IET Image Processing

Research Article

IET Journals
The Institution of Engineering and Technology

Specific category region proposal network for text detection in natural scene

ISSN 1751-9659
Received on 4th June 2019
Revised 12th January 2020
Accepted on 3rd February 2020
E-First on 9th June 2020
doi: 10.1049/iet-ipr.2019.0652
www.ietdl.org

Yuanhong Zhong[1] ✉, Xinyu Cheng[1], Zhaokun Zhou[1], Shun Zhang[1], Jing Zhang[1], Guan Huang[1]

[1]*School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, People's Republic of China*

✉ *E-mail: zhongyh@cqu.edu.cn*

Abstract: Natural scene text usually carries considerable abstract semantic information, which is closely related to the surrounding environment. Thus, natural scene text detection plays a vital role in image content retrieval and understanding. In this study, the authors propose a novel specific category region proposal network (SCRPN) based on maximally stable extremal regions (MSER) and fully convolutional network (FCN) for natural scene text detection. First, FCN for pixel-level recognition is utilised to obtain the text saliency map and MSER is used to obtain oversegmented regions. Then, the multiple features of oversegmented regions and text saliency map are used for region aggregation. Next, single-linkage clustering method is adopted to cluster the segmentation regions to obtain a hierarchical structure of text region proposals. Finally, for the top-ranking region proposals, SCRPN built an end-to-end pipeline for scene text detection directly. Experiments on street view text and international conference on document analysis and recognition (ICDAR) 2013 have demonstrated the effectiveness of SCRPN for generating the text proposals. SCRPN could work with various two-stage text detection networks; thus, faster region convolutional neural network was used as the text detection framework to evaluate the performance of SCRPN in the ICDAR 2015 and MSRA-TD500 benchmarks. The experimental results confirmed that SCRPN makes text detection more robust in complex scenarios.

1 Introduction

Humans perceive surrounding information mainly through vision. Similarly, vision plays a big role in the world of machine perception. Machine vision is a technique of Computer Science and Electronics field, which is useful for many applications such as change detection in synthetic aperture radar images [1], object detection in images [2, 3], analysis pipeline in biomedical images [4], automatic panoramic in unmanned aerial vehicle (UAV) images [5] etc. Moreover, human interaction largely relies on text information; thus, text understanding is a very important research direction in information perception and decision making [6]. Images usually contain textual descriptions such as street names, road signs, building numbers and product descriptions, which often provide key clues for information perception. Thus, scene text understanding in natural image is extremely useful for these fields, such as the direct perception for autonomous driving [7], the image caption for image retrieval [8, 9], the text recognition for automatic translation [10, 11], the text location and recognition for video content analysis [12, 13] etc.

Text understanding in natural image consists of two sub-tasks: text detection and text recognition. Text detection extracts a minimum area containing a text content and a single background so that text recognition task could avoid interference from complex background. Text detection is also a branch of object detection. It plays a vital role in the process of text understanding. However, it faces many challenges: (i) uneven illumination: the environmental factors of shooting process make images a large difference in brightness. (ii) Different forms of writing: natural scene texts are often diverse in terms of colour, scale, font, position and orientation. (iii) Multi-oriented text arrangement: the direction of texts is various and the angle of shooting process will cause affine distortion in the text blocks. (iv) Background similar to text structure: there may be many objects or buildings similar to the structure of natural scene text. Owing to these complex factors, the problem of scene text detection has great challenges both on theory and practice.

Over the last few years, studies on the scene text detection made a remarkable progress. At first, text detection methods [14–17] mainly relied on low-level character features or stroke detection to detect and identify a single character. Then, the detected initial set of candidate characters was divided into phrases based on the spatial constraints or dictionary constraints to implement the text detection. In general, they were insensitive to multi-oriented text detection and had some environmental limitations.

Recently, more efficient text detection models in complex environments have been proposed, such as maximally stable extremal region-based (MSER) methods [18–21], convolutional neural network (CNN) methods [22–29] and so on. Tian *et al.* [27] proposed a connectionist text proposal network to get accurate text line localisation in a natural scene image. Shi *et al.* [28] proposed a rotation region proposal network, which added rotation factors to classical region candidate network for multi-directional text detection. Ghosh *et al.* [29] built an long short-term memory (LSTM)-based soft visual attention model that learned from convolutional features to detect scene text. These text detection frameworks based on CNN [22, 25, 26] mostly adopted general text region proposal methods such as binarised normed gradient (BING) [30] and Edge Boxes [31] features. They adopted a sliding window method, which has high computational complexity and may not get effective result.

Fig. 1 shows the flowchart of text detection. As shown in this figure, the complete process of text detection consists of two main steps: text region proposals and text box predictions. Although the above methods have improved the performance of text detection, the robustness of text detection frameworks in natural scenes has not been effectively solved. Thus, in this work, we focus on how to get text region proposal efficiently and accurately. Specifically, we propose a novel specific category region proposal network (SCRPN) to increase the robustness of the framework of scene text detection. To improve the robustness of the model, we utilise multi-dimensional features, including a few traditional text features and text saliency map based on deep learning. Finally, SCRPN is converted into an end-to-end pipeline. The proposed method is universal and can be used with various two-stage text detection networks. The main contributions of this paper are summarised as follows:

- Inspired by the tremendous achievements of visual attention mechanism in diverse fields of visual processing, we extract text saliency map from text areas in images and use it as a feature of diversification strategies. The addition of text saliency map enables text detection system to recognise out-of-vocabulary words and obtains more accurate results.
- We improve the parameter constraint of MSER to obtain oversegmented regions with a large amount of overlapping information. An effective ranking strategy that considers multiple input channels and multiple complementary similarity metrics is designed to improve the overall detection recall rate.
- SCRPN algorithm is proposed for text region proposals. Extensive experiments on ICDAR 2015 and MSRA-TD500 datasets confirmed the effectiveness of the proposed method, with the framework of faster region CNN (RCNN) [32].

The rest of this paper is arranged as follows. Section 2 discusses the related work. Section 3 describes the proposed SCRPN and the details of network training. Relevant experimental results of SCRPN are shown in Section 4. Conclusive remarks and future work are given in Section 5.

2 Related work

In recent years, natural scene text detection technology has made great progress, and it gradually developed into three solutions: texture-based methods, connected region-based methods and hybrid methods.

2.1 Texture-based methods

Texture-based methods extract the texture features of sliding window in image and use a classifier to predict the probability that the extracted features contain text in each window. Chen and Yuille [15] used the adaptive boost method to fuse the mean and variance of the local grey of the image and locate the region containing the text according to the merged features. Moreover, the optimised Niblack algorithm was used to perform brutalisation operation to obtain the character candidate region in the original image. Wang *et al.* [16] combined the histogram of oriented gradient feature and random ferns classifier to locate the multi-scale text. These methods could keep promising robustness to environmental noise in the image, but the huge amount of computation caused by the sliding window mechanism greatly increased the complexity of the algorithm.

2.2 Connected region-based methods

Connected region-based methods suppose that pixels in the text region have similar characteristic in some aspects, such as colour, brightness, texture etc. Stroke width transform[17, 33] and MSER [18–20, 34] are two typical connected region-based methods. Liu *et al.* [18] proposed a V-MSER algorithm to extract MSERs from multi-channel for text detection. Yin *et al.* [20] treated each MSER region as a vertex processing in a graph and then converted text detection into the graph partitioning problem. In [20], the MSER-based method achieves a 95.2% recall rate in multi-channel processing. Although MSER-based methods have high ability to detect most of text regions in an image, they are usually used for the extraction of a colour-consistent text region such as a cue card or a licence plate. These methods have the characteristics of less computation and higher efficiency when dealing with text location problem. However, they are very sensitive to the environmental noise that may appear in the image and were prone to have a large number of false detections.

2.3 Hybrid methods

In the process of exploring text detection, the combination of texture feature extraction and connected domain partition has been paid much attention. Hybrid methods combine the excellent characteristics of the two methods and consider the detection of text regions more comprehensively. Huang *et al.* [21] proposed a robust text detection and recognition method by employing MSER and support vector machine. Zhao *et al.* [35] presented a robust hybrid method that used learning-based partial differential equation (PDE) to detect texts in natural scene images. Shi *et al.* [28] combined VGG-16 and LSTM to densely slide 3×3 spatial windows through the conv5 layer of VGG-16 model. Continuous windows in each row were connected by bidirectional LSTM, where the convolution characteristics of each window were used as input for 256-dimensional (256D) bidirectional LSTM. This enabled the model to detect text sequences directly in convolutional graphs and avoid the cost of detecting the model through additional CNN. It can be seen that the idea of combining multiple methods has gradually become a prevailing research direction.

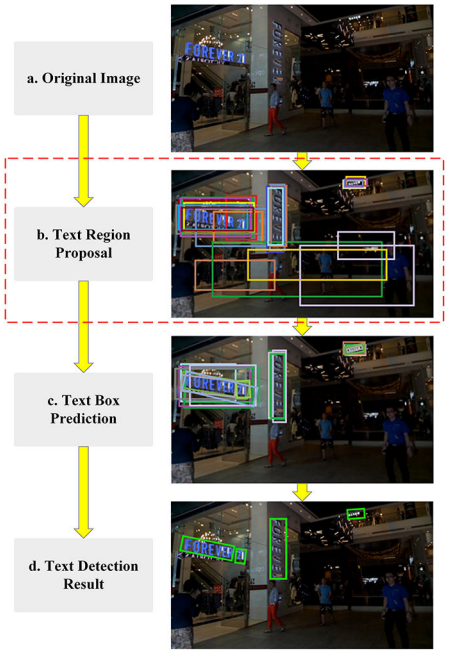

Fig. 1 *Flowchart of text detection*
(a) Original image, ***(b)*** Proposed SCRPN for text region proposals, ***(c)*** Text box prediction, ***(d)*** Text detection result

3 Materials and methods

Fig. 2 shows the framework of text detection, and the proposed method with solid line box is a hybrid method. It includes MSER segment and text saliency map for region generation, weak classifier for region aggregation and proposal sorting for region of interest (ROI) pooling. Details will be delineated as follows.

3.1 Candidate character region generation

As shown in the solid box in Fig. 2, SCRPN is divided into two branches, and the two branches do not interfere with each other. One is the multi-channel region segmentation based on MSER. It could provide a large amount of overlapping region information and does not perform any form of filtering operation on these candidate regions.

The other branch is the network of text saliency map, and it is based on fully convolutional network (FCN). We used FCN to generate candidate saliency regions of the input image. By training FCN in a specific way, the network can capture text saliency regions. Moreover, we will introduce the training details in Section 3.4. It can mimic the human eye's ability to select the region of interest quickly and accurately.

Fig. 3 shows the network of text saliency map. We inherit the first five convolution layers of VGG-16 and replace the fully connected layer with a convolution layer. In this way, the original 1D output vector is converted to a 2D matrix, and the size of original image is restored by deconvolution. The deconvolution layer is composed of a convolutional layer (kernel size is 1×1) and an upsampling layer. Since the ultimate output feature map size is too small, the prediction result which obtained by directly

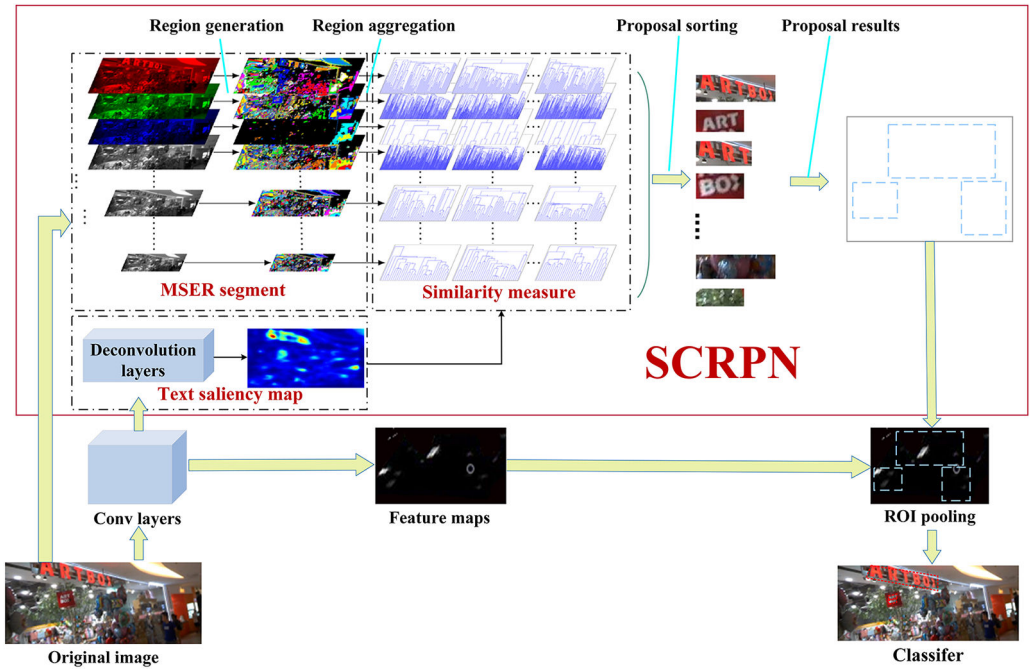

Fig. 2 *Overall architecture of text detection. The solid line box is the proposed SCRPN. SCRPN takes original image and feature maps from faster RCNN as input of MSER segment and text saliency map, respectively. Then, the outputs of MSER and text saliency map are converted into different similarity measures via region aggregation. Next, weak classifiers are used to obtain the possible prediction of each region for proposals sorting. Finally, according to the result of proposals sorting, selection and mapping are carried out in order to obtain (ROI) pooling for text region classification*

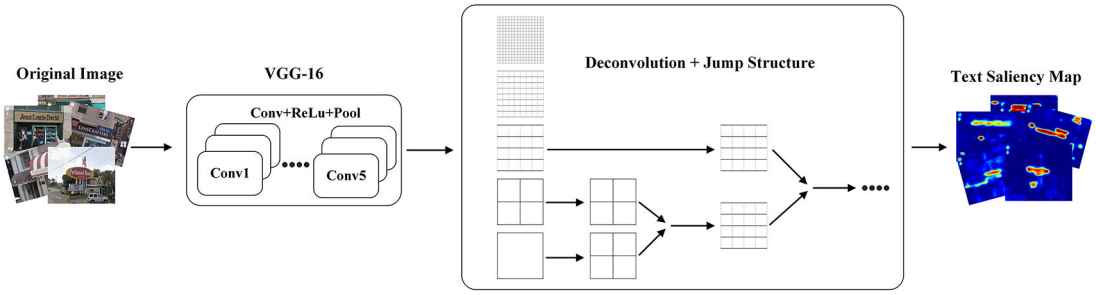

Fig. 3 *Text saliency map structure. Test saliency map can consider the local clues and global context information of image comprehensively, complement each other, thereby enhancing the ability of distinguishing between text and background*

performing a deconvolution operation is very rough. Therefore, we add a jump structure in the form of a stack to fuse the feature maps of diverse stages as shown in Fig. 3. This can thoroughly consider the different scales and different levels of information and combine the regional features that are captured by the shallow network with a deeper abstract features. Finally, the branch can generate a set of candidate regions.

3.2 Candidate character region aggregation

The aggregation process of text candidate regions starts from a set of connected regions, which include MSER segment regions and text saliency map regions, and its aim is to integrate region information of different branches. First, clusters R_c are generated in line with region information. Then, the single-linkage clustering (SLC) is used to iteratively fuse the closest pair of clusters (A, B) until all connected regions are aggregated into one. The SLC can be expressed as below:

$$\min \{d(r_a, r_b): r_a \in A, r_b \in B\}, \tag{1}$$

where r_a, r_b is one region of clusters A and B, respectively, A and B are different clusters of R_c.

The distance measure $d(r_a, r_b)$ is used to describe more abstract similarity relationships between text groupings. It uses a series of complementary features, which have low computational costs, such as the brightness and colour mean of region, stroke width, the brightness and colour mean of external boundary, regional diameter, the mean of boundary gradient, the mean of text saliency map regions etc.

Various complementary features are used independently, and then they are combined with spatial information (coordinates of regional centre). Therefore, the group of interested regions are limited to candidate regions that are spatially close to each other. Next, SCRPN uses the diverse complementary distance metrics $d_i(r_a, r_b)$ for region aggregation. Complementary distance metrics $d_i(r_a, r_b)$ can be expressed as

$$d_i(r_a, r_b) = (f_i(r_a) - f_i(r_b))^2 + (x_a - x_b)^2 + (y_a - y_b)^2, \tag{2}$$

where $i = \{1, 2, \ldots, 8\}$, $\{x_a, y_a\}$ and $\{x_b, y_b\}$ denote centre coordinates of the regions r_a and r_b, respectively, f_i denotes similarity features above. Euclidean distance at the centre of the region is computed to keep the rotation invariant. Consequently, SCRPN can adapt to the complex and changeable outdoor environment.

In addition, we achieve text prior horizontal alignment by adding a parameter $\lambda \in [0, 1]$ to horizontal coordinate items

$$d_i(r_a, r_b) = (f_i(r_a) - f_i(r_b))^2 + \lambda(x_a - x_b)^2 + (y_a - y_b)^2 \tag{3}$$

The more discrepant of aggregation strategy, the more likely to get good quality proposals for a given target text. However, this needs to be done at the cost of increasing the total number of candidate regions. In Section 4, the performance of different combinations is

experimentally analysed to find the best configuration as a compromise between the recall rate and total number of text region proposals. Several diverse strategies that could be combined in different ways are listed below:

- *Diversification of similarity measures*: Different combinations of similarity measurements can be performed by SLC.
- *Diversification of colour channels*: Different colour channels of input images can be used to extract connected domains by MSER.
- *Diversification of spatial pyramid levels*: The three-level spatial pyramid of the input image can be used to extract connected domains by MSER.

The above diversification strategy will generate 96 combinations: a three-level spatial pyramid, four colour channels and eight similarity measurements.

3.3 Candidate character region proposals sorting

In the similarity hierarchy of candidate character regions, each node represents a region proposal. In this case, an effective approach is needed to sort the nodes in the list to pick out the most reliable nodes. Fig. 2 shows SCRPN uses a weak classifier to measure the confidence of the text region proposals and then sorts them according to the confidence that the text region is proposed. Since the classifier needs to assess on each node of the hierarchy, a fast classifier Real AdaBoost [15] with low computational cost is designed as the weak classifier. To train the Real AdaBoost classifier, the coefficient of variation (CV) of the single region feature described above is used as the decision tree streak, such as text saliency map, stroke width, diameter and foreground brightness mean.

3.4 Training and implementation details

We follow the standard practice and explore the faster RCNN model [32] to evaluate the performance of SCRPN. For network training, the training result of VGG-16 in the ImageNet Large Scale Visual Recognition Competition is used to initialise the model parameters. The parameters of network are continuously updated by stochastic gradient descent (SGD). All experiments are made on the same workstation with 32 G RAM, NVIDIA GeForce GTX Titan GPU and Intel Xeon(R) CPU E5–1620 (3.50 GHz). The model is constructed by using the open source Caffe framework [36], and the programming language is Python. The proposed method utilises the strategy in [37] to evaluate all word recognition methods and then perform non-maximum suppression to get the top N text region proposals for ROI pooling.

An end-to-end manner is used to train the network of text saliency map and the ground truth is a binary image, which has the same size as the input image. The pixel with a value of 1 is defined as a positive sample, and the pixel with a value of 0 is a negative sample. In the training phase, we use cross-entropy loss and SGD to train FCN. With the trained model, the network can generate the text saliency map of image. Then, the text saliency map is applied to the aggregation of connected region.

To train a Real AdaBoost classifier, we use the CV of individual region features (e.g. stroke width, diameter, the mean of foreground brightness etc. described in Section 3.2 as decision tree stumps

$$F^i(G) = \frac{\sigma^i}{\mu^i}, \tag{4}$$

where μ^i and σ^i are the mean and standard deviation of the ith feature $\left\{ f^i(r) : r \in G \right\}$ in the specific region group G.

All training samples are from ICDAR 2013 training dataset and ICDAR 2015 training dataset. For each training sample, we utilise all the diversification strategies, which are described in Section 3.2 to create a similarity hierarchy. For each of the 96 hierarchical groups (nodes), we determined the true value of best matching label based on the ratio of the grouping–bounding box on the ground truth. If intersection over union (IoU) > 0.7, the region would be regarded as a positive sample. If IoU < 0.2 and the region does not completely overlap the ground truth, it would be marked as a negative sample. Finally, about 200,000 positive samples and 1,000,000 negative samples are obtained. To balance the training data, 200,000 negative samples are randomly selected.

4 Experiments

We evaluate SCRPN on four text detection benchmarks, named street view text (SVT), ICDAR 2013, ICDAR 2015 and MSRA-TD500. In our experiments, we first verify the efficiency of SCRPN for text region proposal. SVT, ICDAR 2013 and ICDAR 2015 are used for evaluation. SCRPN is general and can work with various text detection networks. Then, we use the faster RCNN [32] to evaluate the performance of SCRPN for natural scene text detection in ICDAR 2015 and MSRA-TD500.

4.1 Benchmark datasets

SVT is a street view text dataset, which is released by Google. It contains about 350 images from Google Street View and has about 904 text regions. The dataset has the characteristics of low resolution, and the text relies on objects that are easy to be searched. It is very suitable for scenic text positioning research.

ICDAR 2013 is the standard dataset used in the scene text detection competition organised by the International Conference on Document Analysis and Recognition (ICDAR) in 2013. The dataset is an improved version of ICDAR 2011 dataset, which fixes some errors. It includes 229 training images and 233 test images.

ICDAR 2015 is the official dataset used by ICDAR in the text detection competition held in 2015, which contains 1000 training charts and 500 test charts. The orientation angles of text in this dataset are different from ICDAR 2013 dataset, which contains horizontally arranged characters. Moreover, because of the smaller texts and blurry images, it is difficult to detect scene text.

MSRA-TD500 contains 300 training images and 200 test images. Image annotations consist of the location and orientation of text instances, and the benchmark can be used to evaluate text detection performance on multi-text instances. What is more, this dataset is extremely challenging as it not only includes both Chinese and English languages, but also is diverse in terms of perspective, size, font, lighting, colour etc.

4.2 SCRPN for text region proposal

The performance of SCRPN with different configurations is evaluated by the recall rate. Table 1 shows the proposal number of text regions for each test picture and the recall rate of different IoU thresholds in several possible combinations.

Diversification strategies include different combinations of colour channels (R, G, B and I), spatial pyramids (P0 is 1:1, P1 is 1:2 and P2 is 1:4) and similarity measurements (D is diameter, F is foreground brightness, B is background intensity, G is gradient, S_{W} is stroke width, Flab is foreground Lab colour, Blab is background Lab colour and S_{FCN} is text salience map).

For scene text detection, we need to get higher-quality proposals, such as IoU > 0.7. On the one hand, because the proposal of IoU > 0.5 may only contain a certain part of the actual text region, this eventually make the detection task complicated. Additionally, a suitable text region proposal does not necessarily need to reach an excessive IoU score, since the description of the 'text object' bounding box label is very vague in some cases. We further analyse the distinctive combination (the combination is shown in bold in Table 1) as a compromise between recall rate and recommended number. This combination of diversification strategies is implemented in subsequent experiments. The selected combination has effective advantages for IoU thresholds of 0.6 and 0.7 compared with other less diversified options. The use of three colour channels [red, green and blue (RGB)] is particularly significant compared with the use of a single colour channel (I). The addition of a secondary space pyramid (P1) does not increase

the computational cost with obtaining a gain. We can see that the increasing of the text saliency map (S_{FCN}) can make the result much better. As analysed in Section 3.1, the text saliency map can be combined with the local cues of the image and the global context information to complement each other, and thereby enhance the discriminating ability.

The performance of the proposed method is also compared with other typical object proposal methods. We use the public code of these methods and their default parameters during the experiment. Table 2 presents the results of the comparison on the SVT dataset and ICDAR 2013 dataset. The graphical performance is shown in Fig. 4.

From Table 2, it can be observed that the proposed method is superior to several other algorithms in the recall rate of text region detection. Since Edge Boxes [31] has achieved valuable results in the end-to-end identification pipeline of scene text, the results compared with it is meaningful. As shown in Fig. 4, for all IoU thresholds, the proposed method provides better area under curve (AUC) features than Edge Boxes. Meanwhile, when the analysis is limited to a relatively small set of proposals (the number of candidate regions is 1000), there is a significant performance improvement.

To evaluate the proposed method in more unconstrained situations, a similar analysis is performed on the ICDAR 2015 dataset as shown in Fig. 5. Although text types in ICDAR 2015 are similar to those in SVT, it contains a large number of non-horizontal and small-sized text instances. In Fig. 5, Edge Boxes has a minor recall rate on the ICDAR 2015, which means that the algorithm is not suitable for detecting non-horizontal and small-sized text blocks.

The proposed method performs roughly the same on the SVT and ICDAR 2015 datasets; hence, it indicates that SCRPN can handle different scene texts more robustly.

Table 1 Different combinations of diversification strategies

Method	#Prop	0.5 IoU	0.6 IoU	0.7 IoU	0.8 IoU	0.9 IoU
P0 + I + D	1605	0.86	0.75	0.53	0.27	0.07
P0 + I + F	1461	0.89	0.79	0.64	0.31	0.09
P0 + I + B	1483	0.83	0.73	0.50	0.23	0.06
P0 + I + S_W	1587	0.80	0.72	0.48	0.20	0.06
P0 + I + S_{FCN}	1607	0.85	0.73	0.52	0.25	0.07
P0 + I + DFBGS$_W$$S_{FCN}$	4588	0.94	0.86	0.72	0.41	0.11
P0 + I + DFBGS$_W$$S_{FCN}$$F_{lab}$$B_{lab}$	5462	0.94	0.86	0.72	0.41	0.11
P0 + RGB + DFBGS$_W$$S_{FCN}$	12,996	0.94	0.91	0.80	0.52	0.20
P0P1 + RGB + DFBGSw	14,663	0.94	0.92	0.83	0.55	0.21
P0P1 + RGB + DFBGS$_W$$S_{FCN}$	16,795	0.95	**0.94**	**0.88**	0.59	0.22
P0P1P2 + RGB + DFBGS$_W$$S_{FCN}$	18,297	0.95	0.94	0.88	0.61	0.25
P0P1P2 + RGBI + DFBGS$_W$$S_{FCN}$	21,663	**0.96**	0.94	0.88	**0.61**	**0.26**

Recall rates are evaluated at different pyramid levels, different colour channels (I and RGB), different distance metrics and text saliency map.
Bold indicates the best results in each evaluation criterion.

Table 2 Comparison of text region proposal methods on SVT dataset and ICDAR 2013 dataset

Method	SVT				ICDAR 2013			
	#Prop	0.5 IoU	0.7 IoU	0.9 IoU	#Prop	0.5 IoU	0.7 IoU	0.9 IoU
BING [30]	2987	0.64	0.09	0.00	2716	0.63	0.08	0.00
Edge Boxes [31]	15,319	**0.94**	0.63	0.04	9554	0.85	0.53	0.08
Randomized prim's (RP) [38]	5620	0.02	0.00	0.00	3393	0.77	0.45	0.08
Geodesic object proposals (GOP) [39]	778	0.53	0.19	0.03	855	0.45	0.18	0.08
SCRPN	16,275	**0.94**	**0.67**	**0.09**	11,628	**0.97**	**0.92**	**0.84**

Bold indicates the best results in each evaluation criterion.

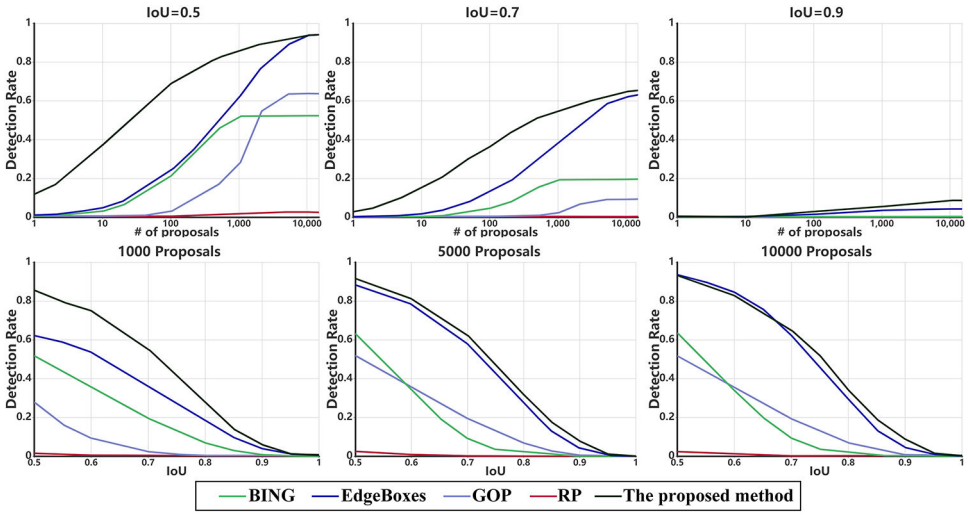

Fig. 4 *Comparison of text region proposal methods on SVT dataset*

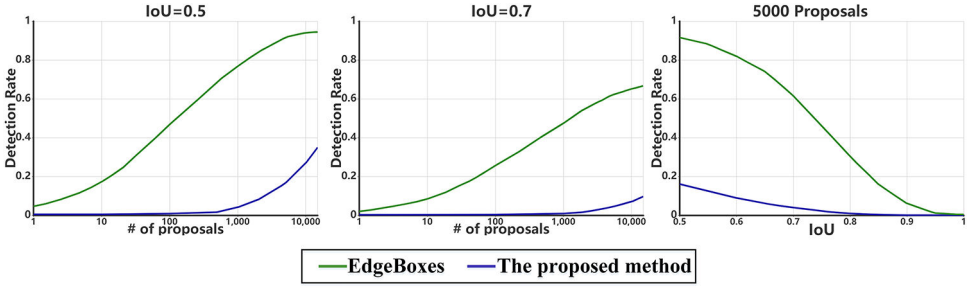

Fig. 5 *Comparison of text region proposal methods on ICDAR 2015 dataset*

Table 3 Comparison of different methods on the ICDAR 2015 dataset

Method	Recall, %	Precision, %	F-measure, %
Rotation region proposal networks (RRPN) [26]	73.23	**82.17**	77.44
SegLink [28]	76.80	73.10	75.00
DMPNet [40]	68.22	73.23	70.64
Connectionist text proposal network (CTPN) [27]	51.56	74.22	60.85
MCLAB_FCN [41]	70.81	43.09	53.58
proposed method	**83.69**	72.81	**77.87**

Bold indicates the best results in each evaluation criterion.

Table 4 Comparison of different methods on MSRA-TD500 dataset

Method	Recall, %	Precision, %	F-measure, %
Efficient and accuracy scene text detector (EAST) [42]	67	**87**	**76**
Zhang *et al.* [43]	67	83	74
Yin *et al.* [44]	63	81	71
Kang *et al.* [45]	62	71	66
Yin *et al.* [20]	61	71	65
RRPN [26]	67	72	69
proposed method	**69**	83	75

Bold indicates the best results in each evaluation criterion.

4.3 SCRPN for text region detection

We evaluate text detection on ICDAR 2015 and MSRA-TD500 datasets with the framework of faster RCNN [1], the evaluation protocols include precision, recall and F-measure. Table 3 shows the performance comparisons of different methods in ICDAR 2015 dataset.

SCRPN is a text region proposal network that comprehensively considers the geometric properties of text objects and the characteristics of different layers of the network. Table 3 shows the recall rate of the proposed method is significantly improved in comparison with other methods. To further evaluate the adaptability of the proposed method, the experiment on the MSRA-TD500 dataset is performed and Table 4 presents the results.

The proposed method achieves a precision of 83%, a recall rate of 69% and an F-measure of 75% in Table 4. We can find that the proposed method can effectively extract multi-language-type text and has special benefits in the recall rate. SCRPN is designed to ensure robustness detection in natural scenes. High recall rate is an indicator for robustness detection. Therefore, we propose a method to make full use of text properties to enhance robustness. Tables 2–4 clearly show a high recall rate. The precision is naturally affected slightly with the recall rate as high as possible. However, only the precision of RRPN [26] and efficient and accurate scene text detector (EAST) [42] obviously outperform SCRPN in ICDAR 2015 dataset and MSRA-TD500 dataset, respectively. Fig. 6 shows some visual detection results for these datasets.

Additionally, we evaluate the time complexity of the method, and it runs at 3 fps. The proposed method includes the operations of the traditional algorithm on the CPU and the CNN on the GPU. The CPU operation which extracts the traditional feature will add to the overall time complexity. We believe that the time complexity of model will be greatly decreased after the traditional algorithm is deployed on the GPU.

5 Conclusion and future work

The text information in the natural scene includes the high-level semantics of image and it has a direct guiding effect on image understanding. Therefore, the extraction technology of natural scene text has broad application prospects in many fields such as image retrieval, image understanding, intelligent transportation and human–computer interaction, thus it has an important research significance in the field of machine vision.

In this paper, we propose SCRPN, which is based on MSER and FCN for natural scene text detection. SCRPN can combine multiple similarity features and cope with various text detection tasks robustly. The experiments on standard datasets confirm their performance.

The future work mainly includes: (i) the precision of the model is slightly lower than other models, so we will try some new feature extraction methods to improve the precision of the model. (ii) The processing speed of model is not fast enough; thus, we will further improve its speed. (iii) The existing algorithms are based on the assumption that the text is arranged in an approximate straight line, which ignores the curved text that may appear in practical applications. Therefore, how to design a universal detection model for more general conditions deserves more attention.

6 Acknowledgments

This research was funded by the National Natural Science Foundation of China (grant nos. 61501069 and 11873001) and the Fundamental Research Funds for the Central Universities (grant no. 106112016CDJXZ168815).

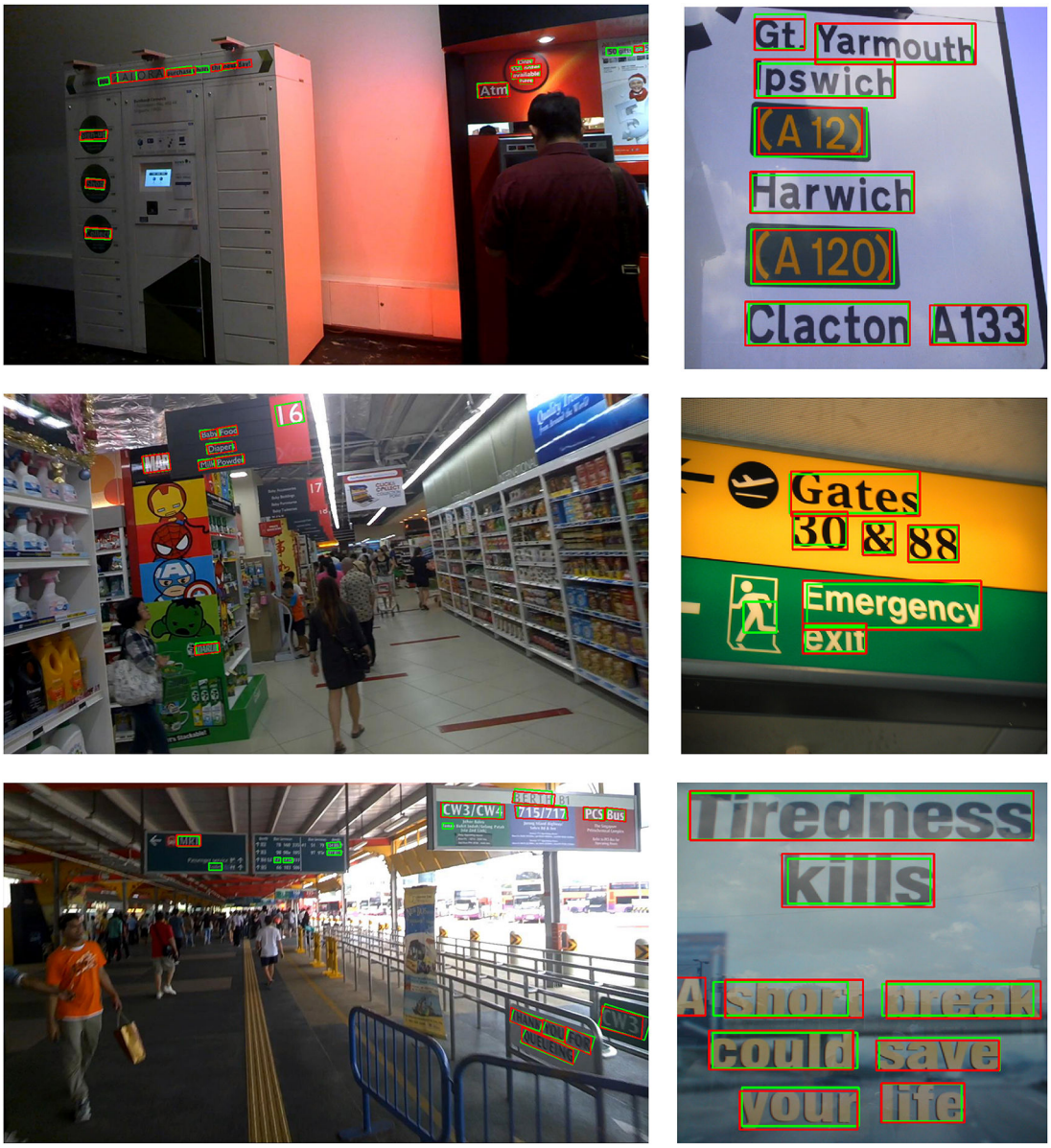

Fig. 6 *Text detection samples of the proposed SCRPN for natural scene text detection. Example images are from different datasets. Red solid boxes are the ground truth and green solid boxes are the detection results*

7 References

[1] Samadi, F., Akbarizadeh, G., Kaabi, H.: 'Change detection in SAR images using deep belief network: a new training approach based on morphological images', *IET Image Process.*, 2019, **13**, pp. 2255–2264

[2] Cheng, G., Han, J., Zhou, P., *et al.*: 'Learning rotation-invariant and Fisher discriminative convolutional neural networks for object detection', *IEEE Trans. Image Process.*, 2019, **28**, pp. 265–278

[3] Cheng, G., Zhou, P., Han, J.: 'Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images', *IEEE Trans. Geosci. Remote Sens.*, 2016, **54**, pp. 7405–7415

[4] Gonzalez, G., Evans, C.L.: 'Biomedical image processing with containers and deep learning: an automated analysis pipeline: data architecture, artificial intelligence, automated processing, containerization, and clusters orchestration ease the transition from data acquisition to insights in medium-to-large datasets', *BioEssays*, 2019, **41**, p. 1900004

[5] Chen, J., Luo, L., Wang, S., *et al.*: 'Automatic panoramic UAV image mosaic using ORB features and robust transformation estimation'. Chinese Control Conf., Wuhan, People's Republic of China, 2018, pp. 4265–4270

[6] Rothkrantz, L.J.M., Wojdel, A.: 'A text-based talking face'. Proc., Berlin, Heidelberg, 2000, pp. 327–332

[7] Ramanishka, V., Chen, Y., Misu, T., *et al.*: 'Toward driving scene understanding: a dataset for learning driver behavior and causal reasoning'. IEEE Conf. Computer Vision Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 7699–7707

[8] Li, X., Xu, C., Wang, X., *et al.*: 'COCO-CN for cross-lingual image tagging, captioning, and retrieval', *IEEE Trans. Multimedia*, 2019, **21**, pp. 2347–2360

[9] Chowdhury, N., Panda, R., Papalexakis, E.: 'Webly supervised joint embedding for cross-modal image-text retrieval', arxiv.org/abs/1808.07793, 2018

[10] Luqman, H., Mahmoud, S.A.: 'Automatic translation of Arabic text-to-Arabic sign language', *Univ. Access Inf. Soc.*, 2019, **18**, pp. 939–951

[11] Protaziuk, G., Kaczyński, M., Bembenik, R.: '*Automatic translation of multi-word labels*' (Springer, Cham, Germany, 2016), pp. 99–109

[12] Muehling, M., Meister, M., Korfhage, N., *et al.*: 'Content-based video retrieval in historical collections of the German broadcasting archive', *Int. J. Digit. Libr.*, 2019, **20**, pp. 167–183

[13] Zhang, L., Zhang, J.: 'Synchronous prediction of arousal and valence using LSTM network for affective video content analysis'. Int. Conf. Natural Computation, Guilin, China, 2017, pp. 727–732

[14] Luká, N., Matas, J.: 'Scene text localization and recognition with oriented stroke detection'. IEEE Int. Conf. Comput. Vis., Sydney, NSW, Australia, 2013, pp. 97–104

[15] Chen, X., Yuille, A.L.: 'Detecting and reading text in natural scenes'. Computer Society Conf. Computer Vision Pattern Recognition, Washington, D.C., USA, 2004, pp. 366–373

[16] Wang, K., Babenko, B., Belongie, S.: 'End-to-end scene text recognition'. Int. Conf. Comput. Vis., Barcelona, Spain, 2011, pp. 1457–1464

[17] Epshtein, B., Ofek, E., Wexler, Y.: 'Detecting text in natural scenes with stroke width transform'. Computer Society Conf. Computer Vision and Pattern Recognition, San Francisco, CA, USA, 2010, pp. 2963–2970

[18] Liu, Z., Li, Y., Qi, X., *et al.*: 'Method for unconstrained text detection in natural scene image', *IET Comput. Vis.*, 2017, **11**, (7), pp. 596–604

[19] Huang, W., Qiao, Y., Tang, X.: '*Robust scene text detection with convolution neural network induced MSER trees*' (Springer, Cham, Germany, 2014), pp. 497–511

[20] Yin, X., Yin, X., Huang, K., *et al.*: 'Robust text detection in natural scene images', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2014, **36**, (5), pp. 970–983

[21] Huang, X., Shen, T., Wang, R., *et al.*: 'Text detection and recognition in natural scene images'. Int. Conf. Estimation, Detection and Information Fusion, Harbin, China, 2015, pp. 44–49

[22] Jiang, Y., Zhu, X., Wang, X., *et al.*: 'R2CNN: rotational region CNN for orientation robust scene text detection', arXiv:1706.09579, 2017
[23] Wang, T., Wu, D. J., Coates, A., *et al.*: 'End-to-end text recognition with convolutional neural networks'. Int. Conf. Pattern Recognition, Tsukuba, Japan, 2012, pp. 3304–3308
[24] Jaderberg, M., Vedaldi, A., Zisserman, A.: '*Deep features for text spotting*' (Springer, Cham, Germany, 2014), pp. 512–528
[25] Yang, L., Hu, H.: 'TVPRNN for image caption generation', *Electron. Lett.*, 2017, **53**, (22), pp. 1471–1473
[26] Ma, J., Shao, W., Ye, H., *et al.*: 'Arbitrary-oriented scene text detection via rotation proposals', *IEEE Trans. Multimedia*, 2018, **20**, (1), pp. 3111–3122
[27] Tian, Z., Huang, W., He, T., *et al.*: '*Detecting text in natural image with connectionist text proposal network*' (Springer, Cham, Germany, 2016), pp. 56–72
[28] Shi, B., Bai, X., Belongie, S.: 'Detecting oriented text in natural images by linking segments'. Conf. Computer Vision Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 3482–3490
[29] Ghosh, S., Valveny, E., Bagdanov, A.D.: 'Visual attention models for scene text recognition'. Int. Conf. Document Analysis and Recognition, Kyoto, Japan, 2017, pp. 943–948
[30] Cheng, M., Zhang, Z., Lin, W., *et al.*: 'BING: binarized normed gradients for objectness estimation at 300 fps'. Conf. Computer Vision Pattern Recognition (CVPR), Columbus, OH, USA, 2014, pp. 3286–3293
[31] Zitnick, C. L., Dollár, P.: '*Edge boxes: locating object proposals from edges*' (Springer, Cham, Germany, 2014), pp. 391–405
[32] Ren, S., He, K., Girshick, R., *et al.*: 'Faster R-CNN: towards real-time object detection with region proposal networks', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017, **39**, pp. 1137–1149
[33] Liu, Y., Yu, F., Ying, C.: 'Scene text localization based on stroke width transform', *J. Chin. Comput. Syst.*, 2016, **37**, (2), pp. 350–353
[34] Chen, H., Tsai, S.S., Schroth, G., *et al.*: 'Robust text detection in natural images with edge-enhanced maximally stable extremal regions'. Int. Conf. Image Processing, Brussels, Belgium, 2011, pp. 2609–2612
[35] Zhao, Z., Fang, C., Lin, Z., *et al.*: 'A robust hybrid method for text detection in natural scenes by learning-based partial differential equations', *Neurocomputing*, 2015, **168**, pp. 23–34
[36] Jia, Y., Shelhamer, E., Donahue, J., *et al.*: 'Caffe: convolutional architecture for fast feature embedding', arXiv:1408.5093, 2014
[37] Jaderberg, M., Simonyan, K., Vedaldi, A., *et al.*: 'Reading text in the wild with convolutional neural networks', *Int. J. Comput. Vis.*, 2016, **116**, pp. 1–20
[38] Manen, S., Guillaumin, M., Van Gool, L.: 'Prime object proposals with randomized Prim's algorithm'. Int. Conf. Computer Vision, Sydney, NSW, Australia, 2013, pp. 2536–2543
[39] Krähenbühl, P., Koltun, V.: '*Geodesic object proposals*' (Springer, Cham, Germany, 2014), pp. 725–739
[40] Liu, Y., Jin, L.: 'Deep matching prior network: toward tighter multi-oriented text detection'. Conf. Computer Vision Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 3454–3461
[41] Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., *et al.*: 'ICDAR 2015 competition on robust reading'. 13th Int. Conf. Document Analysis and Recognition, Tunis, Tunisia, 2015, pp. 1156–1160
[42] Zhou, X., Yao, C., Wen, H., *et al.*: 'EAST: an efficient and accurate scene text detector'. Conf. Computer Vision Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 2642–2651
[43] Zhang, Z., Zhang, C., Shen, W., *et al.*: 'Multi-oriented text detection with fully convolutional networks'. Conf. Computer Vision Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 4159–4167
[44] Yin, X., Pei, W., Zhang, J., *et al.*: 'Multi-orientation scene text detection with adaptive clustering', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015, **37**, (9), pp. 1930–1937
[45] Kang, L., Li, Y., Doermann, D.: 'Orientation robust text line detection in natural images'. Conf. Computer Vision Pattern Recognition, Columbus, OH, USA, 2014, pp. 4034–4041