# Chinese character recognition system

Sun Ziming
*School of Computing*
*Asia Pacific University of Technology and Innovation (APU)*
Kuala Lumpur, Malaysia
tp027269@apu.edu.my

Nowshath K Batcha
*School of Computing*
*Asia Pacific University of Technology and Innovation (APU)*
Kuala Lumpur, Malaysia
nowshath.kb@apu.edu.my

***Abstract*—In today's highly information-based society, how to quickly and efficiently input Chinese characters into computers has become an important bottleneck affecting the efficiency of human-computer communication. Around this issue, people put forward a variety of solutions. At present, the Chinese character input is divided into artificial keyboard input and machine automatic recognition input, manual keyboard input refers to the use of manual keystrokes in accordance with certain rules to enter the Chinese characters into the computer, there are already hundreds of types of programs. However, unlike the typewriter, people need to go through a certain amount of time to learn to master a typing method, more serious is: for a large number of existing documents, the use of manual typing method is slow and labor intensity, will spend a lot of manpower and time. Therefore, the automatic identification of the machine has become a subject to be studied. Automatic recognition input is divided into two kinds of speech recognition and Chinese character recognition, in which Chinese character recognition is the Chinese character dot matrix graphics into electrical signals, and then input to the digital signal processor or computer for processing, according to a certain classification algorithm in the Chinese character set to identify the match with the Chinese characters. Therefore, the purpose of the study of offline handwritten Chinese character recognition is to solve the problem of how to input high-speed Chinese characters, so that information processing can be carried out more conveniently and quickly.**

***Keywords—recognition, convolution, neural***

## I. Introduction

Character recognition is a traditional research field of pattern recognition. Since the 1950s, many researchers have carried out extensive exploration in this field and promoted the development of pattern recognition. Over the past decade, with the rapid development of large-scale integrated circuits and microelectronics technology, the development of computer hardware technology has reached a never-ending speed. At the same time, the rapid decline in the price of computers and their related equipment and the rapid increase in performance make the computer from the previous large laboratory supplies into human society work and life in all aspects, more people can contact and use the computer to complete their work The With the increase in the number of computer users, people are demanding more and more computer, hoping that it cannot only carry out complex numerical calculation, and hope that the computer to a certain extent, imitate human intelligence activities, with similar human intelligence. So, people want the computer to understand the natural language of mankind, to understand people write a variety of words, to understand people's face and even understand people's feelings at the moment. These requirements of human beings have given power to the application of pattern recognition and artificial intelligence. Therefore, speech recognition, face recognition and character recognition have become a hotspot in pattern recognition and artificial intelligence research during this period.

## II. Literature Review

Chinese character recognition originated in 1966. First of all, by Casey and Nagy in IBM for the printing of Chinese characters related to the work [1] they proposed a Chinese character structure based on the matching framework and has been widely adopted. In the same year, the work of online Chinese character recognition has also been carried out, in which MIT's Liu et al [2], proposed a character based on Chinese character strokes based on the direction of Chinese character handwriting. The corresponding weight vector is determined by a learning process similar to that of a perceptron. Inspired by these methods, in the late 1960s, Japanese researchers took the lead in developing a prototype of the printed Japanese language recognition system.

In 1977, Toshiba established the first printed Chinese character recognition system [3], which was able to identify 2,000 Chinese characters. In the late 1970s, Chinese researchers began working on print and online handwriting character recognition. At this point, a variety of different ideas were introduced into the field of printed Chinese character recognition: Stallings [4] provided a coding scheme based on structural analysis; Nakano et al [5] Conducted research on feature projection; Yamamoto et al [6] (For example, Fourier transformations, Rapid transformations) are also used for special extraction of Chinese characters.

During this period, the progress of digital image processing algorithm, but also greatly promoted the research of Chinese character recognition. For example, the Otsu algorithm, until now, is still used to convert grayscale images into binary images. Character regularization [4] and character blur are first proposed, and become the key step of later character extraction of Chinese characters. In addition, the post-processing of the language model has been gradually applied to the field of handwriting recognition after being verified in the field of speech recognition.

In 1985, Fuji produced a Japanese character reading product that could identify a variety of printed fonts. In China, the General Administration of Technology withdrew from GB2312-80 commonly used Chinese character standard set, including a Chinese character set includes 3755 most commonly used Chinese characters, which is used in Chinese character recognition algorithm used in the character set. Due to the similarity between most Japanese characters and

Chinese characters, by drawing on Japan's research experience in the field of word recognition, China's research work on Chinese character recognition in prints has also made exciting breakthroughs and gradually started off-line handwritten Chinese character recognition Related research.

In this period, the statistical classification algorithm has been extensively studied, and the methods of shift similarity, reliability, decision tree and dynamic programming have been successful to some extent, which provides clues for the research of handwritten Chinese character recognition. At the same time, the neural network, the Hidden Markov Model, quadratic discriminant function and other algorithms have also been properly modified to make it suitable for Chinese character pattern classification task, and then some famous models, including the hierarchical neural network, modified quadratic discriminant function (MQDF), Regularized discriminant analysis (RDA), and the Maximum Mutual Information Hidden Markov Models (MMI-HMMs).

During the period of 1990 - 2005, the study of handwritten Chinese character recognition has made great progress, and all aspects of Chinese recognition system have been deeply studied.

In the case of preprocessing, the appearance of 1-dimensional and pseudo-2-dimensional non-linear normalization methods alleviates the problem of character differences caused by handwriting style. While the Elastic Network method solves the problem of adaptive blockade of Chinese characters.

In terms of feature extraction, in order to improve the stability of the directional feature, the directional element feature and the gradient direction feature are presented and used for online and offline Chinese character recognition, respectively. In order to reduce the computational cost, the regularization of character shapes and feature extraction are combined into a step.

In the design of the classifier, a series of statistical methods to a wide range of research, can be divided into Generative Model: such as Hidden Markov Model (HMMs) and Modified Quadratic Discriminant Function (MQDF), Discriminative model including Convolution Neural Network (CNN), Learning Vector Quantization (LVQ) and finally Generative Model of discriminant training like for example Minimum Classification Error Based Hidden Markov Model (MCE-HMMs) and Minimum Classification Error Based Modified Quadratic Discriminant Function (MCE-MQDF).

In addition, the use of multi-classifier integration schemes, both in the small category classification task and on the large category classification task, further enhances the recognition accuracy. In the follow-up processing, the addition of the language model restrictions, in particular the use of N-gram, to identify the results of ambiguity elimination and error correction, effectively enhance the recognition rate. From 2006 to present, a new research boom in Chinese handwriting recognition began in 2006, along with the release of some new handwritten databases. The research direction gradually shifted from isolated Chinese character recognition to more real unrestricted Text recognition, at the same time, due to the development of deep learning technology, the depth of the neural network is applied to the handwritten Chinese character recognition, its good expression ability to accommodate the complex structure of Chinese characters and writing style diversity, so the virtual is very good effect.

Multilayer neural networks have recently ignited much interest in the digital image processing community, thanks to its wide range of applications [7] After 2006, Hinton had a groundbreaking approach to Deep Belief Networks (DBN) [8]. Bengio et al [9] Proposed the Restricted Boltzmann Machine (RBM) and the automatic encoder, Ranzato et al. Used sparse automatic coding [10]. The DBN proposed by Sun Zhijun et al [11]. In 2012 is a typical generation structure. In this structure, the high order correlation of data, the joint probability distribution of observed data and corresponding categories are well described. Unlike the discriminant model which only estimates the posterior probability, the joint probability distribution of the observed data and the classification label can be obtained.

Convolution neural network can be regarded as a structured multi-layer feed forward neural network, which has become a hotspot in the field of speech analysis and image recognition. In general, the first layers of the convolutional neural network are composed of convolutions and subsampling layers alternately. The spatial resolution between the layers is decreasing, and the number of planes in each layer is increased. High spatial and data structure invariance and detect more feature information; Convolution neural network is the traditional layer of the whole connection layer, and multi-layer feed forward neural network function similar to the convolution layer and down-sampling Layer of the feature map, into the full connection layer after a number of non-linear mapping to generate a feature vector, and finally sent to the output layer of the classifier for classification. At present, according to the different tasks, the output layer of the classifier can use logistic. Regression, SoftMax regression, or support vector machine and other forms.

This section has introduced the various works and fields of image processing, neural network, pattern recognition and the algorithms involved in pattern recognition.

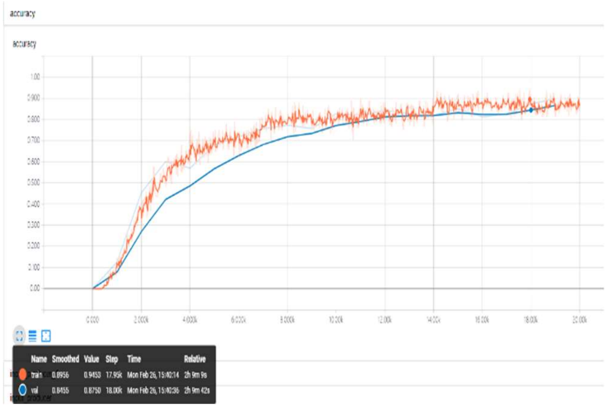
## III. Experimenation

The experimentation is done on multiple stages first need to step is to read images from the data set function, and then will be the design and development of neural networks and begin training. After the first training to build the verification function, through the verification of the structure of the neural network to adjust and do re-training. Python and TensorFlow and with additional libraries needed are used for the experimentation.

There are 3755 Chinese characters in this data set, each character has 450~500 samples. Each Chinese character have a folder which name in order from 00000~03754. This folder name is the index number which the recognize result will return.

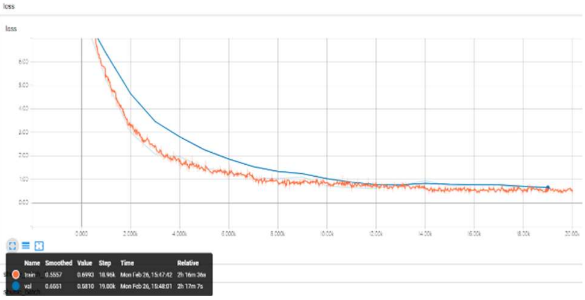
Fig. 1. Sample training data set

First need to get a list of all images path and the corresponding label, and all preprocessing like slice processing and image reading is done. As the image scene is not complicated, just do some basic processing, including image flipping, changing the brightness and so on. Establish convolution neural network and train the organized data. After the training is completed, the model is tested with randomly written word by hand for inference.



| Name | Smoothed | Value | Step | Time | Relative |
|---|---|---|---|---|---|
| train | 0.8956 | 0.9453 | 17.95k | Mon Feb 26, 15:40:14 | 2h 9m 9s |
| val | 0.8455 | 0.8750 | 18.00k | Mon Feb 26, 15:40:36 | 2h 9m 42s |

Fig. 2. Obtained accuracy of the system



| Name | Smoothed | Value | Step | Time | Relative |
|---|---|---|---|---|---|
| train | 0.5557 | 0.6993 | 18.96k | Mon Feb 26, 15:47:42 | 2h 16m 36s |
| val | 0.6551 | 0.5810 | 19.00k | Mon Feb 26, 15:48:01 | 2h 17m 7s |

Fig. 3. Obtained Loss rate of the system

To improve the performance adjust the network structure and repeatedly revise it for enhancement until satisfactory performance is achieved. Tune the performance to achieve less loss rate. Accuracy and Loss achieved by the system is shown in Fig 2. and Fig 3.

## IV. Conclusion and Future Enhancements

For the same algorithm, different handwritten Chinese sample data banks get different accuracy rates. For the identification and recognition of handwritten Chinese characters, the selection and establishment of the Chinese sample database is very important, the next step is to establish a standardized handwritten Chinese sample database. In the handwritten Chinese character handwriting sample image, the interdependence between semantic information, singular information and information carrier needs to be further studied, and the separation method of multiple information needs to be studied and analyzed. In comparing the texture information or characteristic characters of Chinese characters, it is necessary to consider a wide range of information and features, and to combine them effectively, that is, the need for information fusion technology to integrate a variety of handwriting features in order to get better recognition。

## References

[1] R. Casey and G. Nagy, “Recognition of printed Chinese characters,” IEEE Trans. Elect. Cornell. 15, pp.91-101 (1966).

[2] T.H. Hildebrandt and W. Liu, “Optical Recognition of Handwritten Chinese Characters: Advances Since 1980,” Pattern Recognition, vol. 26, pp. 205-225, 1993.

[3] Y. Wang, X. Ding and C.-L. Liu, “MQDF discriminative learning based offline handwritten Chinese character recognition,” International Conference on Document Analysis and Recognition (ICDAR), 2011.

[4] W. Stallings, “Approaches to chinese character recognition,” Pattern Recognition, vol. 8, pp. 87–98, 1976.

[5] H. Fujisawa, Y. Nakano, and K. Kurino, “Segmentation methods for character recognition: from segmentation to document structure analysis,” Proceedings of the IEEE, vol. 80, pp. 1079,1092, July 1992.

[6] K. Yamamoto and Y. Yamazaki, “A network of two-chinese-character compound words in the japanese language,” Physica A: Statistical Mechanics and its Applications, vol. 388, pp. 2555–2560, 2009.

[7] Bengio, Y. (2009). Learning deep architectures for AI,” Now Publishers Inc.

[8] Hinton, G. E., Osindero, S., & Teh, Y. W,“A fast learning algorithm for deep belief nets. Neural computation, 18(7), pp.1527-1554, 2006.

[9] H. Larochelle, M. Mandel, R. Pascanu, and Y. Bengio, “Learning algorithms for the classification restricted Boltzmann,” JMLR, 13:643–pp.669, 2012.

[10] M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun, “Efficient learning of sparse representations with an energy-based model,” In Advances in neural information processing systems, pp.1137–1144, 2007.

[11] Z. Sun, L. Xue, and Y. Xu,“Recognition of SAR target based on multilayer auto-encoder and SNN,” Int. J. Innovative Comput. Inf. Control 9(11), pp.4331–4341 (2013).