Novel receipt recognition with deep learning algorithms

Xie, Dong, Bailey, Colleen

Dong Xie, Colleen P. Bailey, "Novel receipt recognition with deep learning algorithms," Proc. SPIE 11400, Pattern Recognition and Tracking XXXI, 114000B (22 April 2020); doi: 10.1117/12.2558206

SPIE.

Event: SPIE Defense + Commercial Sensing, 2020, Online Only, California, United States

Novel Receipt Recognition with Deep Learning Algorithms

Dong Xie and Colleen P. Bailey

Department of Electrical Engineering, University of North Texas, Denton, TX, USA 76207

ABSTRACT

We propose a new recognition method to extract effective information from receipts by integrating deep learning algorithms from computer vision and natural language processing. Our method consists of three parts. The first part provides effective areas for receipt detection. By removing noise and extracting the gradient of the receipt image, we determine the threshold to crop and reshape the useful receipt area. Detecting text from a receipt image is the second part, we modify and deploy the text detection algorithm connectionist text proposal network (CTPN) to locate the text region in the receipt. In the third part, we import the connectionist temporal classification with maximum entropy regularization as the loss function for updating the convolutional recurrent neural networks (CRNN) to recognize the text detection area, which converts the receipt from an image into the text. Based on our method, the effective information of a receipt can be integrated and utilized. We train and test our system using the data set published by scanned receipts optical character recognition and information extraction (SROIE). The results illustrate that our recognition system is able to identify receipt information quickly and accurately.

Keywords: deep learning, receipt recognition, text detection, text recognition, optical character recognition.

1. INTRODUCTION

With the revolution of artificial intelligence, optical character recognition (OCR) technology is gradually applied to people's daily life.[1] The main purpose of the OCR system is to convert an image into computer characters, which can reduce the storage of image data and easily reuse the recognized characters. In the process of conversion from image to text, two important steps are included: text detection and text recognition. Based on the development of deep learning and the breakthrough of computer chips, many solutions have been proposed, such as non-maximum suppression (NMS), region proposal network (RPN), and connectionist temporal classification (CTC), which create a solid foundation for the improvement of text detection and text recognition.[2–4]

Recently, in the field of image object detection, various algorithms have been introduced with satisfying performance, such as single shot multibox detector (SSD), Faster R-CNN, and Mask R-CNN.[4–6] However, these algorithms do not achieve optimal performance on the text detection task. Therefore, many text detection algorithms work to promote the adoption of image object detection methods. The connectionist text proposal network (CTPN) algorithm, one of the most widely used text detection models, improves the detection accuracy by modifying Faster R-CNN with the convolutional neural network (CNN) and bidirectional long short term memory (BLSTM) model to extract the context features of the image characters.[7] By introducing the rotation region-to-interest method, the rotation region proposal networks (RRPN) provide a new idea for the detection of skewed text.[8] Furthermore, fused text segmentation networks (FTSN) introduce a mask-NMS based on the pixel-level coincidence degree between text instances and have a better detection effect on skewed text.[9] Deep matching prior network (DMPNet) uses a quadrilateral instead of a rectangle to draw the text area boundary for curved text.[10] Moreover, an efficient and accurate scene text detector (EAST) uses a lightweight neural network architecture (PVANET) to extract image features, which speeds up the process of text detection.[11] Based on the classical image detection algorithm SSD, Textboxes, segment linking (SegLink) and Textboxes ++ have been proposed and their produced characters in different scenarios can be detected with high accuracy.[12–14]

Further author information: (Send correspondence to Colleen P. Bailey)
Dong Xie: E-mail: DongXie@my.unt.edu
Colleen P. Bailey: E-mail: Colleen.Bailey@unt.edu

Pattern Recognition and Tracking XXXI, edited by Mohammad S. Alam,
Proc. of SPIE Vol. 11400, 114000B · © 2020 SPIE · CCC
code: 0277-786X/20/$21 · doi: 10.1117/12.2558206

Text recognition with deep neural networks is considered one of the most important branches in the OCR field. The convolutional recurrent neural network (CRNN) is one of the most popular text recognition models at present. Through the organic fusion of CNN, BLSTM, and CTC, the longer text sequence can be well-recognized.[15] Extracting sequence features through the dense convolutional network (DenseNet) algorithm can improve the processing speed of text recognition.[16,17] Because of using the spatial transformer network and the sequence recognition network, a robust text recognizer with automatic rectification (RARE) model can correct the curved text area and directly capture detected text.[18] Moreover, by introducing the abbreviation of focusing attention network (FAN) method, text recognition problems such as resolution and text spacing can be well-handled.[19] The arbitrary orientation network (AON) can better identify the irregular and curved text in any direction by adding the vertical text direction processing BLSTM network with horizontal text processing BLSTM.[20]

Furthermore, many operative end-to-end models can locate and recognize text directly from images. Fast oriented text spotting (FOTS) is configured with a unified model of detection and recognition tasks, which share the features from the convolutional network. The design of FOTS not only saves computing time, but also obtains more image features than the two separate stage training methods.[21] An end-to-end semi-supervised model spatial transform network (STN) is developed in the STN-OCR system to correct the distorted text and improve recognition accuracy.[22] An attentional scene text recognizer with flexible rectification (ASTER) algorithm combines a flexible thin-plate spline transformation rectification network with a sequence-to-sequence recognition model to speed up the recognition of irregular text.[23] Mask TextSpotter improves the Mask RCNN method to obtain the accurate result of text detection and recognition through semantic segmentation.[24]

Inspired by the above research, we propose an end-to-end novel receipt recognition system for capturing effective information from receipts (CEIR). The main contributions of this paper are divided into three parts. First, we develop a preprocessing method for receipt images. By converting the image to grayscale and obtaining the gradient with the Sobel operator, the outline of the receipt area is decided by morphological transformations with the elliptic kernel. Second, we introduce the modified connectionist text proposal network to execute text detection. Third, we combine the convolutional recurrent neural network and the connectionist temporal classification with maximum entropy regularization as a loss function to update the weights in networks and extract the characters from a receipt.[25] We validate our system with the scanned receipts optical character recognition and information extraction (SROIE) database.[26] Compared with another deep learning approach for receipt recognition (DLARR), our method maintains a high accuracy rate.[16] Furthermore, our system has strong robustness and can be extended to a variety of different scenarios beyond receipts.

The rest of this paper is organized as follows. Section 2 introduces the necessary background for our method. In section 3, we provide the structure of our receipt recognition approach CEIR. Section 4 presents the design and evaluation indices of the experiment. Experimental results and analysis are provided to demonstrate the efficacy of our proposed system. Finally, section 5 gives the conclusion of our work.

2. BACKGROUND

The novel CEIR system is based on the connectionist text proposal network and the convolutional recurrent neural network algorithms.

2.1 Connectionist Text Proposal Network

As one of the most popular text detection algorithms, the connectionist text proposal network (CTPN) consists of four parts: a convolutional neural network architecture (VGG16), the bidirectional long short term memory, a fully-connected convolutional neural network, and a modified region proposal network. The four parts organically connected can effectively detect the horizontal distribution of text in various scenarios.

Assuming the number of images input to the CTPN structure is N, after the $conv5\_3$ of the VGG16 processing, the feature map $\mathcal{F}$ with $N \times C \times H \times W$ is obtained where C, H, and W are channel, height, and width respectively. A convolution layer transforms the size of the feature map $\mathcal{F}$ to $N \times 9 \cdot C \times H \times W$. With BLSTM structure processing, the size of the feature map $\mathcal{F}$ is changed to $N \times 256 \times H \times W$. Through another convolution layer, the size of feature map $\mathcal{F}$ is extended to $N \times 512 \times H \times W$. The feature map $\mathcal{F}$ then enters the modified RPN

to obtain text proposals. Finally, a text detector is applied to filter the text proposals with NMS and acquire the supreme detected region.

2.2 Convolutional Recurrent Neural Network

The convolutional recurrent neural network (CRNN) is proposed for recognizing the characters from the image. The CRNN architecture consists of three parts: the convolutional layers, the recurrent layers, and the transcription layer.

Before processing the network, CRNN requires reshaping all images to the same height. Suppose an image with size $(H = 32, W = 100, C = 3)$ is input to CRNN. A standard CNN model with convolutional and max-pooling layers is charged with extracting the feature maps and a convolutional feature matrix with size $(1, 25, 512)$ as the output. According to the matrix, the feature sequence generated is $x = x_1, ..., x_T$, where $T = 25$. The recurrent layers have a deep BLSTM to output 25 characters. In order to obtain the predicted sequence, the transcription layer needs to connect the characters and remove redundancies.

Moreover, the CRNN implements the connectionist temporal classification (CTC) method to compute loss instead of using softmax. One of the most important features of CTC is to solve the align problem. CTC can produce blank characters when there are no characters in some positions under certain circumstances. The structure of the CTC calculation guarantees the gradient can be measured quickly.

3. METHODOLOGY

In this section, we introduce the method of our receipt recognition system for capturing effective information from receipts (CEIR). The CEIR has three steps: preprocessing, text detection, and text recognition. The preprocessing crops the actual receipt text area. The text bounding boxes in the receipt are captured in the text detection part. The text recognition translates the image area to words. Figure 1 shows the CEIR structure, for an example receipt.

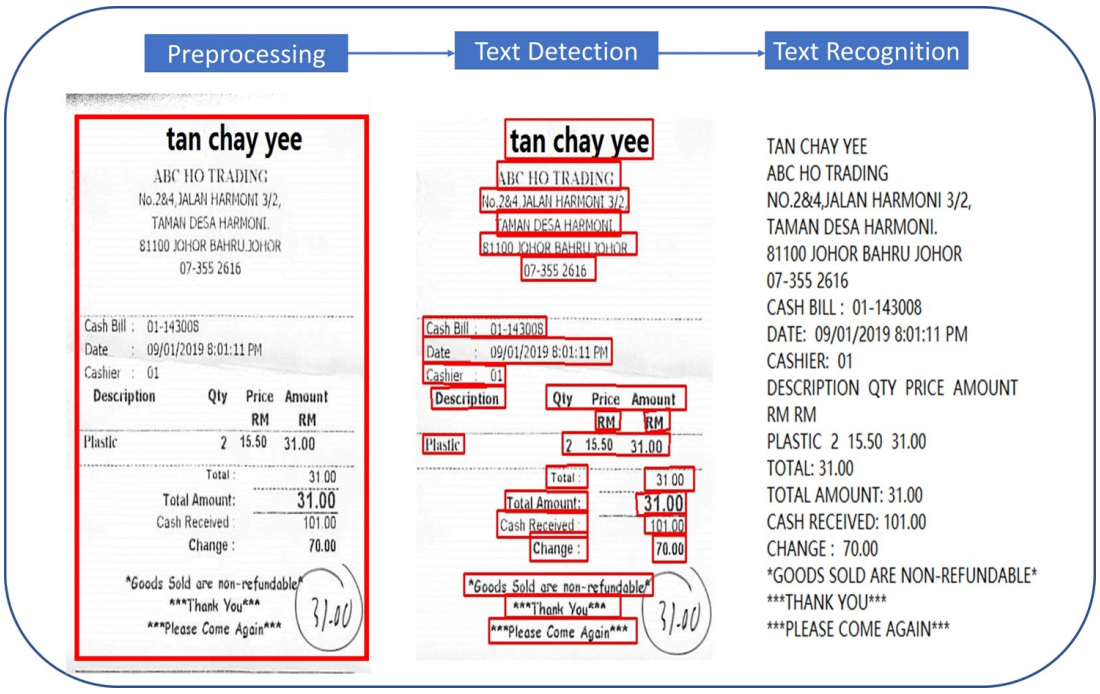

Figure 1. The structure of CEIR.

3.1 Preprocessing

Due to the scanning background or the shape of the image, the size of a receipt background can be extremely large, while the actual receipt area takes up only a small portion of the whole image. Without preprocessing, the text detection will ignore some key features from the receipt and the desired result is not satisfied. In order to provide a better training set for the subsequent text detection, we use classical computer vision methods in CEIR to process the image and crop the effective receipt content.

There are three main steps in preprocessing. After inputting the scanned image, CEIR obtains the gradient with the Sobel operator. Then, depending on the image size and the line space, the elliptic kernel is determined to take the erosion and dilation of morphological transformations and draw the outline of the receipt area. The contour is presented by finding the points of the bounding box and the final actual receipt area is cropped. Figure 2 depicts the preprocessing steps with a receipt image example.

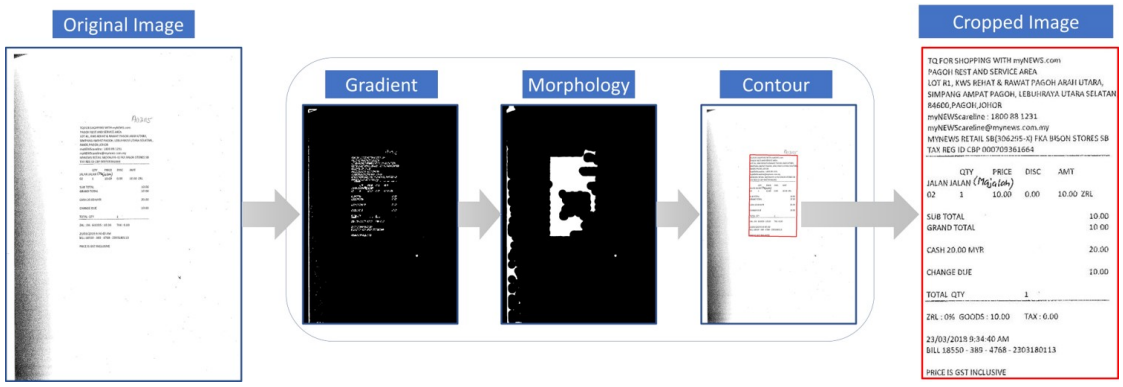

Figure 2. The structure of preprocessing.

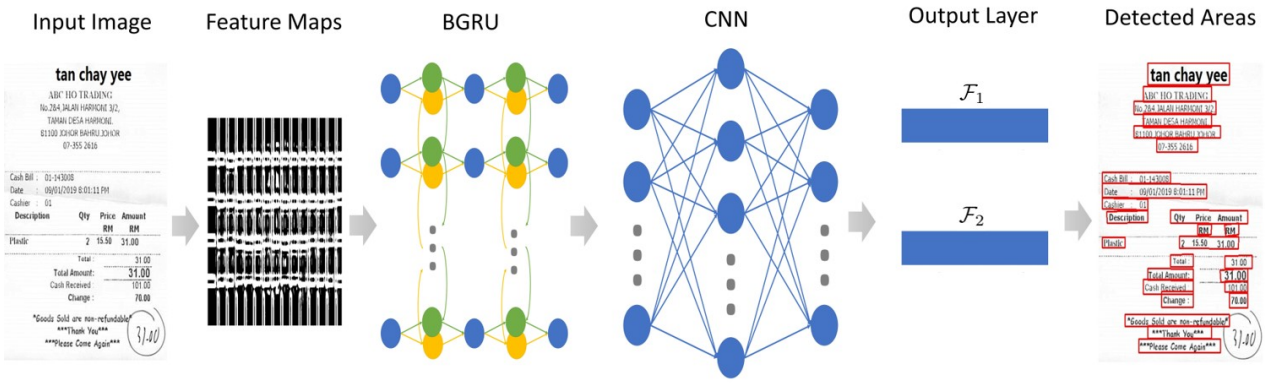

Figure 3. The structure of text detection.

3.2 Text Detection

Our text detection method is based on the CTPN algorithm with a change to the model structure for training the receipt dataset. Our text detection procedure is drawn in Figure 3. The preprocessed image is input to the base net VGG16 to extract the feature map after *conv*5_3 processing. A 3×3 kernel size convolutional network is used to fuse surrounding information and obtain the feature maps. The new feature maps are connected to a bidirectional gate recurrent unit (BGRU) layer with 512 input channels and 128 hidden units. After that, two different basic convolutional structures with the same kernel size 1 are used to generate two new feature maps $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively. The text proposals are calculated with $\mathcal{F}_1$ and $\mathcal{F}_2$ is used to produce scores. Moreover,

a standard non-maximum suppression algorithm is used to filter out the unqualified text proposals. Finally, a text proposal connector is used to merge the obtained text segments into a bounding box.

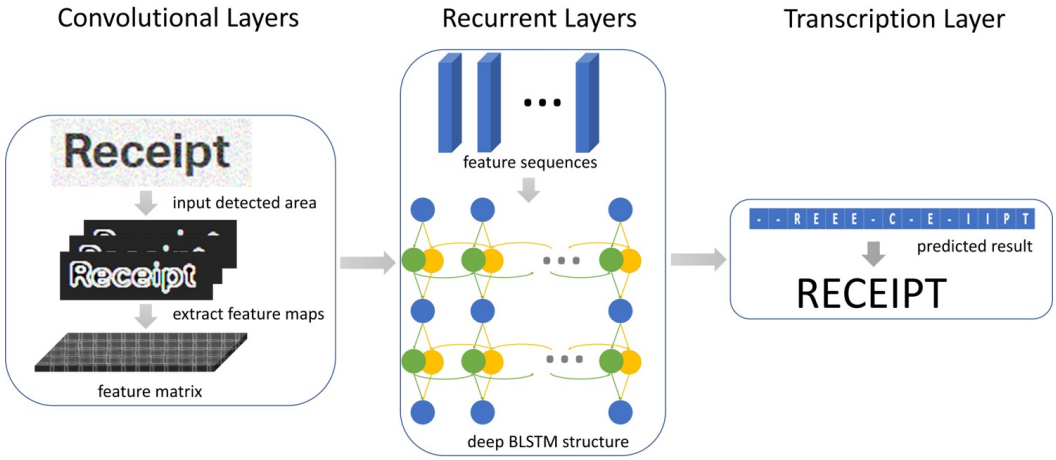

Figure 4. The structure of text recognition.

3.3 Text Recognition

We modify the CRNN model to develop our text recognition structure, as pictured in Figure 4. Our CEIR can recognize 70 characters including letters, numeric digits, special characters and space. The text bounding box image fixed with 32 height first inputs into the convolutional structure with 7 convolutional layers and 4 max pooling layers to get the feature sequences. Then, the feature sequences connect to a deep BLSTM struture with 2 BLSTM layers and 256 hidden nodes. The deep BLSTM ensures the context information is sufficiently used for sequence prediction. In the CRNN model, the weights from convolutional and recurrent layers are updated by a loss function L_{ctc} with equation (1).

$$L_{ctc} = -\log p(l|X_{1:T}), \tag{1}$$

where $p(l|X_{1:T})$ is the conditional probability of target sequence l with an input sequence X from 1 to T.

In our CEIR, we adopt a new loss function $L_{enesctc}$, called connectionist temporal classification with maximum entropy regularization (ENESCTC), to update our network weights. With the process of ENESCTC, the overconfident problem in distributions can be eliminated in our predicted sequence.[25] $L_{enesctc}$ is formulated by the following equation,

$$L_{enesctc} = -\log p_\tau(l|X) - \beta H(p_\tau(\pi|l, X)), \tag{2}$$

where $p_\tau(l|X)$ is the conditional probability with length limit $\tau \in [1, l]$ per segment. $H(p_\tau(\pi|l, X))$ is the entropy of feasible paths of $p_\tau(\pi|l, X)$ for a predicted path π, for example $--REEE-C-E-IIPT$ in Figure 4, and β is a weight to control the entropy. With this loss function, our system CEIR can identify the words in the receipt more accurately.

4. EXPERIMENT

We use the dataset from ICDAR 2019 robust reading challenge on scanned receipts OCR and information extraction (SROIE) to train and test our system.[26] Following previous researchers' work, we randomly pick 500 receipts for training, 63 receipts for validating, and 63 receipts for testing.[16] The text detection model and text recognition model are trained with 2000 episodes using GPU NVIDIA 2070 super. CEIR code and results have been made available at: https://github.com/eadst/CEIR. The evaluation index introduced below is used to examine the accuracy of the result. A comparison with another algorithm is shown in result analysis.

4.1 Evaluation Index

In text detection, call, precision, and harmonic mean are considered metrics to evaluate our experimental results. To have the objectivity of evaluation, the following criteria is employed as the protocol to qualify compatibility. Based on equations (3) and (4), the detected area $A(D)$ is judged by whether or not it matches the ground truth area $A(G)$.

$$Cmatch = \begin{cases} 1, & \frac{A(G \cap D)}{A(G)} > TC \\ 0, & otherwise \end{cases} \tag{3}$$

$$Pmatch = \begin{cases} 1, & \frac{A(G \cap D)}{A(D)} > TP \\ 0, & otherwise \end{cases} \tag{4}$$

where $A(G \cap D)$ is the overlapping area of detected and ground truth areas. TC and TP are the thresholds of call and precision. If $\frac{A(G \cap D)}{A(G)}$ is more than the call threshold, the call match, $Cmatch$, of this detected area will be marked as positive with 1, otherwise it will be 0. Precision match, $Pmatch$, is calculated in the same manner as the call match. The recall, precision, and harmonic mean are presented in the following equations.

$$Recall = \frac{\sum Cmatch}{N_G}, \tag{5}$$

$$Precision = \frac{\sum Pmatch}{N_D}, \tag{6}$$

$$Hmean = 2 \times \frac{Recall \times Precision}{Recall + Precision}. \tag{7}$$

In equation (5), $\sum Cmatch$ is the number of positive detected areas filtered by TC and N_G is the number of ground truth areas. In equation (6), $\sum Pmatch$ is the number of positive detected areas filtered by TP and N_D is the total number of detected areas. In equation (7), $Hmean$ is the harmonic mean of $Recall$ and $Precision$. Supposing we have 100 ground truth areas, if we obtain 200 detected areas where 70 are positive with $TC = 0.8$, the recall is $\frac{70}{100} = 70\%$. If TP is equal to 0.6, the positive detected areas may increase to 90, but the precision is low with $\frac{90}{200} = 45\%$. Therefore, the $Hmean$ is $2 \times \frac{70\% \times 45\%}{70\% + 45\%} = 54.8\%$.

For text recognition, three different evaluation indices are introduced to analyze and evaluate the accuracy of our method. Index recall is the fraction of the positively recognized words over the number of ground truth words. Index precision is the fraction of the positive recognized words over the number of all recognized words. Moreover, the index $F1$ score is defined as the harmonic mean of recall and precision. The evaluation indices are formulated into the following three equations,

$$Recall = \frac{positive\ recognized\ words}{ground\ truth\ words}, \tag{8}$$

$$Precision = \frac{positive\ recognized\ words}{recognized\ words}, \tag{9}$$

$$F1 = 2 \times \frac{Recall \times Precision}{Recall + Precision}. \tag{10}$$

We illustrate the above three equations with an example. Assume we have 1000 ground truth words and our system produces 1200 recognized words, where 900 recognized words are a match with the ground truth set. In other words, the number of positive recognized words is 900. Thus, the recall can be calculated as $Recall = \frac{900}{1000} = 90\%$, and the precision is equal to $\frac{900}{1200} = 75\%$. Moreover, we can obtain $F1 = 2 \times \frac{90\% \times 75\%}{90\% + 75\%} = 81.8\%$.

4.2 Result Analysis

In this section, we present our system (CEIR) result and compare it with a previous researchers' result, a deep learning approach for receipt recognition (DLARR).[16] The results from text detection and text recognition are compared separately. We use the introduced evaluation indices to determine the accuracy of both methods.

The text detection results are provided in Tab. 1. Our CEIR obtains 81.1% in recall and DLARR achieves only 53.9%. CEIR reaches 92.1% in the precision calculation which is better than DLARR precision performance of 77.5%. Therefore, relying on the high accuracy in the recall and precision parts, our CEIR achieves a better harmonic mean of 86.3%, while DLARR only has 63.6%.

Table 1. The result of text detection.

Method	Recall	Precision	Hmean
DLARR	53.9%	77.5%	63.6%
CEIR	81.1%	92.1%	86.3%

In text recognition, the total number of ground truth words is 72357. While 72201 words are recognized in CEIR and 70172 are matched with ground truth words and marked as positive recognized words. Thus, in Tab. 2, we obtain 97.0% and 97.2% in recall and precision, respectively. Meanwhile, DLARR gets 87.6% in the recall and 84.7% in the precision. Moreover, based on the recall and precision data, CEIR achieves 98.2% in F1, while DLARR is only 86.1%. Consequently, our CEIR has better performance in text detection and recognition processing.

Table 2. The result of text recognition.

Method	Recall	Precision	F1
DLARR	87.6%	84.7%	86.1%
CEIR	97.0%	97.2%	98.2%

5. CONCLUSION

A novel receipt recognition system CEIR is proposed in this paper to extract useful data from receipt. CEIR is formed by classical computer vision preprocessing, CTPN text detection with the bidirectional gate recurrent unit layer, and CRNN with maximum entropy text recognition. These three steps cooperate and form the optimal model to capture key receipt information. Under the given evaluation criterion, the experimental results prove that our CEIR system is able to acquire receipt information with high accuracy.

REFERENCES

[1] Mori, S., Nishida, H., and Yamada, H., [*Optical character recognition*], John Wiley & Sons, Inc. (1999).

[2] Neubeck, A. and Van Gool, L., "Efficient non-maximum suppression," *Proc. ICPR* **3**, 850–855, IEEE (2006).

[3] Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J., "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," *Proc. ICML*, 369–376 (2006).

[4] Ren, S., He, K., Girshick, R., and Sun, J., "Faster R-CNN: Towards real-time object detection with region proposal networks," *Proc. NIPS*, 91–99 (2015).

[5] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C., "SSD: Single shot multibox detector," *Proc. ECCV*, 21–37, Springer (2016).

[6] He, K., Gkioxari, G., Dollár, P., and Girshick, R., "Mask R-CNN," *Proc. ICCV*, 2961–2969 (2017).

[7] Tian, Z., Huang, W., He, T., He, P., and Qiao, Y., "Detecting text in natural image with connectionist text proposal network," *Proc. ECCV*, 56–72, Springer (2016).

[8] Ma, J., Shao, W., Ye, H., Wang, L., Wang, H., Zheng, Y., and Xue, X., "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Transactions on Multimedia* **20**(11), 3111–3122 (2018).

[9] Dai, Y., Huang, Z., Gao, Y., Xu, Y., Chen, K., Guo, J., and Qiu, W., "Fused text segmentation networks for multi-oriented scene text detection," *Proc. ICPR*, 3604–3609, IEEE (2018).

[10] Liu, Y. and Jin, L., "Deep matching prior network: Toward tighter multi-oriented text detection," *Proc. ICCV*, 1962–1969 (2017).

[11] Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., and Liang, J., "East: An efficient and accurate scene text detector," *Proc. ICCV*, 5551–5560 (2017).

[12] Liao, M., Shi, B., Bai, X., Wang, X., and Liu, W., "Textboxes: A fast text detector with a single deep neural network," *Proc. AAAI* (2017).

[13] Shi, B., Bai, X., and Belongie, S., "Detecting oriented text in natural images by linking segments," *Proc. ICCV*, 2550–2558 (2017).

[14] Liao, M., Shi, B., and Bai, X., "Textboxes++: A single-shot oriented scene text detector," *IEEE transactions on image processing* **27**(8), 3676–3690 (2018).

[15] Shi, B., Bai, X., and Yao, C., "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 (2016).

[16] Le, A. D., Van Pham, D., and Nguyen, T. A., "Deep learning approach for receipt recognition," *Proc. FDSE*, 705–712, Springer (2019).

[17] Tang, Z., Jiang, W., Zhang, Z., Zhao, M., Zhang, L., and Wang, M., "Densenet with up-sampling block for recognizing texts in images," *Neural Computing and Applications* , 1–9 (2019).

[18] Shi, B., Wang, X., Lyu, P., Yao, C., and Bai, X., "Robust scene text recognition with automatic rectification," *Proc. CVPR*, 4168–4176 (2016).

[19] Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S., and Zhou, S., "Focusing attention: Towards accurate text recognition in natural images," *Proc. ICCV*, 5076–5084 (2017).

[20] Cheng, Z., Xu, Y., Bai, F., Niu, Y., Pu, S., and Zhou, S., "Aon: Towards arbitrarily-oriented text recognition," *Proc. CVPR*, 5571–5579 (2018).

[21] Liu, X., Liang, D., Yan, S., Chen, D., Qiao, Y., and Yan, J., "Fots: Fast oriented text spotting with a unified network," *Proc. CVPR*, 5676–5685 (2018).

[22] Bartz, C., Yang, H., and Meinel, C., "STN-OCR: A single neural network for text detection and text recognition," *arXiv preprint arXiv:1707.08831* (2017).

[23] Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., and Bai, X., "Aster: An attentional scene text recognizer with flexible rectification," *IEEE transactions on pattern analysis and machine intelligence* **41**(9), 2035–2048 (2018).

[24] Lyu, P., Liao, M., Yao, C., Wu, W., and Bai, X., "Mask TextSpotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," *Proc. ECCV*, 67–83 (2018).

[25] Liu, H., Jin, S., and Zhang, C., "Connectionist temporal classification with maximum entropy regularization," *Proc. NIPS*, 831–841 (2018).

[26] "ICDAR 2019 Robust Reading Challenge on Scanned Receipts OCR and Information Extraction." https://rrc.cvc.uab.es/?ch=13 (2019).