A Method of Text Detection and Recognition from Receipt Images Based on CRAFT and CRNN

Xiaohui Wang1,*, Xi Zhang1, Shuya Lei1 and Hongmei Deng2

1 Artificial Intelligence on Electric Power System State Grid Corporation Joint Laboratory(GEIRI), Global Energy Interconnection Research Institute Co. Ltd, Beijing, China
2 State Grid Chongqing Electric Power Co. Customer Service Center, Chongqing China

* wangxiaohui@geiri.sgcc.com.cn

Abstract. Text detection and recognition from paper bank receipts image has a lot of applications in the business field, such as in power system marketing, which is applied in verification and cancellation of electricity charges and automatic archiving. In this paper, we use a text detection method to effectively detect text area in receipt image by exploring each character and affinity between characters. Then put the text sequence into text recognition model which is combined by DCNN and RNN and integrates feature extraction, sequence modeling and transcription. The experiments demonstrate the superiority of the proposed algorithm over the prior arts, which performs well in the task of blurred image-based bank receipt chinese text recognition.

1. Introduction

The electricity system marketing department has the needs of extracting the bank receipt information in the process of checking the account of the electricity charge, as well as the receipt classification and archiving. At present, it mainly relies on manpower, which is time consuming and laborious. They need the accurate and efficient text detection and recognition method of chinese charaters.

The major trend in scene text detection before the emergence of deep learning was bottom-up, such as MSER [1] or SWT [2], usually use handcrafted features. Then text detection method based on deep learning have been proposed by adopting object detection like SSD [3], Faster R-CNN [4], and CTPN [5], or applying semantic segmentation like Pixellink [6]and FCN [7]. Networks in these methods mainly are trained to localize wordlevel bounding boxes. CRAFT [8] localizes the individual character regions and links the detected characters to a text instance. Our method builds a convolutional neural network producing the character region score and affinity score based on CRAFT’s framework with preprocessed bank receipts image as input. And it’s no need to do complex post-processing to get the detected text area.

Image-based sequence recognition is a classic and challenge problem in computer vision, especially scene text recognition. Before deep learning was widely applied in the field of image recognition, pattern matching was a relatively common recognition approach. Specific methods include bayesian decision network, quadratic discriminant function (QDF), support vector machine (SVM) and so on. In recent research, word recognition work was focused on English text recognition. Wang et al. [9] proposed an approach combining a multi-layer CNN with unsupervised feature
learning to train character models, which are used in both text detection and recognition procedures. Bissacco et al. [10] detect and recognize characters with DCNN models trained using labeled character images. Recurrent neural networks (RNN) models use the sequence of image features as input instead of the position of each element in a sequence object image. CRNN was proposed as a model combined DCNN and RNN, which learned from sequence labels (words), requiring no detailed annotations (characters) [11]. Our text recognition approach is based on CRNN, which achieves highly competitive performance on receipt texts recognition than the prior arts.

2. Proposed algorithm

2.1. Preprocess of receipt image

The aim of preprocessing operations on graph is improving image quality and weakening irrelevant backgrounds like seals and table lines to ultimately improve the accuracy of text detection and recognition.

For a three-channel color picture, the seal is generally red. So we convert the original image from the RGB color gamut to the LAB color gamut. While the A channel is a red significant channel, the pixels of the seal image exist in this channel in large quantities and could be easily separated from the picture. Practically, we set the threshold to generate the seal mask based on the separation map of the seal. In the mask, the pixels in the seal area are set to 255, the pixels in other areas are set to 0. Then an or operation on the mask with the original image is performed, setting the pixels in the seal area to 255, and the pixel values of other areas in the original picture are unchanged. This operation can remove a large pixel disturbance for text detection, though remaining the text masked by the seal unable to recover.

For the table lines, we first adopt the Canny edge detection algorithm [11]. Gaussian filtering is used to get the smooth image with noise removed. Then we calculate the gradient intensity and direction of each pixel in the image to detect the horizontal, vertical and diagonal edges. And Non-Maximum Suppression is applied to eliminate the spurious response. The gradient intensity of the current pixel is compared with two pixels along the positive and negative gradient direction. If the gradient intensity of the current pixel is the largest compared with the other two pixels, the pixel point is retained as the edge point, otherwise the pixel point will be suppressed. At last, double-threshold (high threshold and low threshold) detection is applied to determine the real and potential edges. If the gradient value of edge pixel is higher than the high threshold, it is marked as a strong edge pixel. And if the gradient value of the edge pixel is less than the low threshold, it will be suppressed.

After edge information of the image is obtained, we do the hough transform [12] in which line in image space is transformed to points in parameter space. For each point in the edge image, we draw a straight line in the k-b space. We adopt the method of voting to find the slope and intercept of the straight line in origin image. In Practicing, for each straight line, if the line through a point, the vote value of the point plus one. We finally traversal k-b space and find out the local maximum value point \( (k_0, b_0) \).

2.2. Text area detection

We train a deep neural network based on CRAFT method to effectively predict character regions and the affinity between characters.

To train text detect model, we need the pixel level ground truth and character level annotation. There’s few common data sets of character level annotation, and the cost of manual annotation is huge. We constructed the required dataset by generating data and annotation randomly by computer according to the characteristics of ticket image and defined rules. Meanwhile we add text lines including Chinese, English and Numbers, on the background with light changes and some noise to diversify training set.

The sample of generated dataset is shown in Figure 1. The annotation sample is shown in Figure 2 and Figure 3. Figure 2 shows the character area, while Figure 3 shows direct link relationship between two adjacent characters. The probability of the character center is 1, while the probability of other regions
decreases according to gaussian probability distribution. These two kind of annotation images will be used to supervise training of CRAFT network.

![Generated dataset sample](image1)

**Figure 1.** Generated dataset sample. There are a number of randomly generated Chinese characters, English characters, uppercase amount and numbers in the image.

**Figure 2.** Character area annotation

**Figure 3.** Character link annotation

The network architecture is schematically illustrated in Figure 4. ConvBlock is convolution block, including convolution layer, batch normalization layer, activation layer and pooling layer. UpConvBlock is upsample convolution layer, including convolution layer, batch normalization layer, activation layer and upsample layer.

![Character area detection network architecture](image2)

**Figure 4.** Character area detection network architecture

The text detection model is a Fully Convolutional Network (FCN), which consist of convolutional neural network module and Upsample module. The CNN part concludes convolution network layer and the maximum pooling layer. After the operation of convolution kernels, the model output the text image morphological features like color and texture. Mathematically, image features are composed of feature vectors calculated by a large number of convolution kernels. These vectors are further concatenated and finally become the output of convolutional network in the form of feature map. Before upsampling, feature fusion operation between different convolution layers is needed to improve the expression of features. The upsample layer is responsible for the expansion operation of convolution feature map. Linear interpolation algorithm is used to expand the low-scale eigengraph to the higher-scale feature map, which has not much influence on the overall distribution. After four upper sample layers and four consecutive convolution layers, region score and affinity score are outputed through different branches. The region score represents the probability that the given pixel is the center of the character, and the affinity score represents the center probability of the space between adjacent characters.
The objective function is MSE loss function shown in formula (1), \( y_{region} \) and \( y_{affinity} \) are predicted region score and affinity map, while \( \tilde{y}_{region} \) and \( \tilde{y}_{affinity} \) are ground truth annotation. The model parameters are updated by Stochastic Gradient Descent (SGD) algorithm to adjust the weight of the nodes to minimize the objective function.

\[
MSE\ Loss = (y_{region} - \tilde{y}_{region})^2 + (y_{affinity} - \tilde{y}_{affinity})^2
\]  

(1)

In terms of post-processing, text envelop box could be settled by simple threshold operation and connected domain search algorithm based on the score maps instead of Non-Maximum Suppression (NMS).

2.3. Text recognition

We build the CRNN model based on CTC decoding [13] to accomplish text recognition. The model includes the convolutional layers and the recurrent layers. All the images are scaled to the same height before being fed into the CNN network. A sequence of feature vectors is extracted from the feature maps produced by the component of convolutional layers. Then the feature vectors are put into the recurrent layers, which is a deep bidirectional LSTM, capability of capturing long-range contextual information within a sequence. RNN architecture can back-propagates error differentials to its input. Moreover, it also can operate on sequences of arbitrary lengths. And LSTM is a kind of upgraded RNN without the vanishing gradient problem.

Then we convert the per-frame predictions made by RNN into a label sequence by transcription is process. We do the transcription in lexicon-free mode, which means the character of highest probability is the transcription output. The conditional probability could be described as follows:

\[
p(\pi \mid x) = \prod_{t=1}^{T} y_{\pi_t}^{\prime}, \forall \pi \in L^{\prime T}
\]

(2)

Set \( L^{\prime} = L \cup \), where \( L \) contains all labels (all the possible characters). \( \pi = (\pi_1, \pi_2, \ldots, \pi_T) \) is one of the sequences combined by predicted characters within T timestamps, which could be called one path. With \( x \) as the input sequence, \( y_{\pi_t}^{\prime} \) is the probability of having label \( \pi_t \) at time stamp \( t \). The observation probability of each time step from \( t=1 \) to \( t \) is multiplied to get the final probability of the whole path.

In practice, each correct sequence may have multiple paths corresponding to an input sequence. The final conditional probability of input \( x \) can be expressed as the sum probabilities of all the paths. \( B \) is a sequence-to-sequence mapping function, mapping \( \pi \) onto \( l \) by removing the repeated labels, and the blank. We use the negative log-likelihood of this probability as the objective to train the network.

\[
p(l \mid x) = \sum_{\pi : B(x) = l} p(\pi \mid x)
\]

(3)

In the training phase, we adopt stochastic gradient descent (SGD) algorithm with momentum to update the model parameters, and the ADADELTA algorithm to update super parameters, alleviating the problem of huge amount network parameters caused by Chinese characters.

The datasets we used for training consist of Chinese character dataset generated from Chinese corpus, receipt character dataset and datasets of English words, number, amount in words and characters with similar form. We adopt balanced sampling training mode from the five datasets above. We also use the Batch-Wise online data enhancement methods to improve the diversity of training data like Gaussian noise, pixel flips and random noise.

3. Experiment results and evaluation

In the image preprocessing stage, the effect of seal elimination and straight line removal is shown
respectively in Figure 5 and Figure 6. Figure 5 are origin receipt images from Chongqing rural commercial bank. The Chinese characters in these image includes information about date of transaction, time of transaction, name of drawee account, bank of deposit, uppercase amount, name of payee, purpose, postscript, etc.

**Figure 5.** Origin receipt image

**Figure 6.** Seal and straight lines removed

Region score and affinity score of the receipt image after CRAFT detection is shown in Figure 7 and Figure 8.

**Figure 7.** Region score

**Figure 8.** Affinity score

The result example of Chongqing rural commercial bank receipt image text detection and recognition is shown in Figure 9 and Figure 10.

**Figure 9.** Text detection result.

**Figure 10.** Text recognition result.

At present, there are mainly two different evaluation methods of text detection, PASCAL evaluation method based on IoU (Intersection over Union) and DetEval evaluation based on overlap. They have different standards when matching predicted text instances with real labels. The basic idea of the PASCAL evaluation method based on IoU is that if the IoU is greater than the specified threshold, the predicted value and the real annotation box are considered to be matched. We set the threshold to 0.5, then calculate the Precision and Recall rate. DetEval takes the one-to-one, one-to-many, many-to-one case into consideration and it defines the precision and recall of the detection box
based on intersection area. The predicted and the true value are considered to match only when precision and recall are larger than their respective thresholds, which is set as 0.4 and 0.8.

We also use the F1-score named as $H_{\text{mean}}$ to balance the precision and recall rate.

\[
\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} \tag{4}
\]

\[
\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \tag{5}
\]

\[
H_{\text{mean}} = \frac{2 \times \text{precision} \times \text{recall}}{(\text{precision} + \text{recall})} \tag{6}
\]

As for text recognition, edit distance and normalized edit distance are often used for evaluation.

\[
\text{NormalizedEditDist} = 1 - \frac{\sum_{g} \text{EditDist}(P_g, g)}{\sum_{g} \text{Length}(g)} \tag{7}
\]

$G$ is the set of all real labels, $g$ is one of the real labels, $P_g$ is the predicted value corresponding to $g$, $\text{EditDist}(P_g, g)$ is the edit distance between $P_g$ and $g$. $\text{Length}(g)$ is the number of characters in $g$.

In terms of text detection, we compare our method with SWT and CTPN. SWT is a manual feature method and CTPN is a regressive deep learning method. Our CRAFT-based text detection method has significant performance advantages. The accuracy and recall rate of our method and other mainstream text detection methods are above 90% and there is no significant difference for receipt images with clear ink and no complex noise background from banks like Bank of China. In terms of the extremely fuzzy scanned receipts images from some banks, such as chongqing rural commercial bank, our results show strong robustness and availability. The following indicators in Table 1 and Table 2 are compared based on the detection and recognition results of blurred images.

| Table 1. Evaluation of text detection methods |
|-----------------------------------------------|
| method | Recall | Precision | Hmean | Recall | Precision | Hmean |
|-------|--------|-----------|------|--------|-----------|------|
| SWT   | 0.0409 | 0.0336    | 0.0369 | 0.4982 | 0.3376    | 0.4025 |
| CTPN  | 0.2033 | 0.2578    | 0.2273 | 0.2167 | 0.3636    | 0.2715 |
| CRAFT | 0.4794 | 0.6106    | 0.5371 | 0.5229 | 0.5383    | 0.5305 |

In overall scheme for text detection and recognition of the receipt, we compare our system with other Open source or commercial trial tools. The result is shown in Table 1. Our method is the best in recognition and the best in detection in most cases. Baidu OCR shows better score in DetEval because its detection box is larger and its pixel recall rate is more likely to exceed the set threshold.

| Table 2. Performance comparison of different text detection and recognition system |
|-----------------------------------------------|
| method | Recall | Precision | Hmean | Recall | Precision | Hmean | Average | Normalized |
|--------|--------|-----------|------|--------|-----------|------|---------|-----------|
| DetEval |       |           |      | IoU    |           |      |         |           |
| Tesseract | 0.1820 | 0.1791    | 0.1806 | 0.1216 | 0.1748    | 0.1434 | 413.25  | 43.17%    |
| ABBYY  | 0.2003 | 0.2865    | 0.2357 | 0.2023 | 0.3746    | 0.2628 | 306.61  | 29.77%    |
| Baidu  | 0.4945 | 0.5094    | 0.5401 | 0.3526 | 0.5572    | 0.4319 | 274.18  | 39.73%    |
| Huawei | 0.2172 | 0.1081    | 0.1444 | 0.3851 | 0.2489    | 0.3024 | 347.79  | 43.17%    |
| Tencent | 0.4042 | 0.4262    | 0.4149 | 0.4247 | 0.5324    | 0.4724 | 266.82  | 50.12%    |
| Alibaba | 0.2201 | 0.2526    | 0.2352 | 0.5036 | 0.6093    | 0.5515 | 239.31  | 53.10%    |
| Ours   | 0.4794 | 0.6106    | 0.5371 | 0.5409 | 0.7298    | 0.6213 | 213.14  | 53.86%    |

4. Conclusion
In this paper, we have presented a text detection and recognition method based on CRAFT and Convolutional Recurrent Neural Network (CRNN), which is applied in receipt image especially for
blurred one. The text detection part calculates the character region score and the character affinity score which combined to effectively locate the text area, requiring no complicated annotations for each individual element and any further post-processing methods like NMS.

Then we train the detection model with image-based sequence recognition model based on CRNN in an end-to-end fashion, which is combined by convolutional layer, recurrent layer and CTC transcription layer, achieving superior or highly competitive performance. To further improve the robustness, and generalizability of the detection and recognition method we proposed is the next step of our work in the future.

Acknowledgments
This research is supported by research project of State Grid Corporation of China Application Research of Artificial Intelligence Technology in Power Information Collection (5455HJ180014).

References
[1] Matas, J., Chum, O., Urban, M., & Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. Image and vision computing, 22(10), 761-767.
[2] Epshtein, B., Ofek, E., & Wexler, Y. (2010, June). Detecting text in natural scenes with stroke width transform. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (pp. 2963-2970). IEEE.
[3] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.
[4] Ren, S., Girshick, R., Girshick, R., & Sun, J. (2017). Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis & Machine Intelligence, 39(6), 1137-1149.
[5] Tian, Z., Huang, W., He, T., He, P., & Qiao, Y. (2016, October). Detecting text in natural image with connectionist text proposal network. In European conference on computer vision (pp. 56-72). Springer, Cham.
[6] Deng, D., Liu, H., Li, X., & Cai, D. (2018, April). Pixellink: Detecting scene text via instance segmentation. In Thirty-Second AAAI Conference on Artificial Intelligence.
[7] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).
[8] Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H. (2019). Character Region Awareness for Text Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 9365-9374).
[9] Wang, T., Wu, D. J., Coates, A., & Ng, A. Y. (2012, November). End-to-end text recognition with convolutional neural networks. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012) (pp. 3304-3308). IEEE.
[10] Bissacco, A., Cummins, M., Netzer, Y., & Neven, H. (2013). Photoocr: Reading text in uncontrolled conditions. In Proceedings of the IEEE International Conference on Computer Vision (pp. 785-792).
[11] Canny, J. (1986). A computational approach to edge detection. IEEE Transactions on pattern analysis and machine intelligence, (6), 679-698.
[12] Duda, R. O., & Hart, P. E. (1971). Use of the Hough transformation to detect lines and curves in pictures (No. SRI-TN-36). Sri International Menlo Park Ca Artificial Intelligence Center.
[13] Shi, B., Bai, X., & Yao, C. (2016). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE transactions on pattern analysis and machine intelligence, 39(11), 2298-2304.