DEEPTEXT: A NEW APPROACH FOR TEXT PROPOSAL GENERATION AND TEXT DETECTION IN NATURAL IMAGES

Zhuoyao Zhong, Lianwen Jin[+], Shuangping Huang

South China University of Technology, China
[+]lianwen.jin@gmail.com

ABSTRACT

In this paper, we develop a new approach called DeepText for text region proposal generation and text detection in natural images via a fully convolutional neural network (CNN). First, we propose the novel inception region proposal network (Inception-RPN), which slides an inception network with multi-scale windows over the top of convolutional feature maps and associates a set of text characteristic prior bounding boxes with each sliding position to generate high recall word region proposals. Next, we present a powerful text detection network that embeds ambiguous text category (ATC) information and multi-level region-of-interest pooling (MLRP) for text and non-text classification and accurate localization refinement. Our approach achieves an F-measure of 0.83 and 0.85 on the ICDAR 2011 and 2013 robust text detection benchmarks, outperforming previous state-of-the-art results.

Index Terms— text detection, convolutional neural network, region proposal network, natural images

1. INTRODUCTION

Text detection is a procedure that determines whether text is present in natural images and, if it is, to detect and locate each text instance. Text in images provides rich and precise high-level semantic information, which is important for numerous promising applications such as scene understanding, image and video retrieval, and content-based recommendation systems. Consequently, text detection in natural scenes has attracted considerable attention in the computer vision and image understanding community [1–12]. However, text detection in the wild is still a challenging and unsolved problem because of the following factors. First, a text image background is very complex and some region components such as signs, bricks, and grass are difficult to distinguish from text. Second, scene text can be diverse and usually exists in various colors, fonts, orientations, languages, and scales in natural images. Furthermore, there are highly confounding factors, such as non-uniform illumination, strong exposure, low contrast, blurring, low resolution, and occlusion, which pose hard challenges for the text detection task.

In the last few decades, sliding window-based and connected component-based methods have become mainstream approaches to the text detection problem. Sliding window-based methods [1–2] use different ratios and scales of sliding windows to search for the presence of possible text positions in pyramid images, incurring a high computational cost. Connected component based methods, represented by maximally stable extremal regions (MSERs) [3–6] and the stroke width transform (SWT) [7], extract character candidates and group them into word or text lines. Especially, previous approaches applying MSERs as the basic representation have achieved promising performance in the ICDAR 2011 and 2013 robust reading competitions [8–9]. However, MSERs focuses on low-level pixel operations and mainly accesses local component information, which leads to poor performance in some challenging situations, such as multiple connected characters, segmented stroke characters, and non-uniform illumination, as listed in [10]. Further, this bottom-up approach results in sequential error accumulation in the total text detection pipeline, as stated in [11].

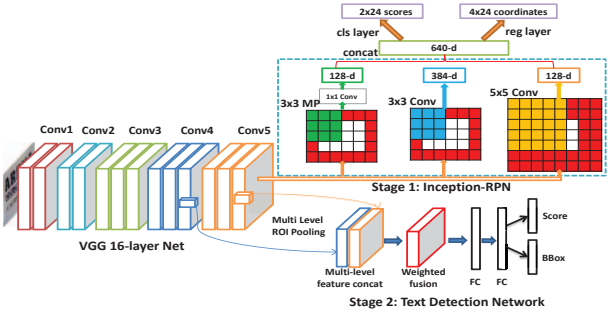

Figure 1: Pipeline architecture of DeepText. Our approach takes a natural image as input, generates hundreds of word region proposals via Inception-RPN (Stage 1), and then scores and refines each word proposal using the text detection network (Stage 2).

Rather than extracting character candidates, Jaderberg *et al.* [12] applied complementary region proposal methods called edge boxes (EB) [13] and aggregate channel features (ACF) [14] to perform word detection and acquired a high word recall with tens of thousands of word region proposals. They then employed HOG features and a random forest classifier to remove non-text region proposals and used a CNN for bounding box regression. They achieved superior text spotting and text-based image retrieval performance on several standard benchmarks. Actually, the region proposal generation step in the generic object detection pipeline has attracted much interest. In recent studies, object detection models based on region proposal algorithms to hypothesize class-specific or class-agnostic object locations have achieved state-of-the-art detection performance [15–18]. However, standard region proposal algorithms such as selective search (SS) [19], multiscale combinatorial grouping (MCG) [20], and EB [13] generate an extremely large number of region proposals. This leads to high recall, but burdens the follow-up classification and regression models and is also relatively

time-consuming. In order to address these issues, Ren et al. [21] proposed region proposal networks (RPNs), which computed region proposals with a deep fully CNN. They generated fewer region proposals, but achieved a promising recall rate under different overlap thresholds. Moreover, RPN and Fast R-CNN can be combined into a joint network and trained to share convolutional features. Owing to the above innovation, this approach achieved better object detection accuracy in less time than Fast R-CNN with SS [17] on PASCAL VOC 2007 and 2012.

In this paper, inspired by [21], our motivation is to design a unified approach for text characteristic region proposal generation and text detection in natural images. In order to avoid the sequential error accumulation of bottom-up character candidate extraction strategies, we focus on word proposal generation. To accomplish this, we incorporate the advantages of the inception module [22] and RPN [21], and propose the novel inception RPN (Inception-RPN). Inception-RPN slides an inception network with multi-scale windows over the top of convolutional feature maps and unites a set of text characteristic prior bounding boxes with each sliding position to generate word region proposals. The multi-scale sliding-window feature can retain local information as well as contextual information at the corresponding position, which helps to filter out non-text prior bounding boxes. Our Inception-RPN enables achieving a high recall with only hundreds of word region proposals. Subsequently, we present a powerful text detection network by introducing extra ambiguous text category (ATC) information and multi-level region of interest (ROI) pooling into the optimization process, which contribute to learn more discriminative information for distinguishing text from complex backgrounds. Our approach achieves an F-measure of 0.83 and 0.85 on the ICDAR 2011 and 2013 robust text detection benchmarks, respectively, outperforming the previous state-of-the-art results.

2. METHODOLOGY

2.1. Text region proposal generation

Our inception-RPN method resembles the notion of RPN proposed in [21], which takes a natural scene image and set of ground-truth bounding boxes that mark text regions as input and generates a manageable number of candidate word region proposals. To search for word region proposals, we apply an inception network to slide over the top of convolutional feature maps (Conv5_3) in the VGG16 model [23] and associate a set of text characteristic prior bounding boxes with each sliding position. The details are as follows.

Text characteristic prior bounding box design. Our prior bounding boxes are similar to the anchor boxes defined in RPN. Taking text characteristics into consideration, for most word or text line instances, width is usually greater than height; in other words, their aspect ratios are usually less than one. Furthermore, most text regions are small in natural images. Therefore, we empirically design four scales (32, 48, 64, and 80) and six aspect ratios (0.2, 0.5, 0.8, 1.0, 1.2, and 1.5), for a total of $k = 24$ prior bounding boxes at each sliding position, which is suitable for text properties as well as incident situations. In the training stage, we assign a positive label to a prior box that has an intersection over union (IoU) overlap greater than 0.5 with a ground-truth bounding box, while assigning a background label to a prior box with an IoU overlap less than 0.1 with any ground-truths.

Inception-RPN. We design Inception-RPN, inspired by the idea of the inception module in GoogLeNet [22], which used flexible convolutional or pooling kernel filter sizes with a layer-by-layer structure to achieve local feature extraction. This method has proved to be robust for large-scale image classification. As depicted in the top half of Fig. 1, our designed inception network consists of a 3×3 convolution, 5×5 convolution, and 3×3 max pooling layers, which is fully connected to the corresponding spatial receptive fields of the input Conv5_3 feature maps. That is, we applied 3×3 convolution, 5×5 convolution, and 3×3 max pooling to extract local feature representations over Conv5_3 feature maps at each sliding position simultaneously. In addition, 1×1 convolution is employed on the top of the 3×3 max pooling layer for dimension reduction. We then concatenate each part feature along the channel axis and a 640-d concatenated feature vector is fed into two sibling output layers: a classification layer that predicts textness score of the region and a regression layer that refines the text region location for each kind of prior bounding box at this sliding position. An illustration of Inception-RPN is shown in the top part of Fig. 1. Inception-RPN has the following advantages: (1) the multi-scale sliding-window feature can retain local information as well as contextual information thanks to its center restricted alignment at each sliding position, which helps to classify text and non-text prior bounding boxes, (2) the coexistence of convolution and pooling is effective for more abstract representative feature extraction, as addressed in [22], and (3) experiments show that Inception-RPN substantially improves word recall at different IoU thresholds with the same number of word region proposals.

Note that for a Conv5_3 feature map of size $m \times n$, Inception-RPN generates $m \times n \times 24$ prior bounding boxes as candidate word region proposals, some of which are redundant and highly overlap with others. Therefore, after each prior bounding box is scored and refined, we apply non-maximum suppression (NMS) [26] with an IoU overlap threshold of 0.7 to retain the highest textness score bounding box and rapidly suppress the lower scoring boxes in the neighborhood. We next select the top-2000 candidate word region proposals for the text detection network in the training phase and the top-300 proposals for testing, respectively.

2.2. Text detection

ATC incorporation. As in many previous works (e.g., [21]), a positive label is assigned to a proposal that has an IoU overlap greater than 0.5 with a ground truth bounding box, while a background label is assigned to a proposal that has an IoU overlap in the range $[0.1, 0.5)$ with any ground-truths in the detection network. However, this method of proposal partitioning is unreasonable for text because a proposal with an IoU overlap in the interval $[0.2, 0.5)$ may probably contain partial or extensive text information, as shown in Fig. 2. We note that promiscuous label information may confuse the

learning of the text and non-text classification network. To tackle this issue, we refine this proposal label partition strategy to make it suitable for text classification. Hence, we assign a positive text label to a proposal that has an IoU overlap greater than 0.5 with a ground truth, while assigning an additional "ambiguous text" label to a proposal that has an IoU overlap with a ground truth bounding box in the range $[0.2,0.5)$. In addition, a background label is assigned to any proposal that has an IoU overlap of less than 0.2 with any ground-truths. We assume that more reasonable supervised information incorporation helps the classifier to learn more discriminative feature to distinguish text from complex and diverse backgrounds and filter out non-text region proposals.

MLRP. The ROI pooling procedure performs adaptive max pooling and outputs a max-pooled feature with the original C channels and spatial extents $H \times W$ for each bounding box. Inspired by [33], to better utilize the multi-level convolutional features and enrich the receptive field information of each bounding box, we perform MLRP over the Conv4_3 as well as Conv5_3 convolutional feature maps of the VGG16 network. Different from [33], we find it good enough to perform ROI pooling over Conv4_3 and Conv5_3 and unnecessary to L2 normalize pooled features from these two layers in our detection network. After MLRP, we obtain two $512 \times H \times W$ pooled features (both H and W are set to 7 in practice). Then, we apply channel concatenation on each pooled feature and encode the concatenated feature with a $512 \times 1 \times 1$ convolutional layer. The 1×1 convolutional layer: (1) combines the multi-level pooled features and learns the fusion weights in the training process and (2) reduces the dimensions to match VGG16's first fully-connected layer. An illustration of MLRP is depicted in the bottom half of Fig. 1. The multi-level weighted fusion feature is then accessed to the bounding box classification and regression model.

2.3. Learning optimization

Both Inception-RPN and the text detection network have two sibling output layers: a classification layer and a regression layer. We minimize a multi-task loss function, as in [15]:

$$L(p, p^*, t, t^*) = L_{cls}(p, p^*) + \lambda L_{reg}(t, t^*), \quad (1)$$

where classification loss L_{cls} is a softmax loss and p and p^* are given as the predicted and true labels, respectively. Regression loss L_{reg} applies the smooth-L_1 loss defined in [17]. Besides, $t = \{t_x, t_y, t_w, t_h\}$ and $t^* = \{t_x^*, t_y^*, t_w^*, t_h^*\}$ stand for the predicted and ground-truth bounding box regression offset vectors, respectively, where t^* is encoded as follows:

$$t_x^* = (G_x - P_x)/P_w, \; t_y^* = (G_y - P_y)/P_h,$$
$$t_w^* = log\left(\frac{G_w}{P_w}\right), t_h^* = log\left(\frac{G_h}{P_h}\right). \quad (2)$$

Figure 2: Example word region proposals with an IoU overlap within the interval $[0.2,0.5)$

Here, $P = \{P_x, P_y, P_w, P_h\}$ and $G = \{G_x, G_y, G_w, G_h\}$ denote the center coordinates (x-axis and y-axis), width, and height of proposal P and ground-truth box G, respectively. Furthermore, λ is a loss-balancing parameter, and we set $\lambda = 3$ for Inception-RPN to bias it towards better box locations and $\lambda = 1$ for the text detection network.

In contrast to the proposed four-step training strategy to combine RPN and Fast-RCNN in [21], we train our inception-RPN and text detection network in an end-to-end manner via back-propagation and stochastic gradient descent (SGD). Furthermore, we apply the iterative bounding box regression scheme mentioned in [18]. The implementation details are as follows: (1) the shared convolutional layers are initialized by a pre-trained VGG16 model for imageNet classification [23]; (2) all the weights of the new layers are initialized with a zero mean and a standard deviation of 0.01 Gaussian distribution; (3) the base learning rate is 0.001 and is divided by 10 for each 40K mini-batch until convergence; (4) the momentum is 0.9 and weight decay is 0.0005; (5) all experiments were conducted in Caffe [24].

3. EXPERIMENTS AND ANALYSIS

3.1. Experimental data

The ICDAR 2011 dataset includes 229 and 255 images for training and testing, respectively, and there are 229 training and 233 testing images in the ICDAR 2013 dataset. Obviously, the number of training image is constrained to train a reasonable network. In order to increase the diversity and number of training samples, we collect an indoor database that consisted of 1,715 natural images for text detection and recognition from the Flickr website, which is publicly available online[1] and free for research usage. In addition, we manually selected 2,028 images from the COCO-Text benchmark [25]. Ultimately, we collected 4,072 training images in total.

3.2. Experimental results on full text detection

We evaluate the proposed DeepText detection system on the ICDAR 2011 and 2013 robust text detection benchmarks following the standard evaluation protocol of ICDAR 2011 [32] and 2013 [9]. Our DeepText system achieves 0.83 and 0.85 F-measure on the ICDAR 2011 and 2013 datasets, respectively. Comparisons with recent methods on the ICDAR 2011 and 2013 benchmarks are shown in Tables 1 and 2. In the tables, we can see that our proposed approach outperforms previous results with a substantial improvement, which can be attributed to simultaneously taking high recall and precision into consideration in our system. The high performance achieved on both datasets highlights the robustness and effectiveness of our proposed approach. Further, qualitative detection results under diverse challenging conditions are shown in Fig. 3, which demonstrates that our system is capable of detecting non-uniform illumination, multiple and small regions, as well as low contrast text regions in natural images.

[1] https://pan.baidu.com/s/1kVRIpd9

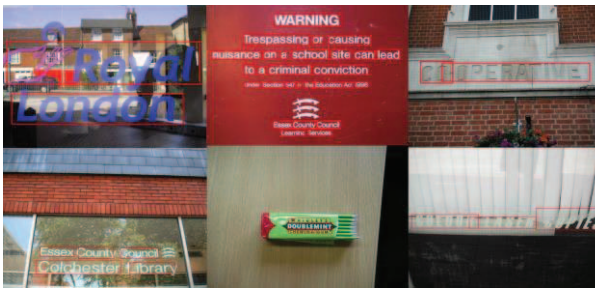

Figure 3: Example detection results of our DeepText system on the ICDAR 2011 and ICDAR 2013 benchmarks.

3.3. Evaluation of Inception-RPN

In this section, we compare Inception-RPN with the text characteristic prior bounding boxes (Inception-RPN-TCPB) to state-of-the-art region proposal algorithms SS [19], EB [13], standard RPN [21] and RPN with TCPB (RPN-TCPB). We compute the recall rate of word region proposals at different IoU overlap thresholds with ground-truth bounding boxes on the ICDAR 2013 testing set, which includes 1095 word-level annotated text regions. In Fig. 4, we show the results of using N= 100, 300, and 500 word region proposals, where the N proposals are the top-N scoring word region proposals ranked in term of these methods. The plots demonstrate that our Inception-RPN-TCPB considerably outperforms RPN-TCPB and standard RPN by 2%-3% and 8%–10% as well as is superior to SS and EB with a notable improvement when the number of word region proposals drops from 500 to 100. Note that our proposed Inception-RPN-TCPB enables achieving a high recall of nearly 90% with only hundreds of word proposals.

3.4. Contributions of proposed methods

In the proposed DeepText system, we mainly employed four methods: Inception-RPN, TCPB, ATC and MLRP. To evaluate the contributions of different techniques, we conducted a series of experiments on the ICDAR 2013 dataset with different settings: standard RPN [21], Inception-RPN with TCPB, Inception-RPN with TCPB and ATC, Inception-RPN with TCPB and MLRP as well as their combination. The results are listed in Table 3. It shows that: (1) Inception-RPN with TCPB rapidly improves the recall rate as stated in section 3.3; (2) ATC considerably enhances the precision rate but discards the recall rate since it removes major false positive proposals while slightly increase the false negative error; (3) MLRP is effective to improve the precision and recall rate by incorporating multi-level pooled feature, which is effective for learning more discriminative features to distinguish text from non-text; (4) the combination of Inception-RPN with TCPB, ATC and MLRP leads to a significant improvement in text detection performance.

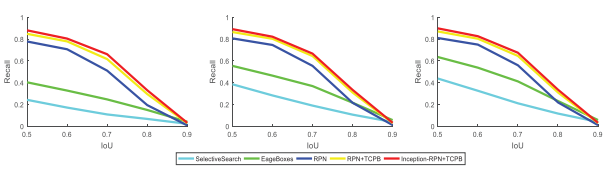

Figure 4: Recall vs. IoU overlap threshold on the ICDAR 2013 testing set. Left: 100 word region proposals. Middle: 300 word region proposals. Right: 500 word region proposals.

Table 1 Comparison with state-of-the-art methods on the ICDAR 2011 benchmark

Method	Precision	Recall	F-measure
DeepText (ours)	0.85	**0.81**	**0.83**
TextFlow[11]	0.86	0.76	0.81
Zhang *et al.* [27]	0.84	0.76	0.80
MSERs-CNN [6]	**0.88**	0.71	0.78
Yin *et al.* [5]	0.86	0.68	0.75
SFT-TCD [28]	0.82	0.75	0.73

Table 2 Comparison with state-of-the-art methods on the ICDAR 2013 benchmark

Method	Precision	Recall	F-measure
DeepText (ours)	**0.87**	**0.83**	**0.85**
TextFlow[11]	0.85	0.76	0.80
Zhang et al. [27]	0.88	0.80	0.80
Neumann [29]	0.82	0.72	0.77
FASText [30]	0.84	0.69	0.77
Yin et al. [5]	0.88	0.66	0.76
Text Spotter [31]	0.88	0.65	0.75

Table 3 Contributions of different proposed methods

Method	Precision	Recall	F-measure
Inception-RPN+TCPB+ ATC+MLRP	0.87	0.83	0.85
Inception-RPN+TCPB+MLRP	0.85	0.81	0.82
Inception-RPN+TCPB+ATC	0.83	0.76	0.80
Inception-RPN+TCPB	0.77	0.79	0.78
RPN [21]	0.75	0.71	0.73

4. CONCLUSIONS

In this paper, we presented a new approach called DeepText for text detection in natural images with a powerful fully CNN in an end-to-end learning manner. DeepText consists of an Inception-RPN with a set of text characteristic prior bounding boxes for high quality word proposal generation and a powerful text detection network embedding ATC and MLRP for proposal classification and accurate localization. Experimental results show that our approach achieves state-of-the-art F-measure performance on the ICDAR 2011 and 2013 robust text detection benchmarks, substantially outperforming previous methods. The example detection results show that our proposed DeepText is capable of detecting low contract, multiple and small, as well as uniform illumination text instances in natural images.

5. ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61472144 and Grant 61502176; in part by GDSTP under Grant 2015B010101004, Grant 2015B010130003, Grant 2015B010131004; and in part by the National Key Research & Development Plan of China under Grant 2016YFB1001405.

6. REFERENCES

[1] T.Wang, D. J.Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," In *Proc. ICPR*, 2012.

[2] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," In *Proc. ECCV*, 2014.

[3] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," In *Proc. ACCV*, 2010.

[4] L. Sun, Q. Huo, and W. Jia, "A robust approach for text detection from natural scene images," *Pattern Recognition*, 48(9): 2906-2920, 2015.

[5] X. Yin, X Yin, K. Huang, and H. Hao, "Robust text detection in natural scene images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(5): 970-983, 2014.

[6] W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolutional neural networks induced mser trees," in *Proc. ECCV,* 2014

[7] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," In *Proc. CVPR*, 2010.

[8] D. Karatzas, S. Robles Mestre, J. Mas, F. Nourbakhsh, and P. Pratim Roy, "ICDAR 2011 robust reading competition," In *Proc ICDAR*, 2011.

[9] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. de las Heras, "ICDAR 2013 robust reading competition," In *Proc ICDAR*, 2013.

[10] S. Zhang, M. Lin, T. Chen, L. Jin, and L. Lin, "Character proposal network for robust text extraction," In *Proc ICASSP*, 2016.

[11] S. Tian, Y. Pan, C. Huang, S. Lu, K. Yu, and C. L. Tan, "Textflow: A unified text detection system in natural scene images," In *Proc ICCV*, 2015.

[12] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *International Journal of Computer Vision*, 116(1): 1-20, 2016.

[13] C. L. Zitnick, and P. Dollár, "Edge boxes: Locating object proposals from edges," In *Proc. ECCV*, 2014.

[14] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(8): 1532-1545, 2014.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," In *Proc. CVPR*, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," In *Proc ECCV*, 2014.

[17] R. Girshick, "Fast r-cnn," In *Proc. ICCV*, 2015.

[18] S. Gidaris and N. Komodakis, "Object detection via a multiregion & semantic segmentation-aware cnn model," In *Proc. ICCV*, 2015.

[19] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders, "Segmentation as selective search for object recognition," In *Proc. ICCV*, 2011.

[20] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," In *Proc. CVPR*, 2014.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," In *Proc. NIPS, 2015.*

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," In *Proc. CVPR*, 2015.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," In *Proc. ICLR,* 2015.

[24] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[25] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images," *arXiv preprint arXiv:1601.07140*, 2016.

[26] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," In *Proc. ICPR 2006.*

[27] Z. Zhang,W. Shen, C. Yao, and X. Bai, "Symmetry-based text line detection in natural scenes," In *Proc. CVPR 2015.*

[28] W. Huang, Z. Lin, J. Yang, and J. Wang, "Text localization in natural images using stroke feature transform and text covariance descriptors," In *Proc. ICCV 2013.*

[29] L. Neumann and J. Matas, "Efficient scene text localization and recognition with local character refinement," In *Proc. ICDAR 2015.*

[30] M. Busta, L. Neumann, and J. Matas, "Fastext: Efficient unconstrained scene text detector," In *Proc. ICCV 2015.*

[31] L. Neumann and K. Matas, "Real-time scene text localization and recognition," In *Proc. CVPR 2012.*

[32] C. Wolf and J. Jolion, "Object count/area graphs for the evaluation of object detection and segmentation algorithms," *International Journal on Document Analysis and Recognition*, 8(4): 280-296, 2006.

[33] S. Bell, C.L. Zitnick, K. Bala and R. Girshick, "Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks," In *Proc. CVPR 2016.*