Multi-orientation Scene Text Detection Leveraging Background Suppression

Xihan Wang[1(✉)], Xiaoyi Feng[1], Zhaoqiang Xia[1], Jinye Peng[2],
and Eric Granger[3]

[1] School of Electronics and Information, Northwestern Polytechnical University,
Xi'an 710072, China
xihanwang@mail.nwpu.edu.cn

[2] School of Information Science and Technology, Northwest University,
Xi'an 710069, China

[3] Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA),
École de technologie supérieure, Université du Québec,
1100, rue Notre-Dame Ouest, Montréal, (QC) H3C 1K3, Canada

Abstract. Most state-of-the-art text detection methods are devoted to horizontal texts and these methods cannot work well when encountering blurred, multi-oriented, low-resolution and small-sized texts. In this paper, we propose to localize texts from the perspective of suppressing more non-text backgrounds, in which a coarse-to-fine strategy is presented to remove non-text pixels from images. Firstly, the fully convolutional network (FCN) framework is utilized to make the coarse prediction of text labeling. Secondly, an efficient saliency measure based on background priors is employed to further suppress non-text pixels and generate fine character candidate regions. The remaining candidates of character regions composite text lines, so that the proposed method can handle multi-orientation texts in natural scene images. Two public datasets, MSRA-TD500 and ICDAR2013 are utilized to evaluate the performance of our proposed method. Experimental results show that our method achieves high recall rate and demonstrates the competitive performance.

Keywords: Scene text detection · Fully Convolutional Network
Background suppression · Multi-orientation texts

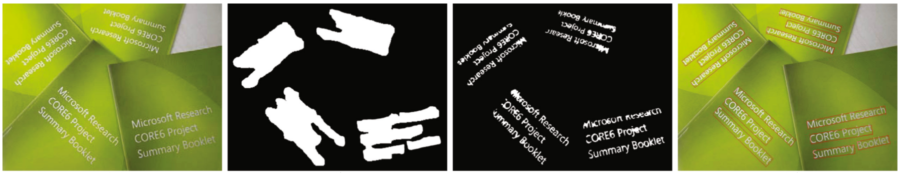
1 Introduction

In recent years, the extraction of information in images has attracted much attention, with the widespread application of video and image acquisition equipments. Scene text provides direct high-level semantic information and plays an important role in a variety of interesting applications [20], such as criminal investigation, visual assistant for the blinds, translators for tourists, automatic driving and navigation. Texts in uncontrolled environments may exhibit in different

Y. Zhao et al. (Eds.): ICIG 2017, Part I, LNCS 10666, pp. 555–566, 2017.
https://doi.org/10.1007/978-3-319-71607-7_49

layout, language, font and size, text-like background objects, non-uniform illuminations, blur and occlusion. Consequently, text localization in natural scene is still a challenging task.

Usually, there exist two main types of conventional methods for scene text detection: sliding window based [24] and connected component (CCs) based [4,12,23]. However, the performance of these methods heavily relies on hand-crafted features. These methods may not work well under severe complexities, such as blur and low resolution. Recently, a new trend has appeared that the deep neural network based algorithms have gradually become the mainstream. State-of-the-art methods for object detection/semantic segmentation have been modified for text detection, and achieve great improvement in this field. However, most methods are specifically designed for horizontal or near-horizontal texts. The deep networks specially designed for object segmentation may not be accurate in word-level or line-level text extraction. When multiple text lines flock together, it becomes difficult to identify each individual text line (See Fig. 1(b)).



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

Fig. 1. Overview of our proposed algorithm: (a) Original image. (b) Prediction result of global background suppression. (c) Character candidates generation by local background suppression. (d) Result of text detection.

To overcome the above challenges, we tackle the problem of text localization from a new perspective. In this paper, we consider the character component extraction from an opposite point of view, i.e., by suppressing more non-text backgrounds. Ideally, better background suppression means better prospects of foreground detection. As illustrated in Fig. 1, an unconventional detection framework for scene text is proposed. In our method, a coarse-to-fine strategy is adopted to remove non-text pixels from global image to local regions. We use the Fully Convolutional Network (FCN) framework [11], which is suitable to generate a pixel-wise text/non-text saliency map. The map provides a powerful guidance for estimating text regions, however, this prediction is regional and it is hard to separate each instance. A saliency measure strategy is utilized to further remove the background a fine level. Then, the foreground is considered as character candidate regions and generates text lines with graph partition.

The contributions of our approach are as follows. First, we cast scene text detection as a background suppression problem and utilize a two-step framework to remove non-text pixel globally and locally. The resulting foreground regions can be considered as character-level proposals extraction. Second, an efficient saliency measure method based on background priors is presented to further suppress the non-text pixel. We use super-pixels substitute single pixels obtaining high accuracy and efficiency. This measure result allows our method to handle multi-lingual text. Third, character-to-line strategy endows the system with the ability to detect multi-oriented and curved texts.

2 Related Work

Scene text detection in natural images has received much attention and many effective methods have been presented. As mentioned, conventional methods heavily rely on manually designed features. Zhang et al. [24] directly extracted text lines from natural images by local symmetry property of character groups. A convolutional neural networks (CNN) with the powerful discrimination ability was trained to eliminate false positives. However, sliding window based methods are usually time-consuming and not suitable for multi-orientation texts. Connected component-based methods usually aim to extract character candidate regions, and then group character candidates into words or text line. Stroke Width Transform (SWT) [4], Maximally Stable Extremal Regions (MSER) [12] are two representative methods for character candidate generation as well as their subsequent works [19,22,23]. Yao et al. [19] proposed a method based on MSER for detecting text of arbitrary directions. Yin et al. [22] presented a learning framework for adaptive hierarchical clustering, in which text candidates are constructed by grouping characters based on this adaptive clustering. Wei et al. [17] captured character regions with the exhaustive segmentation method and presented a learning-based text line grouping method. These CCs-based methods obtained promising performance on a variety of standard benchmark datasets. Nevertheless, these methods may fail when character is not homogeneous or composed of broken strokes in complex backgrounds.

In recent years, deep convolutional neural networks for scene text detection [3,5–7] have been a new trend and achieved superior performance over conventional approaches. These methods use deep convolutional network mainly from two aspects: (1) learn a more robust text representation; (2) take advantage of the powerful classification ability for better eliminating false positives. Different from performing classification on local regions, Zhang et al. [25] utilized both local and global cues, and a Fully Convolutional Network (FCN) model was trained to predict the salient map. Then components are extracted based on MSER for multiple orientations text detection. Some text detection methods treat text words or lines as generic objects. Tian et al. [15] developed a vertical anchors mechanism and constructed a CNN-RNN joint-model to detect extremely challenging text. Zhong et al. [26] attempt to convert Faster-RCNN into text detection. However, those two methods may not be suitable for multi-oriented scene text detection.

The proposed algorithm is inspired by work in saliency detection [16], which aim to measure object level saliency with background priors. In this paper, we tackle the text detection from opposite thought, we focus more on the background suppression instead of text component extraction or window-based classification. Different from [16], the prior information is obtained from pixel-wise prediction maps from FCN.

3 Proposed Method

In this section, details of proposed algorithm are presented (See Fig. 2). The global background suppression is described in Sect. 3.1, character candidates extraction by local background suppression and linking of character candidates into text lines are given in Sects. 3.2 and 3.3, respectively.

3.1 Global Background Suppression

To better distinguish between foreground (text) and background (non-text), we consider this problem as a kind of pixel-labeling task. Following the general design of [18], in this paper, we apply our proposed modification architecture to label text/non-text regions in a holistic way.

Network Architecture. As shown in Fig. 2, necessary modifications are applied for fully using high-level semantic information in upper network layers. The first 5 convolutional stages are adopted from VGG-16 net [13], while the last several stages, including the 5th pooling layer and all the fully connected layers, are cut off. We focus on the last three stages, respectively conv3-3, conv4-3, conv5-3. Each stage is followed by a convolutional layer with 1 × 1 kernel size, and then connected to a deconvolutional layer to ensure that the outputs of all the stages have the same size as the input. The multi-level side-outputs maps are concatenated together to form the fused map with a weight layer (a convolutional layer with 1 × 1 kernel size). There are two reasons that prompt us to omit the first and second stages. Firstly, compared to local structures features from lower layers, our algorithm at this stage is more concerned with high-level

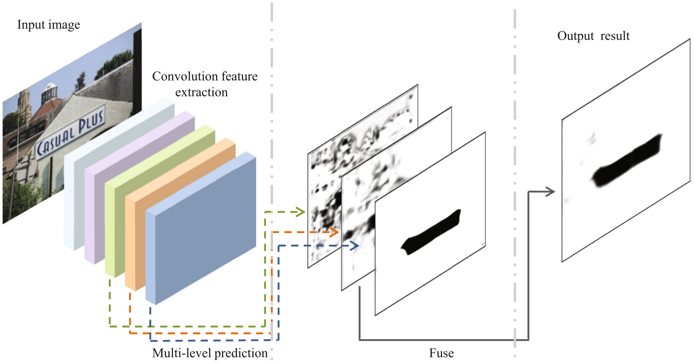


Fig. 2. Network architecture of the proposed algorithm. The first 5 convolutional stages are adopted from VGG 16-layer net, and the architectures are equivalent to making independent predictions from last three stages and fuse the results.

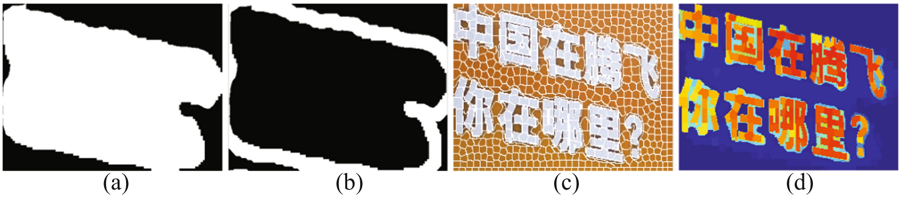


Fig. 3. Illustration of character candidates generation: (a) Prediction results of foreground region by network. (b) Designing background regions by morphological operations. (c) Super-pixel segmentation. (d) Confidence map with background suppression.

semantics semantic information in upper network layers. Secondly, different from the label data in general pixel labeling task, such as semantic segmentation and edge detection, label maps converted by ground truth are mixed with non-text background information and it will cause interference to the lower layers.

Ground Truth and Loss Function. The label maps are generated from ground truth annotations. Pixels inside the bounding box are considered as the positive regions. These pixels are marked as '1' while the background region outside the bounding box are marked as '0'.

In training phase, we use balanced cross-entropy loss to compute the errors between side-outputs and ground truth, which is introduced in [18]. Denote the ground truth for a given pixel as $y_i^* \in \{0, 1\}$ and predicted value as $\hat{y}_i$, L is formulated as

$$L_s = -\beta \sum_{i \in L_s} y_i^* \log \hat{y}_i - (1 - \beta) \sum_{i \in L_s} (1 - y_i^*) \log(1 - \hat{y}_i) \tag{1}$$

where $\beta = |Y_-|/(|Y_+| + |Y_-|)$, $|Y_+|$ and $|Y_-|$ denote the number of text pixels and non-text pixels in the ground truth. In this paper, the multi-level prediction map is used to generate final prediction result through the fusion layer. The loss between the fusion output and ground truth is also using the balanced cross-entropy loss function. A binary map is generated as the result of global background suppression, and the black pixels are considered as background and the text candidate region are marked as white pixels. An example of predicted map is shown in Fig. 3(a).

3.2 Local Background Suppression

After performing global background suppression, the majority of non-text background are removed. For scene text detection, however, it is necessary to differentiate between character instances, for which two-class (text or non-text) semantic predicted are insufficient (e.g. non-text regions adjacent multiple text line are labeled as the same class with characters it make difficult to separate each instance, see Fig. 3(a)). Inspired by the works of [2,16], which aim at saliency object detection problem, two kinds of analysis about text and non-text background in scene images, namely *visual contrast analysis* and *background connectivity analysis*, is utilized to further suppress background information.

People can effortlessly capture text regions in a split second. It illustrates that the vision system is sensitive to contrast in visual signal. The first analysis is based on this rule. In other words, text in scene usually contrast strongly with their background to attract more humans attention. Specifically, the global contrast of text pixel is defined using its color contrast to all other pixels in the background. The confidence value $C(I_i)$ is defined as,

$$C(I_i) = \sum_{\forall I_b \in B} D(I_i, I_b) \tag{2}$$

where I_i is a pixel in image I and background pixel set $B\{I_b, b = 1, ..., N\}$, where N is the number of pixels in the background. $D(I_i, I_b)$ is the color distance between pixels I_i and I_b (measured in $L*a*b$ space), the sum of all distance is mapped to $[0, 1]$ by normalization. Background pixels are based on the predicted result in global background suppression stage.

Based on the contrast and predicted map prior, we further observe that most background regions close to the same character can be easily connected to their surroundings. We call it the *background connectivity*. Benefited from global prediction, the foreground regions in predicted map can be considered to contain both text and background. This suggests that we can define the background confidence of a pixel as the length of its shortest path to the foreground region boundaries, which can be obtained by computing weighted geodesic distances.

We first segment the predicted map into regions (each region contains foreground and considered as mask map). Then the boundary region is obtained with the morphological dilation operation, see Fig. 3(b). For each image region, we build an undirected weighted graph $G = \{V, E\}$. The vertex set V of nodes contains all image pixels in the foreground region and boundary region, $V = \{F\} \bigcup \{B\}$. E is a set of edges, $E \subseteq \{(\upsilon, \nu) | \upsilon, \nu \in V and\, \upsilon \neq \nu\}$. A path from υ to ν is a sequence $P(\upsilon, \nu) = (\upsilon = p_1, p_2, ..., p_n = \nu)$. The edges $(p_i, p_{i+1}) \in E$ connect all adjacent nodes and $i = 1, 2, ..., n-1$. The path weight $\omega(P(\upsilon, \nu))$ is defined as,

$$\omega(P(\upsilon, \nu)) = \sum_{i=1}^{n-1} D(p_i, p_{i+1}) \tag{3}$$

where $D(p_i, p_{i+1})$ is the distance between the color features of two nodes (normalized to $[0, 1]$ in $L*a*b$ space). In order to reduce the interference of similar nodes distance on path weights, a threshold τ is introduced to control the strength of adjacent nodes weight. The τ is taken as mean value of the smallest distances from all nodes with their neighbors. The distances smaller than this threshold will be equal to zero. The minimum of path weight means the shortest path from υ to ν, and it characterizes the connectivity between two nodes. The background connectivity of a pixel I_i accumulates all shortest path weights from I_i to each pixel in boundary region B on the graph G. The function is given in the following manner:

$$S(I_i) = \sum_{\forall I_b \in B} \min\{\omega(P(I_i, I_b))\} \tag{4}$$

We use Johnson's [8] algorithm to find the shortest paths $\min\{\omega(P(I_i, I_b))\}$. The pixel with a lower S value in the foreground has a higher possibility of background.

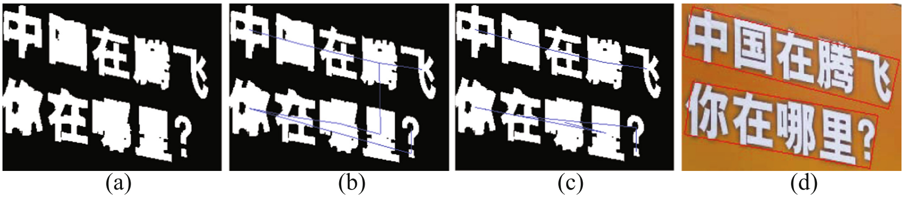
Obviously, using Eqs. 2 and 5 to evaluate the value for each image pixel is an extremely inefficient process, which is too computationally expensive even for medium sized images. The key to speed up the algorithm is the use of super-pixels instead of pixels to reduce the number of pixels in the image. We extract the SLIC super-pixels [1] from image regions. Each region is re-scaled to have maximum dimension of $250 * 250$ pixels with aspect ratio unchanged. Then the mean color of super-pixels is computed and analyzed at the region level. We further incorporate its global contrast and connectivity information to increase the difference between foreground and background. Specifically, for any super-pixel region r_k, we can define $C(r_k)$ and $S(r_k)$ from Eq. 2 and 5. The background suppression result is defined as,

$$S_b(r_k) = \omega(r_k)S(r_k)(C(r_k) + \varepsilon) \tag{5}$$

where $\omega(r_k)$ is the weight of region r_k, and it is defined as $|R_+|/|R|$, $|R_+|$ denotes the count of "1" pixels which marked by predicted map and $|R|$ denotes the total count of pixels. ε controls the strength of global contrast, In our implementation, we use $\varepsilon = 0.4$. The final confidence map is shown in Fig. 3(d).

3.3 Text Lines Formation

The purpose of this stage is to form multi-oriented text lines from the remaining character candidates, which are obtained by connected component analysis. Similar to previous work in [21], we initially establish a fully connected graph G by Delaunay triangulation. In the graph, each vertex represents a character candidate region and the weight of edges model the adjacency relation between pairs of candidates. Based on the graph, a Maximum Spanning Tree is constructed. Different from method [21], we count the direction of each edge and find the main direction of the text line. The edge different from the main direction is removed. Then, the remained edges were merged to construct the original text chains. A straight line is fitted to the centroids of candidates within each chain. The single component which has similar intensities will be re-linked into a chain. The process is iterated until all text candidates have been assigned to a line. An example of text lines formation is shown in Fig. 4. In certain tasks,



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

Fig. 4. Text lines formation: (a) Character candidates. (b) MST construction. (c) Text line partition. (d) Result of detection.

such as ICDAR 2013, word partition is required. Since text in images from these datasets is labelled in word level. We adopted the word partition method in [4] as it has been proven to be simple and effective.

4 Experiments and Discussions

We performed experiments on two standard benchmarks, namely, MSRA-TD 500 and ICDAR 2013 , to evaluate our algorithm, which is compared with other scene text detection methods. The training data are the union of training images from datasets. In training phase, a fixed 300 × 300 sliding window with half window length step is used to corp more patches from scaled images. Then we evenly rotated images to four different angles (with 90° angle interval). In testing, we compute the predicted map by trained model in multi-scale sliding window strategy. The detections of different scales are fused to form the final forecast. All experiments are conducted on a general computer (Intel Core i5, 3.2 GHz 4-core CPU, 8 G RAM, GeForce GTX950 and Windows 64-bit OS). At runtime, all testing images are performed at original dimension and the routine ran on a single GPU.

4.1 Datasets

Two standard datasets used to validate our scene text detection method will be introduced briefly:

MSRA-TD 500: The MSRA-TD 500 is originally proposed in [19] and it is a typical multi-orientation benchmark dataset for assessing detection algorithms. The dataset have 300 training images and 200 test images, all image are high-resolution natural scene images. Text in this dataset Contains varying directions, fonts, mixed languages, and complexity of backgrounds, these factors make the dataset have highly challenging.

ICDAR 2013: The ICDAR 2013 dataset is from the ICDAR 2013 Robust Reading Competition [10]. A total of 229 natural images were used for training, and 233 images for testing. All the text in this dataset are notable horizontal and near-horizontal, and the ground truth annotated in word level.

4.2 Experimental Results and Discussions

We adopted the evaluation method proposed by wei [17] to compare our method with other methods. We first evaluate our algorithm on MSRA-TD 500. Table 1 shows the performance comparison of different algorithms on the MSRA-TD 500 dataset. As can be seen, our method achieves the recall rate of 0.73, precision of 0.75, and f-measure of 0.74. The proposed algorithm achieves the highest recall than other methods on this dataset, the precision slightly lower than other two methods. There are two main reasons: (1) To handle some short, single characters, the global predicted map will bring more false alarms, as well as the

Table 1. Performances of different text detection methods evaluated on MSRA-TD500.

Algorithm	Precision	Recall	F-measure
Proposed	0.75	**0.73**	**0.74**
Zhang et al. [25]	**0.83**	0.67	**0.74**
Yin et al. [22]	0.81	0.63	0.71
Kang et al. [9]	0.71	0.62	0.66
Yin et al. [23]	0.71	0.61	0.66
Yao et al. [19]	0.63	0.63	0.60

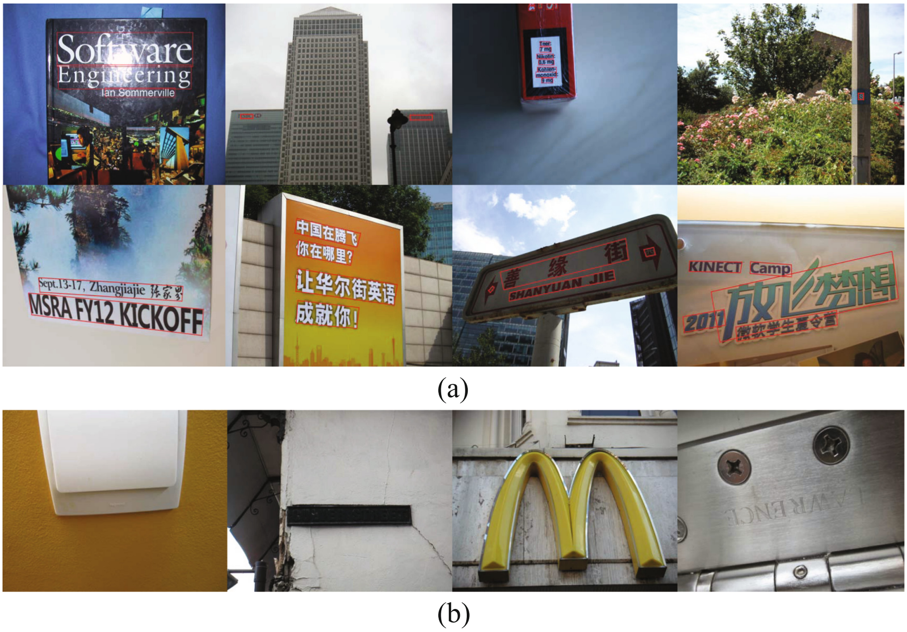
Table 2. Performances of different text detection methods evaluated on ICDAR 2013.

Algorithm	Precision	Recall	F-measure
Proposed	0.82	**0.81**	0.81
Zhang et al. [25]	**0.88**	0.78	**0.83**
Wei et al. [17]	0.84	0.77	0.80
Zhang et al. [24]	0.88	0.74	0.80
Tian et al. [14]	0.85	0.76	0.80
Yin et al. [23]	0.86	0.68	0.76

not-text backgrounds which have very similar structure with text can not be easily distinguish. (2) The character candidates are still mixed with many non text regions, our proposed method currently can not filter these regions.

We then evaluate our method on ICDAR 2013, The performance of proposed algorithm evaluated on the ICDAR 2013 dataset is shown in Table 2. In this experiment, the overall performance of our algorithm is not better than other previous state-of-the-art methods, the proposed method only achieves the highest recall (0.81) among all the methods. Except we discuss in first experiments, the unsuccessful result may be due to the following reasons: (1) text in ICDAR 2013 dataset are notable horizontal and near-horizontal, most text detection algorithms are well-directed, to measure the performance of our system, text line candidates must further partitioned into words. (2) Our method is main focus on the capacity of handling multi-oriented text, the advantages of multi-oriented text detection cannot be reflected in this dataset.

Figure 5(a) shows the successful detection results of the proposed algorithm a number of challenging cases, the qualitative results show our algorithm is able to handle text instances of different orientations, fonts, and languages. The images also show that our system is robust against strong lighting and blur. However, in some certain conditions, our method may fail, for example, characters with art font, low resolution or serious blur. As shown in Fig. 5(b), we believe that our method still has room for performance improvement with the increase of training samples.



(a)

(b)

Fig. 5. Qualitative results of the proposed method: (a) Successful text detection results. (b) Some failure cases

5 Conclusions

In this work, a novel algorithm is presented for multi-oriented text detection in natural scene images. In contrast with a vast majority of the previous methods, we tackled the text extraction problem from a opposite direction. A coarse-to-fine non-text backgrounds suppress strategy was adopted to remove non-text pixels, thus is able to get better prospects foreground detection. Our algorithm can directly handle multi-oriented text from images, while most previous approaches only focused horizontal or near-horizontal text. The experiments on two public datasets demonstrated the competitive performance and effectiveness of the proposed methods.

Acknowledgment. This paper is supported by H3C Foundation of Ministry of Education of China, No. 2017A19050, the National Aerospace Science and Technology Foundation and the National Nature Science Foundation of China (No. 61702419).

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2274–2282 (2012)

2. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.M.: Global contrast based salient region detection. IEEE Trans. Pattern Anal. Mach. Intell. **37**(3), 569–582 (2015)
3. Coates, A., Carpenter, B., Case, C., Satheesh, S., Suresh, B., Wang, T., Wu, D.J., Ng, A.Y.: Text detection and character recognition in scene images with unsupervised feature learning. In: 2011 International Conference on Document Analysis and Recognition (ICDAR), pp. 440–445. IEEE (2011)
4. Epshtein, B., Ofek, E., Wexler, Y.: Detecting text in natural scenes with stroke width transform. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2963–2970. IEEE (2010)
5. Huang, W., Qiao, Y., Tang, X.: Robust scene text detection with convolution neural network induced MSER trees. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 497–511. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_33
6. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Reading text in the wild with convolutional neural networks. Int. J. Comput. Vis. **116**(1), 1–20 (2016)
7. Jaderberg, M., Vedaldi, A., Zisserman, A.: Deep features for text spotting. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 512–528. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_34
8. Johnson, D.B.: Efficient algorithms for shortest paths in sparse networks. J. ACM (JACM) **24**(1), 1–13 (1977)
9. Kang, L., Li, Y., Doermann, D.: Orientation robust text line detection in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4034–4041 (2014)
10. Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L.G., Mestre, S.R., Mas, J., Mota, D.F., Almazan, J.A., de las Heras, L.P.: Icdar 2013 robust reading competition. In: 2013 12th International Conference on Document Analysis and Recognition (ICDAR), pp. 1484–1493. IEEE (2013)
11. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
12. Neumann, L., Matas, J.: A method for text localization and recognition in real-world images. Comput. Vis.-ACCV **2010**, 770–783 (2011)
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
14. Tian, S., Pan, Y., Huang, C., Lu, S., Yu, K., Tan, C.L.: Text flow: a unified text detection system in natural scene images. In: IEEE International Conference on Computer Vision, pp. 4651–4659 (2016)
15. Tian, Z., Huang, W., He, T., He, P., Qiao, Y.: Detecting text in natural image with connectionist text proposal network. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 56–72. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_4
16. Wei, Y., Wen, F., Zhu, W., Sun, J.: Geodesic saliency using background priors. Comput. Vis.-ECCV **2012**, 29–42 (2012)
17. Wei, Y., Zhang, Z., Shen, W., Zeng, D., Fang, M., Zhou, S.: Text detection in scene images based on exhaustive segmentation. Sig. Process.: Image Commun. **50**, 1–8 (2017)
18. Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1395–1403 (2015)

19. Yao, C., Bai, X., Liu, W., Ma, Y., Tu, Z.: Detecting texts of arbitrary orientations in natural images. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1083–1090. IEEE (2012)
20. Yi, C., Tian, Y.: Scene text recognition in mobile applications by character descriptor and structure configuration. IEEE Trans. Image Process. **23**(7), 2972–2982 (2014)
21. Yin, F., Liu, C.L.: Handwritten text line extraction based on minimum spanning tree clustering. In: International Conference on Wavelet Analysis and Pattern Recognition, ICWAPR 2007, vol. 3, pp. 1123–1128. IEEE (2007)
22. Yin, X.C., Pei, W.Y., Zhang, J., Hao, H.W.: Multi-orientation scene text detection with adaptive clustering. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1930–1937 (2015)
23. Yin, X.C., Yin, X., Huang, K., Hao, H.W.: Robust text detection in natural scene images. IEEE Trans. Pattern Anal. Mach. Intell. **36**(5), 970–983 (2014)
24. Zhang, Z., Shen, W., Yao, C., Bai, X.: Symmetry-based text line detection in natural scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2558–2567 (2015)
25. Zhang, Z., Zhang, C., Shen, W., Yao, C., Liu, W., Bai, X.: Multi-oriented text detection with fully convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4159–4167 (2016)
26. Zhong, Z., Jin, L., Zhang, S., Feng, Z.: Deeptext: A unified framework for text proposal generation and text detection in natural images. arXiv preprint arXiv:1605.07314 (2016)