# Research on recognition algorithm of Chinese text image based on Deep Learning

**Qifan Yang [1], Haomin Shao [2], Yi Li[3]**

[1] School of Information and Communication Engineering, Communication University of China, Beijing, China

[2] Nanjing University of Information Science and Technology, Nanjing, China

[3] College of Computer Science and Engineering, ShangDong University of Science and Technology,

**Keywords:** Uneven illumination, local adaptive nonlinear filter, depth learning, text image recognition.

**Abstract:** In order to solve the problems of uneven illumination and low character quality in text image recognition, an image enhancement algorithm and a character recognition model based on convolution cyclic neural network are proposed in this paper. Among them, the image enhancement algorithm uses an improved tone mapping function considering local information to increase the visibility of text in dark areas. The method of background estimation and contrast compensation is used to solve the problem of uneven illumination of the image, and the connected domain method is used to locate the text in the image. The convolution and loop depth neural network model is built based on the text region, and the whole string in the image is taken as the recognition target. In this paper, 30 uneven illumination images are collected for experimental verification, and the experimental results show that the text recognition accuracy of the model in this scene is 98.29%.

## 1. Introduction

Optical character recognition ((OCR)) refers to the automatic extraction of printed text from a picture by a device. Character recognition involves many fields, including artificial intelligence, image processing, computer and so on. The realization of character recognition technology plays a great role in information processing, office automation and other fields[1].

In this paper, taking the VAT invoice character recognition in the scene of uneven illumination as the object, an image enhancement algorithm based on local nonlinear filtering is proposed to process the image, and a new tone mapping function is introduced to generate the enhanced image. Compared with the original image, the text details in the dark area of the enhanced image are improved, and the uneven illumination is weakened. An improved convolutional cyclic neural network is built to build a multi-character recognition model which can avoid character segmentation. In the training stage, the generating learning method is used to create a dataset containing commonly used Chinese characters, English and numbers, and the samples are expanded by deformation, rotation, translation and adding noise to improve the robustness of the model. Comparative experiments show that the low-quality character recognition performance of this system is better than that of the commonly used OCR software in the scene of uneven illumination.

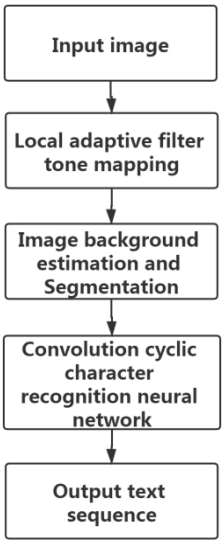


Fig 1 Functional flow chart of recognition algorithm for text images with uneven illumination based on deep learning

## 2. Text Image Enhancement algorithm in uneven Illumination scene

The system flow of this paper is shown in Figure 1, where image enhancement includes local adaptive filter tone mapping, image background and text positioning. The processing is realized based on MATLAB2016a of the Windows system, and the convolutional recurrent neural network is built on the Linux system using Python language[2].

This section uses the VAT invoice in Figure 2 as an example to introduce a low-quality text image enhancement algorithm based on uneven lighting conditions.


Fig 2 Example of uneven Illumination Image

### 2.1. High dynamic range image tone mapping based on global and local processing

The basic idea of tone mapping is to first calculate the average luminance of the scene, then select an appropriate luminance domain according to the average luminance, and finally map the whole scene to the luminance domain. The traditional tone mapping method considers the global average brightness of the image as follows:

$$L_{avg} = \frac{1}{N}\exp(\sum_{x,y}\log(\delta + L(x,y))) \quad (1)$$

In the formula: L(x, y) is the brightness value of the pixel (x, y); N is the total number of pixels in the image; δ is a small constant used to prevent the pixel from pure black. Map the displayable brightness of the new image through equation (2), where the C value is used to control the overall brightness and contrast of the output image:

$$L_d = \frac{L(x,y)}{L(x,y) + CL_{avg}} \tag{2}$$

This paper also considers the global average brightness and local average brightness and proposes a new mapping function to obtain the displayable brightness of each position of the image, α is the coefficient to balance Lbf and Lavg.[3]

$$L_{dm} = \frac{L(x,y)(1+\alpha L_{bf}(x,y)/L_{\max})}{L(x,y) + C[(1-\alpha)L_{avg} + \alpha L_{bf}(x,y)]} \tag{3}$$

In the formula (3): Lbf(x, y) is calculated by the weighted sum of the brightness of the adjacent pixels in the nonlinear filter sliding window block centered on the relevant pixel P.

As shown in Figure 3, the contrast between the light and dark areas of the image mapped by the adaptive nonlinear local filtering algorithm is reduced, and the text details are enhanced.

Fig 3 Image after tone mapping

**2.2. Text image background estimation and segmentation**

Fig 4 Original image gray background

Set the grayscale image of the image to I(x, y). The brightness value of the background pixels in the local area of the unevenly illuminated image is very different from the brightness value of the text area, so the background of the local area can be used for the brighter pixels in the area Point to indicate. The specific algorithm is:

The 21×21 sliding window traverses each pixel in I(x, y) in turn;

Find the 6 pixels with the highest brightness in the window;

Take the average of its 6 pixels as the output Ib(x, y) of the sliding window.

The extracted background is shown in figure 4.

For images with uneven illumination, the local average brightness of the background image varies widely, so it is necessary to normalize the brightness of the background image, the purpose is to change the text image with uneven illumination into an image with uniform illumination.

To remove the background with uneven illumination, it can be simply understood as the original image minus the background image. However, for the uneven illumination image Iu(x, y), the local average brightness variation range of the image background is larger, the darker the background area, the The contrast between the background and the text will be smaller.

The grayscale image after normalizing the image for background segmentation and compensating the contrast is shown in Figure 5.


Fig 5 Image after background segmentation

## 3. Text positioning

Before the text recognition, the text area in the picture needs to be extracted. Because the neural network model in this paper can directly recognize the text image of indefinite length, there is no need to segment the text image, so the method of connected domain labeling is used to locate the text.

### 3.1 Local binarization

The image after image enhancement still has some noise effects. The global binarization will magnify the effect of uneven lighting and generate artifacts that cause broken and missing strokes, as shown in Figure 6. In this paper, the Sauvola algorithm[4] is used to binarize the grayscale image Ie (x, y), and the text pixels are converted to 1 and the background pixels are converted to 0 through the inverse operation. Fig. 6 and Fig. 7 are comparison graphs of the output of the global binarization and Sauvola binarization of Fig. 5 respectively.


Fig 6 Global binary map

Fig 7 Sauvola binary map

### 3.2 Morphological processing and connected domain labeling

Binary images can remove table lines and small noises through graphic morphological operations, and form connected domains of similar characters. Markers can surround the largest rectangle of connected domains, which can achieve the positioning of text. Fig. 8 and Fig. 9 are graphs of connected domains of text obtained after a series of morphological processing and text areas located by filtering.

Fig 8 Connected domain graph

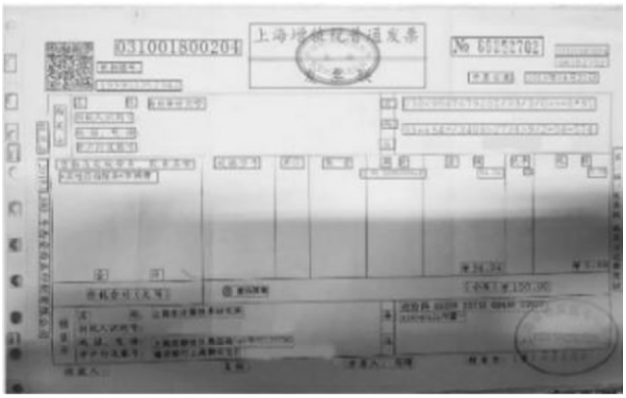

Fig 9 Text area

## 4. Character recognition based on convolutional neural network and short-term memory network

Based on the existing network model of natural scene character recognition, this paper improves it so that it can be applied to commonly used Chinese, English, numeral and punctuation character recognition.[4] The network structure is shown in figure 10, which is mainly composed of convolution layer, loop layer and transcription layer. At the bottom of the network is the ordinary

convolution network, which is used to extract the feature sequence from the input text image. After the convolution network, a circular network layer is constructed by two-way short-term memory units, which is used to predict the feature sequence obtained by the convolution layer. Finally, the transcription layer converts the prediction label of the loop layer into the final recognition result through re-integration.

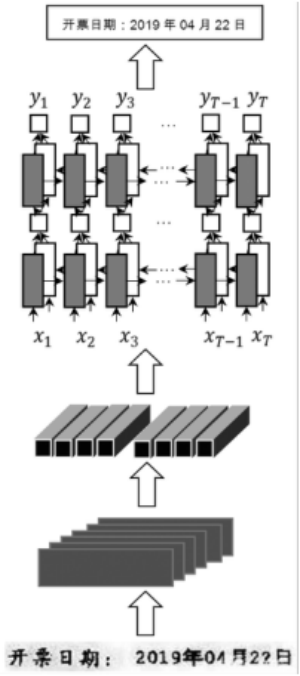


Fig 10 Network structure

### 4.1. Convolutional layer

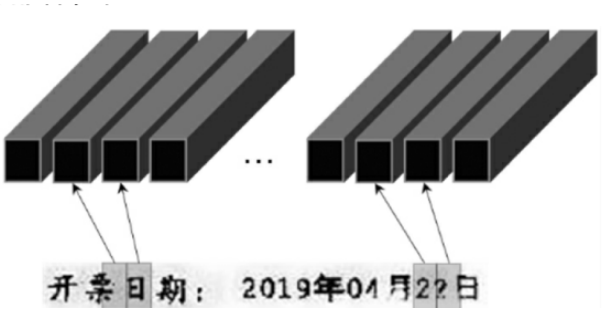


Fig 11 Correspondence between feature sequence and original image

The convolutional layer of the model adopts the classic vgg16 network structure , which sets 7 layers of convolution and pooling structures to extract features. The difference from the standard convolutional neural network is that the fully connected layer is finally removed, thereby greatly reducing training parameters .Since the convolution and pooling structures of the convolutional network have local characteristics, the final output feature sequence of the convolutional layer corresponds one-to-one with the original input image in space. FIG. 11 shows the correspondence between a part of the feature sequence and its associated original image area. The feature sequence can be regarded as an image description of the area. Expressing text images in this way allows the model to process text pictures of indefinite length and retain local feature information of text images.

### 4.2. Circular layer

The function of the recurrent layer network is to predict the feature sequence extracted by the convolutional layer. The feature sequence x = (x1, x2, ..., xt) is input into the deep bidirectional

long-term short-term memory network in a many-to-many manner. The model predicts a label probability distribution yt for each xt of the input sequence.

### 4.3. Transcription layer

The last sequence yt output by the cyclic layer is the C-dimensional vector obtained after the Softmax operation, where C represents the total number of characters to be recognized. Since the cyclic layer performs time series classification, a lot of redundant information will inevitably appear. For example, a character will be recognized twice in succession, which requires a set of mechanisms to remove redundancy, but simply see two consecutive characters and go Redundant methods also have problems. This article uses the CTC[5] decoding method to solve the above problems.

## 5. Experiment and Analysis

### 5.1. Model training

The image enhancement algorithm and text positioning experiment in this paper are based on MATLAB2016a, and the hardware environment is Intel CoreTM i5-8250U CPU @ 1.60 GHz 1.80GHz.

Since there is no fixed data set available for Chinese recognition on the Internet, this paper uses the method of generative learning to automatically generate 3.6 million grayscale images of text with variable length through the program, and the generated images are unified to 280×32 pixels. The image is selected from three sizes of small three, four, small four, five, small five and five, as well as a variety of common Chinese fonts including Song, Hei, Kai, and other common fonts. In order to enhance the robustness of the model, the generated image Rotate, translate, and increase Gaussian random noise. The training data set and the test data set are divided in a ratio of 10:1. The data label contains commonly used 6732 Chinese characters, punctuation marks, 26 English letters, 10 Arabic numerals and a "-" character. The image of part of the data set is shown in Figure 12.

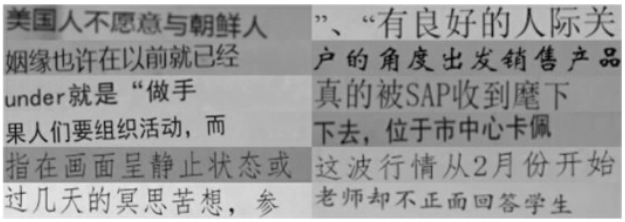


Fig 12 Partial dataset image

In order to find the optimal training parameters of the model and discuss the impact of the batch standardization layer on the model convergence and recognition rate, SGD, monmentum, Adam, and RMSprop methods were used to perform group control experiments. The batch size of the training was fixed at 128. The results are shown in Table 1.

Table 1 Training effects of different optimization methods

| Optimization methods and training parameters | Training effects of different optimization methods | | |
|---|---|---|---|
| | Whether to join the batch normalization layer | Recognition accuracy rate/% | Convergence time/min |
| SGD(α =0. 01) | Y | 95. 84 | 448 |
| | N | 92.43 | 684 |
| Monmentum | Y | 96.52 | 323 |
| | N | 94.66 | 532 |
| Adam | Y | 95.78 | 289 |
| | N | 93.24 | 503 |

| | | | |
|---|---|---|---|
| RMSprop | Y | 98.71 | 263 |
| | N | 96.43 | 427 |

The character recognition model in this paper is compared with the Chinese character recognition method of partial deep learning, and the recognition rate of each model is shown in Table 2. A document cited in this paper uses convolution neural network as the infrastructure by extracting different features of text images, such as local binary pattern (LBP), multi-scale Gabor feature, gradient direction histogram (HOG), general gradient feature and multi-scale gradient feature. When using multi-scale gradient features, the recognition rate can reach 98.36%, but it requires tedious feature extraction and dimensionality reduction preprocessing. In another document, a single character is directly input for training and prediction, and after reducing the process of feature extraction, the recognition rate can still reach 98.33%. In this paper, the text segmentation link is eliminated, and the string image is predicted directly. The text recognition accuracy is up to 98.71%. At the same time, the model can well solve the problem of text compactness and adhesion.

Table 2 Recognition rate comparison

| Method | Recognition rate/% |
|---|---|
| LBP+Convolutional Neural Network | 87.51 |
| GIST+ Convolutional Neural Network | 92.75 |
| HOG+Convolutional Neural Network | 94.50 |
| Gradient feature + convolutional neural network | 95.15 |
| Multiscale gradient feature + convolutional neural network | 98.36 |
| The algorithm proposed in references | 98.33 |
| Algorithm proposed in this paper | 98.71 |

| | | | |
|---|---|---|---|
| proposed | 3107 | 3054 | 98.29 |

Table 3 Comparison between the system proposed in this paper and open source ocr

| System name | Total characters | Recognize the number correctly | Recognition accuracy rate/% |
|---|---|---|---|
| wentong | 3107 | 2884 | 92.82 |
| hanwang | 3107 | 3008 | 96.81 |

Finally, 30 VAT invoice images with uneven illumination noise were collected, and a system comparison experiment was conducted using the system of this paper and the open source OCR text recognition software. The comparison results are shown in Table 3. The comparison results show that the system designed in this paper has the highest character recognition rate under uneven lighting.

## 6. Conclusion

In this paper, based on the difficulty of low-quality text recognition under the condition of uneven illumination, taking the VAT invoice image with uneven illumination noise as an example, an image enhancement algorithm is proposed to enhance the text details in the dark area of the image. an adaptive tone mapping function considering local characteristics is introduced to generate a high dynamic range image, and the text details are enhanced by background segmentation and compensation for contrast. A convolutional cyclic neural network is built, which avoids the step of text segmentation, and CTC is used to decode and predict character sequences. The experimental results show that the image enhancement algorithm proposed in this paper can effectively reduce the uneven illumination noise caused by shooting angle, and the character recognition model based on deep learning can solve the problem of character segmentation errors and recognition difficulties

caused by character adhesion. The application scenarios of the system include license plate recognition, RMB crown number recognition, VAT invoice recognition and so on. In the following research, we consider using deep learning to enhance the image and realize the end-to-end character recognition system.

## References

[1] Impedovo S , Ottaviano L , Occhinegro S . OPTICAL CHARACTER RECOGNITION — A SURVEY[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2011.

[2] Tumblin J , Rushmeier H . Tone Reproduction for Realistic Images[J]. IEEE Computer Graphics and Applications, 1993, 13(6):42-48.

[3] Land E H , Mccann J J . Lightness and Retinex Theory[J]. Journal of the Optical Society of America, 1971, 61(1):1-11.

[4] Sauvola J , Pietik Inen M . Adaptive document image binarization[J]. Pattern Recognition, 2000, 33(2):225-236.

[5] Alex Graves, Santiago Fernández, Faustino Gomez. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks[C]// International Conference on Machine Learning. ACM, 2006

[6]