Research on the character recognition of Tibetan ancient document based on CNN

Yuehui Han, Yusheng Hao, Weilan Wang*

Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, 730030, Lanzhou
School of Mathematics and Computer Science, Northwest Minzu University, 730030, Lanzhou
Corresponding author’s e-mail: wangweilan@xbmu.edu.cn

Abstract: Tibetan ancient documents are treasures among the World Cultural Treasure, which is a valuable historical material for studying Tibetan politics, economy, culture and religion. At present, the research on Tibetan character recognition mainly focuses on printed and handwritten Tibetan, while the research on Tibetan ancient document character recognition lags behind. This paper presents a method of recognizing 562 common characters in Tibetan ancient document based on CNN, and the experimental results show that the recognition rate of this method is 99.85%, which proves the effectiveness of this method.

1. Introductions

Tibetan is a special alphabetic writing, which is composed of 30 consonant letters and 4 vowel letters. The syllable is the most basic word-formation unit in Tibetan, which are separated by syllable nodes. Figure 1 shows the syllable composition of Tibetan, the vertical unit of a syllable is a character, a syllable contains 1 to 4 characters, and a character is superimposed from 1 to 7 components. Therefore, the basic unit of Tibetan syllables is the word, the basic unit of vocabulary is the syllable, and sentences are composed of vocabulary according to certain grammatical structures.

Figure 1. The Composition of Tibetan Syllables

At present, the research of Tibetan character recognition mainly focuses on printing and handwriting, and many researchers have done a lot of research work. Wang Weilan et al. took printed white, black, round, long and bamboo as typeface samples, and achieved high recognition rate after pretreatment, line and line segmentation, feature selection and classification recognition [1]. Zhou Wei proposed a printed Tibetan character recognition method based on geometric shape analysis [2]. Professor Wang Weilan led the team to put forward many excellent on-line handwritten Tibetan recognition methods, such as: on-line handwritten Tibetan recognition methods based on stroke features and MCRNN model [3], HMM classifier [4], 2DPCA and IMLDA [5], and statistical analysis [6]. In addition, the...
difficulties and solutions of Tibetan on-line handwritten recognition [7] and on-line handwritten Tibetan recognition were also proposed. Normalized processing method for Chinese characters in other words [8], Wang Xiaojuan [9] proposed a method of identifying similar characters in Tibetan ancient books based on DNN; Li Zhenjiang [10] using the baseline information of Chinese characters, which significantly improved the recognition efficiency.

In this paper, a recognition method of common characters in Tibetan ancient documents based on CNN is proposed. Tested by 562 common characters in Tibetan ancient documents data set (as shown in Figure 2), and good recognition results are obtained.

2. Introduce of CNN

CNN is a common deep learning framework inspired by the cognitive mechanism of biological natural vision. In 1959, Hubel & Wiesel [11] found that the visual cortex cells of animals were responsible for detecting optical signals. Inspired by this, Kunihiko Fukushima [12] proposed the neocognitron in 1980, which is the predecessor of CNN. In 1990, LeCun et al.[13] proposed LeNet-5 by introducing BP algorithm, which established the modern structure of CNN. And the network structure of LeNet-5 is shown in Figure 3.

LeNet-5 network structure consists of seven layers that excluding input layer and the size of input image is 32×32. The C1 layer contains six 5×5 convolution cores. After convolution processing, six feature maps are obtained, the size of which is 28×28. The S2 layer is the down-sampling layer, which contains six 14×14 feature maps, which are obtained from the down-sampling of the C1 layer. Each neuron of the feature graph is connected to a neighborhood of 2×2 in the layer C1. The C3 layer contains sixteen 5×5 convolution cores. After convolution processing, six feature maps are obtained, the size of which is 10×10. The S4 layer is the down-sampling layer, which contains sixteen 5×5 feature maps, which are obtained from the down-sampling of the C3 layer. The C5 layer is a convolution layer with 120 feature maps. Each feature map has only one neuron, and each neuron is connected with the neighborhood of all feature maps in the S4 layer whose size is 5×5. The F6 layer is a hidden layer containing 84 neurons and is fully connected to C5. Finally, there are 10 neurons in the output layer, which are composed of Radial Basis Function (RBF) units. They are fully connected with F6 layer, and each neuron corresponds to a digital character class.
3. Tibetan character recognition based on CNN

3.1 recognition process
This paper presents a Tibetan character recognition method based on CNN (method process as shown in Figure 4) which consists of two parts. In the first part, the CNN classification model is trained according to the image of training set. In the second part, put the image that to be recognized into the trained model.

![Recognition process](image)

Figure 4. Recognition process

3.2 Preprocessing
The data set used in this paper covers 562 types of commonly used characters in Tibetan ancient documents. Each category contains 5000 images. The size of the image is $128 \times 64$. The example of the original data set image is shown in Figure 5. The input image contains the position information of the characters, which results in more non-text areas in the image and influences the recognition results. Therefore, the input image needs to be preprocessed, and the implementation process is as follows.

Step 1: Remove the redundant parts outside the text area of the image, and the result is shown in Figure 6.

Step 2: Size normalization, which set to $128 \times 64$, as shown in Figure 7.

![Data set images](image)

Figure 5. data set images

![Remove redundant parts](image)

Figure 6. remove redundant parts

![Size normalization](image)

Figure 7. size normalization

3.3 CNN Model Structure for Tibetan recognition
LeNet-5 network has better recognition performance in 10 kinds of handwritten numeral character recognition. However, it cannot be directly used in Tibetan recognition because of too many categories. Therefore, this paper improves the network structure and achieves good recognition results.

The improved network structure consists of nine layers that excluding input layer and the input image size is $128 \times 64$. The C1 layer contains twenty $3 \times 3$ convolution cores. After convolution processing, twenty feature maps are obtained. The C2 layer contains thirty $3 \times 3$ convolution cores. After convolution processing, thirty feature maps are obtained. The S3 layer is the down-sampling layer, which are obtained from the down-sampling of the C2 layer. The C4 layer contains forty $3 \times 3$ convolution cores. After convolution processing, forty feature maps are obtained. The C5 layer contains fifty $3 \times 3$ convolution cores. After convolution processing, fifty feature maps are obtained. The S6 layer is the down-sampling layer, which are obtained from the down-sampling of the C5 layer. The C7 layer is a fully connected layer containing 300 neurons, which converts the output deformation of the S6 layer into a vector. F8 is hidden Layer, which contain 200 neurons and fully connected to C7. Finally, there are 562 neurons in the output layer, which are composed of Radial Basis Function (RBF) units. They are fully connected with F8 layer, and each neuron corresponds to a digital character class.
4. Experiments
In this paper, the data set used contains 562 categories of Tibetan characters, each of which contains 5000 images. Firstly, the image is preprocessed to remove redundant non-text areas and normalized. Then, the image is divided into training set and test set in a ratio of 4:1, and the network performance is trained and tested. The results show that the CNN network structure adopted in this paper has good recognition performance. The training results are shown in the figure 8, table 1 shows the results of partial character recognition separately and the average recognition rate is 99.85%

![Figure 8. Recognition rate](image)

| Number | 1   | 2   | ... | 305 | 306 | ... | 561 | 562 | Average |
|--------|-----|-----|-----|-----|-----|-----|-----|-----|---------|
| Tibetan| མ | མ | ... | མ | མ | ... | མ | མ | 99.85% |
| R-rate | 99.99% | 99.97% | 99.73% | 99.75% | 99.97% | 99.98% | 99.85% |

5. Conclusions
Based on the analysis of LeNet-5 network structure, this paper proposes an improved network structure for identifying 562 common characters in Tibetan ancient documents. The training test on Tibetan ancient document data sets shows that the final recognition rate is 99.85%, which has good recognition performance.

Acknowledgements
This work was supported by the National Science Foundation (No.61772430), the Program for Leading Talent of State Ethnic Affairs Commission, the Fundamental Research Funds for the Central University of Northwest Minzu University (No. 31920170142), the Gansu Provincial first-class discipline program of Northwest Minzu University, the program for innovative research team of SEAC and also supported by postgraduate research innovation project of Northwest Minzu University (No. Yxm2018118).

Reference
[1] Wang WL, Ding XQ, Chen L. (2003) Research on the Recognition of Printing Embodiment for Tibetan Language. J. Computer Engineering, 3: 37-38.
[2] Zhou W, Chen YL, Zeng ZB. (2012) Tibetan Character Recognition Based on Geometric Analysis. J. Computer Engineering and Application, 18: 201-205.
[3] Wang WL. (2008) On-line Handwritten Tibetan Recognition Based on Stroke Characteristics and MCRNN Model. J. Computer Engineering and Application, 14: 91-93.
[4] Liang B, Wang WL, Qian JJ. (2009) Application of HMM-based classifier in online handwritten Tibetan recognition. J. Microelectronics & Computer, 4: 98-101.
[5] Wang DH, Wang WL, Qian JJ. (2010) 2DPCA and IMLDA method of feature extraction for online handwritten Tibetan recognition. In: International Conference on Networking & Digital Society. Beijing. Pp. 563-566.

[6] Wang WL, Qian JJ. (2011) Online Handwriting Recognition of Tibetan Characters Based on the Statistical Method. J. Communications and Computers: Chinese-English Version, 3: 188-200.

[7] Liu HZ, Wang XD, Wang WL. (2005) Difficulties and Solutions of Tibetan Online Handwriting Recognition. J. Journal of Northwest Minzu University: Natural Science Edition, 1: 77-80.

[8] Liu HZ, Wang WL. (2006) On-line Handwritten Tibetan Character Recognition Standardization Processing J. Computer Applied Research, 9: 179-181.

[9] Wang XJ, Wang WL. (2018) A Recognition Method of the Similarity Character for Uchen Script Tibetan Historical Document Based on DNN. In: Chinese Conference on Pattern Recognition and Computer Vision. Gunagzhou. pp. 52-62.

[10] Li ZJ. (2018) Tibetan Historical Document Recognition of Uchen Script using Baseline Information. In: 10th International Conference on Graphics and Image Processing. Chengdu. pp. 134-147.

[11] Hubel D H, Wiesel T N. (1968) Receptive fields and functional architecture of monkey striate cortex. J. Journal of Physiology, 1: 215-243.

[12] Fukushima K. Neocognitron. (1980) A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. J. Biological Cybernetics, 4: 193-202.

[13] Lécu Y, Bottou L, Bengio Y. (1998) Gradient-based learning applied to document recognition. J. Proceedings of the IEEE, 11: 2278-2324.