2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Screen-rendered text images recognition using a deep residual network based segmentation-free method

Xin Xu[1, 2, *], Jun Zhou[1], and Hong Zhang[1, 2]
[1]School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, China, 430065.
[2]Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan, China, 430065.
[*]Corresponding author: xuxin0336@163.com

***Abstract*—Text images recognition has long been known as a research hotspot of computer vision. However, screen-rendered text image pose great challenges to current character or text recognition methods due to its low resolution and low signal-to-noise ratio properties. In this paper, a segmentation-free method utilizing Residual Network (ResNet) and Recurrent Neural Network (RNN)-Connectionist Temporal Classification (CTC) is proposed to recognize Chinese and English texts in screen-rendered images. Text lines are firstly extracted from screen-rendered images to obtain feature sequences. Then, a bidirectional RNN layer is applied to model the contextual information within feature sequences and predict identification results. Finally, a CTC method is employed to calculate loss and yield the results. The proposed method can achieve the best performance on ORAND-CAR-A dataset, ORAND-CAR-B dataset and a generated dataset with the recognition accuracy of 91.89%, 93.79% and 95.67%, respectively. Moreover, experiments on several real screen-rendered text images also demonstrate the effectiveness of the proposed method.**

Keywords—text recognition; screen-rendered image; residual network; recurrent neural network; connectionist temporal classification

I. INTRODUCTION

Recent years, with the thriving development of Deep Convolution Neural Networks, Optical Character Recognition (OCR) has achieved great improvement. As shown in Fig. 1, current character or text recognition methods are mainly focused on handwritten Chinese character recognition (HCCR) [1-4] and scene text recognition [6-12]. Another kind of methods aims to recognize characters from screen-rendered images [13, 14]. However, current methods face great challenges towards Chinese text recognition from screen-rendered images. The reason may lie in two folds: (1) the shapes of Chinese characters are complicated with huge character library; (2) screen-rendered images are featured with low resolution and low signal to noise ratio properties [15-17].

OCR methods can be categorized into segmentation-based recognition method [1-3, 6, 13,14] and segmentation-free recognition method [5, 9, 18]. In segmentation-based recognition methods, each text line in the input image is firstly segmented into a number of independent characters, and then they are classified by means of machine learning algorithms. Finally, recognition results are connected to obtain the final result. Wachenfeld *et al.* [13] first proposed a method for English character recognition of screen rendering images. They used European distance to classify the characters. Rashid *et al.* [14] constructed a classifier using Hidden Markov Models to recognize English text from screen-rendered images. The PhotoOCR [6] system proposed a seven-layer neural network to train character classifier based on the HOG feature of the character image. However, text recognition accuracy of the segmentation-based method depends greatly on character segmentation results.

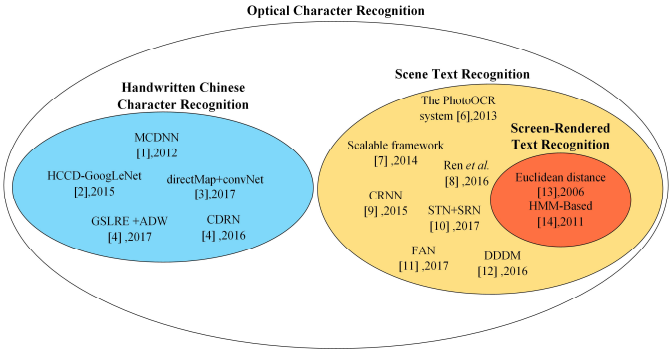


Fig. 1. Overview of the state-of-the-art OCR schemes.

Segmentation-free method avoids character segmenting since it designs a model by establishing a direct mapping from inputs to outputs, and hence attracting extensive attention. Recurrent Neural Networks (RNN)-Connectionist Temporal Classification (CTC) [19] is a basic framework for sequence recognition task. Shi *et al.* [9] put forward a new framework of CRNN by combining CNN with RNN and applied it to English word recognition in natural scenes. Sun et al. [5] applied CRNN to handwritten English text recognition and combined outputs of CNN and RNN, which can facilitate network converge faster. Based on CRNN, Zhan et al. [18] combined ResNet [21] with the RNN-CTC framework to achieve excellent results in handwritten digit sequences.

It is noteworthy that current research focus mainly pays attention to English and digit character or text recognition; few works focus on Chinese character or text recognition from screen-rendered images. On the one hand, it is because Chinese characters are more complex than English in both number and types of strokes; fortunately, there are some literatures that

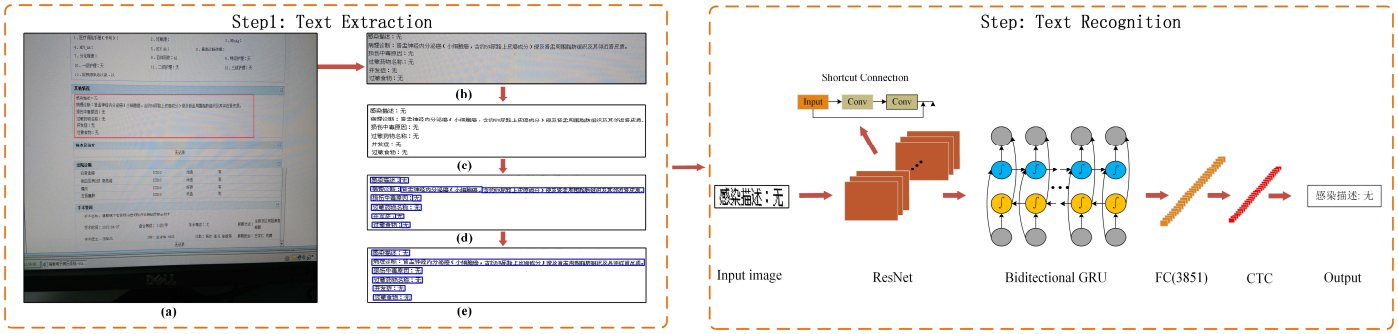


Fig. 2. Overview of the proposed method.

made efforts for HCCR. For example, Zhong et al. [2] designed a streamlined GoogLeNet and combined the Gabor and gradient feature maps with original images as the input of network. Zhang et al. [3] utilized the character shape normalization and direction-decomposed feature maps to enhance the performance of CNN in HCCR. Although the above methods [2-3] have achieved remarkable results in HCCR, they required characters to be segmented before recognition so that they cannot be directly applied to the recognition of text lines. On the other hand, the low resolution and low signal to noise ratio properties of screen-rendered text images may hinder the extraction and recognition processes.

In our previous work [22], an inception deep learning architecture was presented for screen-rendered Chinese-English character recognition. Differently, this method uses vertical projection and word-width fusion method for character segmentation and constructs an inception_v2 module [23] based CNN for character classification. However, the performance of this method is highly depended on the result of character segmentation. In this paper, we further propose a segmentation-free Chinese-English text recognition method for screen-rendered images. In the proposed method, text lines are firstly detected via the OTSU and connected component detection algorithm, and the detected text line is inputted to a ResNet directly to extract discriminative feature sequences. Afterwards, a bidirectional GRU [24] layers with a fully connected layer is used to model the feature sequences. Finally, a standard CTC layer is applied to calculate the loss and output the final result. Experimental results from a synthetic dataset and several real screen-rendered text images demonstrate that the performance of the proposed method can outperform other state-of-the-art methods.

II. Proposed Method

As shown in Fig. 2, the proposed method contains two main stages: text line extraction and text line recognition. In text line extraction, Fig. 2(a) shows a real screen-rendered image, the detail of red box is shown in Fig. 2(b). The OTSU binarization algorithm is firstly performed to preprocess input screen-rendered images; the result is shown in Fig. 2(c). Subsequently, as shown in Fig. 2(d), multiple text candidates in each line are acquired by dilation and connected component detection operation. Finally, connected component confusion method is utilized to guarantee that there is only one text area per line, which is shown in Fig. 2(e). After step1, the obtained text line images are directly used as the input images of step2 to be recognized.

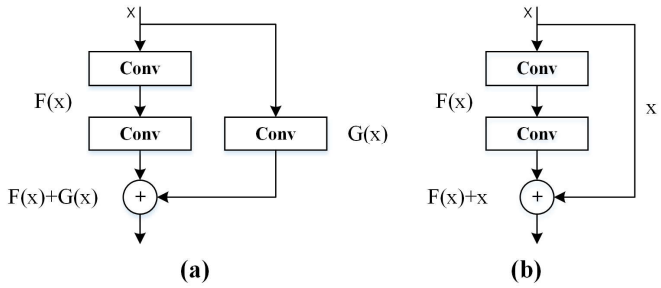


Fig. 3. (a) A shortcut connection with a projection shortcut; (b) A shortcut connection with an identity shortcut.

In text line recognition, a novel network is proposed to recognize the text line images, which consists of three components, including ResNet, bidirectional GRU and a CTC layer, from left to right. ResNet is adopted to automatically extract the feature sequences of each input text line image. In order to efficiently investigate the contextual information in feature sequences, a bidirectional GRU layer is utilized to model the feature sequences and output recognition result. Finally, a standard CTC is applied to calculate loss and yield the final results.

A. Feature Extractor

CNN has achieved tremendous success in many fields of computer vision, which benefited from its efficient and powerful feature extraction capability. Since LeCun put forward the CNN structure for deep learning, recent years have witness the emerging of numerous CNN models, such as AlexNet, VGG and GoogLeNet.

Throughout the development of CNN, it can be observed that the network become deeper and deeper for performance improvement. However, gradient vanishing exposed with the incensement of CNN layers. To address this issue, He et al. proposed a deep residual learning framework, so called ResNet. As shown in Fig. 3(a) and/or Fig. 3(b), a novel shortcut connection that skips one or more layers is introduced into ResNet to deal with the gradient vanishing problem and enable networks to be deeper. The more effective features can be extracted by this kind of structure. In practical applications, the shortcut connection also shows efficient feature extraction ability.

In the proposed method, the shortcut structure is used to build a twelve-layer ResNet without fully connected layers. Inspired by CRNN [9], we use rectangular pooling window after shortcut connections to recognize those characters that have narrow shapes.

B. Sequence Labelling

As a variant of RNN, Long Short-Term Memory (LSTM) adds three memory structures including *input gate*, *forget gate* and *output gate* to solve the vanishing gradient problem. Benefitting from gate mechanism, LSTM is able to store long memory information. However, LSTM require longer time to converge due to the addition of three gates. Based on the LSTM, the GRU [24] combine *input gate* and *forget gate* into *update gate* and introduced *reset gate* to converge faster. The contextual information is essential in text recognition. Therefore, a bidirectional GRU layer is added after ResNet to capture the contextual information of the feature sequences in the proposed method. Unlike the back-propagation in CNN, the back-propagation through time is adopted to update parameters in RNN.

C. Connectionist Temporal Classification

CTC [19] is a kind of output layer, which is usually utilized after the RNN layer to decode the output of RNN layer and calculate loss. For a sequence labelling task, the label is a set A (in this paper, it is a collection of Chinese characters, English characters, digits, and punctuation). A symbol B is added to handle the background in the CTC, a new set $A' = A \cup \{B\}$ is used in reality. The input of CTC is a sequence $y = y_1 y_2 \cdots y_T$, where T is the sequence length. I is denoted as the ground truth sequence label of $\mathbf{y}$. The transformation G is defined on sequence $\pi \in A'^T$. G maps π to I by removing the blank label and repeated characters, such as $G(a-bb--ccc--) = abc$. Then, a conditional probability is defined as the sum of probabilities of all π which are mapped by G to I:

$$p(I \mid y) = \sum_{\pi \in G^{-1}(I)} p(\pi \mid y) \tag{1}$$

where, $p(\pi \mid \mathrm{y})$ represents a conditional probability of π, which is defined as follows:

$$p(\pi \mid y) = \prod_{t=1}^{T} y_{\pi(t)}^{t} \tag{2}$$

where, $y_{\pi(t)}^{t}$ model the probability of the output of RNN equaling to $\pi(t)$ at time t.

Let $S(X, I)$ denoted the training data set, where X is training image, I is ground truth sequence label. The loss function σ of the CTC is defined as the negative log probability of ground truth of all the training examples in training set S:

TABLE I. NETWORK CONFIGURATION SUMMARY. 'K', 'S' AND 'P' STAND FOR KERNEL SIZE, STRIDE SIZE AND PADDING SIZE, RESPECTIVELY.

Type	Configuration	
Input	Input raw image	
Convolution	#maps:64, k:3 × 3, s:2, p:1	
MaxPooling	Window:3 × 3, s:1	
Convolution	#maps:64, k:3x3, s:1, p:1	-
Convolution	#maps:64, k:3x3, s:1, p:1	-
Convolution	#maps:128 k:3x3, s:1, p:1	-
Convolution	#maps:128, k:3x3, s:1,p:1	#maps:128, k:1x1, s:1, p:0
Convolution	#maps:256, k:3x3, s:2, p:1	-
Convolution	#maps:256, k:3x3, s:2, p:1	#maps:256, k:1x1, s:2, p:0
MaxPooling	Window:1 × 2, s:2	
Convolution	#maps:512, k:3x3, s:1, p:1	-
Convolution	#maps:512, k:3x3, s:1, p:1	#maps:512, k:1x1, s:1, p:0
MaxPooling	Window:1 × 2, s:2	
Convolution	#maps:512, k:2 × 2, s:1, p:0	
Bidirectional -GRU	#hidden units:256	
Bidirectional -GRU	#hidden units:256	
Full Convolution	# units:3851	
CTC	-	

$$\sigma = -\sum_{(X,I) \in S} \log P(I \mid y) \tag{3}$$

where, y is the output of RNN layer, I is ground truth sequence label. In addition, the network can be trained end-to-end with backpropagation algorithms.

In many segmentation-free scene text recognition methods, each test sample is associated with a lexicon. The label sequence is recognized by choosing the sequence in the lexicon that has highest conditional probability. It is possible to build such a lexicon for English. However, there are billions of irregular combinations between characters in the Chinese. It is insignificance to establish such a dictionary. Therefore, the output of CTC is used as the final result of the network directly in this paper.

III. EXPERIMENT RESULTS

In this section, the public dataset Computer Vision Lab Handwritten Digit String (CVL HDS) [25] and ORAND-CAR[25] are firstly utilized to demonstrate the performance and robustness of the proposed network. Then, experiments are conducted on real screen-rendered text images to exhibit the practicability of the proposed method.

The network configuration of the proposed method is described in Table I. The input images are resized to 304×32. The proposed network consists of three components including

ResNet, bidirectional GRU and a CTC layer. The ResNet component uses four shortcut connections and two 1×2 pooling layers. Since the input image height is diminutive, a 3×3 convolution kernel is selected in all convolution layers. On the tail of ResNet, a bidirectional GRU layer (There is only a GRU layer in each direction) and a standard CTC are built in the proposed network.

The code is implemented with TensorFlow1.3.0. The experiments are conducted on a TITAN XP graphics card. The network is trained with ADAM optimization algorithm and learning rate is set to 0.001.

A. Result on public dataset

In this subsection, we conduct experiments on two public datasets ORAND-CAR and CVL-HDS to demonstrate the efficiency and robustness of the proposed method.

The first dataset, ORAND-CAR, made up of 11,719 images taken from the Courtesy Amount Recognition (CAR) field of real bank checks. Depending on the different source, the ORAND-CAR images have various types of noise, image qualities and handwriting style. The ORAND-CAR database is classified as two subsets: ORAND-CAR-A and ORAND-CAR-B. ORAND-CAR-A database comprised of 2009 training images and 3784 test images. The ORAND-CAR-B database contains of 3000 training images and 2926 test images. The other dataset is CVL HDS was written by 300 different people, which ensures the high diversity of every sample. The CVL HDS dataset has 7690 images, of which 1262 images are established for training, the rest 6698 images are utilized for testing.

The experimental results are shown in Table Ⅱ. It can be observed that the proposed method achieves desired improvements on both CAR-A and CAR-B datasets compared with other eight state-of-the-art networks. While the segmentation-based method BeiJing [25] obtains better recognition accuracy than the proposed segmentation-free network on the CVL-HDS dataset. The main reason is that the CVL HDS dataset contains 26 different digit strings written by 300 writers, however, only 10 type digit strings are built for the training. For the segmentation-based methods in [25], only 10 numbers need to be identified, so the segmentation-based methods are unaffected by this situation. However, the data distribution absence has a negative influence on segmentation-free methods. What's more, it is exciting that the proposed method is more superior than other segmentation-free methods CRNN [9] and HDSR [18] and other segmentation-based methods except for BeiJing [25] on CVL-HDS dataset. Similar to the proposed method, HDSR also uses the ResNet-RNN-CTC structure. Differently, our method introduced two 1×2 sized rectangular pooling layers into ResNet to recognize some characters with narrow shapes, such as 'i' and 'l'. Besides, the proposed method using more hidden units in the RNN stage to achieve better encode performance. Therefore, proposed method obtained a better performance than HDSR.

B. Result on real screen-rendered text image

The above experiments have demonstrated the efficiency and robustness of the proposed network. In this subsection, experiments are carried out on real screen-rendered text images to exhibit the practicability of the proposed method. Since there is no public training dataset of printed character, we firstly generate 1,000,000 text images as training set, described in section 3.2.1. And then, the proposed text line extraction method is applied to segment a screen-rendered image into multiple text line images. Finally, all text line images are identified by the proposed method.

TABLE II. RECOGNITION ACCURACY OF DIFFERENT MODELS ON THE DATASETS DESCRIBED ABOVE. (THE TOP FIVE METHODS ARE PROPOSED ON THE HDSRC 2014[25], THE FOLLOWING THREE ARE PROPOSED IN NEWEST PAPERS. ESPECIALLY, SAABNI[26] METHOD USES THE ORAND-CAR DATASET AS A WHOLE.)

Methods	ORAND-CAR-A	ORAND-CAR-B	CVL HDS
Tebessa I[25]	0.3705	0.2662	0.5930
Tebessa II[25]	0.3972	0.2772	0.6123
Singapore[25]	0.5230	0.5930	0.5040
Pernambuco[25]	0.7830	0.7543	0.5860
BeiJing[25]	0.8073	0.7013	**0.8529**
CRNN[9]	0.8801	0.8979	0.2601
Saabni[26]	0.858		-
HDSR[18]	0.8975	0.9114	0.2707
Proposed	**0.9189**	**0.9379**	0.6303

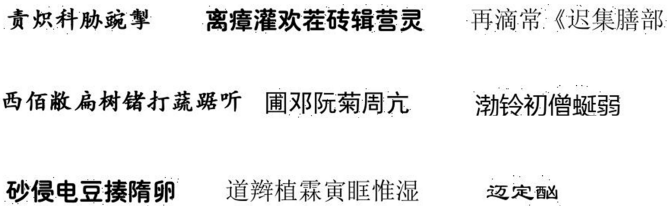


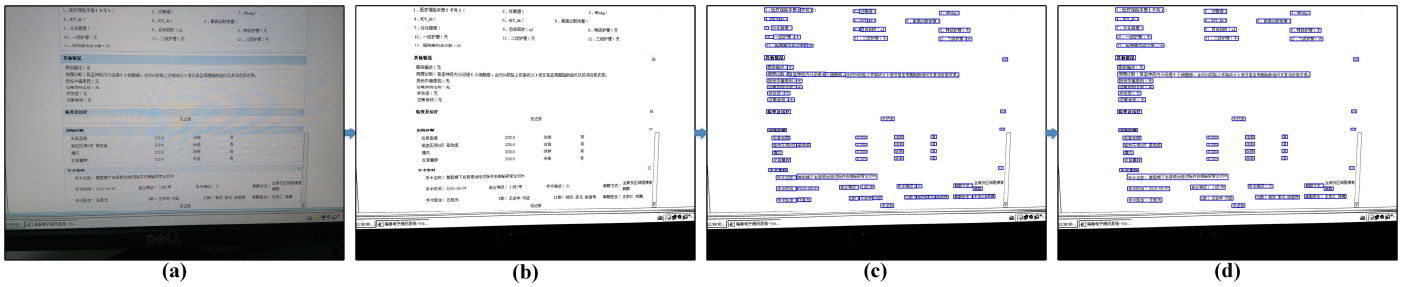
Fig. 4. Sanples of generated images.

1) Generated Dataset

Since there are no public print Chinese-English text dataset, a data generation engine is designed to provide training data for the proposed method.

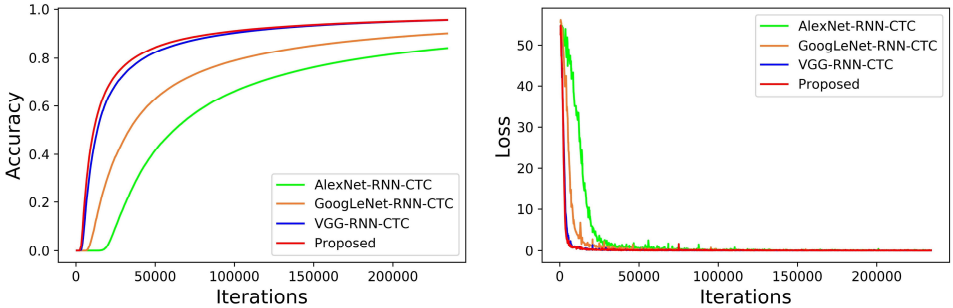
A character set is used in the data generation engine, which contains 3851 characters (3755 primary Chinese characters, 52 English letters, 10 digits, and 34 punctuation marks). To avoid gradient vanishing and make training process faster, the maximum number of characters in each generated image is limited to 10. For each synthetic image, 4-10 characters are selected randomly from the character set. The fonts used in the images are stochastically chose from a font set, which contains 28 fonts. The selected characters are draw to a white image with selected font. Besides, in order to reduce the sensitiveness to noise of the proposed method, a random noise is added to each image. The height of generated images is 32 and width is depending on the number of characters. Finally, we generated a training dataset including 1,000,000 generated images and a validation dataset of 10,000 generated images. The samples of dataset are shown in Fig. 4.

2) Text Extraction Result



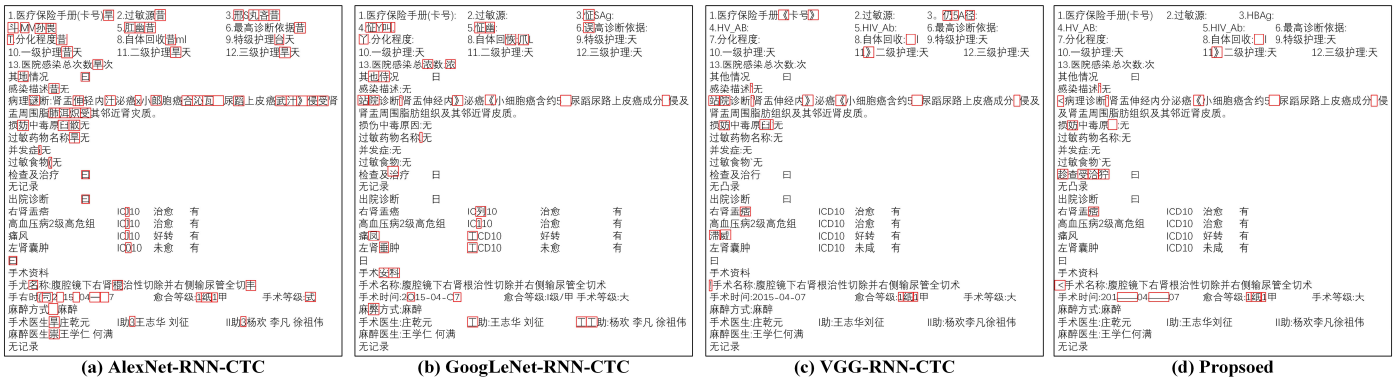
(a) (b) (c) (d)

Fig. 5. Text line extraction results. (a) input image; (b) ~ (d) the result of OTSU binarization, connected component detection and connected component confusion, respectively.



(a) (b)

Fig. 6. The accuracy and loss curves of each network. (a) The accuracy of different net-works; (b) The loss of different networks.



(a) AlexNet-RNN-CTC (b) GoogLeNet-RNN-CTC (c) VGG-RNN-CTC (d) Propsoed

Fig. 7. The recognition results of different network. (a) AlexNet-RNN-CTC; (b) GoogLeNet-RNN-CTC; (c) VGG-RNN-CTC; (d) the proposed method.

In the text extraction stage, the proposed text extraction method only extracts text lines in input image for direct sequence recognition using the proposed network.

Fig. 5 shows text line extraction results of the proposed method. A real screen-rendered text image is shown in Fig 5(a). Fig. 5(b) shows the result of OTSU binarization algorithm. Fig. 5(c) gives the initial text line extraction result by dilation and connected component detection operation. By means of connected component confusion method, there is only one text area per line, as shown in Fig. 5(d). Eventually, the proposed text line extraction method can successfully extraction all text line images from input image.

3) Text Recognition Result

In this subsection, we compare the performance of the proposed method to other three methods: AlexNet, GoogLeNet and VGG (CRNN) combined with RNN-CTC, respectively. All models are trained using the generated training dataset.

Fig. 6(a) and Fig. 6(b) show the accuracy and loss curves of the above network on the train dataset, respectively. As shown in Fig 6(a), the proposed method obtains the higher accuracy with the faster convergence rate. After ten epochs (78130 iterations), the proposed method achieves the accuracy of 89.15% while the best accuracy of the other networks is 87.96%. The final accuracy on the training dataset obtained by AlexNet-RNN-CTC, GoogLeNet-RNN-CTC and VGG-RNN-CTC are 84.16%, 90.19% and 95.59%, respectively. In addition, the proposed method outperforms these methods with an accuracy of 95.67%. Similarly, it can be observed from Fig. 6(b), the proposed method has a strongest ability of feature expression.

Fig. 7 shows the recognition results of a real screen-rendered text image in Fig. 5(a) obtained by AlexNet-RNN-CTC, GoogLeNet-RNN-CTC, VGG-RNN-CTC and proposed method. The red box represents error recognition result. The number of red boxes of AlexNet-RNN-CTC,

TABLE III. AVERAGE ACCURACY OF DIFFERENT NETWORKS

Methods	Average (%)
AlexNet-RNN-CTC	85.74
GoogLeNet-RNN-CTC	91.88
VGG-RNN-CTC	94.16
Proposed	**96.65**

GoogLeNet-RNN-CTC, VGG-RNN-CTC and proposed method are 73, 36, 24 and 21, respectively. Table III gives the recognition accuracy on our testing images. As seen from Fig. 7 and Table III, the proposed method can achieve the best performance on screen-rendered text images than AlexNet-RNN-CTC, GoogLeNet-RNN-CTC and VGG-RNN-CTC.

IV. CONCLUSION

In this paper, we present a novel segmentation-free method based on ResNet and RNN- CTC for screen-rendered Chinese and English text recognition. The experiments on ORAND-CAR and CVL HDS dataset demonstrate the efficiency and robustness of the proposed method. And proposed method shows an excellent performance on screen-rendered images.

Due to the absence of training datasets, there exists a litter of errors in screen-rendered images. In the future, an optimal data generation approach remains to be exploited to fit all scenarios in real-world and train networks. In addition, efforts can be made in combining traditional features with CNN to improve the performance of text recognition.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (61602349, 61440016, 61373109), the Educational Research Project from the Educational Commission of Hubei Province (2016234).

REFERENCES

[1] D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3642–3649.

[2] Z. Zhong, L. Jin, and Z. Xie, "High Performance Offline Handwritten Chinese Character Recognition using GoogLeNet and Directional Feature Maps," In *13th International Confrence on Document Analysis and Recognition*, 2015, p. 5.

[3] X. Y. Zhang, Y. Bengio, and C. L. Liu, "Online and Offline Handwritten Chinese Character Recognition: A Comprehensive Study and New Benchmark," *Pattern Recognition*, vol. 61, pp. 348–360, 2016.

[4] X. Xiao, L. Jin, Y. Yang, W. Yang, J. Sun, and T. Chang, "Building fast and compact convolutional neural networks for offline handwritten Chinese character recognition," *Pattern Recognition*, vol. 72, pp. 72–81, 2017.

[5] Z. Sun, L. Jin, Z. Xie, Z. Feng, and S. Zhang, "Convolutional Multi-directional Recurrent Network for Offline Handwritten Text Recognition," In *15th International Conference on Frontiers in Handwriting Recognition*, 2016, pp. 240–245.

[6] A. Bissacco and M. Cummins, "PhotoOCR : Reading Text in Uncontrolled Conditions," In *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 785–792.

[7] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition," *arXiv preprint arXiv:1406.2227*, 2014.

[8] X. Ren, K. Chen, and J. Sun, "A CNN Based Scene Chinese Text Recognition Algorithm With Synthetic Data Engine," *arXiv preprint arXiv:1604.01891*, Apr. 2016.

[9] B. Shi, X. Bai, and C. Yao, "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*, vol. 39, no. 11, pp. 2298–2304, Nov. 2015.

[10] B. Shi, X. Wang, P. Lyu, and et al. "Robust scene text recognition with automatic rectification," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4168-4176.

[11] Z. Cheng, F. Bai, Y. Xu, and et al. "Focusing Attention: Towards Accurate Text Recognition in Natural Images," In *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5086-5094.

[12] Z. Wang, R. Hu, C. Liang, and et al. "Zero-shot person re-identification via cross-view consistency," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 260-272, 2016.

[13] S. Wachenfeld, H. U. Klein, and X. Jiang, "Recognition of Screen-Rendered Text," *In 18th International Conference on Pattern Recognition*, 2006, pp. 1086–1089.

[14] S. F. Rashid, F. Shafait, and T. M. Breuel, "An evaluation of HMM-based techniques for the recognition of screen rendered text," In *11th International Conference on Document Analysis and Recognition, 2011*, pp. 1260–1264.

[15] Z. Wang, R. Hu, Y. Yu, and et al. "Scale-Adaptive Low-Resolution Person Re-Identification via Learning a Discriminating Surface," In International Joint Conferences on Artificial Intelligence, 2016, pp. 2669-2675.

[16] X. Xu, N. Mu, H. Zhang, and X. Fu, "Salient object detection from distinctive features in low contrast images," In *IEEE International Conference on Image Processing*, 2015, pp. 3126–3130.

[17] X. Xu, N. Mu, X. Zhang, and B. Li, "Covariance descriptor based convolution neural network for saliency computation in low contrast images," *In International Joint Conference on Neural Networks*, 2016, pp. 616–623.

[18] H. Zhan, Q. Wang, and Y. Lu, "Handwritten digit string recognition by combination of residual network and RNN-CTC," In *International Conference on Neural Information Processing*, 2017, pp. 583–591.

[19] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification," In *23rd international conference on Machine learning*, 2006, pp. 369–376.

[20] S. F. Rashid, M. P. Schambach, J. Rottland, and S. von der Nüll, "Low resolution Arabic recognition with multidimensional recurrent neural networks," In *4th International Workshop on Multilingual OCR*, 2013, p. 1.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[22] X. Xu, J. Zhou, H. Zhang, and X. W. Fu, "Chinese characters recognition from screen-rendered images using inception deep learning architecture," In *Proceedings of the Pacific-Rim Conference on Multimedia,* 2017, accepted.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2818–2826.

[24] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[25] M. Diem, S. Fiel, F. Kleber, R. Sablatnig, J. M. Saavedra, D. Contreras, and *et al.*, "ICFHR 2014 Competition on Handwritten Digit String Recognition in Challenging Datasets (HDSRC 2014)," In *14th International Conference on Frontiers in Handwriting Recognition*, 2014, pp. 779–784.

[26] R. Saabni, "Recognizing handwritten single digits and digit strings using deep architecture of neural networks," In *International Conference on Artificial Intelligence and Pattern Recognition*, 2016, pp. 1–6.