recognized samples, which are two added 10% and two added 20% random noise. The column on the right is recognition results. It can be foundin the figure that the network has the capabilityof identifying and classifying the samples with noise, and output the corresponding standard samples. It shows that the network has good anti-interference performance.



(a) (b) (c)

Figure 7: Parts of the training samples and recognition results, (a) Training samples, (b) Noise samples, (c) Recognition results.

6 Conclusion

The BWS feature extraction method based on SOM is proposed in this paper, which simplifies the network structure on the basis of retaining the good classification performance of SOM neural networks and achieve features extraction of 3500 Chinese characters with SOM neural networks of n*n . It breaks through the difficult point that the number of output layer neurons must be more than the number of samples in traditional SOM neural networks. Features of Chinese characters samples are extracted by combining the algorithm of image block and SOM and then assembled to improve recognition rate with samples information. The method is used for extract feature of 3500 Chinese characters in GB2312 Chinese character library， and then the Chinese characters are recognized by Euclidean distance. The recognition accuracy reaches 99.897% and 99.897% respectively in the cases of adding 10% and 20% random noise.

Since SOM neural networks was proposed, it has been a hot research topic in the field of neural networks, the network structure of SOM has much closer to the actual structure of the brain on processing information than other neural networks.Image block algorithms which can extract minutiae of image have been widely applied in the field of face recognition and so on. A novel method—BWS feature extraction with the combination of image block and SOM neural networks is proposed in this paper to recognize large character set of Chinese characters. The proposed method has a high anti-interference ability than optical character recognition (OCR), and it also has simpler structure and a larger capacity than other neural networks on characters recognition. Therefore, the method has a good self-learning ability and versatility, which provides a new idea for wide application of SOM neural networks in

Chinese characters recognition.

Acknowledgements

This work was funded by the National Natural Science Foundation of China (Grant No. 21227008).

References

[1] Kohonen T.(1995),Self-organizing maps. *Springer*,30(4), 266 - 270.

[2] Kaski S. (1997). Data exploration using self-organizing maps, *Acta Polytechnica Scandinavia Mathematics &Computing*, 82(12), 580-8.

[3] Vesanto J and Vesanto J. (2002), Data exploration process based on the self-organizing map, *Acta Polytechnica Scandinavica Mathematics& Computing*.

[4] Yusob B, Shamsuddin S M H, and Hamed H N A. (2013),Spiking self-organizing maps for classification problem, *Procedia Technology*, 11(1), 57-64.

[5] JunhaoRen, PeiqiJi and YueGeng. (2011), Application and improvement of SOM network in remote sensing image classification,*APPLICATION RESEARCH OF COMPUTERS*, 2011.

[6] Binmei Liang. (2009). Study on improvement and application of serf-organizing map neural network,*COMPUTER ENGINEERING AND APPLICATIONS*, 45(31),134-137.

[7] Jianhua Zhao and Weihua Li. (2012), Application of Supervised SOM Neural Network in Intrusion Detection, *Computer Engineering*, 38(12), 110-111.

[8] Astudillo C A and Oommen B J.(2013), On achieving semi-supervised pattern recognition by utilizing tree-based soms, *Pattern Recognition*, 46(1),293-304.

[9] Abdelsamea M M, Gnecco,G and Gaber, M M. (2014), A Concurrent SOM-Based Chan-Vese Model for Image Segmentation, *Advances in Self-Organizing Maps and Learning Vector Quantization*.

[10] Abdelsamea M M, Gnecco G and Gaber M M. (2014), A Survey of SOM-Based Active Contour Models for Image Segmentation, *Advances in Self-Organizing Maps and Learning Vector Quantization*.

[11] Hua Sun and Hang Zhang. (2010), Survey on Chinese Character Recognition Method, *COMPUTER ENGINEERING*, 36(20), 194-197.

[12] Hu J, Lim S G and Brown M K. (2000), Writer independent on-line handwriting recognition using an hmm approach, *Pattern Recognition*, 33(1), 133-147.

[13] Cheng Luo and Yueheng Sun. (2009), Handwriting Digit Recognition Based on Enhanced Bayes Classification, *MICROPROCESSORS*, 30(03), 77-79.

[14] Li Y, Yang H, Xu J, He W and Fan, J. (2007), Chinese Character Recognition Method Based on Multi-features and Parallel Neural Network Computation. *Advanced Intelligent Computing Theories and Applications, With Aspects of Theoretical and Methodological Issues, Third International Conference on Intelligent Computing, ICIC 2007, Qingdao, China, August 21-24, 2007, Proceedings* (Vol.4681, pp. 1103-1111).

[15] Odate, R., &Goto, H. (2015, September), Fast and accurate candidate reduction using the multiclass LDA for Japanese/Chinese character recognition, In *Image Processing (ICIP), 2015 IEEE International Conference on* (pp. 951-955). IEEE.

[16] Tao, D., Lin, X., Jin, L., & Li, X. (2016), Principal component 2-D long short-term memory for font recognition on single Chinese characters, *IEEE transactions on cybernetics*, *46*(3), 756-765.

[17] Ding, X., Chen, L., & Wu, T. (2007), Character independent font recognition on a single chinese character, *IEEE Transactions on pattern analysis and machine intelligence*, *29*(2), 195-204.

[18] Zhang S, Jin L and Lin L. (2016), Discovering similar chinese characters in online handwriting with deep convolutional neural networks,*Document Analysis & Recognition*, 1-16.

[19] Bengio Y, De Mori R, Flammia G, Kompe R, Moody J E and Hanson S J,et al. (1992), Neural network - gaussian mixture hybrid for speech recognition or density estimation,*Advances in Neural Information Processing Systems Nips*, 175-182.

[20] Lin Zhang, Shan Wang, Xiaoyu Qin and Lianming Wang. (2014), Musical instrument recognition based on the bionic auditory model,*Journal of Northeast Normal University(Natural Science Edition)*, 46(01),75-79.

[21] Pentland A, Moghaddam B and Starner T. (1994), View-based and modular eigenspaces for face recognition, *IEEE International Conference on Computer Vision \& Pattern Recognition*. (pp.84 - 91).

[22] XipengLan, Xuefeng Tong and GuorongXuan. (2011), Feature Extraction Method of Face Image Partition and Color Separation, *Computer Engineering*. 37(16), 164-166.