Handwriting Text Recognition Based on Faster R-CNN

*Junqing Yang
College of Electrical Engineering and Automation
Shandong University of Science and Technology
Tsingtao, China
yjqings2005@163.com

Peng Ren
College of Electrical Engineering and Automation
Shandong University of Science and Technology
Tsingtao, China
rpeng2058@gmail.com

Xiaoxiao Kong
College of Electrical Engineering and Automation
Shandong University of Science and Technology
Tsingtao, China
xxkong6@163.com

***Abstract*—Handwriting text recognition is one of the emphases in computer vision. The traditional Optical Character Recognition (OCR) technology requires the text writing neatly and handwriting clearly, but in fact, the handwriting text always fail to meet such states. In this paper, a novel handwriting text recognition algorithm based on deep learning is presented to improve the problems. In this paper, the method based on an object detection algorithm (Faster R-CNN) finds a new dimension to study the problem. The algorithm sets two steps: First, preprocessing the handwriting character based on Faster R-CNN, second, character recognition based on the Convolutional Neural Networks. The correctness of this method is better than the traditional OCR by the testing data. Experimental results show that the recognition algorithm in this paper is effective and illuminating.**

Keywords—Optical Character Recognition; Faster R-CNN; Deep learning; Convolutional Neural Networks

I. INTRODUCTION

Deep Learning is popular in Machine Learning due to the big data and the improvement of computing power, it develops rapidly in the fields of computer vision (CV), natural language processing (NLP), recommender system and so on. Traditional machine learning needs handcrafted feature extractor, relies on the expertise of individuals experts but sometimes people studies one problem as one-sided, so the actual effect is not good in some specific areas. Deep Neural Network (DNN) gets more feature function by defining a big function, finds the best feature function based on data. By visualizing the output of each hidden layer in the Convolutional Neural Networks (CNNs), every hidden layer represents some different feature extraction methods, CNNs can be regard as the combination of the low-level of and the high-level of feature extraction [1]. So CNNs is a modular process in training data, tunes the different levels of feature extraction function (each hidden layer) to optimizing the total network. A complex problem can be split into many simpler sub-problems, then using specific method to solve different sub-problems, it can get better results through a continual process of balance and optimization.

In this paper, the method based on CNNs and Faster region-based convolutional neural networks (Faster R-CNN) [2] is proposed to improve the problem about complex handwriting text recognition. CNNs is an exciting version of DNN, it has demonstrated excellent performance into image classification tasks such as handwriting character recognition. Faster R-CNN has high performance in object detection, it can be simply regarded as the combination of Region Proposal Networks (RPNs) and Fast R-CNN [3]. Faster R-CNN achieves end-to-end training specifically for the task by using novel RPNs, it is further promoted on the basis of region proposal methods and Fast R-CNN.

The way of handwriting text recognition mainly consists of two parts: First, finishing character segmentation. Second, character recognition. The accuracy rates of recognition based on Traditional OCR is relatively poor to these two parts for complex conditions, so it is very desirable to solve this problem based on deep learning.

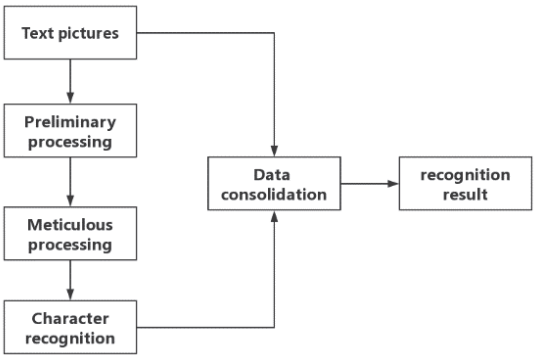

Fig.1. Flow chart of the Proposed Algorithm

For the task of complex handwriting text recognition in this paper, the four main steps can be detailed as follow, and the flow diagram is shown in Fig.1.

1) Preliminary processing; Character segmentation of the words based on Faster R-CNN, obtaining the image of each word as the output to the next network structure, meanwhile the spatial information of words that in the text can be obtained.

2) Meticulous processing; Character segmentation of the letters based on Faster R-CNN, getting the pictures of each letter as the output to the next network structure, recording the spatial coordinate connection of each letter.

978-1-7281-4094-0/19/$31.00 ©2019 IEEE

3) Character recognition; Classification and recognition of the input each letter image based on CNN. Output the recognition results as easy-to-use data, save data to next step.

4) Data consolidation; Data consolidation based on the spatial information of word in the text, finishing visualization of results data in the text image through covering the original text area.

After completing the steps above, the information of one handwriting text image is extracted, then processing information according to targets based on their spatial coordinates relations, such as translation (e.g. English to Chinese). Finally integrating results to generate new images, achieving the task of complex handwriting text recognition.

The rest of this paper is arranged as follows. Section □ introduces about proposed method. Section □ describes the three different datasets that be used in this paper. Section □ demonstrates about the Experiments and Results. Section □ concludes the paper.

II. Proposed Method

A. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a category of Deep Neural Networks (DNN) that have shown very effective in specific fields such as image recognition and classification, it is inspired by the natural visual perception mechanism of the living creatures.

LeNet is one of the very first CNNs which is created by Yann LeCun and is used mainly for character recognition tasks such as letters, digits etc. And now, CNNs have been extensively used to many difficult fields of deep learning such as Natural Language Processing tasks.

CNNs derive their name from the convolution operator. The primary purpose of Convolution is extracting features from the input image. Convolution preserves the spatial relationship between pixels by learning image features using small squares of input data.

Spatial pooling (also is called downsampling) reduces the dimensionality of each feature map. Removing unnecessary redundant information of feature map, but retaining the most important information. Spatial Pooling can be of different types: Max pooling, Average pooling, Sum pooling etc.

Therefore, CNNs has unique advantages in image processing with its special structure of local weight sharing. Spatial sharing is the prior knowledge introduced by convolutional neural networks. Weight sharing reduces the complexity of the network and avoids the complexity of data reconstruction in feature extraction and classification [6].

In a nutshell, CNNs consists of two main steps: feature extraction and the classification. The feature extraction part consists of one or several convolution layers and pooling layers. Feature map that is extracted becomes the input to the fully connected layers to classify.

In this paper, first convolution layer and pooling layer are designed to extract features. Then to eliminate the influence of size on recognition results, multiple different kernal sizes are used to capture different concepts in different ranges at one time, and the network can choose the desired features by itself. Finally, the feature map that is extracted as the input to full connection layers to classify. The CNNs architecture shown in Fig.2.

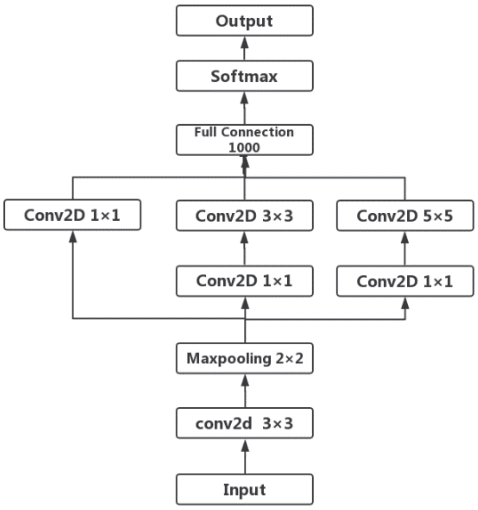

Fig.2. Network Structure of CNN

B. Faster R-CNN

After the accumulation of R-CNN and Fast R-CNN [3,4], Ross B. Girshick proposed a new Faster R-CNN in 2016. In general, Faster R-CNN has integrated feature extraction, proposal extraction, bounding box regression, classification into a network, achieving end-to-end training in a real meaning.

Faster R-CNN integrates Region Proposal Networks(RPNs) that a novel network structure instead of Selective search(SS) [7] into deep network, which not only solves the problem that long computation time due to SS is implemented by CPU, but also shares the convolutional layer parameters with overall deep networks. While maintaining the accuracy, the detection speed is greatly improved, so it is an effective method for object detection. Faster R-CNN can be mainly divided into four aspects:

a) feature extraction; As an object detection method of CNNs, Faster R-CNN extracts feature maps of image using a set of basic convolution and pooling operation.

b) Region Proposal Networks; RPNs is used to generate region proposals. This network scores anchors belong to foreground or background by softmax, and then modifies anchors to obtain accurate proposals by bounding box regression.

c) Region of interest pooling(RoI pooling); For every RoI from the input, it takes a section of the input feature map that corresponds to it and scales it to fixed size. Processing input feature maps and proposals, summing these information and extracting proposal feature maps, determining the object category by sending to the full connection layer.

d) Classification regression; Obtaining object classification and bounding box based on the classification information and location information.

Generally, firstly scaling the input image, and entering it into the convolution layers and pooling layers to extract the feature map. Secondly the feature map is fed into the RPNs to generate a series of possible candidate boxes for the target.

Thirdly, the original feature map and all candidate boxes of RPNs output are fed into the RoI pooling layer for extracting and collecting the proposal, then calculating the feature maps with a fixed size of 7*7, which are fed into the full connection layer, finishing output object classification and coordinate regression. The flow chart of Faster R-CNN as shown in Fig.3.

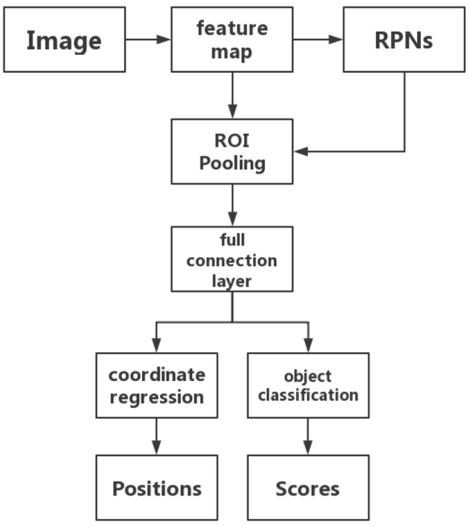

Fig.3. Flow chart of the Faster R-CNN

In this paper, the algorithm sets another way for dealing with the problem. Character segmentation is transformed into binary classification that is divided into object region and background region.

In preliminary processing, the whole text is divided into word region and background region, then finishing character segmentation based on Faster R-CNN, saving the word region image and sending to the next step. Similarly, in meticulous processing, the word region image is divided into character region and background region based on Faster R-CNN, recognizing the letters based on CNNs. Finally, combining the results of character recognition with their corresponding spatial coordinates, storing it for arrange letters to word, then it is integrated into text recognition result.

III. Dataset

In this paper, the algorithm needs three different datasets for three tasks. They are Word dataset, Letter dataset and EMNIST [8]. Word-dataset is used to train model for preliminary processing. Letter-dataset is used to train model for meticulous processing. Finally, training a classifier by EMNIST for character recognition.

A. Word-dataset

Due to the complexity of handwritten text, it is necessary to complete preliminary processing. Using LabelImg (a useful software for labelling data) to label the word of each text. A total of 1000 training data has been collected. Each data needs to label all words.

At the same time, preprocessing data for saving computing resources and getting better effect of feature extraction. Firstly, It does not change the spatial information of the image by scaling an image, so scaling of data image pixels to 320*240 together. Secondly, the median filter algorithm is deployed to preprocess the word's images. Thirdly, graying the image to reduce the scale of data processing. In addition, expansion of data to 4000 by data enhancement for increasing training data.

Finally, 4000 data and xml files that stored their word's label coordinates has got. In actual operation, training the network based on tensorflow, so it needs to convert the Word-dataset to the format of tfrecord for training, then randomly selecting about 3000 of them as training data, the rest as a verification data, the min batch size of each iteration is 60.

B. Letter-dataset

After preliminary processing, words are separated from the text. Every word would be segment to letter images for character recognition, so creating a dataset to train model for letter segmentation.

Letter-dataset is similar as Word-dataset. First, collecting 500 word pictures as training data, each picture has about 5 to 10 letters. Second, removing noise points from pictures using median filter algorithm, using the Roberts gradient, Sobel operator and Laplacian sharpening for preprocessing data, comparing the results, choosing Laplacian algorithm as sharpening algorithm. Thirdly, increasing samples to 2000 by data enhancement. Finally, graying samples and scaling it to 120*80. Labelling each letter of the word image, Letter-dataset has a total of 2000 data, about 11000 letters.

For easy training, the type of dataset is converted to the format of tfrecord, then randomly selecting 1500 of them as training data, the rest as a verification data, the min batch size of each iteration is 80.

C. EMNIST

The EMNIST dataset is a set of handwritten character digits derived from the NIST Special Database 19, it is converted to a 28x28 pixel image format, and the dataset structure that directly matches the MNIST dataset. The dataset have been size-normalized and centered in a fixed-size image. It is a useful dataset for researchers who want to try training model for character recognition, it will help human costs minimal efforts on preprocessing and formatting.

There are six different classes provided in this dataset, including ByClass, ByMerge, Balanced, Letters, Digits and MNIST. In this paper, CNNs used the class of EMNIST-Letters as dataset for training.

EMNIST-Letters including 145,600 characters and 26 balanced classes, finally randomly selecting 100,000 of them as training data, the rest as a verification data, the min batch size of each epoch is 200.

IV. Experiment Results And Analysis

In this paper, the Faster R-CNN uses VGG-19 [9] as pretraining model. CNNs has powerful feature extraction ability, the feature extraction part by training of classification task can be also used for various tasks such as location detection [4]. Different tasks can be finished by changing the last layers of the network, it is unnecessary to training the parameters of the whole network from scratch. It mainly regards the front part of the network as feature extraction layer, then different tasks share this feature extraction layer [10].

VGG-19 is very deep CNNs, which has been trained from ImageNet, it has powerful feature extraction function. And the pretraining model not only reduces the training time, but also

ensures the accuracy rate of recognition. Therefore, choosing VGG-19 as pretraining model.

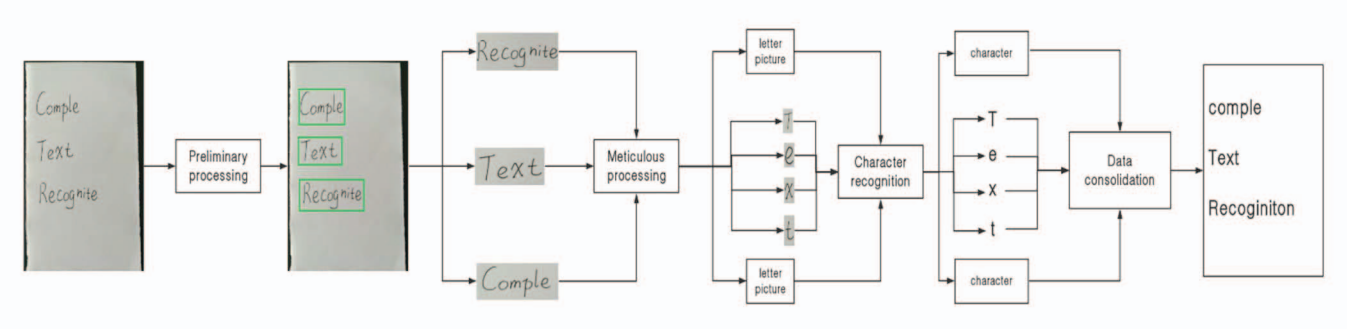

Fig.4. Integral Recognition System

The experimental environment with: GeForce CUDA7.0, memory 6GB. Results of the test sample is shown in Fig.4.

In the first two steps, the training process of Faster RCNN is similar to an iteration process, but only two cycles:

1) Training RPN networks on pretraining models(VGG-19).

2) Collecting proposals from the trained RPN network to training fast R-CNN.

3) RPNs is retrained using the Fast R-CNN.

4) Fast R-CNN is retrained using RPNs of step.3.

For preliminary processing and meticulous processing, the loss function decreases with the iteration increases. Loss function curve of the two networks tend to converge after 40,000 and 36,000 iterations, recording the loss value every 100 times. For preliminary processing, the min loss value is 0.112, and meticulous processing's min loss value is 0.204, this is a relatively good result. After testing, the model can achieve better recognition effect. The trend change of loss function is shown in the Fig.5 and Fig.6.

In Fig.5 and Fig.6, the total loss(total) is the aggregation of the goal classifier(object class), the loss of RPNs (bounding box), classification for whether the box contains the object (box is object), and the regression for the bounding box coordinates (coordinates).

The network structure of CNNs in Section □ CNNs, its curve tends to converge quickly after 10,000 times, each 100 times to record an error rate verification results, the min loss value of it is 0.01.

In order to test this method accuracy, randomly selecting 200 new images as testing data. Under the condition of pure text, the average character segmentation rates of the word can reach 99%, the average character segmentation rates of the letter can reach 95%, and average character recognition can reach 97%, the method accomplishing character recognition is better than traditional OCR.

Experimental result shows that the algorithm has better accuracy and robustness. For a recognition task, character segmentation of word can be done well based on Faster R-CNN. The error is recognizing two words as one due to they are too near. For this problem, a bag of words can be set up as insurance, or using a novel method, processing the final result to overcome this problem [11].

The average character segmentation rates of letter can reach 95%, the mainly problem is the linking of each letter. Many people can easily write in tandem when they writing, traditional OCR is difficult to finish character segmentation for this condition. But this algorithm based on Faster R-CNN is better to deal with this problem.

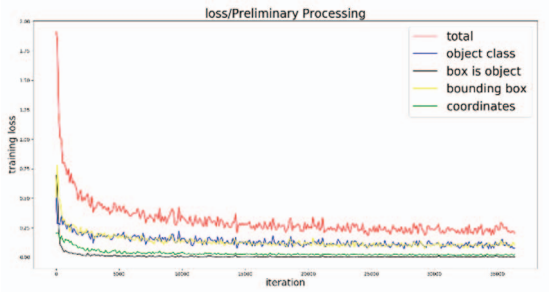

Fig.5. The loss of preliminary processing

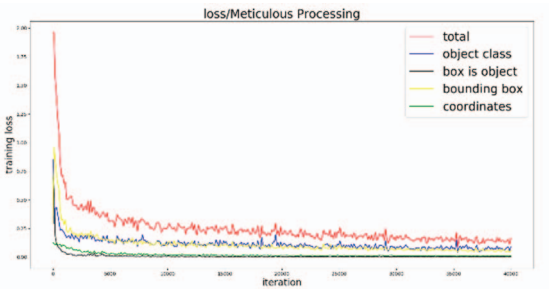

Fig.6. The loss of meticulous processing

Due to the handwriting text background is relatively complex condition, everyone writing style is different, this method is modularizing the recognition task and solving this problem step-by-step. Traditional OCR needs set feature function artificially, the problem of poor generalization ability still exists, and the final actual effect can't meet the realistic needs. In this paper, the algorithm based on Faster R-CNN can finish character segmentation with high accuracy, then completing character recognition based on CNNs. For different tasks of text recognition, the method works well.

Comparing with the traditional OCR, this algorithm is relatively stable and intelligent, and the accuracy is relatively well.

V. Conclusion

In this paper, the novel handwriting text recognize method based on Faster R-CNN is presented, the experimental results show that it is better than the traditional OCR. The difficult of

handwriting text recognition is character segmentation, by transforming the problem of character segmentation into the object detection, successful finishing character segmentation with high accuracy, so the method based on Faster R-CNN is used to overcome this difficult. Then finishing character recognition based on CNNs for each character.

In this paper, the method finishes handwriting text recognition through opening a new dimension, transforming the problem into many sub-problems by modular design. Comparing with traditional OCR, this algorithm shows well performance on complex handwriting text recognition. Meanwhile, method has the generalization ability for similar recognition problems, such as Chinese text recognition. So this method is enlightening and intelligent.

REFERENCES

[1] Zeiler M.D., Fergus R. (2014) Visualizing and Understanding Convolutional Networks. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8689. Springer, Cham

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," in *International* Conference on Neural Information Processing Systems, 2015, pp. 91–99.

[3] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1440-1448.

[4] Girshick R , Donahue J , Darrelland T , et al. Rich feature hierarchies for object detection and semantic segmentation[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2014.

[5] Krizhevsky A , Sutskever I , Hinton G . ImageNet Classification with Deep Convolutional Neural Networks[C]// NIPS. Curran Associates Inc. 2012.

[6] Gu J , Wang Z , Kuen J , et al. Recent Advances in Convolutional Neural Networks[J]. Computer Science, 2015.

[7] Uijlings J R R , K. E. A. van de Sande⋯. Selective Search for Object Recognition[J]. International Journal of Computer Vision, 2013, 104(2):154-171.

[8] Cohen, G., Afshar, S., Tapson, J., & van Schaik, A. (2017). EMNIST: an extension of MNIST to handwritten letters. Retrieved from http://arxiv.org/abs/1702.05373

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

[10] Sermanet P , Eigen D , Zhang X , et al. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks[J]. Eprint Arxiv, 2013.

[11] Sueiras J , Ruiz V , Sanchez A , et al. Offline continuous handwriting recognition using sequence to sequence neural networks[J]. Neurocomputing, 2018:S0925231218301371.