2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

# Symmetry-based Object Proposal For Text Detection

Xuelei Zhang†*,Zheng Zhang*,Chengquan Zhang*,Xiang Bai*
†State Key Laboratory of Digital Publishing Technology
Beijing, China 100000
*School of Electronic Information and Communications
Huazhong University of Science and Technology
Wuhan, Hubei, China 430074
Email: {xleizhang,macaroniz1990,zchengquan}@gmail.com,xbai@hust.edu.en

***Abstract*—Scene text detection and recognition have become active research topics in computer vision. In this paper, we focus on the detection of text proposal from wild images. Text proposals attempt to generate a relatively small set of bounding box proposals that are most likely to contain text. Different from previous methods that merge similar region based on property of individual region, we assumed that text word bare strong symmetry property. We propose a new algorithm that exploit the symmetry property to directly generate word-level proposals. Proposals generation process using the region features, and rank process making use of the symmetry structures in text groups. Experiments on two standard datasets demonstrate that the proposed algorithm has achieve the state-of-the-art performance, especially in the case of smaller proposal number.**

## I. Introduction

With the explosive growth of images in the internet and mobile devices, text detection and recognition, as a core technology for a variety of computer vision applications, have become an active research direction in this community. In the past two decades, a rich body of approaches have been proposed, and make impressive progress. These methods can be roughly divided into two categories: texture-based methods [1], [2], [3], [4], and component-based methods [5], [6], [7], [8], [9], [10], [11], [12], [13]. In particular, Maximally Stable Extremal Regions (MSER) [8], as a typical basic representation of the later category, attracts more attention in the filed of scene text detection. However, these approaches are still far away from the real application, due to three major challenges: 1) the diversity of text 2) the complexity of background and 3) expensive time cost.

Recently, a new approach [11] of generic object detection by using deep learning makes a great impact. This method adopts a two-step framework to detect objects: first, a fast but coarse detector is applied to find potential object regions and extract a thousands of proposals with high detection recall; Then, another powerful detector revisits the object proposals, and false positives are removed in this step. Because this two-step framework has the advantages of efficient and accurate, it has become a mainstream in the filed of generic object detection. Inspired by [11], [9] proposed a novel proposal based text detection and recognition approach, and achieved an impressive performance. This approach utilized a generic object proposal algorithm on text to generate word-level proposals. Then, a well trained CNN [14] classifier was applied for removing non-text proposals. [13] proposed a class specific algorithm to generate text proposals in natural scenes. However, both of these two methods have a poorly performance with the small number of proposals. Different from the above methods that generate word-level proposals, [15] proposed a novel approach to generate line-level proposals by exploring the symmetry property of text lines. However, the line-level proposals usually need a complex post-processing to split the words.

In this paper, we exploit the symmetry property to directly generate word-level proposals. Our basic idea is grouping connected component to generate word level proposals at first. Then, the symmetry feature designed for word-level proposals, and the regions features illustrated by [13] are extracted. Finally, an efficient classifier is used to scoring the word proposals by combining the word-level symmetry feature and the region features. Unlike the previous methods exploit the symmetry property at characters [5] or local patches in a sliding window manner [15], we extend the symmetry to the bounding boxes of arbitrary size.

The contributions of this paper can be concluded in two aspects: first, we explore symmetry feature to ranking text proposals. Second, we achieve the state-of-the-art performance on two standard scene text datasets.

In the remainder of this article, we briefly review previous works that focused on the fields of scene text in Section II. Then we present our methods in detail in Section III, including the proposal generation process, distinctive strategy for extracting symmetry feature and feature in text hypothesis. In Section IV, we demonstrate the effectiveness of the proposed algorithm and make comparisons with the other related works. Finally, we conclude the whole work and promising research directions in Section V.

## II. Relate work

Text, as a carrier of human thoughts and emotions, conveys high-level semantic information and acts an important role in many real-world applications, such as image understanding, image retrieval, and navigation system. Text detection in natural images and videos have received much attention in recent years. The comprehensive surveys about scene text reading can be found in [16], [1]. In summary, text detection methods can be divided into two categories: texture-based [1],

978-1-5090-4847-2/16/$31.00 ©2016 IEEE

[2], [3], [4], [17] and component-based [5], [6], [7], [8], [9] [10], [11], [12], [13].

The texture-based methods scan all potential locations and scales in the image. Scanned patches are classified to text or non-text by their texture or local structure properties. Chen et al. [1] built a cascade classifier to distinguish text/non-text patches by combining multiple features. Wang et al. [3] trained a character classifier by using Random Ferns [18], and detected text regions in a sliding window manner. In [17], Neumann et al. combined the advantages of sliding-window and connected component methods, and seek scene text characters at a fine granularity (parts or strokes). The main weakness of these texture-based methods is low efficiency. Moreover, these methods are limited to the detection of a single language for which they have been trained on.

Region-based methods generate class-independent text hypotheses in a bottom-up manner: One image is first divided into a set of segmentations, according to its inherent structure, then are classified into text or non-text. [5] assumed that characters are consisted of one or several connected components and presented a local image operator Stroke Width Transform (SWT). [7] extend the SWT and proposed a low-level filter called Stroke Feature Transform (SFT), and realized better inter-component separation and intra-component connection. [8] extracted MSER based on extremal property of the intensity function in the region and on its outer boundary, and it is the basis of our work. While Yin et al. [10] proposed a multi-stage clustering algorithm for grouping character components to detect text. [11] used the deep convolutional neural network to classifier the character proposals obtained by MSER.

Inspired by selective search method [19] on generic object detection, Gomez et al [12] proposed a fast hierarchical clustering method and adopted a probabilistic measure as the stopping rule. While [9] leverage the convolutional structure of the CNN to generate text saliency maps, and utilized a generic object proposal algorithm on text to generate word-level proposals. Finally, Gomez et al. [13] proposed a text-specific selective search algorithm and ranked the obtained hypotheses by trained classifier.

On the other hand, the proposed algorithm is inspired by previous researches on symmetry detection [5], [15], [20], where symmetry structures can used to distinguish between text regions and non-text regions. Tsogkas et al. [20] developed a learning-based approach to detect symmetry axes in natural images. While [15] employed symmetry property on character group level and devise a multi-scale template in a sliding-window manner.

In this paper, we adopt the same framework proposed in [13] and improve it from two aspect: 1) A orientation consistency constraint is considered in proposal generation stage to reduce the proposals number; 2) We explore the symmetry feature in ranking stage, capturing the text property of a whole text group.

## III. METHODOLOGY

In this section, we will describe the proposed method for text proposal generation in detail. Fig.1 illustrates the pipeline of our method that consists of two major steps: proposals generation and proposals ranking. First, character components are extracted by MSER. Then, the components are merged together by an agglomerative clustering method to form a set of text proposals; Finally, the symmetry feature and the region feature within a group are extracted to rank the proposals by a pre-trained classifier.

### A. Text proposals generation

MSER is widely used in text detection as it is insensitive to variations in scales, positions, and languages [21]. [4] has proved that a majority of character components can be captured by MSER by combining the result from multiple channels. In our method, we use MSER to get the initial low-level text parts $R_c$ from channel c.

In [13], proposals are generated in an bottom-up manner with the single linkage criterion (SLC), in which similar regions are merged together based on complementary features in individual region and diversity distance metrics. We develop this algorithm by considering spatial configuration property of text in natural scenes: characters components from same text line or word are approximate arranged in a straight line. We adopt the SLC with an orientation consistency constraint between two clusters. More specifically, two regions $r_a$ and $r_b$ will be grouped together only when their orientation angle $\gamma(r_a, r_b)$ is under a threshold $T$. This strategy helps us filter out the most of background noises as well as reduce the proposals quantity without any time cost.

We adopt the similar diversity strategy to [13], to detect text as more as possible in any case. We diversify our agglomerative process by following aspects: 1)MSER are extracted from a variety of color channels (i.e. Red, Green, Blue, Gray) and spatial pyramid levels, 2) different distance metrics when apply SLC. In this step, we do not combine any feature cues. Noticed that both of the clustering algorithm and the subsequent rank process are performed in a single color space and scale.

The similarity between two obtained cluster $r_a$ and $r_b$ is defined as:

$$\begin{aligned} d^{(i)}(r_a, r_b) = ||f^{(i)}(r_a) - f^{(i)}(r_b)||^2 \\ +(x_a - x_b)^2 + (y_a - y_b)^2 \\ s.t. \gamma(r_a, r_b) < T \end{aligned} \tag{1}$$

where $f(r)$ is the feature of region $r$, and $\{(x_a - x_b)^2 + (y_a - y_b)^2\}$ is a spatial constraint term between the centers point $(x_a, y_a)$,$(x_b, y_b)$ of the clusters $r_a$ and $r_b$ to ensure adjacent regions can be merged firstly. For a fair comparison, we use the same features as [13]: mean gray value of the region, mean gray value in the immediate outer boundary of the region, regions major axis, mean stroke width, and mean of the gradient magnitude at the regions border.

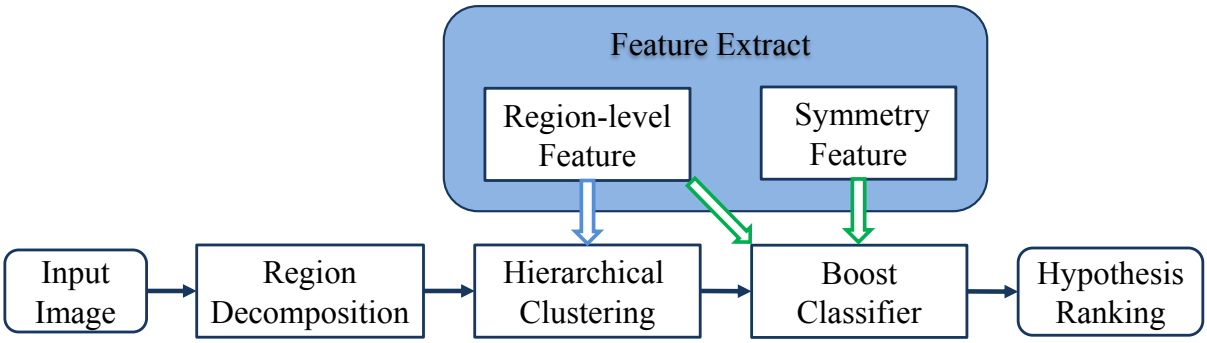


Fig. 1. Pipeline for text detection algorithm: start from region decomposition and extract region-level feature from individual region, then the regions are merged together to form text proposals using hierarchical clustering. Merge the region-level features and symmetry feature from text proposals to train a boost classifier and rank the text proposals

### B. Ranking

Thousands of text proposals are generated in the previous step. Same as the other proposal methods [22], [19], [13]. Now we describe how to score a given proposal by combining the symmetry feature and the region features.

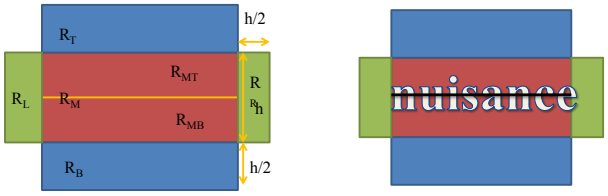


Fig. 2. Left: The middle rectangle $R_M$ denote the proposal generated from the agglomerative clustering process, its height and width is $h$ and $w$ respectively, the height and the width of the top $R_T$ and bottom $R_B$ rectangles are $h/2$ and $w$, the two besides rectangle $R_L$,$R_R$ are $h$ and $h/2$. Right: The contents within the two middle rectangles $R_{MT}$, $R_{MB}$ are similar to each other but dissimilar to the contents of the four surrounding rectangles. Therefore, the symmetry response can distinguish between text region and non-text region

*a) Feature for ranking:* Previous works [15], [20] have demonstrated that symmetry structures in the image can be utilized to locate text regions in nature images, as text regions bear similar color and structure, and have a strong contrast to its local surrounding. In [15], they exploit the symmetry structure at text lines level. They devise a symmetry template to extract features by centering the template at every location and multiple scales. Multiple symmetry probability maps are estimated by Random Forest [23] and line-level bounding boxes are further estimated. The mainly difference of the proposed method is that our symmetry feature is designed for word level. Specifically, beside the symmetry properties on the vertical context of text, we also consider the horizontal context. As shown in Fig.2, the middle rectangle denoted as $R_M$, is the resulting proposal, the other four rectangles denoted by $R_T, R_B, R_L, R_R$ respectively are local background. The height of $R_T, R_B$ is defined as half of the proposals, and the width of $R_L, R_R$ is defined as half of its height.

Following [15], we collect the histgram of low-level cues c in each each rectangle $R_P(P \in T, B, L, R, MT, MB)$ as features, denoted as $h^c(R_P)$. Since adjacent characters share similar structure, text group has strong self-similarity. We define a similarity function $S^c(R_{MT}, R_{MB})$ to characterize the self-similarity of text word.

$$S^c(R_{MT}, R_{MB}) = \chi^2(h^c(R_{MT}), h^c(R_{MB})) \quad (2)$$

where $\chi^2(.)$ is the $\chi^2$-distance function. Meanwhile, the context in a text region is different from its background, they can be regard as another kind of symmetry feature. Since the three inner rectangles $(R_{MT}, R_{MB}, R_M)$ both different from the four outer rectangles $(R_T, R_B, R_L, R_R)$, we define four contrast functions

$$C_l^c(R_M, R_L) = \chi^2(h^c(R_M), h^c(R_L)) \quad (3)$$

$$C_r^c(R_M, R_R) = \chi^2(h^c(R_M), h^c(R_R)) \quad (4)$$

$$C_t^c(R_T, R_{MT}) = \chi^2(h^c(R_T), h^c(R_{MT})) \quad (5)$$

$$C_b^c(R_B, R_{MB}) = \chi^2(h^c(R_B), h^c(R_{MB})) \quad (6)$$

We use four kinds of features: brightness, color, texture and gradient. We create brightness histogram and color histogram by converting the image to LAB color space, and quantize the pixel values from brightness channel L and the color channels a and b into 16 bins respectively. For texture, we use the method proposed in [24]. We calculate the gradient magnitude map, and quantize the values into 16 bins to get gradient histogram. These four kinds of features are used separately to calculate the self-similarity and the four differences, and are combined producing a 25 dimensional feature vector to represent a proposal. The features of an image can be easily calculated in the way of integral channel, thus, thousands of proposals will be processed in a second.

*b) Classifier Training:* Finally, we build a classification model $F$ to scoring the proposals. A higher score means that the proposal is more likely to contain text, and the lower score means that the proposal is more likely to be a background. We use a Real AdaBoost classifier with decision stumps for its efficiency. The features used in our classifier consist of two parts, symmetry features and region-level features. The symmetry features are described as above. The region-level features are the mean $\sigma^i$ and deviation $\mu^i$ of the region features


Fig. 3. Examples of some text proposals obtained using the proposed algorithm in ICDAR013 dataset. Red bounding boxes are ground truth, the green are the proposals with maximumal IoU

$f^i$ in a particular group G ($f^i(r) : r \in G$). We merge the region-level features to further improve the performance and it will be discussed in the experiment section. To deal with some redundant text candidates, we perform Non-Maximal Suppression (NMS) to the ranked list of proposals.

## IV. Experiment

In this section, we evaluate the performance of the proposed algorithm on two standard datasets: ICDAR2013 [25] and SVT [2]. Fig.3 illustrates several detection proposals of the proposed algorithm in the ICDAR2013 dataset.

### A. Dataset

**ICDAR2013**. This dataset is a horizontal text database, including 229 training images and 233 testing images. We follow the standard protocol that evaluate the quality of proposals and analyse the detection recall under certain condition as [13].

**SVT**. The dataset is collected from Google Street View Image, which consists of 100 images for training and 249 images for testing. Text detection is a challenging task in SVT as the images in the dataset exhibits high variability and low resolution. We use the same evaluation protocol as in ICDAR2013.

### B. Implementation Details

We choose the threshold of the MSER algorithm as 13, the orientation angle threshold $T$ is set to $T = \pi/6$. All training samples are collected from the training set of ICDAR2013 and SVT datasets. In training phrase, we first generate proposals of the training images, and the proposals whose IoU with ground truth is larger than 0.7 are collected as positives. The proposals whose IoU is less than 0.1 are collected as negative. We train a AdaBoost classifier with three times hard negative mining.

All of the following experiments were carried out on a regular computer (2.4 GHz 8-core CPU, 64G RAM and Red Hat 4.8,64-bit). All the comparison methods adopt the code provided by authors and the parameters setting are set to default. We follow the "FULL" version and "FAST" version in [13] and propose our corresponding "FULL" and "FAST".

TABLE I
Recall rate using the first ranked 1000 proposals and time performance comparison with state of the art in the ICDAR2013 dataset at different settings. We follow the color channels and cues setting in "fast" [13] to generate proposals, (B)rightness),(C)olor,(G)radient feature for ranking as our "fast","full" in [13] and B,C,G,(T)exture,(R)egion level feature as our "full"

| Method | IoU=0.5 | IoU=0.7 | Time(s) |
|---|---|---|---|
| BING [22] | 0.38 | 0.08 | 1.75 |
| EdgeBoxes [26] | 0.71 | 0.48 | 2.63 |
| RP [27] | 0.62 | 0.38 | 14.32 |
| GOP [28] | 0.45 | 0.18 | 5.41 |
| Object-FAST [13] | 0.82 | 0.69 | **0.96** |
| Object-FULL [13] | 0.83 | 0.72 | 2.55 |
| Ours-FAST | 0.87 | 0.78 | 1.25 |
| Ours-FULL | **0.88** | **0.80** | 4.06 |

### C. Comparison with state of the art

In this section, we compare our proposed algorithm with the state of the art generic object proposals methods [22], [26], [27], [28] and [13].

Quantitative comparison with the other methods in ICDAR2013 are illustrated in Table I and Fig.4. As can be seen in Table I, our proposed algorithm demonstrates higher recall rate using smaller proposals than other methods. It is worth noticing that [22], [26], [27], [28] have poorly performance at large IoU thresholds, while [13] and our method still have promising performance.

The text images in SVT demonstrates a wide variety of fonts and lighting conditions. As demonstrated in Table II and Fig.4, EdgeBoxes is slighter better than [13], while our algorithm still have better performance than other methods at all IoU setting. It is apparent that our method can handle more cases.

TABLE II
DETECTION RATE USING THE FIRST RANKED 1000 PROPOSALS AT DIFFERENT SETTINGS AND TIME PERFORMANCE, COMPARISON WITH STATE OF THE ART IN THE SVT DATASET

| Method | IoU=0.5 | IoU=0.7 | Time(s) |
|---|---|---|---|
| BING [22] | 0.28 | 0.06 | 1.21 |
| EdgeBoxes [26] | 0.61 | 0.39 | 3.46 |
| RP [27] | 0.03 | 0.01 | 13.07 |
| GOP [28] | 0.53 | 0.19 | 5.18 |
| Object-FAST [13] | 0.62 | 0.36 | **0.81** |
| Object-FULL [13] | 0.58 | 0.21 | 2.95 |
| Ours-FAST | 0.78 | 0.46 | 1.20 |
| Ours-FULL | **0.80** | **0.48** | 4.53 |

### D. Evaluation of text feature

Firstly, we analyse the contribution of different features presented in Section III. Table III shows the the recall rate and time cost at different settings. When we use the four text cues to extract the symmetry feature, there is a significant improvement (from 0.72 to 0.78) at 0.7 IoU. When we merge the symmetry feature and the region features, recall rate can further improved to 0.79. It shows that these features are indeed complementary. On the other hand, all features except texture can be extracted within little time, we use the color feature, brightness feature and gradient feature with less proposals as our "FAST", and combine the texture, region feature and whole proposals as "FULL" to further improve the performance.

TABLE III
CONTRIBUTION OF DIFFERENT TYPES OF FEATURES IN THE ICDAR2013 DATASET. WE GENERATE PROPOSALS IN THE FULL VERSION OF OBJECT PROPOSAL [13]. WE INDICATE THE USE OF FEATURE:C(OLOR), B(RIGHTNESS) AND (T)EXTURE, (G)RADIENT IN SYMMETRY FEATURE.(R) DENOTES THE REGION-LEVEL FEATURE IN [13]

| Method | IoU=0.5 | IoU=0.7 | IoU=0.9 | Time(s) |
|---|---|---|---|---|
| R | 0.83 | 0.72 | 0.47 | **2.55** |
| T | 0.77 | 0.60 | 0.36 | 3.58 |
| G | 0.80 | 0.59 | 0.32 | 3.30 |
| B + C | 0.84 | 0.72 | 0.48 | 3.61 |
| B + C + G | 0.85 | 0.75 | 0.48 | 3.63 |
| B + C + T + G | 0.87 | 0.78 | 0.56 | 4.06 |
| B + C + T + G + R | **0.87** | **0.79** | **0.59** | 4.06 |

### E. Evaluation of orientation consistency

We evaluate the impact of orientation consistency constraint $T$ as illustrated in Section III, Fig.5 shows the performance in our "FAST" pipeline at 0.7 IoU. Recall rate is 0.75 when $T = \pi/2$, that means we do not imposed any text orientation constrains in the SLC process. When varying $T = \pi/6$, we get a slighter better result 0.78. It demonstrates that constrain char components in a line can provide smaller and higher quality proposals.

## V. CONCLUSION

In this paper, we presented a novel representation for text proposals utilizing symmetry property. Different from generic selective search methods focus on component grouping algorithms, we explore text specific property at group level. The superior performance over other relate methods demonstrate that symmetry is a promising direction that can be further explored. In the future, we can extend our algorithm to multi-oriental scenario and recognition framework.

## VI. ACKNOWLEDGEMENTS

This work was mainly supported by National Natural Science Foundation of China (NSFC) (No.61222308, No.61573160) and the Opening Project of State Key Laboratory of Digital Publishing Technology.

## REFERENCES

[1] X. Chen and A. L. Yuille, "Detecting and reading text in natural scenes," in *Proc. of CVPR*, 2004, pp. II–366.
[2] K. Wang and S. Belongie, "Word spotting in the wild," in *Proc. of ECCV*. Springer, 2010, pp. 391–405.
[3] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *Proc. of ICCV*, 2011, pp. 1457–1464.
[4] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *Proc. of CVPR*, 2012, pp. 3538–3545.
[5] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proc. of CVPR*, 2010, pp. 2963–2970.
[6] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Proc. of CVPR*, 2012, pp. 1083–1090.
[7] W. Huang, Z. Lin, J. Yang, and J. Wang, "Text localization in natural images using stroke feature transform and text covariance descriptors," in *Proc. of ICCV*, 2013, pp. 1241–1248.
[8] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and vision computing*, vol. 22, no. 10, pp. 761–767, 2004.
[9] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," in *Proc. of ECCV*. Springer, 2014, pp. 512–528.
[10] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, "Robust text detection in natural scene images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 5, pp. 970–983, 2014.
[11] W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolution neural network induced mser trees," in *Proc. of ECCV*. Springer, 2014, pp. 497–511.
[12] L. G. i Bigorda and D. Karatzas, "A fast hierarchical method for multi-script and arbitrary oriented scene text extraction." *CoRR*, 2014.
[13] L. Gomez and D. Karatzas, "Object proposals for text extraction in the wild," in *Proc. of ICDAR*, 2015, pp. 206–210.
[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[15] Z. Zhang, W. Shen, C. Yao, and X. Bai, "Symmetry-based text line detection in natural scenes," in *Proc. of CVPR*, 2015, pp. 2558–2567.
[16] Y. Zhu, C. Yao, and X. Bai, "Scene text detection and recognition: Recent advances and future trends," *Frontiers of Computer Science*, vol. 10, no. 1, pp. 19–36, 2016.
[17] L. Neumann and J. Matas, "Scene text localization and recognition with oriented stroke detection," in *Proc. of ICCV*, 2013, pp. 97–104.
[18] M. Ozuysal, P. Fua, and V. Lepetit, "Fast keypoint recognition in ten lines of code," in *Proc. of CVPR*, 2007, pp. 1–8.

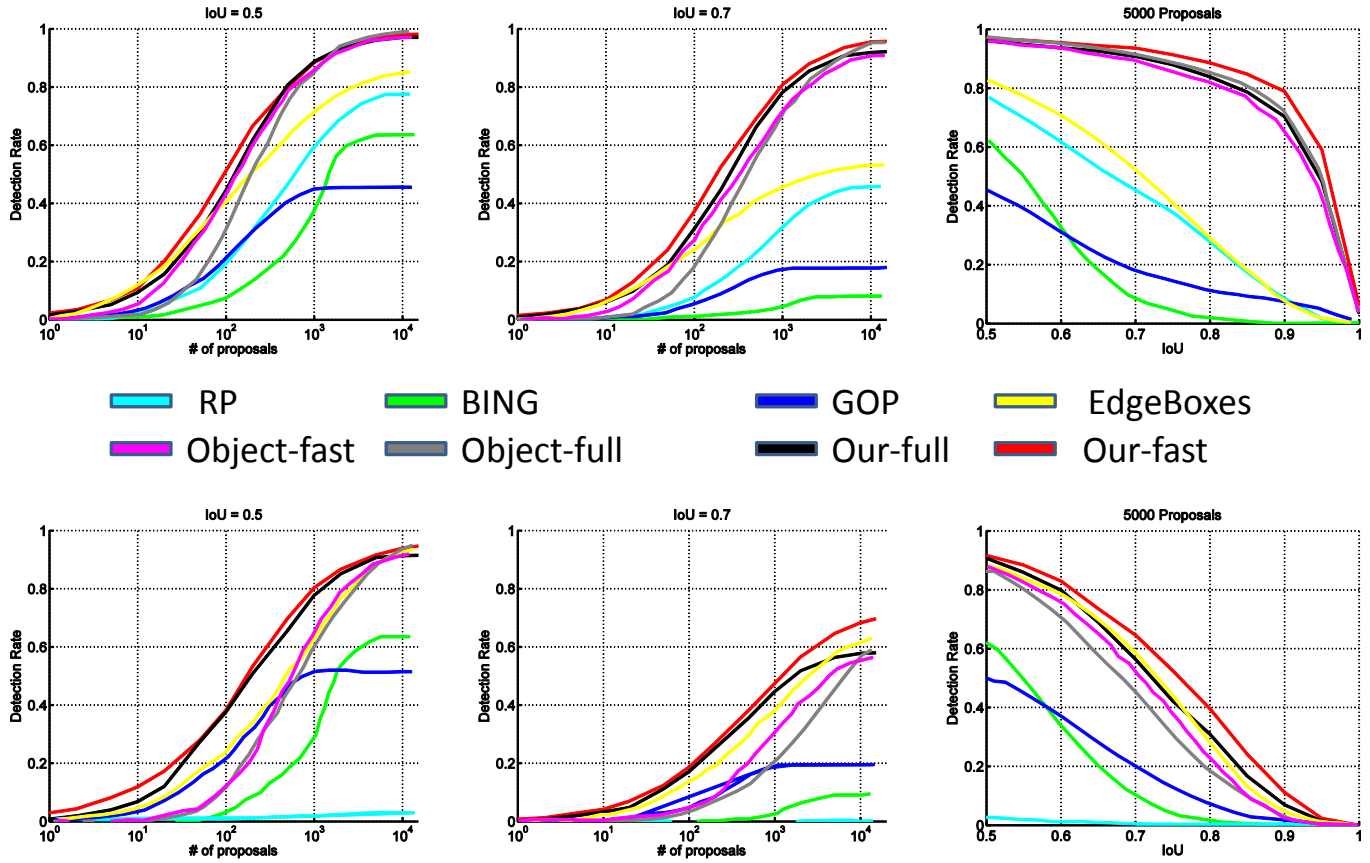


Fig. 4. A comparison of various leading object proposals methods in the ICDAR2013 (top) and SVT (bottom) datasets. (left and center) Detection rate versus number of proposals for various intersection over union thresholds. (right) Detection rate versus intersection over union threshold for various fixed numbers of proposals

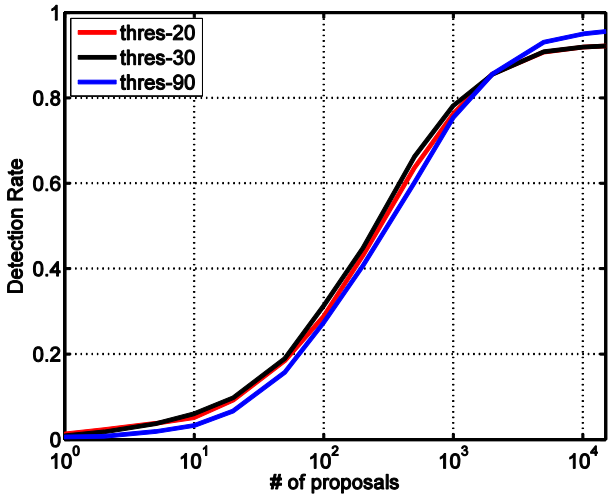


Fig. 5. Impacts of orientation consistency using our "FAST" version at 0.7 IoU in ICDAR013 dataset


[19] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[20] S. Tsogkas and I. Kokkinos, "Learning-based symmetry detection in natural images," in *Proc. of ECCV*. Springer, 2012, pp. 41–54.

[21] L. Gomez and D. Karatzas, "Multi-script text extraction from natural scenes," in *Proc. of ICDAR*, 2013, pp. 467–471.

[22] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *Proc. of CVPR*, 2014, pp. 3286–3293.

[23] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[24] D. R. Martin, C. C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 5, pp. 530–549, 2004.

[25] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. Gomez i Bigorda, S. Robles Mestre, J. Mas, D. Fernandez Mota, J. Almazan Almazan, and L.-P. de las Heras, "Icdar 2013 robust reading competition," in *Proc. of ICDAR*, 2013, pp. 1484–1493.

[26] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. of ECCV*. Springer, 2014, pp. 391–405.

[27] S. Manen, M. Guillaumin, and L. Gool, "Prime object proposals with randomized prim's algorithm," in *Proc. of ICCV*, 2013, pp. 2536–2543.

[28] P. Krähenbühl and V. Koltun, "Geodesic object proposals," in *Proc. of ECCV*. Springer, 2014, pp. 725–739.