This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2016.2625259, IEEE Transactions on Multimedia

A Convolutional Neural Network Based Chinese Text Detection Algorithm via Text Structure Modeling

Xiaohang Ren, Yi Zhou, Jianhua He, *Senior Member, IEEE,* Kai Chen *Member, IEEE,* Xiaokang Yang, *Senior Member, IEEE,* and Jun Sun, *Member, IEEE,*

***Abstract*—Text detection in natural scene environment plays an important role in many computer vision applications. While existing text detection methods are focused on English characters, there is strong application demands on text detection in other languages, such as Chinese. As Chinese characters are much more complex than English characters, innovative and more efficient text detection techniques are required for Chinese texts. In this paper, we present a novel text detection algorithm for Chinese characters based on a specific designed convolutional neural network (CNN). The CNN model contains a text structure component detector layer, a spatial pyramid layer and a multi-input-layer deep belief network (DBN). The CNN is pre-trained via a convolutional sparse auto-encoder (CSAE) in an unsupervised way, which is specifically designed for extracting complex features from Chinese characters. In particular, the text structure component detectors enhance the accuracy and uniqueness of feature descriptors by extracting multiple text structure components in various ways. The spatial pyramid layer is then introduced to enhance the scale invariability of the CNN model for detecting texts in multiple scales. Finally, the multi-input-layer DBN is used as the fully connected layers in the CNN model to ensure that features from multiple scales are comparable. A multilingual text detection dataset, in which texts in Chinese, English and digits are labeled separately, is set up to evaluate the proposed text detection algorithm. The proposed algorithm shows a significant 10% performance improvement over the baseline CNN algorithms. In addition the proposed algorithm is evaluated over a public multilingual image benchmark and achieves state-of-the-art results for text detection under multiple languages. Furthermore a simplified version of the proposed algorithm with only general components is compared to existing general text detection algorithms on the ICDAR 2011 and 2013 datasets, showing comparable detection performance to the existing algorithms.**

***Keywords*—Chinese Text Detection, Unsupervised Learning, Text Structure Detector, Convolutional Neural Network**

I. Introduction

WITH increasing penetration of portable multimedia recording devices (such as smart phones and tablets), multimedia contents proliferate in image and video sharing websites, e.g. Youtube and Flickr. Extracting text information from those natural images and videos are conducive to a wide range of applications such as image classification, scene recognition and video retrieval. Although traditional optical character recognition (OCR) systems have achieved good performance in extracting text information from scanned documents, their performance on natural images and videos could drop significantly. The biggest challenge of using OCR systems in natural environment is detecting text regions, as the background in natural images and videos is much larger in size and much more complex in texture. To quantify and track the progress of text location in natural images, several competitions, including four ICDAR Text Location Competitions in 2003, 2005, 2011 and 2013 [1], [2], [3], [4] have been held in recent years. However, even the best performing algorithm reported in ICDAR 2013 can localize only 66% of words in the dataset [4], which clearly shows that there is still a large room for performance improvement.

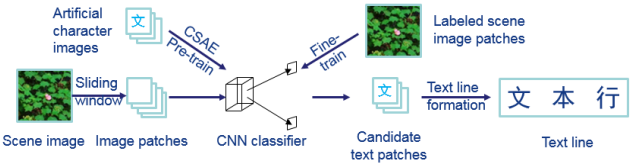


Fig. 1: The flowchart of the algorithm.

The challenges in detecting texts from natural images come from the variations of texts in font, size and style, complex backgrounds, noise, unconfirmed lighting conditions (like using flash lamps), and geometric distortions [5], [6], [7], [8], [9], [10]. As video contains additional time sequence information, effective utilization of text motion estimate technique is vital in video text detection and tracking [11], [12], [13], [14], [15]. Moreover, due to the widespread usage of smart phones, the limited computational ability also becomes a main challenge of text detection [16], [17]. The existing text detection algorithms can be roughly classified into two major categories: region-based methods and texture-based.

Region-based approaches detect texts by analyzing local features in extracted image regions. Those local features are unique in representing scene texts and ensure most text regions can be detected. However, as some complex background regions also have similar texture as text regions, it is very challenging to design filtering rules or classifiers. Texture-

X. Ren, Y. Zhou, K. Chen, X. Yang and J. Sun are with the Department of Electronic Engineering, Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, Y, Zhou is the corresponding author, e-mail: xiaomu@sjtu.edu.cn; zy_21th@sjtu.edu.cn; kchen@sjtu.edu.cn; xkyang@sjtu.edu.cn; junsun@sjtu.edu.cn.

J. He is with School of Engineering and Applied Science, Aston University, Birmingham, United Kingdom, email: j.he7@aston.ac.uk.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2016.2625259, IEEE Transactions on Multimedia

based approaches analyze global texture features in the entire image to localize texts. Global texture features of text and background regions are clearly distinguishable, thus the background regions are rarely mistaken for text regions. Among text regions, the global text features also vary significantly due to the various scene conditions of texts and hence cause a large number of missed detected texts.

Most of the above text detection algorithms use one or several manually designed features such as HOG or SIFT to extract text regions using a discriminative classifier or some heuristic rules. Those features are designed for universal image description instead of specific usage, which leads to difficult optimization problem and weak adaptability. In contrast to those traditional algorithms, recently some deep learning model based text detection algorithms [18], [19] report significant performance improvement. Deep learning algorithms employ original image pixels to detect candidate text regions by extracting strongly adaptable features. Convolutional neural network (CNN) is one of the most widely used deep networks in text detection. A large labeled dataset is needed to train a responsible CNN but labeled scene text datasets have only limited sizes. And as the size of feature maps becomes larger, which is essential in extracting text features, the similarity of features also becomes higher.

It is noted that the above reported works are mainly focused on extracting English text from natural images, while few research works on Chinese text extraction have been reported in the literature. Chinese characters are more complex than English characters. Most Chinese characters contain more than 5 strokes, while the most complex English character "W" has only 4 strokes (we split a line into strokes by the turn point). In addition, there are more than 30 different types of Chinese strokes, while only 10 different types of strokes exist in English. Therefore, for English text detection algorithms, analyzing the relationship between the English characters such as words is more important than character-level detection. On the contrary, the complexity of Chinese characters requires the detection algorithms to focus more on the inner relationship of strokes.

In this paper we propose a Chinese scene text detection algorithm based on CNN (the structure of our CNN is shown in Fig.2), making a number of key contributions.

Our main contribution is a novel Chinese text structure feature extractor, which is a special layer in CNN called text structure component detector (TSCD) layer. In the TSCD layer, Chinese text characters are modeled in different ways as multiple text structure components by the TSCDs. By analyzing the structures of Chinese characters, the Chinese text structure component types can be effectively classified to several easily distinguishable groups based on their aspect ratios. For each text structure component group, a specific TSCD is designed to extract its feature, which has its unique feature map shape. The multi-shape feature maps in the TSCD layer also limit the similarity of features when the feature map size expands thus the requirement of training set is reduced. Extensive simulations demonstrate the TSCD is effective in improving Chinese text detection performance.

Our second contribution is a novel unsupervised learning method, named convolutional sparse auto-encoder (CSAE), for complex and abstract Chinese texts. As the availability of public scene Chinese text datasets is very limited, applying an unsupervised learning method to pretrain a CNN model is important in avoiding overfitting. The CSAE is designed by combining the convolutional layer in CNN and the sparse coding method. Apart from the optimization functions of sparse coding, we add another optimization function to enhance the ability of complex feature representation in our unsupervised learning method.

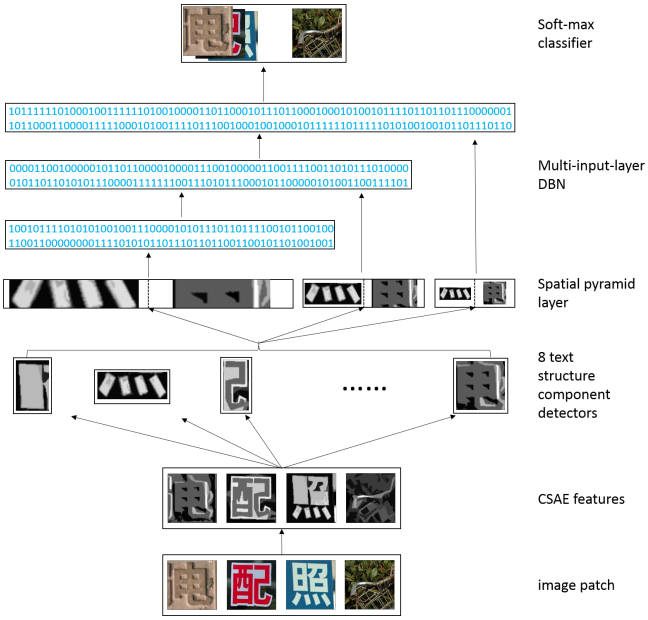


Fig. 2: The overview of the CNN model.

Our third contribution is on the application of a spatial pyramid layer (SPL) and designing a multi-input-layer deep belief network (DBN) as the fully connected layer in our model. The SPL improves the scale invariability of CNN, which is vital to detect various scale texts in natural. With the multi-input-layer DBN, the scale features extracted by SPL and the text features extracted by TSCD can be combined effectively.

Our fourth contribution is setting up a new multilingual text detection dataset for training and evaluation. Different from the public multilingual dataset in [20], our dataset labels Chinese, English and digits separately for both training set and testing set to evaluate text detection algorithm which detects one specific type of language text more accurately with appropriate evaluate method.

The rest of the document is organized as follows. In Section II, we introduce the related works. In Section III, we describe the proposed CNN model based text detection algorithm. In Section IV, we present the experimental evaluation setting up, results and discussions. The paper is concluded in Section V.

II. Related works

Traditionally, text detection algorithms can be roughly classified into two major categories: region-based and texture-based.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2016.2625259, IEEE Transactions on Multimedia

Region-based approaches, such as the traditional sliding window based approach, limit the detection and feature extraction route to a subset of image rectangles. For example, Wang et al. [5] use Random Ferns to classify the sliding windows in the images by some chosen features, then use non-maximal-suppression to detect the text regions. Shivakumara et al. [11] segment an image to a number of blocks and detect text blocks by applying several edge descriptors in different block contrast. Li et al. [6] apply a stroke filter to extract feature maps from images, and then classify the feature maps in a sliding window fashion to detect text regions. On the other hand, connected component (CC)-based approach is another type of region-based approach, which extracts regions from images and uses a set of rules to filter out non-text regions. Following this line of researches, Jung et al. [7] apply a stroke filter for the canny edge map of images, and generate CC regions to detect text regions with several additional features. Epshtein et al. [8] propose a CC extractor named stroke width transform, which is generated by shooting rays from canny edges in the gradient direction, and filter out non-text regions by geometric constraints. Shivakumara et al. [13] filter the input images with Fourier-Laplacian, and compute the text string straightness and edge density to exclude non-text regions.

Texture-based approaches detect texts by their special texture structures and usually use machine learning methods to distinguish texts from background by extracting certain features. As a typical example, Chen et al. [9] design several weak classifiers by using joint probabilities for feature responses and use Adaboost machine learning algorithm to build a strong classifier for detecting texts. Ye et al. [10] use multi-scale wavelet transform to extract features and an SVM classifier is applied to identify text lines from scene images.

Recently deep learning based text detection algorithms have been ever more reported. Deep learning based approaches train a deep network to extract features in replace of the manually designed feature extractors, which are hard to optimize for text detection. Convolutional neural network (CNN) is one of the most popular deep learning models for text detection. The work in [18] trains a five-layer CNN model to detect text regions in natural images by using a supervised learning method. Huang et al. [19] also train a CNN model with two convolutional layers to detect text regions in natural images. The first convolutional layer is pre-trained with an SVM classifier. Maximally Stable Extremal Region (MSER) is used as a candidate text region extractor to reduce the number of background regions before the CNN model.

III. The Proposed Text Detection Algorithm

A. Overview of the proposed algorithm

The proposed Chinese scene text detection algorithm consists of three parts: image patches extraction, CNN based classifier and text line formation method. The CNN based classifier is the core in the proposed algorithm. The flowchart of the algorithm is shown in Fig.1.

The functionality of image patch extraction model is to extract patches from scene images, in which a multi-scale sliding window method is used to guarantee all the texts in the image can be detected with full range of text scales.

The functionality of the CNN based classifier is to classify the candidate text patches obtained from the image patch extraction model with a 5-layer CNN model and a linear classifier. The overview of the CNN model is shown in Fig.2. The first convolutional layer of the CNN is pre-trained by convolutional sparse auto-encoder (CSAE), which is an unsupervised learning method designed for CNN, to extract Chinese text features effectively. The CSAE is to be introduced with more details in Section III.B. The second convolutional layer is replaced by a text structure component detector layer to enhance the accuracy and uniqueness of feature description, which can extract different text structure components in different ways. The text structure component detector layer is presented with more details in Section III.C. The extracted features of text structure component detector layer are input to a spatial pyramid layer to generate scale property, which enhances the scale invariability of the CNN model and has advantages in detecting texts with various sizes. A multi-input-layer deep belief network (DBN) is designed for analyzing the features with properties, which is used as the fully connected layer in the CNN model. The design of spatial pyramid and multi-input-layer DBN is described in Section III.D.

The functionality of the text line formation method is to merge candidate text patches to text lines based on the scale information and several other geometric and heuristic rules.

B. Convolutional Sparse Auto-Encoder (CSAE)

CNN was first introduced in 1980 [21] and becomes one of the most popular deep learning models. In 1990s, following the discovery of human visual mechanisms, local visual field is designed to make CNN model deep and robust. In a standard CNN structure, convolutional layers and pooling layers are connected one by one and trained by supervised learning method with labeled data. CNN is usually used as a strong feature extractor and has achieved great success in image processing fields. The feature extracting ability of CNN is highly correlated with the quantity of training data. However, as the research reports about Chinese text detection are very few, the quantity of labeled Chinese text data is not adequate for supervised learning method. Recently some works [22], [23], [24] introduced several unsupervised learning methods to train CNN with unlabeled data. The features extracted by unsupervised learning CNN have better performance in applications. However, those unsupervised learning methods cannot extract Chinese character features effectively, because Chinese characters are much more abstract than other natural objects. Thus, we need to specifically design unsupervised learning methods for Chinese texts, which is introduced below.

Convolutional layer determines the ability of the CNN model to extract useful features from image data, which is suitable for processing image data due to its convolutional operation. It is the most important part in a CNN model. A single convolutional layer is formed as follows:

$$f(x; W, b) = h = \{h_k\}_{k=1\ldots n}, \tag{1}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2016.2625259, IEEE Transactions on Multimedia

$$h_k = \tanh(x \otimes W_k + b_k), \tag{2}$$

where $x \in \mathbb{R}^{p\times q\times q}$ is an input data matrix. $W \in \mathbb{R}^{n\times p\times m\times m}$ is a set of filters where each filter is represented by $W_k \in \mathbb{R}^{p\times m\times m}$. $b \in \mathbb{R}^n$ is a bias for each filter output. $\otimes$ operator represents convolution operator that applies on a single input and a single filter. The output $h \in \mathbb{R}^{n\times q-m+1\times q-m+1}$ is a set of feature maps extracted by the convolutional layer.

Sparse coding is a popular unsupervised learning method used in many deep learning models. The general formulation of sparse coding is a reconstruction model with a sparse penalty term:

$$z^* = \arg\min_z \|x - Dz\|_2^2 + \lambda s(z), \tag{3}$$

where z^* is the optimal sparse representation corresponding to the input $x \in \mathbb{R}^m$ and the coefficients $z \in \mathbb{R}^n$, $D \in \mathbb{R}^{m\times n}$ is an overcomplete dictionary ($m > n$), $s(.)$ is a sparse penalty function and λ is a penalty parameter. Here we use the $\|.\|_1$ norm penalty, which has the same weight on all the elements in z. The aim is to minimize the function (3) to obtain the optimal z^* .

The CSAE merges a single convolutional layer and the sparse coding method, which enables more effective unsupervised training for CNN models. Considering the functions of convolutional layer and sparse coding, the optimization function of CSAE is defined as follows:

$$h^* = \arg\min_{D,h} \|x - \sum_k D_k \otimes h_k\|_2^2 + \lambda\|h\|_1, \tag{4}$$

where h^* is the optimal feature map correspond to the input $x \in \mathbb{R}^{p\times q\times q}$ and convolutional feature map $h \in \mathbb{R}^{n\times q-m+1\times q-m+1}$, D is a dictionary of filters with the same size as W, λ is a penalty parameter. Note that function (4) contains two variables D and h, so one variable needs to be fixed when optimizing the other variable. We first optimize the sparse feature map h using the FISTA method proposed in [25] with fixed D. Then the stochastic gradient decent (SGD) method is used to update the dictionary D for an efficient estimation of the gradient with the optimal h. Finally the convolutional parameters W and b are optimized using SGD in the following function:

$$(W^*, b^*) = \arg\min_{W,b} \|h^* - f(x; W, b)\|_2^2, \tag{5}$$

where h^* is the optimal feature map.

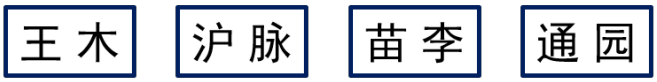
In CSAE, the convolutional parameters are updated more than once by one set of training samples as the feature maps of Chinese characters are so complex that the parameters need more updating to extract them accurately. The optimization goal is defined as follows:

$$\|h^* - f(x; W, b)\|_2^2 \le \theta \tag{6}$$

where θ is the parameter of optimization goal, which decreases as the iteration of CSAE increases. The optimization procedure is sketched in Algorithm 1.

Algorithm 1 Convolutional Sparse Auto-Encoder

function $CSAE(x, D, P, \{\lambda, \beta\}, \{W, b\}, \eta)$
  Initialize $:z = \emptyset$, D, W and b randomly
  repeat
    Minimize function 4 wrt h using FISTA.
    Update D using SGD in function (4).
    repeat
      Update W using SGD in function (5).
      Update b using SGD in function (5).
    until function (6) is satisfied
  until convergence
  Return $\{D, W, b\}$
end function



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

Fig. 3: (a) Left-right structure. (b) Top-bottom structure. (c) Inner-outer structure. (d) Single character.

C. Text structure component detector (TSCD)

1) Analysis of text structure feature extraction: Chinese characters are a kind of pictographs, which contain a large number of radicals and structures. To detect Chinese texts, an efficient method is to analyze Chinese character structure, which is the most remarkable feature of Chinese characters. Chinese character structure is abstracted form natural objects. After a long time use of the Chinese characters, their structures evolved to be more and more abstracted. Modern Chinese character structures are largely different from natural object structures. In [26], Chinese character structures are classified into four basic types: left-right structure, top-bottom structure, inner-outer structure and single character (examples are shown in Fig.3). There are many complex Chinese character structures based on the basic structures, such as top-middle-bottom structure. Chinese character structure component is the most basic constituent element of Chinese characters. Each Chinese character is constituted by one or more structure components. Therefore, the Chinese character structure component is considered as one of the most important features for Chinese text detection and recognition.

The large quantity and valid formation of Chinese character structure components can make a big difference among the structure component features. However, CNN model has difficulty in learning many features with large difference in

Fig. 4: Visualized features in a single convolutional layer.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2016.2625259, IEEE Transactions on Multimedia

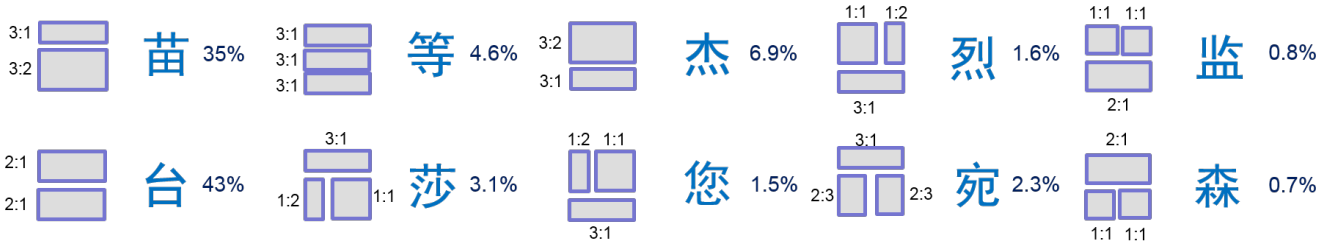


Fig. 5: The statistical result of top-bottom structure characters.

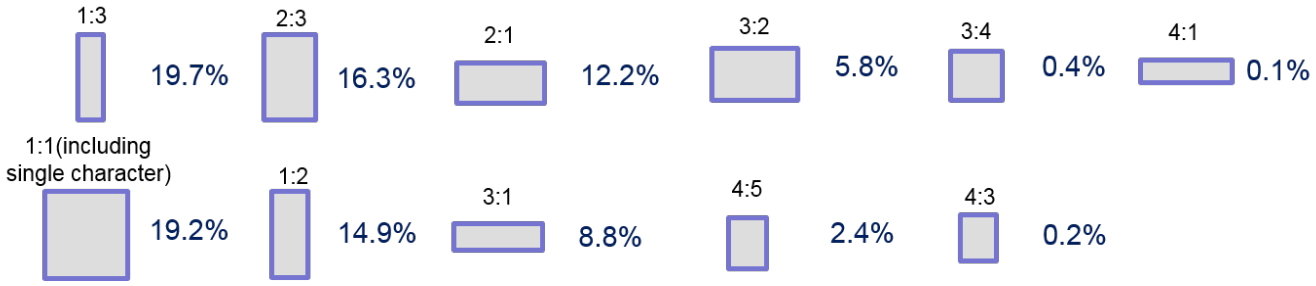


Fig. 6: The statistical result of Chinese character structure component aspect ratio types.

one convolutional layer. The learning methods of the convolutional features in the same layer are the same. The final difference of the convolutional features is determined by the initial parameter values. In many cases, due to the strong learning ability of CNN, the learned convolutional features have big difference even with similar initial parameter values. However, in one convolutional layer, structure component features are too valid to be learned with the initial parameter values. In Fig.4, a set of features in a single convolutional layer are visualized. It shows that each convolutional feature corresponds to one type of image features. It can be observed that some convolutional features are very similar. With larger number of convolutional features, the similar cases occur more frequently. Therefore, a convolutional layer needs to be very large to extract most Chinese text structure component features as the features are large in quantity and valid in formation. An efficient CNN model demands more initial differences rather than the initial parameter values alone to extract Chinese text structure component features.

2) Design of text structure component detector layer: In order to initialize a convolutional layer properly for Chinese text structure component features, we analyze the structure components of some commonly used Chinese characters. The work in [27] studies the utility of Chinese characters. They propose a Chinese character utility function based on the fitted model of character occurrence rate, which is presented below:

$$f(n) = \frac{1}{748.814^{0.487}\Gamma(0.487)} n^{-0.512} e^{-n/748.814}, \quad (7)$$

where $f(n)$ is the utility of the n^{th} most commonly-used Chinese character, and $\Gamma(x)$ is the gamma function. We analyze the structure components of the most commonly-used 1290 Chinese characters because the utility drops below 10^{-4} when n is larger than 1290. Among the most commonly-used characters, 46% of the characters are formed with left-right structure, 26% are with top-bottom structure, 11% are with inner-outer structure and 17% are with single character. Each basic structure type can be further divided into several sub-structures based on the character statistical analysis. The statistical result of top-bottom structure characters is shown in Fig.5. More than 95% top-bottom structure characters can be classified into the 10 sub-structures as shown in Fig.5. It is noted that although the character structures have diverse forms and many types of components, the aspect ratios of the structure components are highly clustered. There are three main aspect ratio types of the structure components with the top-bottom structure characters: 3:1, 3:2 and 2:1. There are also three secondary aspect ratio types in the top-bottom structure characters: 1:2, 1:1 and 2:3. Thus, the aspect ratio type is an important property of character structure component, which makes the structure components easier to be distinguished and classified into several groups. The aspect ratio type is therefore used as an initial difference to enable the CNN model to extract more Chinese text structure features with acceptable complexity.

A TSCD is a convolutional feature extractor for the character structure components with a particular aspect ratio type. In the TSCD the aspect ratio of the normal convolutional window with an aspect ratio of 1:1 is adjusted to be the same aspect ratio type of target character structure components by fixing the length of the longer edge. For example, a TSCD adjusts the aspect ratio of convolutional window to 2:1 to detect character structure components whose aspect ratio is also 2:1. The convolutional window determines which part of the feature map the convolutional operation will be performed on. A convolution window with a 2:1 aspect ratio uses all the information of the structure components with a 2:1 aspect ratio and less information of structure components is used with other aspect ratio. Thus, the TSCD is sensitive to the structure

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2016.2625259, IEEE Transactions on Multimedia

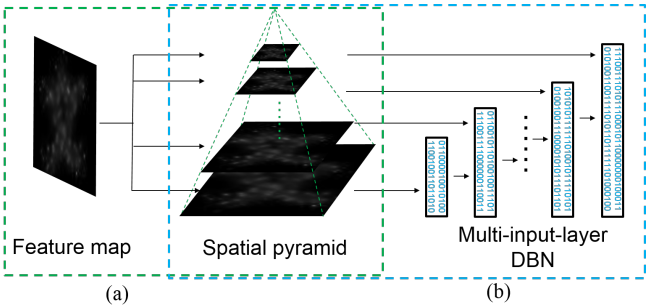


Fig. 7: (a) The structure of spatial pyramid layer. (b) The structure of multi-input-layer DBN.

components with a 2:1 aspect ratio. In this way, TSCDs for extracting features of the structure components with different aspect ratio can be designed. To ensure the completeness of the text structure component features, we analyze the character structure components of all the most commonly-used 1290 Chinese characters focusing on their aspect ratios. It is found from the statistical result (as shown in Fig.6) that there are 11 aspect ratio types of structure components. The share of the most common eight structure component aspect ratio types is over 99%. The number of text structure component detectors is set up based on the proportion of the structure component aspect ratio types. In the TSCD layer, all the TSCDs extract features in parallel. Therefore the quantitative distribution of TSCDs corresponds to the quantitative distribution of natural structure components, which ensures most of the Chinese character structure components can be detected in the TSCD layer. The CNN can extract Chinese text structure component features more accurately and comprehensively by using the TSCD layer than normal convolutional layer with the same number of feature maps.

D. Spatial Pyramid Layer (SPL) and Multi-input-layer Deep Belief Network (DBN)

The texts in natural images have many manifestations, including various sizes, fonts and colors. The CNN model has strong invariability in extracting features from natural text regions with different fonts and colors. However, when extracting features from natural text regions with different sizes, the invariability of the CNN model is very weak. Generally, if a text region is small in the input image, or an input image cuts off a small part of one text region, the CNN cannot extract the text feature accurately. In order to generate image patches for every text region with proper size, the most popular method is to extract image patches by multi-scale sliding window algorithm. However, there are two major problems with the multi-scale sliding window algorithm. Firstly, because of the wide range of recognizable text sizes in natural images, a large number of sliding window scales are needed to generate image patches. Thus there are a large number of image patches, which will significantly increase the computational complexity. Secondly, the source image needs to be magnified to generate image patches for small-scale texts, which will also boost the image noise and reduce the accuracy of the extracted text features.

A spatial pyramid layer (SPL) with a structure shown in Fig.7(a), is designed to solve the problem of various text sizes in natural images. Several features with scale properties are generated based on the extracted features without scale properties by the SPL. The spatial pyramid is named after the feature map pyramid in which each layer represents the size of the feature map with one scale type. The scale invariability of the CNN model is enhanced by adding the SPL to generate scale properties for extracted features. Taking the advantages of the SPL, the features of small texts can be extracted accurately from much bigger scale image patches. Because the number of small scale image patches is much larger than the big scale image patches, the computational complexity of CNN is significantly reduced for the small scale image patches are not essential. And the accuracy of small text features is significantly increased as the interference of magnified image noise is decreased. Meanwhile, the text structure component features with scale properties are used to analyze the relationship of the text structure components as the scales of the text structure components in a single character are not always the same.

The text structure component features with scale properties have two feature dimensions: structure component dimension and scale dimension. They are extracted by the text detection CNN with TSCD layer and SPL. The learning ability of a normal fully connected layer, which is usually a deep neural network, is limited in learning complex two dimension features. Deep Belief Network (DBN) is a generative graphical model, which generates a joint probability distribution of observational data and their labels. A DBN can learn both $P(Observation \mid Label)$ and $P(Label \mid Observation)$, while a deep neural network can only learn the latter. DBN is proposed by Hinton in 2006 [28], which is composed of several hidden layers and one visible layer. The hidden layers in the DBN are pre-trained by restricted Boltzmann machines (RBMs) layer by layer. The first layer is pre-trained as a RBM using the input training data as the visible layer. Then the second is pre-trained as a RBM using the output of the first layer as the visible layer. After all the hidden layer are pre-trained in this way, they are connected to the visible layer with a classifier and fine tuned by the training data for classification. The learning ability of DBN is much stronger than the normal fully connected network used in CNN, which has large advantage in learning complex two dimension features. Thus, a DBN based network is used as the full-connect layer in the text detection CNN.

It is noted that the sizes of feature maps with scale properties vary with different scales. The size of the feature maps in the lower spatial pyramid layer is much larger than that in the upper layers. When the feature maps are input to a DBN together, the small scale features have much smaller influence to the output as the large scale features, which may weaken the effects of scale properties. In order to solve the influence problem, we designed a modified DBN model, named multi-input-layer DBN, which is suitable for learning features with different map size. The multi-input-layer DBN learns the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2016.2625259, IEEE Transactions on Multimedia

TABLE I: A description of proposed methods

Methods	Usage in CNN	Modeling
CSAE	Pretraining	Chinese specialized
TSCD layer	Convolutional layer	Chinese specialized
SPL	Down-sampling layer	General
Multi-input-layer DBN	Fully connected layer	General

TABLE II: A description of the various datasets evaluated on

Datasets	Texts	Label	#Train	#Test
ICDAR 2011	English, digits	-	229	255
ICDAR 2013	English, digits	-	229	233
Pan's dataset	Chinese, English, digits	Mixed	248	239
Our dataset	Chinese, English, digits	Separately	194	200

features with scale properties half-jointly, which guarantees the influences of features with different size are similar and retains the correlation of neighbor scale features. The structure of the multi-input-layer DBN is shown in Fig.7(b). The number of hidden layers in a multi-input-layer DBN equals the number of spatial pyramid layers. The feature maps in the lowest spatial pyramid layer are input to the first hidden layer of the multi-input-layer DBN. Besides the first layer, the input of each hidden layer combines the output of its former hidden layer and the features in the corresponding spatial pyramid layer. In order to guarantee the appropriate influences of features with different sizes and retain the correlation of neighbor scale features, the number of hidden units is set as the size of feature maps in the corresponding spatial pyramid layer of its next hidden layer.

E. Text Line Formation Method

The text line formation method merges the candidate text patches selected by the CNN model into text lines. The texts in a text line usually share similar scales while the neighbor background text-alike regions unlikely share similar scales. Thus, a candidate text region to form a text line should contain more candidate text patches with similar scale than those with the average value of the image. The filtered candidate text regions in similar horizons are merged if they satisfy several geometric and heuristic rules such as similar colors and horizontal distances, which is similar to the work in [8]. Finally, the boundary of the text line is delimited by the text line scale as the candidate text patches with larger scale are usually half outside the text line boundary.

IV. Experiments

In this section the proposed Chinese text detection algorithm is evaluated. The algorithm is first evaluated with two multilingual text detection datasets. The first one is proposed by this paper for training and testing the proposed algorithm. The multilingual text detection dataset is a separate-labeled dataset in which different language texts are labeled separately. The second dataset is mix-labeled in which different language texts are labeled without language information [20]. The Chinese text detection performance of the proposed algorithm is evaluated with the separately labeled dataset, while the text detection performance under multiple language is evaluated with the mix-labeled dataset.

It is noted that the proposed algorithm has a number of components, including CSAE, TSCD layer SPL and multi-input-layer DBN. Some of them are specially designed for Chinese texts to improve text detection performance, while some are general and can be used to detect other language texts such as English texts. The functionalities of the major components and their generalities for text detection are shown in Table I. Apart from the overall performance evaluation of the whole Chinese text detection algorithm on Chinese text datasets, it is also interesting to know how the individual algorithm components perform and contribute to the performance improvements, and how the proposed Chinese text detection algorithm can perform on general text dataset for English text detection to have a reasonable comparison to general text detection algorithms. With this performance evaluation in mind, we have an additional set of experiments on evaluating a slightly simplified algorithm in which the components specifically designed for Chinese texts processing are removed and replaced by the general ones. The simplified version of the proposed algorithm is evaluated on the two most recent ICDAR text detection datasets [3], [4], which are the most widely used datasets for evaluation of scene text detection algorithms.

Next we present the proposed multilingual text detection dataset in Section IV.A. In section IV.B, the training method of our text detection deep network is introduced. Then the experiment results with the separately labeled dataset, mix-labeled dataset and two ICDAR datasets are presented and discussed.

A. Multilingual Text Detection Dataset

To enable effective evaluation of the new CNN model based Chinese text detection algorithm, we set up a multilingual text detection dataset for our algorithm training and evaluation, which labels Chinese, English and digits separately for both training set and testing set. The dataset has similar size of the most popular ICDAR text detection datasets. The size and image categories of our dataset are both similar to the most popular ICDAR 2011 text detection dataset and another multilingual text detection dataset set up by Pan et al. in [20]. In the ICDAR datasets text contents and positions are labeled for all text regions in training set and testing set, while Pan's dataset only labels text position for text regions in testing set. We not only label text contents and positions for all text regions in training set and testing set, but also label Chinese, English texts and digits separately, which is suitable for evaluating a text detection algorithm used to detect one specific type of language text. In this way our dataset can be designed to evaluate Chinese text detection algorithm as well as general English text detection algorithm.

The multilingual text detection dataset contains a training set and a testing set. For training text detection artificial network, the training set needs to simulate common application usages, in which various noises, backgrounds, text types, light conditions etc. are all need to be included. Thus we set up the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2016.2625259, IEEE Transactions on Multimedia

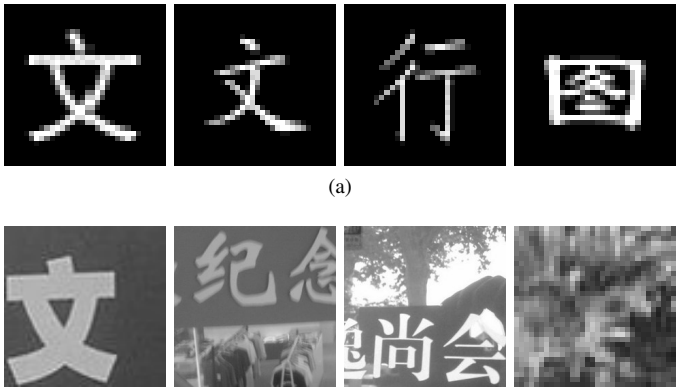


Fig. 8: Examples of training image.(a) The artificial Chinese character examples. (b) The training image patch examples. The first is a simple text sample, the second and third are complex text samples, the last is a background sample.

training set by taking text pictures in open fields with different weathers and indoor with different lights from various objects by a camera. The training set contains 194 scene pictures which have two different sizes: 1632×1224 and 1224×1632. There are 597 text regions in the training set, including 457 Chinese text regions, 102 English text regions and 38 digits regions. For evaluating text detection algorithm, the testing set needs to simulate not only application usages but also some special usages. Thus we set up the testing set by adding the image sources such as processed images from the Internet. The testing set contains 200 scene pictures, which have various sizes because of their various sources. There are 531 text regions in the testing set including 391 Chinese text regions, 115 English text regions and 25 digits regions.

Compared to the wide-used text detection datasets ICDAR 2011, ICDAR 2013 and Pan's dataset, which contains 255, 233 and 239 test images, our dataset contains a bit less (200) test images. However, as the labels in our dataset are more comprehensive, it is expected using our dataset can be more effective than the other datasets above for evaluation of our proposed Chinese text detection algorithm. A summary of the datasets is shown in Table II.

B. Training Details

1) Training Samples: The training samples for the CSAE to pre-train the convolutional layer are set up artificially (examples are shown on Fig.8 (a)). As the number of Chinese characters is very large, we collect all the characters in the training set of the separate-labeled dataset and 50% of the most common Chinese characters which are not included in the training set. The artificial images have white character and black background. The characters are generated by the 15 most popular Chinese fonts such as Sun, Kai and Hei to enrich the descriptions of characters, which enhances the feature learning ability of the CSAE.

The training samples to train the 5-layer CNN model with a text structure component detector layer are extracted from the training set of the separate-labeled dataset (examples are shown on Fig.8 (b)). There are approximately 25,000 training samples used to train the CNN model. The training samples are composed of about 3000 simple text samples, 6000 complex text samples and over 16000 background samples. The percentage of text region in a simple text sample is over 80%, while in a complex text sample is 25% to 66% and in a background sample is less than 10%. The characters in a text sample needs to be brighter than their nearby background, otherwise the sample will be inverted.

As the ICDAR 2011 and ICDAR 2013 share the same training set, the training samples to train the 5-layer CNN model for general text detection are only extracted from the training set of ICDAR 2011 text detection dataset. There are 25,000 training samples are used to train the CNN model. 60% of them is background samples. In the other 40% text samples, simple and complex text samples have similar amount (about 5000) because many characters in the training set are too large to extract complex samples.

2) Deep Network Parameters: The CNN model in our Chinese text detection algorithm has five layers including a convolutional layer, a down-sampling layer, a text structure component detector layer, a spatial pyramid layer and a fully connected layer. The convolutional layer is pre-trained by using the CSAE to learn a dictionary $D \in \mathbb{R}^{n \times m \times m}$ and the convolutional parameters from the training samples for the CSAE with size of 32×32. N is the number of filters in the convolutional layer and $m \times m$ is the size of the convolutional window, which are 64 and 9×9 in our experiment. The learning rates of the dictionary and the convolutional parameters are different for their different convergence speeds, which are 1.5×10^{-3} and 5×10^{-4}, respectively. The maximum times of updating the convolutional parameters in an epoch are set to 50 to avoid overfitting. The down-sampling layer is max-pooling with pool-size of 2×2. The text structure component detector layer is composed of eight text structure component detectors corresponding to most common eight structure component aspect ratio types. The total number of the filters in the text structure component detector layer is 256, which is distributed to the eight text structure component detectors based on their proportions of the total text structure components (as shown in Fig.6). The spatial pyramid layer has three scale layers, whose pool-sizes are 2×2, 4×4 and 8×8, to generate features with scale properties. The output feature maps are fully connected to a 1024 dimension feature vector by the fully connected layer, which is a three-input-layer DBN. The learning rate for pre-training the three-input-layer DBN is 1×10^{-4}. The final output of the CNN is input to a softmax classifier to learn the parameters in the model by using the stochastic gradient decent (SGD) method from the train samples with size of 64×64 after the convolutional layer is pre-trained. The learning rate for all the parameters is 1×10^{-3}. Totally there are approximately 32 thousand convolution parameters and 132 million fully connected parameters in our Chinese text detection CNN model.

The general text detection CNN model has five layers including two convolutional layer, a down-sampling layer, a spatial pyramid layer and a fully connected layer. The structure of this model is the same as the Chinese text detection model

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2016.2625259, IEEE Transactions on Multimedia

expect the following two differences. First, the TSCD layer is replaced by a normal convolutional layer, which has 256 filters with 9x9 window size. Second, no convolutional layer in the general text detection model is pertained. Totally there are approximately 48 thousand convolution parameters and 81 million fully connected parameters in our general text detection CNN model.

3) Training Steps: First, the convolutional layer is pre-trained by the CSAE unsupervised learning method using the artificial training samples. The learning processing stops when the average reconstruction error in function 4 decreases to less than 1×10^{-5} and the average network error in function 5 decreases to less than 1×10^{-7}, which guarantees the accuracy of trained parameters and controls the computational complexity.

Second, the text detection CNN model is trained by replacing the multi-input-layer DBN with a normal two-layer fully connected DNN after the convolutional layer is pre-trained. The parameters in the text structure component detector layer are randomly initialized. The simplified text detection CNN model is trained 500 rounds by SGD method using the natural training samples to generate the multi-scale feature maps for pre-training the multi-input-layer DBN.

Third, the multi-input-layer DBN is pre-trained using the multi-scale feature maps generated in the former step. The unit number of each hidden layer is set as the input feature map size of its next hidden layer to guarantee the appropriate influences of features with different sizes and retain the correlation of neighbor scale features. Every hidden layer are pre-trained 1000 rounds.

Finally, the whole text detection CNN model is fine-tuned. The convolutional layer and text structure component detector layer are initialized by the parameters trained in step 2 and the multi-input-layer DBN is initialized by the parameters pre-trained in step 3. The text detection CNN model is fine-tuned 300 rounds by SGD method using the natural training samples to improve the classification accuracy.

C. Experimental Results on the Separate-Labeled Dataset

The proposed text detection algorithm is designed for detecting Chinese text regions in scene image. Most text detection evaluation methods, including the ICDAR evaluation methods, evaluate all-language text detection performance. In order to evaluate the Chinese text detection performance of our proposed algorithm on the separately-labeled dataset, the text detection evaluation method in ICDAR 2011 [3] is modified. The modified evaluation method is also composed of a two-part measurement precision p and recall r and an overall measurement $f-measure$:

$$p = \frac{\sum_i^N \sum_j^{|D^i|} M_D(D_j^i, G^i)}{\sum_i^N |D^i|}, \tag{8}$$

$$r = \frac{\sum_i^N \sum_j^{|Gc^i|} M_{Gc}(Gc_j^i, D^i)}{\sum_i^N |Gc^i|}, \tag{9}$$

$$f-measure = \frac{1}{\frac{\alpha}{p} + \frac{1-\alpha}{r}}, \tag{10}$$

TABLE III: The peculiarities of the CNN models in this paper

Models	CSAE	TSCD	SPL	Multi-input-layer DBN
CNN	○	○	○	○
CNN-C	●	○	○	○
CNN-T	○	●	○	○
CNN-CT	●	●	○	○
CNN-CTS	●	●	●	○
CNN-CTSD	●	●	●	●
CNN-S	○	○	●	○
CNN-SD	○	○	●	●

● donates the method is applied and ○ donates the method is not

where N is the total number of images in a dataset. $|D^i|$ and $|Gc^i|$ are the number detection and Chinese ground truth regions in the i^{th} image. α represents the relative weight between the two measures. In our evaluation α is typically set to 0.5, which gives equal weight to precision and recall.

$M_D(D_j^i, G^i)$ and $M_{Gc}(Gc_j^i, D^i)$ are the matching scores for detection region D_j and Chinese ground truth region Gc_j. They are described in function 11 and 12,

$$Match_{Gc}(Gc_j, D) = \begin{cases} 1 & \text{if } Gc_j \text{ matches one } d \\ 0 & \text{if } Gc_j \text{ matches no } d \\ f_{sc}(k) & \text{if } Gc_j \text{ matches some}(k)\ d \end{cases} \tag{11}$$

$$Match_D(D_j, G) = \begin{cases} 1 & \text{if } D_j \text{ matches against one } g_c \\ 0 & \text{if } D_j \text{ matches against no } g \\ f_{scn}(l) & \text{if } D_j \text{ matches against one } g_{uc} \\ f_{sc}(k) & \text{if } D_j \text{ matches some}(k)\ g, \text{ one is } g_c \\ f_{scn}(l) \cdot f_{sc}(k) & \text{if } D_j \text{ matches some}(k)\ g, \text{ none is } g_c \end{cases} \tag{12}$$

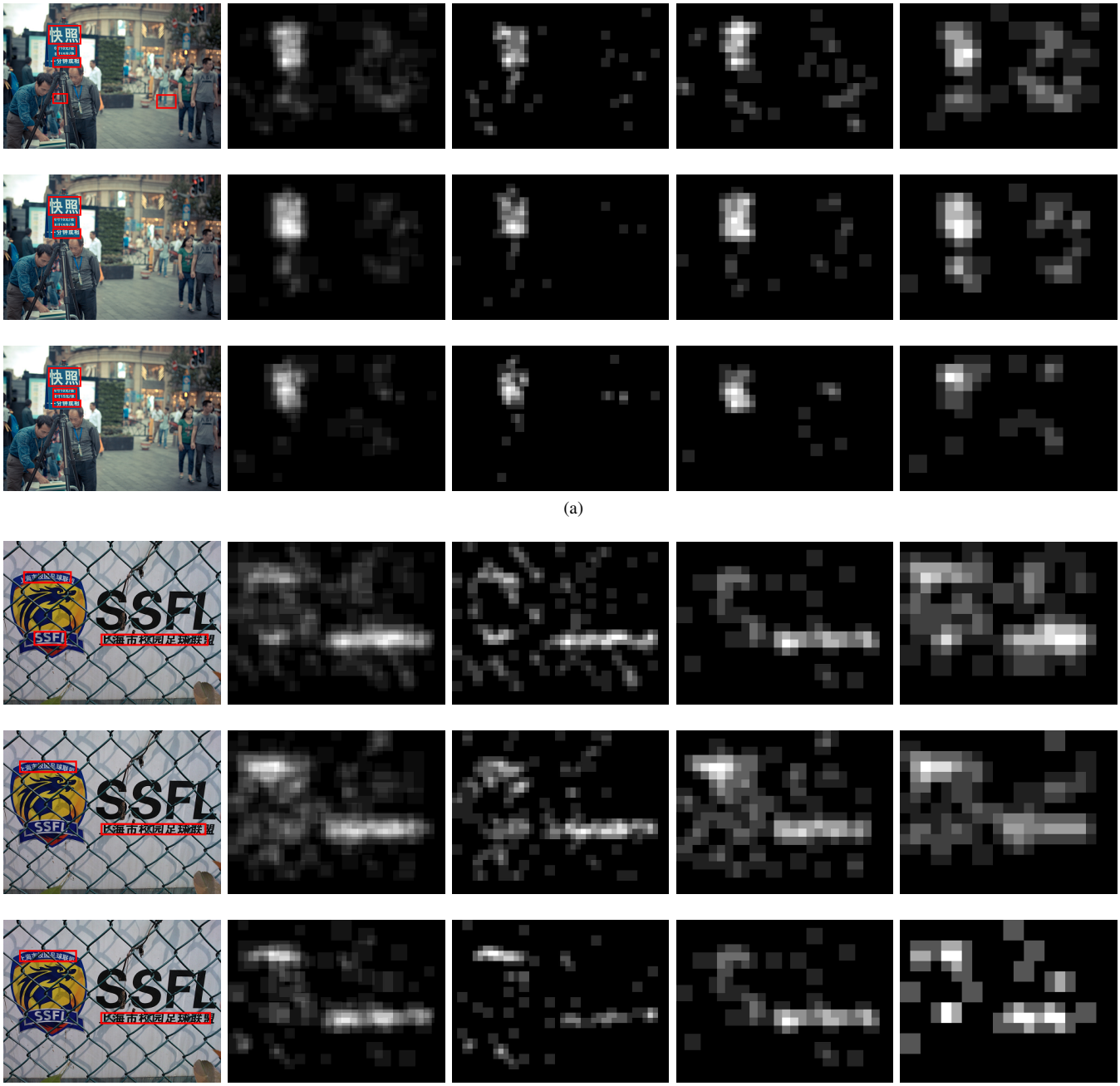
where d represents one or some of the regions in D and g represents one or some of the regions in G. g_c represents one or some of the regions in G_c and g_{uc} represents one or some of the regions in G but not in G_c. $f_{sc}(k)$ is a punishment function for matching against more than one text regions. In our experiment we set it equal to the punishment function of the ICDAR 2011 text detection evaluation method. $f_{scn}(l)$ is a punishment function for matching against non-Chinese text regions. In our experiment we set it to a constant value of 0.8 for both English and digits regions.

TABLE IV: **Text detection results with different models.**

	$precision$	$recall$	$f-measure$
CNN	0.73	0.74	0.73
CNN-C	0.79	0.78	0.78
CNN-T	0.81	0.76	0.78
CNN-CT	0.85	0.81	0.83
CNN-CTS	0.82	0.80	0.81
CNN-CTSD	0.86	0.83	0.84

Table IV presents the evaluation results of the proposed Chinese text detection algorithm with different CNN models. The proposed CNN model with all the previously introduced components (named CNN-CTSD) is compared to the models with only partial components. A normal CNN model has five

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2016.2625259, IEEE Transactions on Multimedia



(a)

(b)

Fig. 9: The image rows of both (a) and (b) from up to bottom are the results of text detection algorithm based on CNN-C, CNN-CT and CNN-CTSD. In each image row, the images from left to right are the text detection result, origin intermediate result map, small scale intermediate result map, middle scale intermediate result map and large scale intermediate result map.

layers in the following order: a convolution layer, a down-sampling layer, a convolution layer, a down-sampling layer, a fully connected layer. The CNN models are configured with various component combinations in the normal CNN model, aiming to identify the potential impact of the different components. The major components included in these CNN models are listed in Table III. Note that if the CSAE component is used, the first convolutional layer is pre-trained. The second convolutional will be replaced by the TSCD layer if it is applied. The SPL is in replacement of the second down-sampling layer. And the multi-input-layer DBN is used as a more effective fully connected layer.

In the evaluations of the proposed text detection algorithm with different CNN models, the proposed Chinese text detection CNN model achieves the best performance of precision 0.86 and recall 0.83. The experiment results show that using CSAE to pre-train the CNN model can effectively improve its Chinese text detection accuracy no matter if the TSCD layer is applied. In both CNN and CNN-T based models, the CSAE has similar improvements to precision and recall

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2016.2625259, IEEE Transactions on Multimedia

(in CNN/CNN-T, precision: 6%/4%, recall: 4%/5%). With suitable pre-training, the model can extract more accurate text features, which has balanced effects to precision and recall improvements. It can be noted that the improvements of the text structure component detectors are different to precision and recall. In both CNN and CNN-C based models, the precision has more improvements than the recall (in CNN/CNN-C, precision: 8%/6%, recall: 2%/3%). This is because the text structure component detectors extract more accurate and unique features by detecting the Chinese character structure components in the image. Less effective Chinese character structure component features can be extracted from background regions than Chinese text regions, and the Chinese character structure component features vary significantly between background regions and Chinese text regions. Therefore, the text structure component detector layer has bigger improvement to precision than recall. The performance of CNN-CTS is even worse than CNN-CT (precision is decreased from 0.85 to 0.82 and recall is decreased from 0.81 to 0.80). The results indicate that a normal fully connected DNN is not suitable for the features with scale properties generated by the text structure component detector layer and the spatial pyramid layer. It is because the features with scale properties contain two feature dimensions: structure component dimension and scale dimension while a normal fully connected DNN has limited learning ability to learn such complex features with two feature dimensions. However, the two-dimension features can be learned effectively by the multi-input-layer DBN due to its strong learning ability and input layers designed for every scales. The evaluation of CNN-CTSD shows that both precision (0.86) and recall (0.83) are improved when using multi-input-layer DBN as the fully connected layer.

Fig.9 shows the intermediate result maps, including the origin result map and three main scales maps, of CNN-C, CNN-CT and CNN-CTSD. It can be noted that the result of CNN-CT contains less non-Chinese-text regions than the result of CNN-C, which demonstrates the text structure component detector layer has advantage in detecting Chinese texts than a normal convolutional layer. The scale results of CNN-CTSD contain more Chinese text regions of appropriate scale and less of other scales than the scale results of CNN-CT. For example, the regions corresponding to small texts in the small scale result (the third image) of CNN-CTSD are much brighter (which means the regions are more likely to contain texts) than those corresponding to larger texts. However, in the small scale result of CNN-CT, the difference is much unclear. It can be indicated that the spatial pyramid layer and multi-input-layer DBN can effectively improve the scale invariance of CNN.

D. Experimental Results on the Mix-Labeled Dataset

In order to evaluate our proposed text detection algorithm in wider language environment and compare to more text detection algorithms, the mix-labeled dataset set up by Pan [20] is used as benchmark to evaluate our proposed text detection algorithm. Most of the text regions in the mix-labeled dataset are Chinese text regions, which limits the performance influence. The evaluation method is an all-language text detection evaluation method as described in [20] in this experiment for the text regions are labeled without language information. In this experiment, the proposed Chinese text detection algorithm is based on the CNN-CTSD model.

TABLE V: **Text detection results with different algorithms.**

	$precision$	$recall$	$f-measure$
The proposed algorithm	0.82	0.72	0.77
Pan's algorithm [20]	0.65	0.66	0.65
Yin's algorithm [29]	0.83	0.69	0.75
Liu's algorithm [30]	0.63	0.67	0.65

Table V summarizes the evaluation results of different text detection algorithms. It can be noted that although the dataset contains a number of non-Chinese regions, which has negative effects on the recall measurement, the proposed algorithm achieves the best recall and the state-of-art result. Liu's algorithm is a text detection algorithm designed for detecting Chinese texts presented recently and the best Chinese text detection algorithm on the mix-labeled dataset. The result shows that the proposed algorithm has better performance than Liu's algorithm especially on precision.

E. Experimental Results on ICDAR Datasets

As mentioned before, some components of the proposed Chinese text detection algorithm are generally designed and can be used for general text detection. To assess the effectiveness of these general components and compare them with general text detection algorithms, we evaluate the general text detection algorithm on ICDAR 2011 and 2013 text detection datasets, which are the most commonly used text detection datasets. In this set of experiments, three variants of the proposed algorithm (namely CNN, CNN-S, CNN-SD with only general components) are evaluated.

TABLE VI: **ICDAR 2011 text detection results.**

	$precision$	$recall$	$f-measure$
Proposed CNN	0.74	0.64	0.69
Proposed CNN-S	0.79	0.66	0.72
Proposed CNN-SD	0.78	0.67	0.72
Zhang's algorithm [31]	0.84	0.76	0.80
Huang's algorithm [19]	0.88	0.71	0.78
Yao's algorithm [32]	0.82	0.66	0.73
Tsai's algorithm [33]	0.73	0.66	0.69
Neumann's algorithm [34]	0.73	0.65	0.65

TABLE VII: **ICDAR 2013 text detection results.**

	$precision$	$recall$	$f-measure$
Proposed CNN	0.76	0.63	0.69
Proposed CNN-S	0.81	0.66	0.73
Proposed CNN-SD	0.81	0.67	0.73
Zhang's algorithm [31]	0.88	0.74	0.80
Yin's algorithm [4]	0.88	0.66	0.76
Neumann's algorithm [35]	0.88	0.65	0.74
Bai's algorithm [36]	0.79	0.68	0.73

Table VI and VII summarize the evaluation results of different text detection algorithms in ICDAR 2011 and 2013 datasets. It can be observed that the performance of CNN-SD is similar to CNN-S on the ICDAR datasets, which indicates

that the multi-input-layer DBN has no obvious advantage in analyzing the features with only scale properties than a normal fully connected DNN. It is also noted that the state-of-the-arts (e.g., Zhang's algorithm) perform better than our proposed general text detection algorithm (with partial components). The results are not surprising as the leading algorithms all have specialized modelings for English text detection while our proposed one does not have. For example, Zhang et al. [31] design a symmetry detector to extract symmetry features, which are special for English characters. Huang et al. [19] propose an error-connected component splitting method, which is special for English texts, to improve text detection performance. However, compared to the other algorithms that consider only text generality, our algorithm achieves either similar or better results. Thus, we believe the general components in our proposed text detection algorithm are performing well for detection of both English and Chinese texts.

V. Conclusion

In this paper, we present a novel text detection algorithm for Chinese texts based on CNN, which contains a text structure component detector layer, a spatial pyramid layer and a multi-input-layer deep belief network (DBN). The CNN model is pre-trained via a convolutional sparse auto-encoder (CSAE) in an unsupervised way to help extracting complex Chinese text features from natural images and enlarging the training set. The text structure component detector (TSCD) layer, which contains several text structure component detectors, is specifically designed for extracting Chinese text structure features. Each of the text structure component detectors is designed to extract the unique features of certain types of Chinese character structure components. The spatial pyramid layer is then introduced to enhance the scale invariability of the CNN model by generating features with scale properties. In order to learn the text structure component features with scale properties, a multi-input-layer DBN is used as the fully connected layer. The multi-input-layer DBN ensures the features from multiple scales are comparable by inputting different scale features to different hidden layers. Experimental results demonstrate that the proposed algorithm is effective in Chinese scene text detection and significantly outperforms the existing algorithms. The pre-trained CNN model has advantages in extracting complex Chinese text features. It is also observed that the unique Chinese character structure component features extracted by the TSCD layer are suitable for identifying text regions. The text structure component features with scale properties, which are generated by the spatial pyramid layer for different scale texts, can be learned effectively by the multi-input-layer DBN, which has strong learning ability with input layers designed for each scale.

Acknowledgment

This work was supported in part by the National Key Research and Development Program of China (2016YFB1001003), National Natural Science Foundation of China (61521062, 61527804), Shanghai Science and Technology Committees of Scientific Research Project (Grant No.14XD1402100, 15JC1401700), and the 111 Program (B07022).

References

[1] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "Icdar 2003 robust reading competitions," in *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*. IEEE, 2003, p. 682.

[2] S. M. Lucas, "Icdar 2005 text locating competition results," in *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*. IEEE, 2005, pp. 80–84.

[3] A. Shahab, F. Shafait, and A. Dengel, "Icdar 2011 robust reading competition challenge 2: Reading text in scene images," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 1491–1496.

[4] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, L. P. de las Heras *et al.*, "Icdar 2013 robust reading competition," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 1484–1493.

[5] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1457–1464.

[6] X. Li, W. Wang, S. Jiang, Q. Huang, and W. Gao, "Fast and effective text detection," in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*. IEEE, 2008, pp. 969–972.

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

[8] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2963–2970.

[9] X. Chen and A. L. Yuille, "Detecting and reading text in natural scenes," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2. IEEE, 2004, pp. II–366.

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

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

[12] Y. Liu, Y. Song, Y. Zhang, and Q. Meng, "A novel multi-oriented chinese text extraction approach from videos," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 1355–1359.

[13] P. Shivakumara, T. Q. Phan, and C. L. Tan, "A laplacian approach to multi-oriented text detection in video," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 2, pp. 412–419, 2011.

[14] X. Liu and W. Wang, "Robustly extracting captions in videos based on stroke-like edges and spatio-temporal analysis," *Multimedia, IEEE Transactions on*, vol. 14, no. 2, pp. 482–489, 2012.

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

[16] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, "Accurate and robust text detection: a step-in for text retrieval in natural scene images," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2013, pp. 1091–1092.

[17] K. L. Bouman, G. Abdollahian, M. Boutin, and E. J. Delp, "A low complexity sign detection and text localization method for mobile applications," *Multimedia, IEEE Transactions on*, vol. 13, no. 5, pp. 922–934, 2011.

[18] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 3304–3308.

[19] W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolution neural network induced mser trees," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 497–511.

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

[21] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[22] R. Chalasani, J. C. Principe, and N. Ramakrishnan, "A fast proximal method for convolutional sparse coding," in *Neural Networks (IJCNN), The 2013 International Joint Conference on*. IEEE, 2013, pp. 1–5.

[23] H. Bristow, A. Eriksson, and S. Lucey, "Fast convolutional sparse coding," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 391–398.

[24] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 3626–3633.

[25] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[26] 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 *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 1044–1048.

[27] X. Zhou and Y. Li, "A research of chinese character utility function (in chinese)," in *Linguistic Research*, 2009.

[28] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[29] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, "Robust text detection in natural scene images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 5, pp. 970–983, 2014.

[30] X. Liu, Z. Lu, J. Li, and W. Jiang, "Detection and segmentation text from natural scene images based on graph model," *WSEAS Transactions On Signal Processing*, 2014.

[31] Z. Zhang, W. Shen, C. Yao, and X. Bai, "Symmetry-based text line detection in natural scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2558–2567.

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

[33] S. Tsai, V. Parameswaran, J. Berclaz, R. Vedantham, R. Grzeszczuk, and B. Girod, "Design of a text detection system via hypothesis generation and verification," in *Proc. Asian Conf. Comp. Vis*, vol. 12, 2012, pp. 13–37.

[34] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3538–3545.

[35] ——, "On combining multiple segmentations in scene text recognition," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 523–527.

[36] B. Bai, F. Yin, and C. L. Liu, "Scene text localization using gradient local correlation," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 1380–1384.

[37] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," in *Workshop on Deep Learning, NIPS*, 2014.

[38] X. Zhao, K.-H. Lin, Y. Fu, Y. Hu, Y. Liu, and T. S. Huang, "Text from corners: a novel approach to detect text and caption in videos," *Image Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 790–799, 2011.

Xiaohang Ren received the B.S. degree in electronic engineering from Zhejiang University, Hangzhou, China in 2011. She is currently pursuing the Ph.D. degree in the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include text information extraction, deeplearning network and image retrieving.

Yi Zhou received the Ph.D. degree from Shanghai Jiaotong University in 2010 in China. She is working in Computer Science Department of Shanghai Jiaotong University, China. Her major research includes object recognition and big data mining. Her project of Chinese Characters Reconginition is supported by the National Science Foundation .

Jianhua He is a Lecturer at the School of Engineering and Applied Science, Aston University, UK. He received a BEng and MEng degree on Electronic Engineering from Huazhong University of Science and Technology (HUST), China, and a Ph.D. degree on Communications Engineering from Nanyang Technological University, Singapore, in 1995, 1998, and 2002, respectively. He joined HUST in 2001 as an Associate Professor. He worked at University of Bristol from 2004 to 2006 and at University of Essex in 2007. He was a Lecturer at Swansea University, UK from 2007 to 2011. His main research interests include protocol design and modelling for wireless network, Internet of Things and data mining. He has authored or co-authored over 100 technical papers in major international journals and conferences. He is an associated editor of International Journal of Communication Systems, KSII Transactions on Internet and Information Systems, and International Journal of Smart Home, and was associated editor of Wireless Communication and Mobile Computing. He served as TPC co-chair of SNA 2008, ICAIT 2009 and ICAIT2010.

Kai Chen received the Ph.D. degree from Shanghai Jiaotong University in 2003 in China. He work in Institute of Image Communication and Network Engineering, Shanghai Jiaotong University, China. His major research includes information retrieving, object recognition and big data mining. He is the key member of his institute on network engineering research. He has been working for several key nation projects and hosted lots of IAR (Industry-Academia-Research) projects.

Xiaokang Yang 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 Centre 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 of the School of Electronic Information and Electrical Engineering, and the Deputy Director of the Institute of Image Communication and Information Processing with Shanghai Jiao Tong University, Shanghai. He has authored over 200 refereed papers, and has filed 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 Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society, a member of Multimedia Signal Processing Technical Committee of the IEEE Signal Processing Society, the Chair of Multimedia Big Data Interest Group of Multimedia Communications Technical Committee of the IEEE Communication Society. He was a member of Editorial Board of Digital Signal Processing. He is also an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS and the Series Editor of Communications in Computer and Information Science (Springer).

Jun Sun received his B.S. in 1989 from University of Electronic Sciences and technology of China, Chengdu, China, and a Ph.D. degree in 1995 from Shanghai Jiao Tong University, all in electrical engineering. He is currently a professor and Ph.D. advisor of Shanghai Jiao Tong University. In 1996, he was elected as the member of HDTV Technical Executive Experts Group (TEEG) of 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 has published over 50 technical papers in the area of digital television and multimedia communications and received 2nd Prize of National Sci. & Tech. Development Award in 2003, 2008. His research interests include digital television, multimedia communication, and video encoding.