A New Type Method of Adhesive Handwritten Digit Recognition Based on Improved Faster RCNN

2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP) | 978-1-7281-6896-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICSIP49896.2020.9339270

Zuo Huahong
Science and Information Center
Jingzhou Tobacco Company
Jingzhou, China
e-mail: zuohua@163.com

Tang Junyi
School of Information Engineering
Wuhan University of Technology
Wuhan, China
e-mail: 14358631@qq.com

Han Ping*
School of Information Engineering
Wuhan University of Technology
Wuhan, China
e-mail: hanping@whut.edu.cn

***Abstract*—Aiming at the low recognition accuracy of the traditional machine learning algorithm which is susceptible to digital writing quality, inter-digital adhesion, random noise background and other factors in the process of adhesion handwritten digit recognition, an new method based on improved fast regional convolutional neural network(Faster RCNN) of adhesion handwritten digit recognition is proposed. Firstly, the NIST19 dataset is used as the basic dataset, and a mixed dataset is created by setting different hand-to-hand ratios with different degrees of overlap, and then randomly add salt and pepper noise and Gaussian noise in the experimental images. Secondly, aiming at the problem of a large number of overlapping objects in the handwritten digital images, a model based on improved Faster RCNN network is built and trained with the above data sets. Finally, the average accuracy of the model is evaluated. The experimental results show that the average detection accuracy of the proposed model is good. Compared with the original Faster RCNN and YOLO models, the improved model not only reduces the scale of parameters, but also ensures high recognition accuracy, and realizes the accurate and efficient recognition of handwritten adhesive digits.**

Keywords-deep learning; image processing; adhesion handwritten digital recognition; image quality evaluation

I. INTRODUCTION

The traditional recognition algorithm of continuous digit string mainly extract different types of hand-made features, and use the traditional learning algorithm to train effective classifiers for detection and recognition. Most of these methods are limited to specific data sets, and each component is optimized separately, hence the handwritten strings under complex conditions can't be recognize accurately.

In recent years, with the rise of deep learning methods, especially the deep convolution neural network [1], [2], it provides new ideas for image recognition and object detection. Aiming at handwritten adhesive digit recognition, some related papers use the idea of deep learning. A graph transformer networks [3] replaces the original handwritten numeral location, segmentation and recognition module, so that the system can use gradient based training method in the global range, so as to reduce the overall performance index. A dense convolution network [4], which connects each layer of the network architecture with all its subsequent layers, solves the problem of over fitting and trains the network accurately at a faster speed. The algorithm has achieved the most advanced results on the famous benchmark datasets(CIFAR-100, SVHN, Imagenet). In addition, Hochuli [5] proposed a framework for multi-length digit string recognition without the use of segmenta-tion algorithm and general classifier. The segmentation algorithm is successfully replaced by two CNN classifiers trained for specific tasks, and good experimental results are obtained on NIST19 dataset.

Most of the existing detection systems based on neural network require the experimental images to keep noise free background or specify specific background. In order to ensure the recognition accuracy, it is necessary to add the corresponding pretreatment process for different problems before the formal experiment. These preprocessing will lead to work inefficient and accurate identification, and the recognition ability of model on non-experimental data set will be greatly reduced. In this paper, deep learning technology is applied to the recognition of handwritten digits. A recognition framework based on the improved faster RCNN [6] model is proposed, and a series of experiments are conducted on NIST19 data set to evaluate its recognition ability. The structure of this paper is arranged below: section I introduces the research condition, section II builds the recognition model and gives the algorithm, section III analyzes the performance of the algorithm by experiments, and section IV is conclusion.

II. THE RECOGNITION MODEL OF ADHESIVE HANDWRITTEN DIGITS

The network model based on deep learning is shown in Fig. 1. The model consists of 8 convolution layers, 4 pooling layers, 4 fully connected layers and 8 relu activation functions. The specific network parameters are shown in Table I. The network is divided into three parts: the first part extracts the features of the original input image through five convolution layers and three pooling layers; the second part feeds the feature map extracted from the first part into a 3x3 RPN area to generate a network (conv6_3). According to nine different sizes of anchor points designed in advance, the target suggestion frame is obtained. Two convolution operations are carried out to screen out the extraction box which may contain the target, and the first bounding box regression is performed to obtain the anchor position that may contain the detection object for the first time.

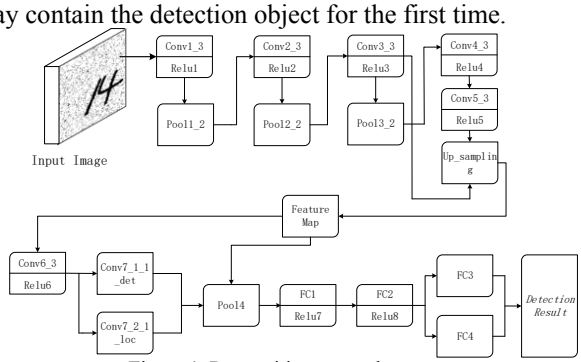

Figure 1. Recognition network structure

TABLE I. NETWORK MODEL PARAMETER CONFIGURATION

Layer	Input Size	Kernel Size	Stripe	Output Size
Conv1	400× 400× 3	3× 3× 32	2	200× 200× 32
Pool1	200× 200× 32	2× 2× 32	2	100× 100× 32
Conv2	100× 100× 32	3× 3× 64	1	100× 100× 64
Pool2	100× 100× 64	2× 2× 64	2	50× 50× 128
Conv3	50× 50× 128	3× 3× 128	2	25× 25× 128
Pool3	50× 50× 128	2× 2× 128	2	25× 25× 128
Conv4	25× 25× 128	3× 3× 256	1	25× 25× 256
Conv5	25× 25× 256	3× 3× 512	1	25× 25× 512
Up_sampling	25× 25× 512			50× 50× 512
Conv6	50× 50× 512	3× 3× 512	1	50× 50× 512
Conv7_1	50× 50× 512	1× 1× 18	1	50× 50× 18
Conv7_2	50× 50× 512	1× 1× 36	1	50× 50× 36
Pool4				7× 7× 512
FC1	7× 7× 512	4096		1× 1× 4096
FC2	1× 1× 4096	4096		1× 1× 4096
FC3	1× 1× 4096	11		1× 1× 11
FC4	1× 1× 4096	44		1× 1× 44

Because the size and shape of the target suggestion box generated by the second part are different after modification, in order to further accurately modify it, the suggestion box generated by RPN network and the feature map generated by the first part are input into the ROI pooling layer(pool4) to obtain the mapping area of the suggestion box on the feature map. Specifically, for the suggestion box of *M*n* size, firstly, it is mapped back to the (*M*/8)*(*n*/8) size feature map scale, and then it is cut into 7*7 pieces, and each copy is processed by Max pooling [7]. After this processing, the input boxes of different scales can achieve the output of fixed length. Finally, calculate the specific category of each suggestion box(i.e. the ten numbers of 0-9) through the full connection layer and softmax. At the same time, more accurate target detection frames are obtained by using bounding box regression, and the final detection results are output [8].

Since the frame size of the number in the original image in NIST database is fixed in a certain range, which is not conducive for neural network to learn the characteristics of different levels of numbers, and also has certain interference on the selection of anchor box size. Therefore, after obtaining the digital string images with different intersection and merge ratios, the numbers on the image are appropriately reduced without changing the original size of the image. In order to simulate the writing habits in the actual situation, different numbers in the same image keep the same scaling condition.

After processing the experimental data set, the reasonable anchor box size is set for the region generation network to generate the corresponding detection frame. Since more than 20000 anchor frames in NIST19 contain many overlapped frames, Non Maximum Suppression[9] (NMS) is needed to coarse screen them, and the detection frame with the highest local score is selected from the overlapped boxes whose intersection ratio exceeds the set threshold and is retained. This step only classifies whether the detection box contains recognition objects(i.e. whether it contains handwritten digits). The classification model with different numbers is located in the FC4 full connection layer of the network. This layer uses the error between the calculated result and the actual value of the softmax loss function, and optimizes the network parameters through the back-propagation algorithm until it converges.

Because there are a large number of adhesive detection targets in this experiment, if the number of suggestion frames with the highest score is directly set to zero, a large number of FN(False Negatives) will be generated. That is to say, the positive samples are wrongly identified as negative samples, thus affecting the recognition accuracy of the model.

To solve this problem, Soft-NMS[10] algorithm is used to replace the original NMS algorithm in the second screening process. Different from NMS algorithm, soft NMS algorithm attenuates the detection score of non-highest score detection frame instead of removing it completely. The attenuation degree of soft NMS algorithm is positively related to the overlap between the detection frame and the adjacent highest score detection frame. The larger the overlap, the lower the score. This method can effectively avoid the problem that the detection frames with lower scores in overlapping targets are eliminated directly. The score reset function of Soft-NMS algorithm is as follows:

$$s_i = \begin{cases} s_i & IOU(\mu, b_i) < N_t \\ s_i(1 - iou(\mu, b_i)) & IOU(\mu, b_i) \geq N_t \end{cases} \quad (1)$$

It can be seen from the above formula that when the overlap area between the detection frame and the adjacent highest score detection frame exceeds the threshold, the detection score of the detection frame decreases linearly, while the detection frame whose overlap area is less than the threshold value will not be affected. Because the overlapping degree of the function changes suddenly when the threshold value is, the fraction in the set of detection frames is faulted. In order to solve this problem, the following formula is adopted as the scoring standard of Soft-NMS:

$$s_i = s_i e^{-\frac{iou(\mu, b_i)^2}{\delta}}, \forall b_i \notin D \tag{2}$$

III. EXPERIMENTAL ANALYSIS

All experiments in this paper use NVIDIA Tesla K80 GPU for training and calculation. The running environment is ubuntu16.04, and the training framework is tensorflow. Firstly, the filtered images are put into the constructed network for training. The maximum number of iterations is 10000, the momentum factor is 0.9, and the initial network learning rate is 0.001. After 5000 iterations, the learning rate is adjusted to 0.0001. In the training process, the image is zoomed while the original aspect ratio is maintained. At the same time, horizontal flipped images are added to enhance the ability of model feature extraction.

A. Sample Set Construction

The purpose of this paper is to study the recognition ability of the improved convolution neural network model for the adhesive handwritten digits. In order to compare and evaluate the existing recognition algorithms, NIST19 is selected as the basic data set for training and testing. The data set contains 10 kinds of numbers from 0 to 9, and the size of each single digit image is 128*128. In order to meet the needs of model training and testing in this paper, the nist19 data set is processed in the preprocessing stage.

In order to study the accuracy of the improved model for digital recognition with different degree of adhesion, this paper simulated the digital images with different intersection and merging ratio(*IOU*). The *IOU* is the ratio of the Intersection area and the Union area between the ground truths in different single digital images:

$$IOU_{snd}^{fst} = \frac{area(box(fst) \cap box(snd))}{area(box(fst) \cup box(snd))} \tag{3}$$

Two single digital images are randomly selected from nist19 database. According to the coordinate box value of the digital border in the label file, the images are cut and overlapped in different degrees. The size of the composite image is set to 200*200 pixels. Fig. 2 is an example of adhesion of two numbers under two different *IOU*. In the process of constructing the sample, it is found that if the *IOU* is greater than 0.2, there will be a large number of numbers that do not conform to the human eye recognition logic. Therefore, in this experiment, the synthetic pictures with *IOU* of 0.1 and 0.2 are selected as the training and testing objects.

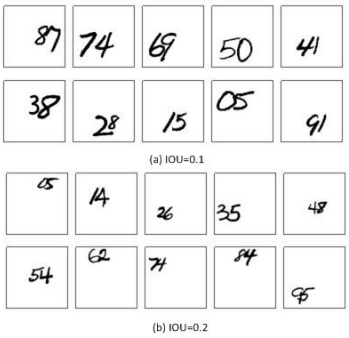

Figure 2. Example of adhesion between different cross-sections.

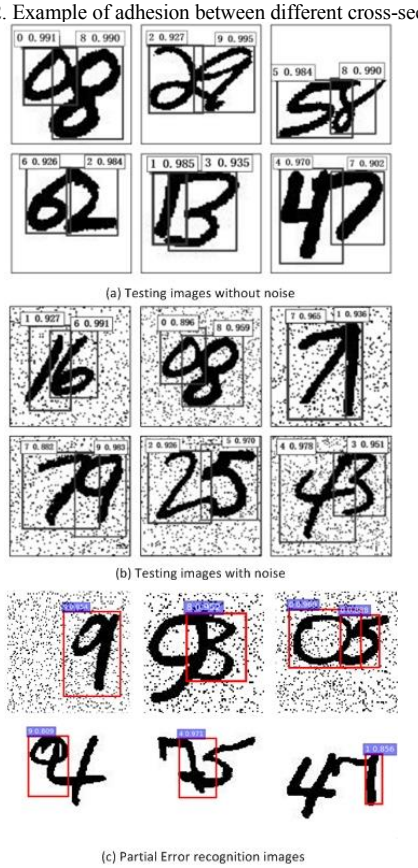

Figure 3. Example of Testing result

B. Model Testing

The number of images used in this experiment is shown in Table II.

TABLE II. VARIOUS TYPES OF EXPERIMENTAL IMAGES

Images	Training IOU=0.1	Testing IOU=0.1	Training IOU=0.2	Testing IOU=0.2
Images without noise	2000	1500	2300	1800
Images with Gauss noise	1800	1000	2000	1200
Images with pepper noise	1800	1000	2000	1200
Summary	5600	3500	6300	4200

As seen from Fig. 3, model recognition errors mainly focus on misjudging the part of the real number or the connecting part of two numbers into a new number. From the perspective of human vision, there is also a high possibility

of misjudgment for the images with irregular handwritten digits and over overlapping in the original data set.

C. Comparative Analysis

In order to further test the performance of the network model designed in this paper, the faster RCNN, YOLO [11] algorithm and the algorithm in this paper are used to test on the verification data set. The P-R (precision vs. recall) curve is drawn to compare the performance of each algorithm. P-R curve is a comprehensive evaluation of model precision and recall, where *P* represents the proportion of real positive samples in all positive samples detected in the test process, and *R* represents the proportion of positive samples detected by results in all positive samples in the test process. Their definition formula is as follows:

$$P(Precision) = \frac{TP}{TP + FP} \tag{4}$$

$$R(Recall) = \frac{TP}{TP + FN} \tag{5}$$

TP(True Positives) is the number of correctly identified positive samples, FP(False Positive) is the number of negative samples identified as positive samples by this error, and FN(False Negatives) is the number of positive samples that are wrongly identified as negative samples. The P-R curves of three different algorithms obtained by adjusting the classification threshold are shown in Fig. 4.

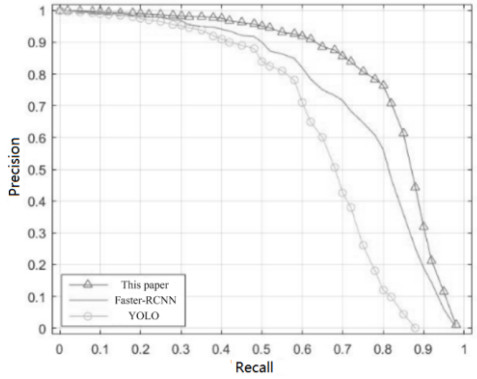

Figure 4. Comparison chart of different algorithms P-R curves

From the Fig.4, compared with the existing two recognition models, the detection model designed in this paper has more accurate recognition ability on the data set. At the same time, the network model designed in this paper uses less parameters, so it can achieve accurate and fast near real-time detection effect. Table III shows the recognition time of different algorithms on the testing dataset, the model proposed in this paper has good timeliness, so it can be well applied to practical systems.

TABLE III. TIME OF DIFFERENT MODELS ON TESTING DATASET

Model	Recognition time(ms)
This paper	147
Faster-RCNN	326
YOLO	182

IV. CONCLUSION

In this paper, a new method of handwritten numeral recognition and detection using deep learning technology is proposed. Specifically, according to the characteristics of handwritten digit string images, the current mainstream target detection algorithm faster RCNN framework is extended to design a network model which is consistent with the characteristics of the detection object, and corresponding improvements are made to obtain good detection results. In addition, on the basis of NIST19 data set, a series of experiments are made to verify the recognition ability of the framework. Compared with other types of target detection, such as face recognition, adhesive digit recognition has more obvious features, and it can have a high recognition rate through less training, which is also verified by comparative experiments. Experiments show that the GPU computing unit can be used to recognize and locate the handwritten digits in near real time and end-to-end. This will lay a good foundation for the subsequent intelligent digital detection. On the basis of this experiment, the next step is to study the application of the model in the actual scene, that is, the recognition ability of adhesive handwritten numeral images in different backgrounds.

REFERENCES

[1] X. Y. Li, L. Chen, "Multi-step segmentation method of online handwritten Chinese characters based on SVG," Application Research of Computers, vol. 11, 2017, pp. 3364-3372.

[2] Y. L. Zhu, Recognition of merged text-based CAPTCHA. Nanjing University of Science & Technology, 2017.

[3] F. Wu, Image Restoration Technology Based on Variational Partial Differential Equation. Peking University Publication, 2008, pp. 120-152.

[4] Huang G, Liu Z, Maaten L V D, et al. Densely Connected Convolutional Networks[C], 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 1, pp. 2261-2269..

[5] Hochuli A G, Oliveira L S, Jr A S B, et al. Handwritten Digit Segmentation: Is it still necessary?[J]. Pattern Recognition, 2018, 78. pp. 1-11.

[6] Y. M. Chen, S. Levine, and M. Rao, "Variable Exponent, Linear Growth Functionals in Image Restoration," SIAM Applied Mathematics, vol. 4, 2006, pp. 1383-1406.

[7] Z. X. Xie, X. F. Wang, X. L. Xiong, and Q. Hu, "Color image quality assessment based on noise model of human vision perception and color image quality optimization," Journal of Image and Graphics, vol. 10, 2010, pp. 1454-1464.

[8] J. Q. Yan, Text Detection and Recognition in Complex Scene of Image and Video. Xidian University, 2014.

[9] A. B. Petro, C. Sbert, and J. M. Morel, "Multiscale Retinex," Image Processing On Line, 2014, pp. 71–88. https://doi.org/10.5201/ipol.2014.107

[10] M. C. Hanumantharaju, M. Ravishankar, and D. R. Rameshbabu, "Color Image Enhancement Using Multiscale Retinex with Modified Color Restoration Technique," International Conference on Emerging Applications of Information Technology, Feb. 2011, pp. 93-97. doi:10.1109/EAIT.2011.64.

[11] Redmon J, Farhadi A . [IEEE 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Honolulu, HI (2017.7.21-2017.7.26)] 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - YOLO9000: Better, Faster, Stronger[J]. 2017, pp. 6517-6525.