Received February 15, 2017, accepted February 26, 2017, date of publication March 3, 2017, date of current version March 28, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2676158

A Novel Text Structure Feature Extractor for Chinese Scene Text Detection and Recognition

XIAOHANG REN[1], YI ZHOU[1], ZHENG HUANG[2], JUN SUN[1], (Member, IEEE), XIAOKANG YANG[1], (Senior Member, IEEE), AND KAI CHEN[1], (Member, IEEE)

[1]Department of Electronic Engineering, Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[2]Institute of Information Security and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Corresponding author: Y. Zhou (zy_21th@sjtu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001003, in part by the National Natural Science Foundation of China under Grant 61521062 and Grant 61527804, in part by the Shanghai Science and Technology Committees of Scientific Research Project under Grant 14XD1402100 and Grant 15JC1401700.

ABSTRACT Scene text information extraction plays an important role in many computer vision applications. Most features in existing text extraction algorithms are only applicable to one text extraction stage (text detection or recognition), which significantly weakens the consistency in an end-to-end system, especially for the complex Chinese texts. To tackle this challenging problem, we propose a novel text structure feature extractor based on a text structure component detector (TSCD) layer and residual network for Chinese texts. Inspired by the three-layer Chinese text cognition model of a human, we combine the TSCD layer and the residual network to extract features suitable for both text extraction stages. The specialized modeling for Chinese characters in the TSCD layer simulates the key structure component cognition layer in the psychological model. And the residual mechanism in the residual network simulates the key bidirectional connection among the layers in the psychological model. Through the organic combination of the TSCD layer and the residual network, the extracted features are applicable to both text detection and recognition, as humans do. In evaluation, both text detection and recognition models based on our proposed text structure feature extractor achieve great improvements over baseline CNN models. And an end-to-end Chinese text information extraction system is experimentally designed and evaluated, showing the advantage of the proposed feature extractor as a unified feature extractor.

INDEX TERMS Text structure feature, Chinese text, deep learning, residual network, unified model.

I. INTRODUCTION

Text, an abstract presentation of artificial information, is scattered throughout the human society in this succinct presenting age. As portable digital recording devices are rapidly in fashion among ordinary people, natural images and videos contents proliferate in image and video sharing websites, e.g. YouTube and Flickr. By extracting text information, which carries high-level semantics, natural media contents can be effectively understood and used. It is crucial for a wide range of applications such as image classification, scene recognition and automatic navigation in urban environments.

While text recognition in scanned documents has been well studied and successfully deployed in real-world applications, the detection and recognition of texts in uncontrolled environments still remains an open issue. Generally, text information extraction can be divided into two stages: text detection and text recognition. Traditionally, most works place the two components in consecutive stages, where text detection algorithms commit to tackle the challenges of the variations of text font, size and style, complex backgrounds, noise, and unconfirmed lighting conditions (like using flash lamps) [1]–[8] and text recognition algorithms of the huge variations of text layouts, orientations, geometric distortions and partial occlusions [9]–[13]. These challenges have spawned various feature designs to meet the needs of one stage. Recently, a variety of end-to-end text information extraction algorithms [14]–[18] are proposed that unify the features for both stages to remove the error propagation between them, leading to significant improvements.

However, existing works still have certain limitations. In particular, most of them only focus on English texts, which are relatively easy to define and recognize due to the simplicity in strokes and structures. In this age of globalization, the recognition of multilingual texts attracts increasing interests. Among them, the logographic text is the most special text type, which presents both pronunciation and meaning in its appearance. As a typical logographic text, Chinese reveals significant different properties than English, which is a typical Latin-based text and has been well studied, from the following aspects:

1. Number of strokes. Most Chinese characters contain more than five strokes, while the most complex English character only has four.
2. Type of strokes. There are more than 30 different types of Chinese strokes, while only 10 different types of strokes exist in English.
3. Style of characters. It is widely acknowledged that most Chinese characters are picto-phonetic, while English uses abstract characters.
4. Intra-character structures. Chinese characters are more complicated than English, where certain structures may exist within a character.

Due to the aforementioned differences between Chinese and English texts, the study of Chinese text recognition is of much value in both theoretical and practical perspectives. Unfortunately, current state-of-the-art text information extraction algorithms for English texts cannot be easily deployed to Chinese. Because of the complexity of Chinese texts including the strokes and intro-character structures, it is extremely difficult for an algorithm to combine text detection and recognition in a unified framework.

In this paper, a Chinese text structure feature extractor is designed, trained and applied to both text detection and recognition, making a number of key contributions.

Our main contribution is a novel Chinese text structure feature extractor. Motivated by the three-layer Chinese text cognition model of human in psychology, we combine the text structure component detector (TSCD) layer and the residual network to simulate the two key mechanisms in the three-layer model. The TSCD layer has specialized modeling for Chinese structure components that are the bridge between strokes and characters and play key role in the three-layer model. The residual network has its unique residual mechanism that is an effective bidirectional information transmission between the upper and lower layers. It is highly similar to the key bidirectional connection in the three-layer model. By reconstructing the components in the TSCD layer into a TSCD block (shown in Fig. 1) refer to the design idea of residual network, the organic combination is established and the three-layer Chinese text cognition model is well simulated. Therefore, the extracted features are applicable to both stages in Chinese text information extraction.

Our second contribution is unifying Chinese structure feature extractor in both text detection and recognition with sharing trained parameters. In deep learning models, both text detection and recognition are regarded as classification tasks. In our detection and recognition models, the feature extractor is trained in the recognition model, which focuses more on the structure features in Chinese characters. The structure features are unique and rarely seen in background regions,

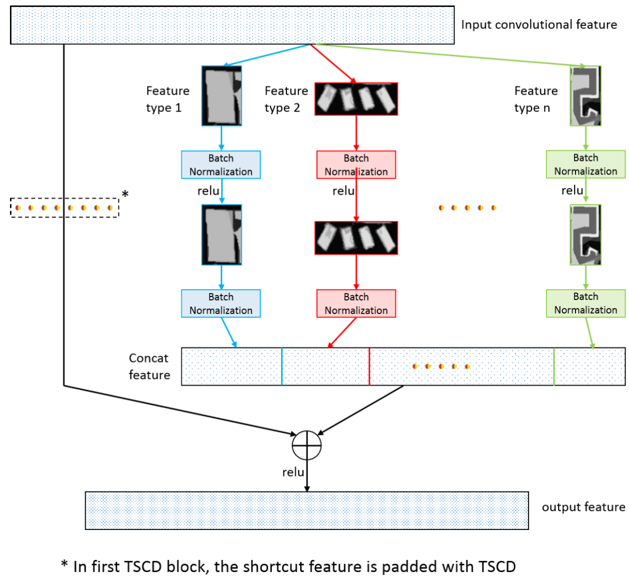

FIGURE 1. The TSCD block.

thus they are applicable to distinguish text and background regions without fine-turning the extractor parameters.

Our third contribution is a synthetic data engine. As the size of public available Chinese scene text datasets is not enough to train deep models, it is essential to expand the training data with artificial samples. The synthetic data engine is composed of three stages, each of them simulates one scene text characteristic. Thus, the generated samples can partially substitute the scene text images and become an essential data component in training the Chinese text structure feature extractor.

The rest of the document is organized as follows. In Section II, we introduce the related works of our work. In Section III, we describe the proposed Chinese text structure feature extractor. In Section IV, we present the experimental evaluation setting up, results and discussions. The paper is concluded in Section V.

II. RELATED WORKS

Generally, text information extraction is divided into two stages: text detection and text recognition. As text recognition in scanned documents is well studied and many OCR systems have achieved quite good performance, most existing researches focus on the text detection stage.

Text detection approaches concern how to discover and locate the regions containing texts from scene images [19]. Traditionally, there are two major categories in text detection algorithms: region-based and texture-based. The symbolic component of region-based algorithms is the image region extraction that limits the feature analyzation on image patches. The features in these algorithms are always unique in scene text regions. Sliding window based algorithms and connected component (CC)-based algorithms are the main region-based algorithm types. Sliding window based algorithms extract numerous overlapping image rectangles for feature extraction [2], [14], [20]. CC-based algorithms extract connected image regions with uncertain shapes and use a set of rules to identify scene text regions [3], [4], [21]–[23]. The symbolic component of texture-based algorithms is the global texture features in the entire scene image. By analyzing the distribution of the global texture features, text regions are extracted from the scene image and connected into text lines. Machine learning methods are the most popular tools in texture-based algorithms, which have strong capability to extract text regions with the global texture features [1], [5], [7]. As scene texts can be in various orientations, multi-orientation approaches provide a practical solution to scene text detection. Component analyzing is the core in recent multi-orientation approaches, including component aggregating, clustering and linking [24]–[26].

Text recognition aims to extract text content information from the cropped text images. As different language texts are in great disparities, most existing scene text recognition algorithms focus solely on recognizing one language texts. Among them, English text recognition is widely studied and can be classified into two major categories: character based recognition and word based recognition [27]. Character based recognition recognizes the characters in the image by a character classifier and combine the recognize results into words. As English character types are very limited, most common classifiers are competent to the task with appropriate features [10], [12], [28]. Word based recognition directly recognizes the word in the text image. As there are numerous English words, common classifiers or models are not competent to recognize words directly. So in some algorithms, the efficient components of character based recognition are still adopted [9], [13], [29].

The strategy of performing text detection and text recognition in separate and independent stages may have certain problems. As the goal of text detection and the requirement of text recognition occur discrepancies in many aspects, it is argued that additional challenges may turn up when exploiting the text detection results for text recognition. Therefore, a variety of end-to-end text information extraction algorithms [14], [17], [18], [30] are proposed to combine both text detection and recognition stages by a unified framework. The key mechanism in the unified framework is unified features which are applicable to both text detection and recognition. Due to the higher complexity of Chinese texts than English texts, the design of unified features is much more difficult. In our previous work [31], a TSCD layer is designed to extract Chinese structure features in CNN based model. The structure features are essential in Chinese characters, thus they are applicable for both Chinese text detection and recognition.

Recently residual network proposed by He *et al.* [32] is under the spotlight in deep learning researchers. They design several convolutional blocks with residual compute and feedback routes, which enable the top parameters in a very deep model learned effectively by current training method. Benefit from the convolutional blocks, very deep convolutional neuronal networks, e.g. 152 layers and 1001 layers, can be easily designed. And they have achieved huge improvements in classification, detection and segmentation tasks. Furthermore, the convolutional blocks also quicken the bidirectional information transmission route between layers. Thus the design of convolutional block is instructive for other tasks that require bidirectional information transmission.

Synthetic data engines for artificial text image generation achieve great success in deep learning based scene English text information extraction as the size of public available English text dataset is not enough for training a deep model. In [16], numerous simple synthetic text images are generated and used to train text detection and recognition models. A complete synthetic data engine is designed in [33], which simulates English word regions in scene images. In [34], a fast and scalable synthetic data engine that naturally blends text in existing natural scenes is proposed to generate scene images with text regions. As there are fewer public available Chinese text datasets and Chinese texts are more complex, it is in need of a synthetic data engine for Chinese texts.

III. CHINESE TEXT STRUCTURE FEATURE EXTRACTOR

The design of the proposed Chinese text structure feature extractor is motivated by the three-layer Chinese text cognition model of human, which is described in Section III.A. The text structure component detector (TSCD) layer and the residual network simulate the key mechanisms in the model: structure component layer, which is the bridge in the model, and the bidirectional information transmission between the upper and lower layers, respectively. Based on their design ideas, a TSCD block is constructed with various structure feature extractors in residual information transmission route, which is described in Section III.B. Replacing a set of convolutional blocks in a residual network model with TSCD blocks, Chinese text structure feature can be extracted accurately to meet the requirements of both text detection and recognition. In order to enhance the unity of the text detection and recognition models, the Chinese text structure feature is applied as unified features in the two models, whose sharing structure is described in Section III.C. The network parameters of the unified feature extractor are learned solely in recognition model with a training set consisting of scene and artificial samples. A considerable number of samples in the training set are artificial text images generated by a synthetic data engine proposed in Section III.D. The engine contains three generation stages, which cover every essential process from putting a text in the scene to taking a text image, to ensure the effectiveness of the generated samples in training.

A. CHINESE TEXT COGNITION MODEL OF HUMAN

When human extract text information from vision, text detection and recognition stages are not independent and sequential. They work at the same time and complement each other: detection process assists recognition, and recognition processing assists detection at the meanwhile. Thus we believe that the extracted features in human cognition system for text are applicable to both detection and recognition. By analyzing and simulating the Chinese text cognition model of human, we can design a feature extractor for both text detection and recognition.

Based on the Chinese text recognition multi-layer activation model proposed by Taft and Zhu [35], Shen *et al.* [36] proposed a three-layer Chinese text cognition model of human, which has been widely accepted by the psychology community. In the cognition model (shown in Fig. 2), the Chinese text cognitive process can be divided into three layers: text layer in the top, text structure component layer in the middle and stroke layer in the bottom. Psychological experiments show that the information transmission between two adjoining layers is bidirectional, which has top-down and bottom-up two transmission directions (represented by the arrow directions in Fig. 2). Furthermore, the bidirectional information transmission can be excitatory and inhibitory (represented by the solid and dashed arrows in Fig. 2).

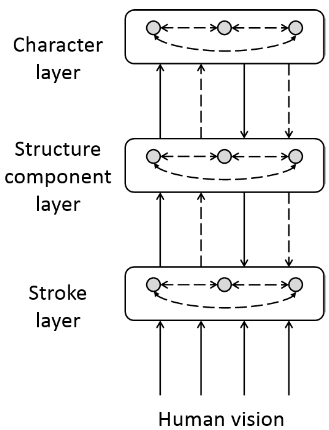

FIGURE 2. The cognition model.

It can be seen from the model, as the middle layer of the three-layer model, the text structure component has direct connection and interaction to upper text and lower stroke. So the structure component is the bridge and key structure in the Chinese text cognition model of human. And the bidirectional connections strengthen the speed and flexibility of information transmission. So the bidirectional information transmission is the key transmission in the Chinese text cognition model of human. Therefore, the key to simulate the human text cognition feature lies in the simulation of the two key mechanisms: structure and information transmission.

B. TEXT STRUCTURE COMPONENT DETECTOR (TSCD) BLOCK

Chinese character is a kind of pictographs, which contains a large number of radicals and structures. Compared with other non-Latin language scripts, i.e. Japanese, Koran and Arabic scripts, Chinese script has richer character structures, which makes the structure features more important in Chinese script than other language scripts. There are many structure types in modern Chinese characters, which contain four basic types: left-right structure, top-bottom structure, inner-outer structure and single character [37] and their sub-types such as left-middle-right, top-left-right and so on. Each Chinese character is comprised of several structure components according to its structure type. Following the utility function proposed in [38], Chinese character structure components are extracted from the most commonly-used 1290 Chinese characters based on their basic structure types. Although Chinese characters contain a wide variety of the structure components in appearance, their aspect ratio types are limited by their structures. Thus the structure components are able to be classified into a few types according to their aspect ratio types. An example of the statistical result of top-bottom structure is shown in Fig. 3, in which structure components extracted from top-bottom structure characters are classified into 10 types. It can be seen from the statistical result that the top-bottom structure characters contain three main structure component types(3:1, 3:2 and 2:1) and three secondary types(1:2, 1:1 and 2:3).

According to the aspect ratio types of Chinese character structure components, several text structure component detectors are designed in convolutional neural network. According to its target character structure component type, the aspect ratio of the convolutional window in a text structure component detector is set as the same. The longer edge of the convolutional window is fixed. Convolutional window determines what kind of feature the convolutional feature extraction is adapt in. For example, a 2:1 convolutional window is much more sensitive to structure components with 2:1 aspect ratio than other convolutional windows and is insensitive to structure components with other aspect ratios. Thus it will focus on extracting accurate features from 2:1 structure components after effective training. Those text structure component detectors with various convolutional window sizes extract structure component features in parallel and form a TSCD layer. Thus, the TSCD layer well simulates the structure of structure component layer in the three-layer Chinese text cognition model of human. More details of TSCD layer are proposed in our previous work [39].

In residual network [32], the basic units are residual convolutional blocks. A residual convolutional block contains several convolutional layers, batch normalization layers and a shortcut route. The input features are separated to the shortcut route and the layers, whose functionality is to compute the residual of the feature. And the input features and the residual features are merged in the end. The information transmission in a residual convolutional block is quick and flexible due to the shortcut route and the residual computing and merging. And because of forward and backward propagations in deep network models, the information transmission is bidirectional. Thus, the residual network well simulates the information transmission in the three-layer Chinese text cognition model of human.

The TSCD layer and the residual network both simulate one of the two key mechanisms in the three-layer Chinese text cognition model of human. By effectively combining

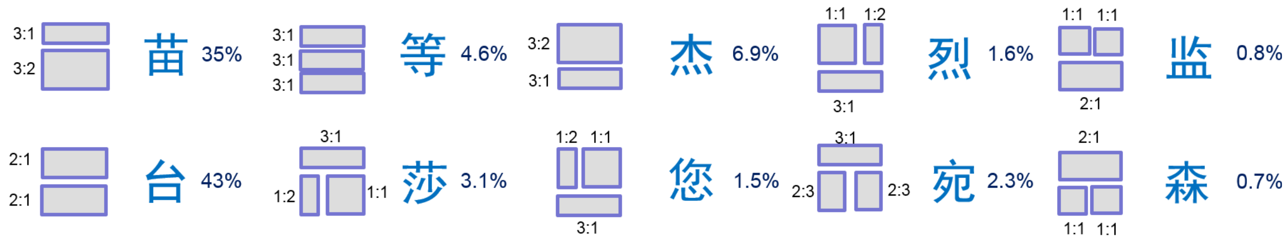

FIGURE 3. The statistical result of top-bottom structure characters.

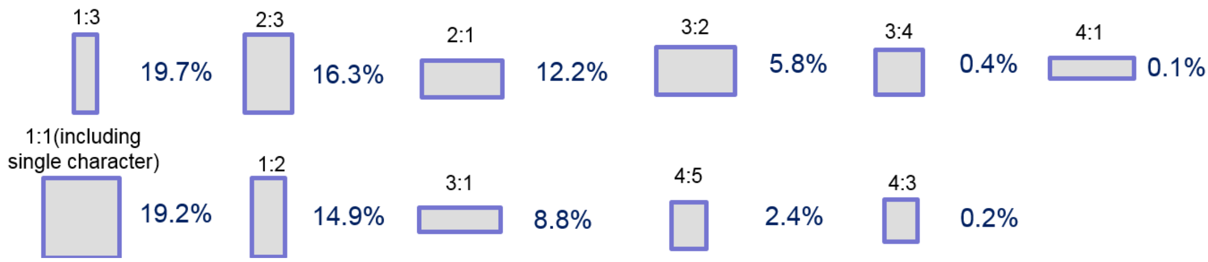

FIGURE 4. The statistical result of Chinese character structure component aspect ratio types.

them, the three-layer model can be well simulated and image features applicable to both Chinese text detection and recognition can be extracted accurately. However, though they are both based on CNN, their design directions are completely different. Simply connecting them into a CNN model cannot equip the model with the two key mechanisms. Thus, we combine them in design idea, to construct a convolutional structure that not only has a feature extraction capacity of various structure component types as the TSCD layer, but also has a quick and flexible information transmission as the residual network. The design idea is to reconstruct the convolutional block of residual network with more convolutional window types and place their residual computations in parallel, which extracts various Chinese text structure features exclusively and keeps their independence for accuracy.

Based on the above design ideas, we propose a TSCD block to completely combine the TSCD layer and residual network. The TSCD block is a special layer combination in CNN model, whose structure overview is shown in Fig. 1. In the TSCD block, the various components in the TSCD layer are broken up and reconstructed in parallel with batch normalization and activation to compute their residual independently. The activation function is Rectified Linear Units (ReLu), which simulates the sparse activation of human neurons. Using linear activation function also alleviates the Vanishing Gradient Problem in training. As there is no correlation among the structure component types, the residuals of each feature types should be computed individually to keep their independence, which also reduces the computational complexity. After the residual computations are completed, they are connected into undivided features, which has the same size as the input features, using concat function. In this way, the influences among different component types in residual computations are avoided. Finally, the input features and the residual features are merged into the output features as they do in a convolutional block in residual network. It should be noted that if the input features are stroke features extracted by a convolutional block rather than structure features extracted by a TSCD block, there will be a TSCD layer in the shortcut route to extract structure features. The feature type order and sizes in this TSCD layer are exactly the same as those in the residual computing route of the TSCD block, which ensures the input features and the residual structure features are correctly matched according to their structure types before merging.

Through the statistical result (as shown in Fig. 4), the Chinese text structure component types to be detected and extracted in the TSCD block can be determined. Those structure component types are extracted from the most commonly-used 1290 Chinese characters and classified by their aspect ratios. More than 99% structure components can be classified in to the 11 types shown in the statistical result. Among them, the least three structure component types accounted for less than 1% of the total. Thus, we choose the most common eight structure component types to be detected and extracted features in the TSCD block. To balance the feature influences of different structure component types in the extracted features, the feature sizes have the same proportions as their corresponding structure component types. With such a design, the quantitative distribution of text structure features matches that of natural Chinese text structure component types, which grantees the most Chinese character structure components can be detected and extracted features in the TSCD block.

Our proposed Chinese text structure feature extractor is constructed by two convolutional block sets and one TSCD block set. The two convolutional block sets are applied to extract accurate stroke features from input image. The TSCD block set is applied to extract specified Chinese text structure features from the stroke features with bidirectional information transmission. In this way the two lower layers in three-layer Chinese text cognition model of human are simulated in a deep learning model based on their two key mechanisms: structure and information transmission. Thus, the extract features are applicable to both text detection and recognition.

C. UNIFIED FEATURES IN TEXT DETECTION AND RECOGNITION MODELS

In current deep learning model design, it is difficult and costly to get detection and recognition results from an output layer. So we do not unify the Chinese text detection and recognition models from top to bottom, but we input the features into different fully connected and output layers after the unified feature extraction.

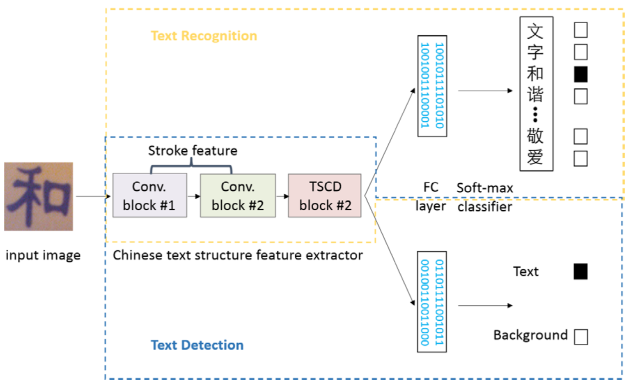

FIGURE 5. The unified Chinese structure feature extractor in text detection and recognition models.

Fig. 5 shows the overview of the unified Chinese structure feature extractor in text detection and recognition models. In the top, the proposed Chinese text structure feature extractor is applied to extract the unified feature from input image patches. Then the features are flattened and inputted to two different analyzing structures. One is the text detection structure, which contains a fully connected layer to analyze the features for text detection and a binary softmax classifier to output the detection result of whether the input image patch is a text patch or a background patch. The other is the text recognition structure, which contains a fully connected layer to analyze the features for text recognition and a multi-class softmax classifier to output the recognition result of which Chinese character the input image patch represents. In the view of text information system, these structures can be regarded as sharing feature extraction in text detection and recognition models.

As text recognition requires more accurate and unique text structure features than text detection, the shared Chinese text structure feature extractor is trained in the text recognition model with text images. Benefit from the unique text structure features in Chinese, the extracted features for text recognition work well in text detection model. So the feature extractor trained for text recognition has not to be fine-tuned for text detection, which simplifies the training process and reduces the computational complexity of detection and recognition. Due to the sharing mechanism, the model unity is significantly improved than that in [31], which has unified structure of feature extraction.

D. SYNTHETIC DATA ENGINE

Chinese text is more complex than English text in stroke, style and structure, which imposes higher requirements on the quantity and quality of training images to train credible deep models. However, the size of public available scene Chinese scene text datasets is much smaller than English text and far from enough to satisfy the requirement. Inspired by the successful use of artificial text images in English text information extraction deep model [18], we design a synthetic data engine to generate high quality artificial Chinese character images.

Considering the high similarity of Chinese fonts, characters are generated by several representative Chinese fonts rather than all the available fonts. Meanwhile, as Chinese characters are all square shaped, we randomly assign the character size in generation. Our synthetic data engine consists of three stages: character image generation, image transformation, camera influence generation. Their details are as follows:

1. Character image generation: The character image are generated as we usually do when printing characters in nature. There are three steps in this stage. Firstly, characters are generated by Chinese fonts. As many Chinese fonts are highly similar, we collect 32 Chinese fonts, which includes three basic fonts Kai, Song and Hei, 24 derived fonts such as FangSong, XiHei and five common special fonts, to generate characters. (When generating characters, 30% of them are basic fonts, 60% are derived fonts and 10% are special fonts.) Secondly, the character color and background color are randomly selected form the most common 27 colors in scene text images. Finally, borders with random thickness and color are added randomly to the characters.
2. Image transformation: To simulate the characters in scene environment, some natural transformations are added to the generated character images. There are three steps in this stage. Firstly, shadows with random orientation, thickness and intention are added randomly to the characters, which simulates nature light. Secondly, full-projective transformation method are used the transform the image with random parameters, which simulates different viewpoints of observer. Finally, scene background image patches are blended to the characters with random blend intensity and blend mode, which simulates the reflection in nature.
3. Camera influence generation: In this stage, some types of image noise, which is common in taking pictures from natural, are added to the image. In our engine,

we add Gaussian noise and Gaussian blur to the character image with random intensity.

The representative Chinese character images generated by our synthetic data engine with random parameters perfectly simulates the text regions in scene images. And they become essential components in the training set to train our proposed Chinese text structure feature extractor.

IV. EXPERIMENTS

In this section the proposed Chinese text structure feature extractor is evaluated with scene text detection and recognition models. The feature extractor is first evaluated with text recognition model as it is trained in the model. Then it is evaluated with text detection model. Finally the detection and recognition models are connected and the proposed feature extractor is evaluated as a unified feature extractor. Three multilingual text datasets [6], [40], [41] are used in the evaluations, one of them is used in all the evaluations and the other two are used in either text detection or recognition evaluation.

In Section IV.A, the datasets used in evaluations are presented. Then the details of training samples, deep models, training methods are introduced in section IV.B. In section IV.C, the text recognition evaluation results are presented and discussed. The text detection evaluation results are presented and discussed in section IV.D. The evaluation results of connected models are presented and discussed in section IV.E.

A. DATASETS

Three multilingual datasets are used to evaluate our proposed Chinese text structure in different models. Among them, Ren's dataset [6] is used in all the evaluations. Zhou's dataset [40] is used in the evaluation of text recognition and Pan's dataset [41] is used in the evaluation of text detection. A summary of the datasets is shown in Table I.

TABLE 1. A description of the various datasets evaluated on.

Datasets	Text Label	Images	#Train	#Test
Ren's dataset	Separately	Natural	194	200
Zhou's dataset	Separately	Natural & Artificial	483	484
Pan's dataset	Mixed	Natural	248	239

Ren's dataset is a multilingual scene text detection and recognition dataset proposed by us in 2015. There are 194 images in the training set and 200 images in the testing set. All the training images and half of the testing images are taken from open fields with different weathers and indoor with different lights from various objects by a camera. The other half of the testing images are taken by various photographers on the Internet. All of them are scene images and taken by various shutterbugs in natural. And all the text regions have language labels to improve the evaluation accuracy of particular language types.

Zhou's dataset is a multilingual text detection and recognition dataset proposed in 2015. The training set contains 483 images and the testing set contains 484 images. It is noted that not all the images in the dataset are scene images. A considerable amount of images are artificial text images including Internet advertisements and screenshots.

Pan's dataset is a multilingual scene text detection dataset proposed in 2011. It is popular in multilingual text detection algorithm evaluations. There are 248 images in the training set and 239 images in the testing set. It is noted that the text regions in the dataset do not have language labels. Thus, it can only evaluate the general multilingual text detection performance with all language types.

Compared with other image detection and recognition datasets, scene text detection and recognition datasets are smaller, in which training set and testing set generally contain only a few hundred images. However, because many text and text regions are contained in a scene image, hundreds of training and test images are sufficient to comprehensively and effectively evaluate text detection and recognition algorithms. Compared with the most widely used text detection dataset ICDAR 2011 [42], which contain 229 train images and 255 test images, the above datasets contain similar or more scene text images. Note that although Ren's dataset contains a bit less (194, 200) images, it is expected that using our dataset can be more effective for evaluation of our proposed Chinese text detection and recognition models as the labels in our dataset are more comprehensive.

(a) (b)

FIGURE 6. The training samples. (a) The examples of the artificial image part. (b) The examples of the natural image part.

B. TRAINING DETAILS

1) TRAINING SAMPLES

The training samples to train the proposed Chinese text structure feature extractor in the text recognition model are composed of two parts (examples are shown on Fig. 6). One part is the artificial Chinese character scene images generated by our proposed synthetic data engine (Section III.E). There are 96,000 artificial character images generated with 1500 Chinese characters(64 images each) which are composed of the most commonly used characters and the characters in the multilingual datasets. The other part is the scene character images extracted from the training sets of the multilingual datasets. The models evaluated in different datasets are trained with different scene training sets. There are 6000 character images extracted in Ren's dataset and 7000 character images extracted in Zhou's dataset.

The training samples to train the text detection model with the proposed Chinese text structure feature extractor are the image patches extracted by the multilingual datasets using multi-scale sliding window method. In Ren's dataset, there are approximately 25,000 training samples being extracted and selected. The number of simple text samples, complex text samples and background samples is 3000, 6000 and 16000, respectively. The text region percentage in a simple text sample, complex text sample and background sample is over 80%, 25% to 66% and less than 10%, respectively. In Pan's dataset, there are approximately 35,000 training samples being extracted and selected, in which 5000 are simple text samples, 9000 are complex text samples and 21,000 are background samples.

2) DEEP MODELS

The proposed Chinese text feature extractor is mainly composed of one convolutional layer, two convolutional block sets and one TSCD block set in sequential. Each block set contains three blocks. The convolutional layer has 16 filters with 3×3 window size. The convolutional blocks in the first set all have 32 filters and 3×3 window size, while those in the second set all have 64 filters and the same window size. The output features of each convolutional block set are down-sampled by 2×2 max-pooling. Each layer in the TSCD blocks has 128 filters in total, which are distributed to eight text structure component types according to their proportions. The output features of the TSCD block set are down-sampled by average-pooling with pool-size of 4×4 and output as the extracted features of the proposed Chinese text feature extractor.

The proposed Chinese scene text recognition model is composed of the Chinese text feature extractor, a fully connected layer and a softmax classifier in sequential. The fully connected layer has 2048 units. The softmax classifier has 1500 classification categories corresponding to the 1500 Chinese characters.

The proposed Chinese scene text detection model is also composed of the Chinese text feature extractor, a fully connected layer and a softmax classifier in sequential. The fully connected layer has 512 units. The softmax classifier has two classification categories corresponding to text region and background region.

3) TRAINING STEPS

The proposed Chinese text feature extractor is trained in the scene text recognition model. And the fully connected layer and the softmax classifier in the recognition model in the recognition model are trained at the same time. Firstly the recognition model is trained by using all the character images in the artificial part. Then it is further trained by using all the scene character images and 7500 artificial character images, which are randomly selected from the artificial part and each character has five samples. The two-step training balances the large amount of artificial character samples and the smaller amount but more realistic scene character samples to ensure the recognition accuracy.

Then the trained Chinese text feature extractor is applied in the scene text detection model to train the fully connected layer and the softmax classifier of in the model. During the model training, the parameters in feature extractor are fixed and only the fully connected parameters are updated. In this way, the features extracted by our proposed Chinese text feature extractor can be used in both text detection and recognition models as a unified feature.

C. CHINESE SCENE TEXT RECOGNITION

The proposed Chinese text structure feature extractor is completely trained in text recognition model. So the performance of the Chinese scene text recognition model is the key evaluation of the proposed feature extractor. And the trained feature extractor will be directly applied to the Chinese scene text detection model, the quality and accuracy of its extracted features almost determine the performance of the text detection model. Therefore, we firstly conduct several experiments to evaluate effectiveness of the proposed feature extractor in text recognition.

The synthetic data engine we designed in Section III.D generates numerous high quality artificial Chinese character images for training the text recognition model. Its effectiveness is evaluated by training the text recognition model with different training set formations.

The Chinese text recognition model with the proposed Chinese text structure feature extractor is evaluated on two text recognition datasets. One of them is an all natural text image dataset [6] in which all the text images are taken from natural. The other dataset [40] is a partly natural text image dataset which contains a number of artificial text images like Internet ADs. Their details are shown in Section IV.A.

TABLE 2. Chinese text recognition results.

Model		Ren's dataset	Zhou's dataset
TSCD layer	S	0.38	0.45
	A	0.42	0.70
	A+S	0.80	0.81
Residual network	S	0.32	0.35
	A	0.42	0.71
	A+S	0.79	0.78
Proposed feature extractor	S	0.31	0.32
	A	0.44	0.72
	A+S	0.88	0.86
ABBYY		0.40	0.41

Table II summarizes the Chinese scene text recognition evaluation results with different training set formations and deep models. The ABBYY, which is a well-known OCR system, is evaluated as baseline. "S," "A" and "A+S" represent different training set formations. "S" represents only the scene character images extracted from the respective training sets are used in training and "A" represents only the artificial character images generated by the synthetic data engine are used. "A+S" represents the artificial and scene character images are all used to train the deep model. The TSCD layer model is the Chinese scene text recognition CNN model with TSCD layer in our previous work [39]. The residual network model is a normal 20-layer model with three convolutional block sets, each set contains three blocks. The proposed feature extractor model is the scene text recognition model with our proposed Chinese text structure feature extractor, whose detail is shown in Section IV.B.

In the evaluation results, the text recognition model our proposed Chinese text structure feature extractor achieves better results in both datasets than other models and the OCR system. The proposed feature extractor shows its advantage in extracting accurate features for recognition.

It is observed from the results that the text recognition models trained with "S" performs much worse than the leading results on both datasets, which demonstrates the scene characters images extracted from available datasets are far from enough to train a creditable deep text recognition model. And the models with residual design perform worse than the shallow TSCD layer based model when trained with "S." These results show that the residual design in deep model is prone to over-fitting with small training set.

When using the artificial character images, there is a huge gap between the results on the two datasets. It is because Zhou's dataset contains a number of artificial text images, which is similar to the generated images. So its performance has huge improvement. And as the images in Ren's dataset are all scene images, the "A" models perform better than the "S" models but still much worse than the leading results. Models trained with "A" training set formation have similar recognition performances. The results of TSCD layer model and residual network model are almost equal. The proposed model achieves better recognition accuracies in both dataset, which preliminary shows effectiveness of the proposed Chinese text structure feature extractor.

The text recognition models trained with "A+S" have considerable improvements than "A," especially in Ren's dataset, which shows the importance of adding scene character images to train a creditable deep model. Among the models, TSCD layer model and residual network model achieve similar results on both datasets. The TSCD layer model performs slightly better than the residual network model, which indicates priori knowledge structure has larger influence than very deep network in text recognition. The model with our proposed Chinese text structure feature extractor achieves much better accuracies than the other two models in "A+S," which demonstrates that the TSCD layer and the residual network are organically combined in the proposed feature extractor.

D. CHINESE SCENE TEXT DETECTION

The scene text detection model with our proposed Chinese text structure feature extractor is evaluated on two multilingual text datasets. Ren's dataset contains language labels for every text lines, thus the Chinese text detection model can be evaluated more accurately by focusing on the Chinese text lines. Pan's dataset does not contain language labels but has been widely used in many text detection algorithm evaluations. On Pan's dataset, the text detection model is evaluated in wider language environment and compared to more text detection algorithms.

TABLE 3. Text detection results on Ren's dataset.

	$precision$	$recall$	$f-measure$
CNN	0.73	0.74	0.73
CSAE	0.79	0.78	0.78
Residual network	0.78	0.81	0.79
TSCD layer	0.81	0.76	0.78
CSAE + TSCD layer	0.85	0.81	0.83
Proposed	0.87	0.82	0.84

Table III summarizes Chinese scene text recognition evaluation results with different deep models. The CNN model is a baseline model contains two convolutional layers, two downsampling layers, and a fully connected layer. The CSAE model uses an unsupervised learning method, which is proposed by us in [6], to pretrain the CNN model. The TSCD layer model has a TSCD layer in place of the second convolutional layer in the CNN model. And the CSAE + TSCD layer model uses the CSAE unsupervised learning method to pretrain the TSCD layer model. The residual network model is a normal 20-layer model with three convolutional block sets, each set contains three blocks. The proposed model is the scene text detection model with our proposed Chinese text structure feature extractor, whose detail is shown in Section IV.B. The evaluation method in this experiment is the same one we used in our previous work [39], which is based on the evaluation method of ICDAR 2011 and specially designed for one-language text detection evaluation.

In the Chinese scene text detection evaluation results of different deep models, the text detection model with our proposed Chinese text structure feature extractor achieves the best performance of precision 0.87 and recall 0.82. The CSAE unsupervised learning method is applied to extract more accurate stroke features by pretraining the model. The residual network is also applied to extract more accurate stroke features by computing feature residuals. Their evaluation results are similar, but the deconvolution computing in the CSAE is much more complex than the residual computing in the residual network model. The TSCD layer model also achieves similar overall measurement as the CSAE and residual network models do. And it can be noted that it has larger improvement in precision than recall, which demonstrates the uniqueness of Chinese text structure features extracted by the TSCD layer. Compared with the model that combines the CSAE method and the TSCD layer, the model with our proposed Chinese text structure feature extractor achieves better results, even in the case where the feature extractor is not fine-tuned in the detection model. It indicates that the design of the proposed feature extractor has good combination of the TSCD layer and residual network. The results also demonstrate that the extracted Chinese structure features are applicable to both text detection and recognition.

The proposed Chinese scene text detection model is evaluated on Pan's dataset which contains many Chinese text regions and has been widely used in multilingual text detection. It enables the proposed model to be compared with more text detection algorithms in wider language environment. As most text regions in Pan's dataset are Chinese text regions,

FIGURE 7. The Chinese text information extraction result examples.

the performance degradation of the proposed model is limited. In this evaluation, the proposed Chinese text structure feature extractor in the text detection model is also trained in the text recognition model in Ren's dataset and not fine-tuned.

TABLE 4. Text detection results with different algorithms.

	$precision$	$recall$	$f-measure$
The proposed algorithm	0.82	0.72	0.77
Pan's algorithm [41]	0.65	0.66	0.65
Yin's algorithm [43]	0.83	0.69	0.75
Tian's algorithm [44]	0.85	0.78	0.81
Liu's algorithm [45]	0.63	0.67	0.65

Table IV summarizes the evaluation results of the proposed text detection model and some text detection algorithms on Pan's dataset. Although many text regions in the dataset do not contain Chinese texts, which has negative effects on both of the measurements, the proposed model achieves the second best result. Compared to Tian's algorithm, our model has obvious weakness in recall, which is resulted by the unique Chinese structure we used in the model would cause many English text regions to be detected as background regions. The result implies that the uniqueness of the features extracted by proposed Chinese text structure feature extractor would not be as effective in other language environments. Compared with Liu's algorithm, which is the best Chinese text detection algorithm on Pan's dataset, our proposed model achieves better results on both measurements, especially on precision. The results indicate that the Chinese text structure feature has advantages in text detection.

E. CHINESE SCENE TEXT INFORMATION EXTRACTION

We design a Chinese scene text information extraction system, which could be regarded as an embryonic form, by simply connecting the text detection and recognition models. In this system, a text image is firstly divided to many image patches by a multi-scale sliding window method. Secondly, the image patches are classified by the text detection model. Then the image patches are merged into text lines by several geometric and heuristic rules such as similar colors and horizontal distances. After the text lines are detected, they are divided to image patches again by a single-scale sliding window method. Then the characters in the text line are recognized individually by the text recognition model. Finally, the characters are connected into recognized text lines according to their positions and recognize scores.

In the evaluation, three types of the embryonic system, which are constructed with CNN models, TSCD layer models and the proposed models, are evaluated on Ren's dataset [6] as it contains real and well labeled scene text regions. The system with the proposed models achieves the best result P/R/F of 0.67/0.58/0.62, compared to the result of 0.64/0.52/0.57 with the TSCD layer models and 0.48/0.39/0.43 with normal CNN models. Some examples of successful results are shown in Fig. 7. It can be noted that using text structure features (the TSCD layer and the proposed feature extractor), the system performances are significantly improved, which shows the potential of text structure features in unifying text detection and recognition. And the system performers better with the proposed feature extractor than the TSCD layer, which demonstrates the features extracted by the proposed feature extractor are more accurate than those extracted by TSCD layer based models.

V. CONCLUSION

In this paper, we present a novel Chinese text structure feature extractor for both text detection and recognition. It well simulates the key mechanisms in the three-layer Chinese text cognition model of human, which enables human detect and recognize texts at the same time. In the Chinese text structure feature extractor, text structure component detector layer, which simulates the structure component layer in the psychological model, and residual network, which simulates the bidirectional information transmission in the psychological model, are combined by a TSCD block. The design of TSCD block incorporates the design ideas of the TSCD layer and the residual network, which not only integrates them into a whole, but also enables them to complement each other. Therefore, the features extracted by the TSCD block based Chinese text structure feature extractor are applicable to both Chinese text detection and recognition as we human do. Experimental results demonstrate that the proposed Chinese text structure feature extractor is effective in both Chinese scene text detection and recognition due to its high accuracy and completeness in Chinese text structure feature extraction. It is also observed that adding artificial training samples is essential in training the proposed feature extractor in a text recognition model as the existing training samples are limited. The feature extractor trained in the text recognition model is effective in a text detection model even it is not fine-tuned for detection. And an embryonic form of Chinese scene text information extraction system using the feature extractor in both stages achieves promising results. Therefore, the proposed feature extractor is suitable to Chinese text information extraction algorithms as a unified feature extractor for both text detection and recognition.

REFERENCES

[1] X. Chen and A. L. Yuille, "Detecting and reading text in natural scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2. Jul. 2004, pp. II-366–II-373.

[2] X. Li, W. Wang, S. Jiang, Q. Huang, and W. Gao, "Fast and effective text detection," in *Proc. 15th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2008, pp. 969–972.

[3] C. Jung, Q. Liu, and J. Kim, "A stroke filter and its application to text localization," *Pattern Recognit. Lett.*, vol. 30, no. 2, pp. 114–122, 2009.

[4] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2963–2970.

[5] Q. Ye, Q. Huang, W. Gao, and D. Zhao, "Fast and robust text detection in images and video frames," *Image Vis. Comput.*, vol. 23, no. 6, pp. 565–576, Jun. 2005.

[6] X. Ren, K. Chen, X. Yang, Y. Zhou, J. He, and J. Sun, "A new unsupervised convolutional neural network model for chinese scene text detection," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process. (ChinaSIP)*, Jul. 2015, pp. 428–432.

[7] W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolution neural network induced MSER trees," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 497–511.

[8] L. Wu, P. Shivakumara, T. Lu, and C. L. Tan, "A new technique for multi-oriented scene text line detection and tracking in video," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1137–1152, Jan. 2015.

[9] A. Mishra, K. Alahari, and C. Jawahar, "Scene text recognition using higher order language priors," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2012, p. 1.

[10] C. Yao, X. Bai, B. Shi, and W. Liu, "Strokelets: A learned multi-scale representation for scene text recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 4042–4049.

[11] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet, (2013). "Multi-digit number recognition from street view imagery using deep convolutional neural networks," [Online]. Available: https://arxiv.org/abs/1312.6082

[12] C. Shi, C. Wang, B. Xiao, Y. Zhang, S. Gao, and Z. Zhang, "Scene text recognition using part-based tree-structured character detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Apr. 2013, pp. 2961–2968.

[13] T. Novikova, O. Barinova, P. Kohli, and V. Lempitsky, "Large-lexicon attribute-consistent text recognition in natural images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 752–765.

[14] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *Proc. Comput. Vis. (ICCV)*, 2011, pp. 1457–1464.

[15] L. Neumann and J. Matas, "On combining multiple segmentations in scene text recognition," in *Proc. 12th Int. Conf. Document Anal. Recognit. (ICDAR)*, Oct. 2013, pp. 523–527.

[16] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Nov. 2012, pp. 3304–3308.

[17] C. Yao, X. Bai, and W. Liu, "A unified framework for multioriented text detection and recognition," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4737–4749, Jun. 2014.

[18] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *Int. J. Comput. Vis.*, vol. 116, no. 1, pp. 1–20, 2016.

[19] Y. Zhu, C. Yao, and X. Bai, "Scene text detection and recognition: Recent advances and future trends," *Frontiers Comput. Sci.*, vol. 10, no. 1, pp. 19–36, 2016.

[20] P. Shivakumara, W. Huang, T. Q. Phan, and C. L. Tan, "Accurate video text detection through classification of low and high contrast images," *Pattern Recognit.*, vol. 43, no. 6, pp. 2165–2185, Jun. 2010.

[21] P. Shivakumara, T. Q. Phan, and C. L. Tan, "A Laplacian approach to multi-oriented text detection in video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 412–419, Feb. 2011.

[22] H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk, and B. Girod, "Robust text detection in natural images with edge-enhanced maximally stable extremal regions," in *Proc. 18th IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2011, pp. 2609–2612.

[23] H. Cho, M. Sung, and B. Jun, "Canny text detector: Fast and robust scene text localization algorithm," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Mar. 2016, pp. 3566–3573.

[24] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Oct. 2012, pp. 1083–1090.

[25] L. Kang, Y. Li, and D. Doermann, "Orientation robust text line detection in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Mar. 2014, pp. 4034–4041.

[26] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, (2016). "Multi-oriented text detection with fully convolutional networks," [Online]. Available: https://arxiv.org/abs/1604.04018

[27] Q. Ye and D. Doermann, "Text detection and recognition in imagery: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1480–1500, Jul. 2015.

[28] C.-Y. Lee and S. Osindero, (2016). "Recursive recurrent nets with attention modeling for OCR in the wild," [Online]. Available: https://arxiv.org/abs/1603.03101

[29] B. Shi, X. Wang, P. Lv, C. Yao, and X. Bai, (2016). "Robust scene text recognition with automatic rectification," [Online]. Available: https://arxiv.org/abs/1603.03915

[30] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 512–528.

[31] X. Ren, K. Chen, X. Yang, Y. Zhou, J. He, and J. Sun, "A novel text structure feature extractor for chinese scene text detection and recognition," in *Proc. 23nd Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2016, pp. 36–45.

[32] K. He, X. Zhang, S. Ren, and J. Sun, (2015). "Deep residual learning for image recognition," [Online]. Available: https://arxiv.org/abs/1512.03385

[33] M. Jaderber, K. Simonyan, A. Vedaldi and A. Zisserman. (2014). "Synthetic data and artificial neural networks for natural scene text recognition." [Online]. Available: https://arxiv.org/abs/1406.2227

[34] A. Gupta, A. Vedaldi, and A. Zisserman, (2016). "Synthetic data for text localisation in natural images," [Online]. Available: https://arxiv.org/abs/1604.06646

[35] M. Taft and X. Zhu, "The representation of bound morphemes in the lexicon: A Chinese study," in *Proc. Morphol. Aspects Lang. Process. (MALP)*, 1995, pp. 293–316.

[36] M. Shen, Z. Li, and Z. Zuxiang, "The influence of component startup on the recognition of fitting Chinese characters," *Psychol. Sci.*, vol. 3, no. 5, pp. 206–211, 1997.

[37] J. Liu, S. Zhang, H. Li, and W. Liang, "A Chinese character localization method based on intergrating structure and CC-clustering for advertising images," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2011, pp. 1044–1048.

[38] X. Zhou and Y. Li, "A research of chinese character utility function (in chinese)," *Linguistic Res.*, vol. 1, no. 10, pp.62–65, 2009.

[39] X. Ren, Y. Zhou, J. He, K. Chen, X. Yang, and J. Sun, "A convolutional neural network based Chinese text detection algorithm via text structure modeling," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 506–518, Mar. 2017.

[40] X. Zhou, S. Zhou, C. Yao, Z. Cao, and Q. Yin, (2015). "Icdar 2015 text reading in the wild competition," [Online]. Available: https://arxiv.org/abs/1506.03184

[41] Y.-F. Pan, X. Hou, and C.-L. Liu, "A hybrid approach to detect and localize texts in natural scene images," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 800–813, Mar. 2011.

[42] A. Shahab, F. Shafait, and A. Dengel, "ICDAR 2011 robust reading competition challenge 2: Reading text in scene images," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Mar. 2011, pp. 1491–1496.

[43] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, "Robust text detection in natural scene images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 970–983, Oct. 2014.

[44] S. Tian, Y. Pan, C. Huang, S. Lu, K. Yu, and C. L. Tan, "Text flow: A unified text detection system in natural scene images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4651–4659.

[45] X. Liu, Z. Lu, J. Li, and W. Jiang, "Detection and segmentation text from natural scene images based on graph model," in *Proc. WSEAS Trans. Signal Process.*, 2014, p. 10.

XIAOHANG REN received the B.S. degree in electronic engineering from Zhejiang University, Hangzhou, China, in 2011. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include text information extraction, deep learning network, and image retrieving.

YI ZHOU received the Ph.D. degree from Shanghai Jiao Tong University, China, in 2010. She is currently with the Computer Science Department, Shanghai Jiaotong University, China. Her project of Chinese characters recongintion is supported by the National Science Foundation. Her major research includes object recognition and big data mining.

ZHENG HUANG received the Ph.D. degree in computer application technology from Shanghai Jiao Tong University, China, in 2003. He was a Computer Application Technology Professional Doctorate with Shanghai Jiao Tong University in 2003, where he is currently an Associate Professor with the Institute of Information Security and Engineering and also with the School of Information Security Engineering. His research area is information security and research interests include cryptography and machine learning, information security, cryptography, and machine learning.

JUN SUN (M'06) received the B.S. degree in electrical engineering from the University of Electronic Sciences and Technology of China, Chengdu, China, in 1989, and the Ph.D. degree in electrical engineering from Shanghai Jiao Tong University, in 1995. In 1996, he was elected as a member of the HDTV Technical Executive Experts Group, China. Since then, he has been acting as one of the main technical experts for the Chinese government in the field of digital television and multimedia communications. In the past five years, he has been responsible for several national projects in DTV and IPTV fields. He is currently a Professor and Ph.D. Advisor with Shanghai Jiao Tong University. He has authored or co-authored over 50 technical papers in the area of digital television and multimedia communications. His research interests include digital television, multimedia communication, and video encoding. He received the Second Prize of the National Science and Technology Development Award in 2003 and 2008.

XIAOKANG YANG (SM'14) received the B.S. degree from Xiamen University, Xiamen, China, in 1994, the M.S. degree from the Chinese Academy of Sciences, Shanghai, China, in 1997, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, in 2000. From 2000 to 2002, he was a Research Fellow with the Center for Signal Processing, Nanyang Technological University, Singapore. From 2002 to 2004, he was a Research Scientist with the Institute for Infocomm Research, Singapore. From 2007 to 2008, he visited the Institute for Computer Science, University of Freiburg, Germany, as an Alexander von Humboldt Research Fellow. He is currently a Distinguished Professor with the School of Electronic Information and Electrical Engineering and also the Deputy Director of the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University. He has authored or co-authored over 200 refereed papers. He holds 40 patents. His current research interests include visual signal processing and communication, media analysis and retrieval, and pattern recognition. He is a member of the Asia-Pacific Signal and Information Processing Association, a member of the Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society, a member of the Multimedia Signal Processing Technical Committee of the IEEE Signal Processing Society, the Chair of the Multimedia Big Data Interest Group of Multimedia Communications Technical Committee of the IEEE Communication Society. He was a member of the Editorial Board of the *Digital Signal Processing*. He is also an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS and the Series Editor of the *Communications in Computer and Information Science* (Springer).

KAI CHEN (M'09) received the Ph.D. degree from Shanghai Jiao Tong University, China, in 2003. He is currently with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University. His major research includes information retrieving, object recognition, and big data mining. He is the Key Member of the Institute on Network Engineering Research. He has been involved with several key nation projects and hosted many Industry-Academia Research projects.