TIBETAN AND CHINESE TEXT IMAGE CLASSIFICATION BASED ON CONVOLUTIONAL NEURAL NETWORK

Jincheng Li[1], Penghai Zhao[1], Yusheng Hao[2], Qiang Lin[2], Weilan Wang[1*]

[1]Key Laboratory of China's Ethnic Languages and Information Technology
(Northwest Minzu University), Ministry of Education,
Lanzhou, P. R. China
[2]College of Mathematics and Computer Science, Northwest Minzu University,
Lanzhou, P. R. China

ABSTRACT

The first stage of Tibetan-Chinese bilingual scene text detection and recognition is the detection of Tibetan- Chinese bilingual scene text. The detection results are mainly divided into three categories: successfully detected regions of Tibetan text and Chinese text, non-words regions with failed predictions. If the detected text image results are accurately classified, then the non-text images should be filtered in the recognition phase, meanwhile the Tibetan and Chinese text images can be identified by using different classifiers, such procedure can reduce the complexity of classification and recognition of two different characters by one recognition model. An accurate classification of Tibetan and Chinese text images is mattered. Therefore, this paper conducts a research on the classification of Tibetan, Chinese and non-text images by using convolutional neural networks. We perform a series of exploration about the classification accuracy of Tibetan, Chinese text images and non-text images with convolutional neural networks in different depths, and compare the accuracy with the classification results based on the transfer learning then analyze it. The results show that for the classification of Tibetan, Chinese and non-text images in the scene, using 7-layer convolutional neural network has reached saturation, and increasing the network depth does not improve the results, which provides reference values for Tibetan-Chinese text image classification.

KEYWORDS

Convolutional Neural Network, Tibetan-Chinese scene text image, image classification, transfer learning

1. INTRODUCTION

Image classification is one popular direction of the computer vision, which also provides a vital foundation for the application of object detection[1,2], face recognition[3,4], pose estimation[5,6], etc. Therefore, image classification technology has high value in academic research and applied value of science technology[7]. As AlexNet[8] surpassed the traditional methods at the Large Scale Visual Recognition Challenge 2012 and achieved remarkable results in the classification task, the following convolutional neural networks(CNN) model such as VGG[9], GoogLeNet[10], ResNet[11] were proposed. These networks made the CNN-based deep learning technology become the mainstream of the classification task[12-15]. Compared to the general neural networks, the basic structure of the convolutional neural network includes two layers, one is the

David C. Wyld et al. (Eds): SAIM, SIPM, ACSIT, FCST, ICITE - 2020
pp. 59-67, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.100807

feature extraction layer, the other is the feature mapping layer. And from a perspective of model features, convolutional neural network has two particularities which can reduce the complexity of the model, one is the sparse connectivity, the other is the shared weights. With convolutional neural network constantly improving and optimizing by researchers, various excellent convolutional neural network models [16-20] were presented and had achieved acceptable results in the classification task.

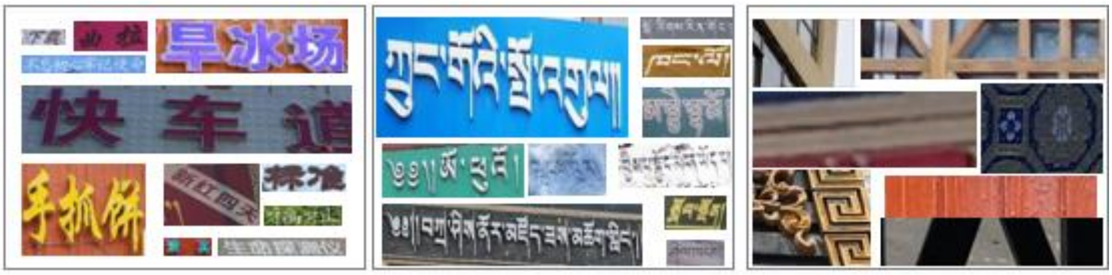
In the Tibetan areas of China, almost all the textual information in various scene contains both Tibetan and Chinese characters. The objective of text detection for this kind of scene is to locate the position of Tibetan and Chinese characters, and the follow-up is to put the segment of two detected text regions and non-text regions into the trained classifier, thus scanned images with text is converted into computer-readable data. This paper provides a feasible scheme for Tibetan-Chinese scene text recognition by classifying Tibetan text images, Chinese text images, and non-text images. That means using different classifiers to recognize Tibetan text image and Chinese image text respectively, so as to make Tibetan-Chinese scene text recognition more simple and effective.

Traditional image classification methods are generally divided into two steps: First, calculating artificially designed features from the input image. Second, training a classifier based on the extracted features. The effect of this classification method depends on the artificially designed features, thus it has great uncertainty. For the above situation, this paper conducted an exploratory study on the classification of Tibetan, Chinese, and non-text images by using convolutional neural networks. Regarding the constructed data set as experimental data, the features of text images were extracted by employing convolutional neural networks with different depths. Then applying softmax to classify and comparing the classification results with the pre-trained VGG16 model on which transfer learning method is implemented. The results show that the seven-layer deep convolutional neural network has achieved a 98.28% classification accuracy, and increasing the depth of the network has no significant improvement on the classification accuracy, which indicates that for our classification task, the seven-layer network has reached saturation.

2. DATASET AND EVALUATION PROTOCOL

2.1. The Dataset

For the Chinese text image dataset, 5000 pieces of text were selected from the existing dataset[21] and the images taken in Tibetan area. For the Tibetan text image dataset, the same amount of text was obtained by intercepting from the text images of Tibetan area or synthesizing. For non-text image dataset, we randomly cropped 5000 images from the real background image captured by camera. Then it is divided into three categories: Tibetan text, Chinese text and non-text. Figure 1. shows some samples of Tibetan, Chinese, and non-text images. It can be seen from the figure that the Chinese text image and the Tibetan text image have distinct shape of characters. Comparing with Chinese one, Tibetan text image has more complex backgrounds, various colors, and different scales, image sizes. Non-text images also have various textures, colors, and backgrounds which increase the difficulty of classification. During the training, these data of samples are randomly augmented to improve the generalization ability of the model.



(a) Chinese text images (b) Tibetan text images (c) non-text images

Figure 1. Scene image examples

2.2. Evaluation Methods

We randomly divided the Tibetan-Chinese text image dataset into training set and test set at a ratio of 7:3, and then evaluated the experimental results with the following two evaluation methods.

1) *Precision*: Number of correctly identified test samples as a percentage of total test samples, they are given by:

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

Where TP is the number of positive test samples that are correctly classified as positive samples. FP is the number of negative test samples that are incorrectly classified as positive samples. $TP + FP$ is all test samples that are classified as positive samples.

2) *F1 Score*: It is an index used to measure the accuracy of classification models in statistics. It takes into account both the accuracy and recall of the classification model. They are given by:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

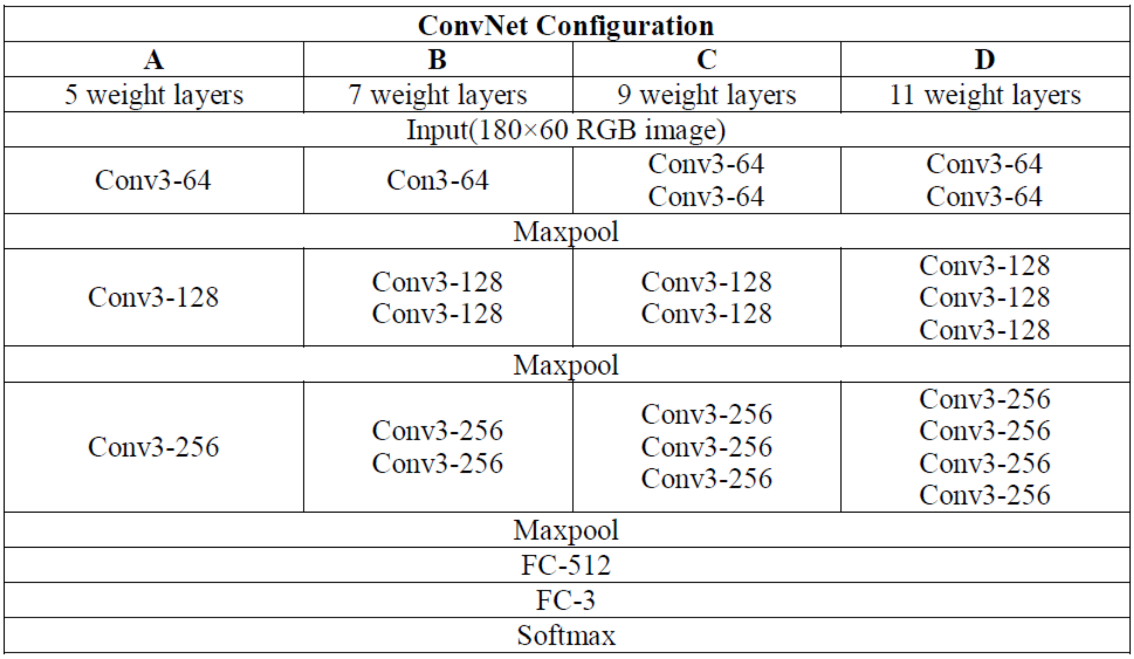
Where FN is the number of positive test samples that were misclassified as negative samples; $TP + FN$ is all the positive test samples; Recall is the recall rate.

3. The Architecture

3.1. Convnet Configurations

The structure of our convolutional neural network is shown in Table 1. The configurations of four different depth are: 5-layer, 7-layer, 9-layer, and 11-layer (excluding pooling and softmax layers). With image size of input layer being 180 × 60, the network carries out the convolution and subsampling operation alternately three times, and sends the data to the two followed fully connected layers. All convolution manipulations use a size of 3 × 3 kernel, while subsampling manipulations use a size of 2 × 2 max-pooling kernel of which step size is 2, and the convolution layer depths are 64, 128, and 256 respectively.

Table1. Network Configuration



ConvNet Configuration			
A	B	C	D
5 weight layers	7 weight layers	9 weight layers	11 weight layers
Input(180×60 RGB image)			
Conv3-64	Con3-64	Conv3-64 Conv3-64	Conv3-64 Conv3-64
Maxpool			
Conv3-128	Conv3-128 Conv3-128	Conv3-128 Conv3-128	Conv3-128 Conv3-128 Conv3-128
Maxpool			
Conv3-256	Conv3-256 Conv3-256	Conv3-256 Conv3-256 Conv3-256	Conv3-256 Conv3-256 Conv3-256 Conv3-256
Maxpool			
FC-512			
FC-3			
Softmax			

The network configuration in Table 1 includes the use of convolutional layer, pooling layer, BNlayer and Dropout. The detailed descriptions of each deep network structure are as follows：

1) Input: First, the training images of different sizes are scaled to 180 × 60 by using a bilinear interpolation algorithm. In order to improve the generalization ability of the classification model and avoid overfitting, the training data were augmented by reversal, mirroring, translation and perspective transformation to increase the amount of data. Then send 32 training images in each batch to the network for training.

2) Convolution Layer: The convolutional layer includes convolution, activation functions, batch normalization, and max-pooling. The 5-layer, 7-layer, 9-layer, and 11-layer depth network contains three, five, seven, and nine convolutional layers respectively, and each convolutional layer obtains feature maps with different sizes. Each convolution operation uses a 3 × 3 convolution kernel. The advantage of this method is that it can reduce the number of parameters and implement more nonlinear mapping at the same time. After each convolution, the ReLU function is used to activate the feature map F_i. The generation process of F_i is given by:

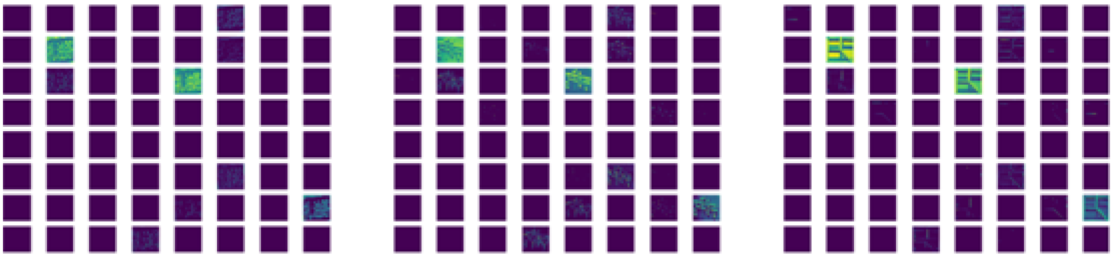
$$f(x) \ = max(0, x) \tag{4}$$

$$F_i = f(F_{i-1} * W_i + b_i) \tag{5}$$

Where W_i is the weight of convolution in layer i , b_i is the offset in layer i , $*$ is the convolution operation, f is the nonlinear excitation function of ReLU.

In order to accelerate the convergence of the network and prevent the overfitting of the network, each feature map F_i is followed by the batch normalization(BN) layer, that is, batch standardization, then proceed to the next layer. As the feature map passed through one to four convolutional layers, max-pooling layer operates on each standardized feature map independently. The purpose of this procedure is reducing the dimension of the feature map and to remaining invariant to changes in scale or rotation.

Obviously, for each depth network, the main features of the image are obtained by convolution operation, nonlinear excitation function, and batch standardization. Figure.2 shows the visual effect of the first convolution operation of each image category. It can be seen that the first convolution layer mainly extracts the edge, corner and color feature information about the image



(a) Chinese image feature maps (b) Tibetan image feature maps (c) no-text image feature maps

Figure. 2 After the first convolution feature maps

3) Output Layer: After multiple convolution and subsampling layers, two full connected layers are connected. The full connection layer is to integrate the local information output from the subsampling layer. The dropout strategy is used in two full connection layers to prevent the overfitting and improve the generalization ability of the model. It will set the output of neurons in the full connection layer to 0 with a probability of 0.6, and these neurons output as 0 will no longer carry out forward propagation and back propagation. At last, the prediction results of the classification output by applying softmax are used to realize the image classification.

3.2. Loss Function

During the training, we use the cross entropy loss function to calculate the loss for the classification results. Specifically, the loss function L of cross entropy is given by:

$$J = - \sum_{i=1}^{N} y_i \cdot \log(p(y_i)) \tag{6}$$

Where N is the number of categories, y represents the label (0 or 1). If the category is the same as the sample, it is 1, otherwise, it is 0. $p(y_i)$ is the predicted probability that the observation sample belongs to category i .

4. Experimental Results and Analysis

We adopted Windows10 + Python3.7 + Keras2.2 as the deep learning framework and conducted experiments on this basis. To prepare the training samples, we use the datasets containing 15000 samples from section II.A, 5000 samples for each category. There are 10500 images in the training set and 4500 images in the test set.

4.1. Analysis of the Influence of Different Depth Network Structures on Classification Results

We perform experiments on the network of four different depths of 5-layer, 7-layer, 9-layers, and 11-layer to analyze the impact of different depths on the classification results. For each network, the training image is scaled to 180 × 60 by using bilinear interpolation at the first beginning. The training data is augmented by random horizontal flipping, mirroring, and other operations. Among them, the batch size is 32, and each convolution layer is followed by a batch normalization layer. The initial learning rate is set to 0.0001 and subject to exponential decay every 20 epochs. The Adam optimization algorithm is used and the image is scaled to 180 × 60 during testing. With the same implementation details, the overall average accuracy and F1 evaluation of different depths of the network after 200 epochs are shown in Table 2.

Table 2. Results of networks at different depths

Network layers	**5-layer**	**7-layer**	**9-layer**	**11-layer**
Precision(%)	97.50	98.28	98.25	98.15
Recall(%)	97.46	98.28	98.24	98.15
F1-score(%)	97.47	98.28	98.25	98.15

In Table 2, from the perspective of precision, while using convolutional neural network with only five layers, the classification accuracy is about 97.5%. By using a 7-layer convolution neural network, the accuracy of correct classification has reached 98.3%. The accuracy of 9-layer and 11-layer networks is similar to that of 7-layer network, which shows that the depth has a great influence on the performance of the convolution neural network, but with the increase of the depth, the network will gradually reach saturation. At the same time, it also reveals that for the Tibetan and Chinese text image classification, the 7-layer network is basically saturated, increasing the depth of the network does not significantly improve the results.

4.2. Comparative Analysis of Experimental Results

According to the analysis in section 4.1, we can know that the optimal classification accuracy can be obtained when the depth of the convolutional neural network is seven. In order to verify the effectiveness of the network depth, a comparison experiment is performed with the transfer learning method. Using the pre-trained VGG16[9] model for transfer learning, the parameter values of the convolutional layer (feature layer) are fixed, but the last fully connected layer is retrained from scratch, and let the number of output neurons be consistent with the number of categories of the dataset. The data uses the same set of training samples and test samples, regarding 10500 images as the training set and 4500 images as the test set. The experimental results are shown in Table 3.

Table 3. Results of transfer learning

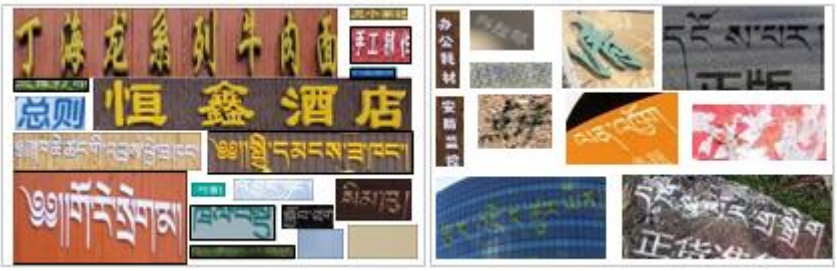
Network Model	**Pretrain VGG16**	**Depth 7-layer**
Precision(%)	97.46	98.28
Recall(%)	97.43	98.28
F1-score(%)	97.43	98.28
speed(s/epoch)	19	23

It can be seen from Table 3. that the classification result of the convolutional neural network with a depth of 7 is compared with the pre-trained VGG16 model for transfer learning. Although transfer learning can turn the learned model parameters to the new model and thus speed up and optimize the learning efficiency of the model, but essentially, VGG16 is a 16-layer convolutional network, the network depth is more than twice that of the 7-layer network, besides, its accuracy is lower than the latter. This further illustrates that deepening the network depth does not improve the accuracy of our classification task, and verifies the effectiveness of the 7-layer network.

4.3. Discussion

Through the classification of Tibetan, Chinese and non-text images on convolutional neural networks of different depths, with the increase of the network depth, the classification accuracy is constantly improved. When the network depth is 7, the best accuracy is achieved, and the accuracy will decrease if the depth continues to increase.This could be that the deeper the network, the smaller the size of the feature map, which losses a lot of information, and the phenomenon of gradient disappearance will become more and more obvious, so the classification accuracy will be reduced.

In the process of text image classification, when the text in the image is written horizontally or with a single background, it can be accurately classified, as shown in Figure.3(a). When the text in the image is written in the vertical direction, the non-text background of the image is diverse or other text appears, there will be misclassification occurring, see Figure.3(b). For the problem of vertical text image misclassification, the main reason could be that there are few vertical text images in the training samples, when two types of text appear at the same time in one text image, the misclassification will happen, which is not the result we need obviously.



(a) Examples of correctly classified (b) Examples of misclassification

Figure.3 Classification results

5. Conclusions and Future Work

Based on the in-depth study of CNN, we use the advantages of convolutional neural networks to conduct an exploratory study of multi-layer deep neural networks to extract and classify text image and non-text image features of scenes in different languages, analyze the impact of different depths on classification results, and simultaneously compare with the classification results of the transfer learning methods. The results show that the network depth has an impact on the classification results. As the network depth increases, the overall classification accuracy increases first and then decreases. Therefore, an appropriate convolutional network depth should be selected for our Tibetan, Chinese, and non-text image classification. For the classification of our task, the shortcomings are that the various non-text background and the paper with low opacity will make the network misclassify. Future studies will focus on exploring the comparison of different deep convolutional neural networks such as ResNet to avoid the gradient disappearing along with network deepening. At the same time, more Tibetan and Chinese text images and non-text image data are collected to train the network, so that it can classify more complex background text images, and final to be able to apply Tibetan, Chinese and non-text image classification to Tibetan- Chinese bilingual scene text detection and recognition.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (No.61772430) and by Program for Innovative Research Team of SEAC ((2018)98). The Gansu Provincial First-Class Discipline Program of Northwest Minzu University (No.11080305). The Northwest Minzu University for Nationalities' Central University Fundamental Research Service Fund Special Funds Support Postgraduate Research and Innovation Projects (Yxm2020108) supports the work also.

REFERENCES

[1] Ouyang W, Zeng X, Wang X, et al. DeepID-Net: Object Detection with Deformable Part Based Convolutional Neural Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(7): 1320-1334.

[2] Ali Diba, Vivek Sharma, Ali Pazandeh, et al. Weakly super-vised cascaded convolutional networks[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 5131-5139.

[3] HU G, YANG Y X, YI D, et al. When face recognition meets with deep learning: an evaluation of convolutional neural net-works for face recognition[C]. International Conference on Computer Vision, 2015: 142-150.

[4] LAWRENCE S, GILES C L, TSOI A C, et al. Face recognition: a convolutional neural-network approach[J]. IEEE Transactions on Neural Networks, 1997, 8(1): 98-113.

[5] Cao Z, SIMON T, WEI S, et al. Realtime multi-person 2D pose estimation using part affinity Fields[C]. IEEE Conference on Computer Vision and Pattern Recognition,2017: 1302-1310.

[6] TOSHEV A, SZEGEDY C. DeepPose: human pose estimation via deep neural networks[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2014: 1653-1660.

[7] SuFu, Lv Qin, LuoRenze, Review of image classification based on deep learning[J]. Telecommunications Science, 2019,35(11):58-74.

[8] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2012. 1097–1105.

[9] Simonyan, Karen, Zisserman, Andrew. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.

[10] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[J]. arXiv preprint arXiv:1409.4842, 2014.

[11] He KM, Zhang XY, Ren SQ, et al. Deep residual learning for image recognition[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2015. 770–778.

[12] Wang JZ, Yang Y, He YH. Pornographic Images Recognition Framework Based on Multi-Classification and ResNet[C]. Computer Systems and Applications, 2018, 27(9): 100-106.

[13] Zhao B, Li P, Dai MR, Ma XN. Research on Optimization Method of Railway Image Scene Classification Based on Deep Learning Method. Computer Systems and Applications, 2019, 28(6): 228-234.

[14] Fang HW, Shi HJ. Satellite Image Recognition and Classification Method Based on Deep Learning. Computer Systems and Applications, 2019, 28(10): 27-34.

[15] Yang Bing, Chen Hao-yue, Wang Xiao-hua, YaoJin-liang. Chinese Painting Image Classification Based on Convolution Neural Network[J]. Software Guide,2019, 18(01):11-14.

[16] Bai Cong, Huang Ling, Chen jia-nan, Pan Xiang, Chen Shengyong. Optimization of Deep Convolutional Neural Network for Large Scale Image Classification[J]. Journal of Software, 2018,29(04):1029-1038.

[17] Liu Wanjun, Liang Xuejian,Qu Haicheng. Learning performance of convolutional neural networks with different pooling models[J]. Journal of Image and Graphics, 2016,21(9):1178-1190.

[18] Wang Min, Liu Kexin, Liu Li, Yang Runling. Super-Resolution Reconstruction of Image Base on Optimized Convolution Neural Network[J]. Laser&Optoelectronics Progress, 2017, 54(11): 111005.

[19] Li Ming, Zhang Hong. Image classification based on convolution neural network of iterative optimization[J]. Computer Engineering and Design, 2017, 38(1): 198-202.

[20] Guo ST, Luo YX, Song YZ. Random forests and VGG-NET: An algorithm for the ISIC 2017 skin lesion classification challenge. arXiv preprint arXiv:1703.05148, 2017.

[21] Jaderberg M, Simonyan K, Vedaldi A, et al. Synthetic data and artificial neural networks for natural scene text recognition[J]. arXiv preprint arXiv:1406.2227, 2014.

AUTHORS

Jincheng Li received a bachelor's degree in 2018. He started his master's degree in computer technology at Northwestern University for nationalities in 2018. His research interests include image processing, pattern recognition and artificial intelligence. Contact him at ljicher@gmail.com

Penghai Zhao received a bachelor's degree in 2019. He started his master's degree in computer technology at Northwestern University for nationalities in 2019. His research interests include image processing, pattern recognition.

Yusheng Hao is a lecturer in the School of Mathematics and Computer Science, Northwest Minzu University. He is also a member of CCF and currently pursuing his PhD at the Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University. His research interests include signal processing, document analysis and idetification, thangka image processing. Contact him at haoyusheng@xbmu.edu.cn

Qiang Lin was born in 1979. He received the Ph.D. degree in computer science and technology from Northwestern Polytechnical University in 2014. Now he is an associate professor and M.S. supervisor at Northwest Minzu University. His research interests include pervasive computing, intelligent information processing, data stream mining, etc.

Weilan Wang received a B.S degree in Mathematics from Northwest Normal University, Lanzhou. China, in 1983. She was a visiting scholar with the Sun Yat-sen University, Guangzhou, China, in 1987. From 2001 to 2002, she was a visiting scholar with Tsinghua University, Beijing, China. From 2006 to 2007, she was a visiting scholar with Indiana University, Bloomington, USA. She is currently a Professor with the College of Mathematics and Computer Science, Northwest Minzu University, Lanzhou City, China. Her research interests include image processing, pattern recognition, Tibetan information processing and so on.