SCENE TEXT DETECTION BASED ON SKELETON-CUT DETECTOR

Xiang He , Yonghong Song , Yuanlin Zhang
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, China

ABSTRACT

As the structural information of an object can be well descripted by edge pixels, we observe that the greatest challenge for locating text edges on scene image is how to handle the edge-adhesion problem. In this paper we propose the Skeleton-cut Text Detector, which take text-specific edge cues such as a novel presentation *skeleton* into account to hunt text efficiently with improved recall rate. To address edge-adhesion problem, skeleton-junctions detection and elimination are performed first to cut candidate text out of the edge map. Then the candidates are verified through a two-stage classifier based on properties like concentration ratio. Finally iteratively local refinement (IRL) is applied to enhance the overlap of proposals. Experimental results on public benchmarks, ICDAR 2013 and MSRA, demonstrate that our algorithm achieves state-of-the-art performance. Moreover in severe scenarios, our proposed method shows stronger adaptability to texts by exploiting skeleton compared to conventional presentations like MSERs.

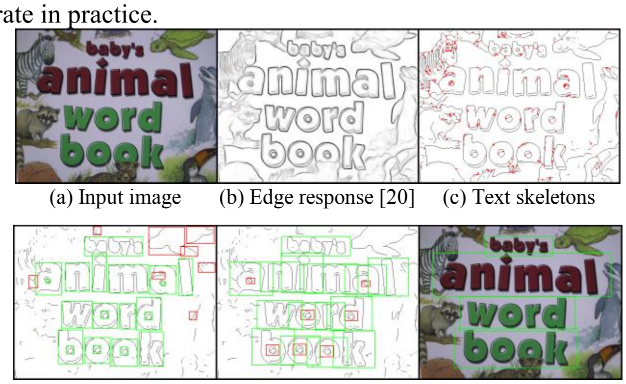
Index Terms— Scene text detection, Edge adhesion, Text-specific representation, Skeleton-cut strategy, IRL

1. INTRODUCTION

Text could convey valuable informations in scene images. Scene text detection and recognition have become an active direction in community driven by the huge demand for a variety of practical computer vision techniques such as automotive assistance, multilingual translation and image retrieval. Though studied extensively in recent years [1, 2, 3, 4, 5, 6, 7, 8], text detection in natural scenes is still quite challenging due to complex backgrounds, various texts and interferential factors [9].

For detecting text in scene images, previous works have utilized the connected component analysis (CCA) [2, 5, 10, 11, 12, 13, 14, 15, 16] and sliding window approach [1, 17, 18, 19] as two classes of mainstream methods. The latter category includes exhaustive search based methods which could achieve high recall rates by shifting a window through every locations of a given scene image in multiple scales to detect texts. However, a large number of candidates generated from the through scanning of windows would give rise to intolerably heavy computations, and may result in a great deal of false positives in addition.

On the other hand, in those methods based on CCA, character candidates are extracted from the input scene image first, and refined to suppress non-text candidates subsequently. In particular, the connected component based methods which employ Stroke Width Transform (SWT) or Maximally Stable Extremal Region (MSER) as their basic representations have achieved outstanding performances on ICDAR competitions due to the representations' efficiency and stability. But even these representations might perform poorly under some severe conditions (e.g. low resolution, blur or non-uniform illumination), leading to low detection rate in practice.



(a) Input image (b) Edge response [20] (c) Text skeletons
(d) Skeleton cutting (e) Text-line refinement (f) detection result

Fig.1. The pipeline of Skeleton-cut text detector.

To overcome the drawbacks mentioned above, in this paper we tackle the problem of text proposals in challenging scenarios where SWT and MSER based approaches may fail by developing a new text descriptor namely Skeleton-cut detector to hunt text skeletons from natural images directly. The main contributions of our work include: (1) We design the text-specific representation *skeleton*, which could be applied to handle edge-adhesion problem via Skeleton-cut strategy including skeleton-junctions detection and elimination ; (2) Two-stage filtering using novel features like concentration ratio is applied to ensure comparably good precision rate , and iteratively local refinement for better detection accuracy. Specially, the overall process of the skeleton-cut text detector is shown in Fig 1. At the input of the pipeline, structured forest edge response [20] is efficiently computed from the image (see Fig. 1(b)) and then segmented through a range of intensity thresholds to get several edge maps. Subsequently in each map the corresponding text skeletons are constructed, on which

He Xiang is also a visiting researcher with Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen 518060, China.

skeleton-junctions are also detected (illustrated as red points in Fig. 1(c)). As partial text skeletons obtained from the edge maps with low segmentation thresholds are disordered and not suitable for Skeleton-cut operation, Euler-number calculation is applied to select the most distinct skeleton. As the next step, most of the background skeletons adjoined to text skeletons via skeleton junctions are eliminated using geometric constraints (from Fig 1(c) to 1(d)). Moreover, remaining false-skeletons (contained in red box in Fig. 1(d)) are removed by a CNN classifier [21]. Finally, NMS and text line refinement are performed (see Fig. 1(e)), giving the detection result.

2. THE SKELETON-CUT TEXT DETECTOR

In the first step of the proposed approach, we employ the Structured-forest edge detector described in [20] on the input image I （see Fig 2(a)）to obtain the sparse edge response R (see Fig 2(b)). Each pixel of the response R is labeled with a value indicating the probability of the pixel belonged to an edge. While the structural information of text could be well represented by edge response, it is difficult to extract text edges from response R for two challenges:

a) How to solve the ***edges-adhesion problem***: In the response, some text edges which are connected to background edges are hard to identify.

b) What is the ***essential difference*** between text edges and non-text edges?

2.1. Text Skeletons Construction

To solve the ***edges-adhesion problem***, we first segment the edge response through incremental thresholds to get a series of binary edge map M (for instance, one of the maps is illustrated as Fig 2(c)). And then skeleton-junctions detection and elimination are performed on these maps to cut text-edges out from background.

2.1.1. Skeleton-junctions Detection

Suppose that $x_1, x_2, ..., x_8$ are the values of neighbors of pixel p in M, starting with the east neighbor and numbered in counter-clockwise order. Given the map M, we construct the presentation skeletons by deleting pixel p if and only if conditions G_1, G_2, and G_3 are satisfied in the first sub-iteration.

Condition G_1: $X_H(p)=1$, where $X_H(p)=\sum_{i=1}^{4} b_i$, and

$$b_i=\begin{cases}1, \text{ if } x_{2i-1}=0 \text{ and } (x_{2i}=1 \text{ or } x_{2i+1}=1)\\ 0, \quad\quad\quad\quad\quad\quad\quad\quad \text{otherwise}\end{cases}$$

Condition G_2: $2 \le \{n_1(p), n_2(p)\} \le 3$, where

$n_1(p)=\sum_{k=1}^{4} x_{2k-1} \vee x_{2k}$, and $n_2(p)=\sum_{k=1}^{4} x_{2k} \vee x_{2k+1}$

Condition G_3: $(x_2 \vee x_3 \vee \bar{x}_8) \wedge x_1 = 0$

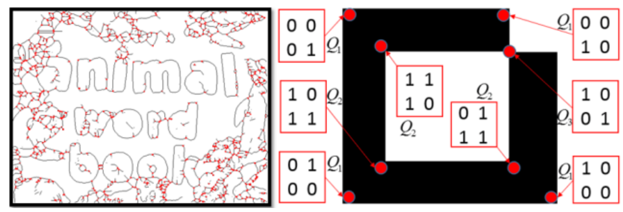
In the second sub-iteration, pixel p is deleted if and only if conditions G_1, G_2, and G_3' are satisfied.

Condition G_3': $(x_6 \vee x_7 \vee \bar{x}_4) \wedge x_5 = 0$

Each iteration of the text-skeletons-construction algorithm is made up of the two sub-iterations. And the iterations are repeated until the skeletons stops changing. The resulting skeletons S are shown as the minimally connected 1-pixel-width strokes or rings in Fig. 2(d).

(a) Image (b) Response (c) Map (d) Skeleton

Fig.2. The procedure of Skeleton-junctions detection



(a) Disordered edge skeleton (b) Quads detected on skeleton

Fig.3. Select the most distinct skeleton based on Euler-number calculation.

After text skeletons construction, a number of skeleton-junctions J could be easily obtained on skeletons S using the following equation:

$$J=\left\{p \mid n(p)=\sum_{k=1}^{8} x_k > t\right\} \quad (1)$$

Where $n(p)$ is the sum values of the pixel p ' neighbors, t is a threshold value used for extracting skeleton-junctions and is experimentally set as $t=3$. The resulting skeleton-junctions J are illustrated as red points in Fig. 2(d).

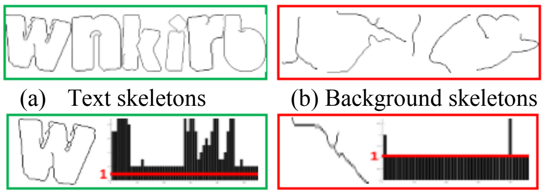
2.1.2. Text Skeleton-cut operation

Consider the case in which part of the skeletons obtained from the edge response with low segmentation thresholds are disordered and not suitable for skeleton-junctions detection and elimination. To avoid getting the disordered skeletons as shown in Fig. 3(a), we employ Euler-number calculation to select the most distinct skeleton *skel* from S.

Euler number, a scalar that specifies the number of objects in the region minus the number of holes in those objects, could indicate whether a skeleton is disordered or not. We first detect quads $Q_1, Q_2, and\ Q_3$ on *skel* (see Fig. 3(b)), and then calculate the Euler number of *skel* using Equation (2).

$$\eta(skel)=\tfrac{1}{4}(n(Q_1)-n(Q_2)+2n(Q_3)) \quad (2)$$

Where quads $Q_1, Q_2, and\ Q_3$ are 2×2 pixel patterns, and $n(Q)$ is the number of Q. Next, only *skel* with moderate Euler number η is preserved. This selection not only reduce the computational cost introduced by constructing the disordered skeleton, but also prevent *skel* from being segmented into too-small fragments. Finally we eliminate the skeleton-junctions of selected *skel* to cut text skeletons out from background and therefore address the edge-adhesion problem.



(a) Text skeletons (b) Background skeletons

(c) Concentration ratio of text skeleton and background skeleton

Fig.4. Text has higher cohesion degree than background.

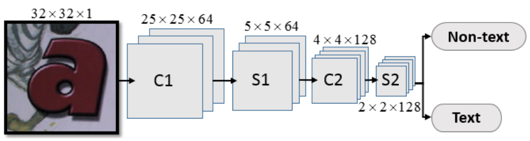


Fig.5. The architecture of single-scale CNN classifier

2.2. Text Skeletons Filtering

In the second step of Skeleton-cut detection, we filter out non-texts using two-stage classifier. Now let us consider the ***difference*** between text skeletons and other skeletons. As we can see in Fig 4(a-b), background skeletons are relatively straight, open and diffused, while text skeletons are rather curve, close and agglomerate. Based on this observation, most of non-texts could be easily eliminated by geometric constraints on properties like concentration ratio and so on. Remaining text proposals are hard to distinguish, thus need to be classified by a CNN classifier inspired by [21].

2.2.1. Geometric Filtering

Majority of non-text skeletons are removed in the first stage by geometric constraints as follows, ***Concentration ratio***: the ratio of the $skel_{ol}$'s length to the $skel$'s length. We form this constraint mathematically as,

$$C(skel) = length(skel_{ol}) / length(skel) \qquad (3)$$

Where $skel_{ol} = skel(sum_{ver}(skel) > 1)$ is defined as the overlapped part of $skel$, and $Sum_{ver}(skel)$ is a vector obtained by adding up the values of $skel$ in the vertical direction (see Fig 4(c)). Text often has higher cohesion degree $C(skel)$. Besides, we also employ other simple constraints such as aspect ratio, eccentricity, solidity and extent to implement geometric filtering.

2.2.2. Single-scale CNN Classifier

For further precise classification, we train a convolutional neural network including two convolutional layers with $n_1 = 96$ and $n_2 = 256$ filters respectively, two pooling layers and one fully-connected layer as illustrated in Fig 5. We first resize the cropped character to 32×32 pixel training sample, then perform convolution to obtain a $25\times25\times n_1$ response in the first layer. Next, average pooling with 2×2 window is applied over response, which reduce the response's dimension to $5\times5\times n_1$. Similarly we execute another convolution and pooling in the second layer to obtain $2\times2\times n_2$ response, which are fully connected to the classification layer. Finally the network is trained discriminatively by squared hinge loss.

In the testing stage, we first use Skeleton-cut detector to generate the text proposals P and their bounding boxes B. Instead of computing *full-image*-responses pyramid by *multi-scale*-sliding-window-based detection, in this paper we perform convolutional operation only on P by single scale s, where $s = height(B)/32$. Then the responses of the last layer are fed into the SVM classifier to score each proposal p in P. Finally false proposals with low confidence are rejected, giving the reliable results.

2.3. Text Line Refinement

Though our text detection and filtering approach give comparably high recall, the overlap of text proposals P can be still unsatisfied. To meet the *compactness* criteria, we adjust the proposed bounding boxes B's coordinates within a refinement framework as follows,

Algorithm : Iterative Local Refinement (ILR)

Input: the proposal's bounding box set B; MSER set M.
Output: the refined bounding box set R.

Divide M into 2 category: strong set S which is inside the set B, and weak set W outside.
for $b \in B$ **do**
Expand b to $b_{\exp}$ horizontally;
Calculate the mean features ms of S in b;
repeat
$ws := \{w \in W \mid |yc_{ms} - yc_w| < T_{yc}, ol(b_{\exp}, w) > T_{ol}\}$;
$rs := \{w \in ws \mid Dist_{h,w,sw,c}(ms, w) < T_{h,w,sw,c}\}$;
$S := S \cup rs,\ W := W / rs$;
until $rs = \varnothing$
$R := B \cap S$.
end for

Here, $Dist_f(s,w)$ denotes the normalized L2 distance between s_f and w_f. And f is the feature of a candidate MSER: sw,c refer to stroke width, color of HSV and gray channels respectively. yc means center coordinate in y-axis. Moreover function ol is applied to calculate the overlap ratio of two bounding boxes. And T denotes threshold. As illustrated in Fig. 6, compared to original text proposals B (red rectangle on the top image), the refined bounding boxes R' overlap and recall ratio have been enhanced to a large extent (yellow dashed rectangle on the bottom image). Finally, we employ word-level segmentation using beam search in [22], giving the final detection result.

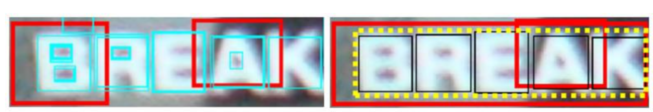


Fig.6. The text line refinement (from left to right).

3. EXPERIMENT

3.1. Datasets and Evaluation Protocol

To evaluate our text detector's performance, we run it on the widely used benchmark ICDAR 2013, which consists of 462 fully-annotated scene text images (229 images for training, and 233 images for testing) . This dataset are quite challenging for varieties of fonts, colors, scales (ranging from 459×405 to 3888×2592) and illuminations. Object count/area graphs [23] protocol is applied to measure recall, precision and f-measure as follows:

$$recall = \frac{\sum_{i=1}^{|G|} Match_G(G_i, D, t_r, t_p)}{|G|}, \quad precision = \frac{\sum_{j=1}^{|D|} Match_G(D_j, G, t_r, t_p)}{|D|}$$

And f is the harmonic mean of precision and recall.

Besides, MSRA-TD500 dataset is applied to validate the proposed method on multiple orientations and languages.

3.2. Experimental Results

As seen in table 1, our proposed method achieve 86.79% in precision, and 76.21% in recall under the scheme of ICDAR 13. Our recall rate outperforms all results from the text detection competition much owes to the well-designed skeleton proposal strategy. In the meanwhile, we employ text classification and refinement model to ensure state-of-the-art precision rate. It is worth mentioning that the only approach whose f-score win against ours is a sliding window based method, which would bear much higher computational cost than the proposed method. In addition, f-score would drop 3.3% when the ILR section is removed, which indicates that iteratively local refinement dose enhance the detection performance both on recall and precision rate.

Table 2 shows the quantitative comparison of character detection rate on ICDAR 2011 benchmark with Sung et al. [24] which propose the state-of-the-art ERs extraction method. The test set includes 255 scene images and 6309 characters. Unlike the refined ERs of Sung et al. that need further classification, our final detection results have eliminated more than 90% of false candidates while preserving comparably high recall rate. This demonstrates that the representation *skeleton*'s adaptability is stronger than that of ERs-based representations.

As shown in table 3, our proposed approach achieves recall 0.64, precision 0.82 and f-score 0.72 on MSRA-TD, which outperform other methods significantly. In summary, our method could successfully detect text in natural images with multiple orientations and mixture of multi-language.

Table 1: Text Localization Performance on ICDAR 2013(%).

Method	*Recall*	*Precision*	*F-score*
Text_CNN [25]	72.89	92.79	81.65
Proposed	***76.21***	***86.79***	***81.16***
without ***ILR***	73.28	84.80	77.83(-3.3)
Tian [26]	75.89	85.15	80.25
Zhu [28]	75.00	85.00	79.00
FCRN+multi_fit[27]	66.30	94.80	78.00
Yin [29]	65.11	83.98	75.89
Neumann [30]	64.84	87.51	74.49

Table 2: Evaluation of character detection rate on ICDAR 2011.

Method	*No. of candidates*	*Detection Rate*
All ERs	6,051,331	0.966
MSERs [31]	39,762	0.539
Sung et al. [24]		
Initial ERs	1,729,833	0.896
Refined ERs	93,357	0.877
Our method		
Initial Skeletons	608,394	0.936
+ Classifier	6,034	0.842
+ Classifier + ILR	8,231	0.879

Table 3: Performance comparisons on MSRA-TD500 benchmark.

Method	*Recall*	*Precision*	*F-score*
Proposed	0.64	0.82	0.72
Yin [32]	0.63	0.81	0.71
Kang [33]	0.62	0.71	0.66
Yin [29]	0.61	0.71	0.65
Yao [5]	0.63	0.63	0.60

3.3. Analysis

Our Skeleton-cut text detector has been shown to be robust and efficient for scene text detection. The good performance of the proposed algorithm is mainly due to the Skeleton-cut proposals strategy which introduce a novel presentation skeleton and accordingly address the edge-adhesion problem via skeleton-junctions elimination, and is partially due to the iteratively local refinement which ensures the text candidates' high recall as well as the two-stage filtering which gives a comparably good precision rate.

4. CONCLUSION

In this paper, we have presented an efficient framework for scene text detection. Different from previous works, which concentrate on scanning through sliding windows or hunting strokes、ERs via CCA, this proposed method is able to cut texts out straightforward from natural images using the novel presentation *skeleton* and concise properties like *con-centration ratio*. On multi-oriented and robust text detection datasets, our approach represents superior performance over other prevalent methods, which indicates that making use of text-specific edge cues for text localization is a worthy-of-being-studied direction. In the future, we will combine our skeletons proposal strategy with faster-RCNN framework to improve detection rate. Moreover, another direction that is worthy of exploring is extend our proposed method to an end-to-end system for text recognition.

5. ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (91520301) and the National High-tech R&D Program of China (2012AA010904).

6. REFERENCES

[1] Chen X, Yuille A L. Detecting and reading text in natural scenes[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2004:366-373.

[2] Epshtein B, Ofek E, Wexler Y. Detecting text in natural scenes with stroke width transform[C]// Computer Vision and Pattern Recognition. IEEE Xplore, 2010:2963-2970.

[3] Neumann L, Matas J. A Method for Text Localization and Recognition in Real-World Images[M]// Computer Vision – ACCV 2010. Springer Berlin Heidelberg, 2010:770-783.

[4] Wang K, Babenko B, Belongie S. End-to-end scene text recognition[C]// IEEE International Conference on Computer Vision. IEEE, 2011:1457-1464.

[5] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu. Detecting texts of arbitrary orientations in natural images. In *Proc. of CVPR*, 2012.

[6] Neumann and J. Matas. Scene text localization and recognition with oriented stroke detection. In *Proc. of ICCV*, 2013.

[7] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven. PhotoOCR: Reading text in uncontrolled conditions. In *Proc. of ICCV*, 2013.

[8] Ye Q, Doermann D. Text Detection and Recognition in Imagery: A Survey[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(7):1480-1500.

[9] C. Yao, X. Zhang, X. Bai, W. Liu, Y. Ma, and Z. Tu. Rotation-invariant features for multi-oriented text detection in natural images. *PLoS One*, 8(8), 2013.

[10] H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk, and B. Girod. Robust text detection in natural images with edge-enhanced maximally stable extremal regions. In *Proc. ICIP 2011*, pages 2609–2612, 2011

[11] C. Yi and Y. Tian. Localizing text in scene images by boundary clustering, stroke segmentation, and string fragment classification. *IEEE Trans. Image Processing*, 21(9):4256–4268, 2012.

[12] C. Yi and Y. Tian. Text extraction from scene images by character appearance and structure modeling. *Computer Vision and Image Understanding*, 117(2):182–194, 2013.

[13] W. Huang, Z. Lin, J. Yang, and J. Wang. Text localization in natural images using stroke feature transform and text covariance descriptors. In *Proc. CVPR 2013*, pages 1241–1248, 2013.

[14] L. Neumann and J. Matas. On combining multiple segmentations in scene text recognition. In *Proc. ICDAR 2013*, pages 523–527, 2013.

[15] C. Shi, C. Wang, B. Xiao, Y. Zhang, and S. Gao. Scene text detection using graph model built upon maximally stable extremal regions. *Pattern Recognition Letters*, 34(2):107–116, 2013.

[16] A. Zamberletti, L. Noce, and I. Gallo. Text localization based on fast feature pyramids and multi-resolution maximally stable extremal regions. In *Proc. Int'l Workshop on Robust Reading (in ACCV 2014)*, pages 91–105, 2014.

[17] K. I. Kim, K. Jung, and J. H. Kim. Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm. *IEEE Trans. Pattern Analysis Machine Intelligence*, 25(12):1631–1639, Dec 2003.

[18] S. M. Hanif and L. Prevost. Text detection and localization in complex scene images using constrained adaboost algorithm. In *Proc. ICDAR 2009*, pages 1–5, 2009.

[19] J.-J. Lee, P.-H. Lee, S.-W. Lee, A. L. Yuille, and C. Koch. Adaboost for text detection in natural scene. In *Proc. ICDAR 2011*, pages 429–434, 2011.

[20] Dollar P, Zitnick C L. Fast Edge Detection Using Structured Forests[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(8):1558-1570.

[21] Wang T, Wu D J, Coates A, et al. End-to-end text recognition with convolutional neural networks[C]// International Conference on Pattern Recognition. IEEE, 2012:3304-3308.

[22] S. J. Russell, P. Norvig, J. F. Candy, J. M. Malik, and D. D. Edwards. *Artificial intelligence: a modern approach*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1996.

[23] Wolf C, Jolion J M. Object count/area graphs for the evaluation of object detection and segmentation algorithms[J]. Document Analysis & Recognition, 2006, 8(4):280-296.

[24] M.-C. Sung, B. Jun, H. Cho, and D. Kim. Scene text detection with robust character candidate extraction method. In *Proc. ICDAR 2015*, pages 426–430, 2015.

[25] He T, Huang W, Yu Q, et al. Text-Attentional Convolutional Neural Network for Scene Text Detection[J]. IEEE Transactions on Image Processing, 2016, 25(6):2529-2541.

[26] S. Tian, Y. Pan, C. Huang, S. Lu, K. Yu, and C. L. Tan. Text flow: A unified text detection system in natural scene images. In *Proc. ICCV 2015 (to appear)*, 2015.

[27] Gupta A, Vedaldi A, Zisserman A. Synthetic data for text localisation in natural images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2315-2324.

[28] Zhu R, Mao X J, Zhu Q H, et al. Text detection based on convolutional neural networks with spatial pyramid pooling[C]// IEEE International Conference on Image Processing. IEEE, 2016:1032-1036.

[29] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao. Robust text detection in natural scene images. *IEEE Trans. Pattern Analysis Machine Intelligence*, 36(5):970–983, May 2014.

[30] L. Neumann and J. Matas. Real-time scene text localization and recognition. In *Proc. CVPR 2012*, pages 3538–3545, 2012.

[31] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust widebaseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10):761–767, 2004.

[32] Yin X C, Pei W Y, Zhang J, et al. Multi-Orientation Scene Text Detection with Adaptive Clustering.[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(9):1930.

[33] Kang L, Li Y, Doermann D. Orientation Robust Text Line Detection in Natural Images[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2014:4034-4041.