WORDFENCE: TEXT DETECTION IN NATURAL IMAGES WITH BORDER AWARENESS

Andrei Polzounov[1], *Artsiom Ablavatski*[2], *Sergio Escalera*[3], *Shijian Lu*[2], *Jianfei Cai*[4]

[1]Universitat Politècnica da Catalunya, [2] A*STAR Institute for Infocomm Research,
[3]Universitat de Barcelona and Computer Vision Center, [4]Nanyang Technological University

ABSTRACT

In recent years, text recognition has achieved remarkable success in recognizing scanned document text. However, word recognition in natural images is still an open problem, which generally requires time consuming post-processing steps. We present a novel architecture for individual word detection in scene images based on semantic segmentation. Our contributions are twofold: the concept of WordFence, which detects border areas surrounding each individual word and a novel pixelwise weighted softmax loss function which penalizes background and emphasizes small text regions. WordFence ensures that each word is detected individually, and the new loss function provides a strong training signal to both text and word border localization. The proposed technique avoids intensive post-processing, producing an end-to-end word detection system. We achieve superior localization recall on common benchmark datasets - 92% recall on ICDAR11 and ICDAR13 and 63% recall on SVT. Furthermore, our end-to-end word recognition system achieves state-of-the-art 86% F-Score on ICDAR13.

***Index Terms*—** CNN, segmentation, word detection

1. INTRODUCTION AND RELATED WORK

Detection and recognition of text in natural images has long been an outstanding challenge in the computer vision and machine learning communities. Text recognition in the wild can provide context and semantic information for scene understanding, object classification or action recognition in images or video. The task has attracted the interest of many researchers [1, 2, 3, 4, 5, 6]. Due to the difficulty of text detection in natural images, even state-of-the-art systems struggle with word localization because of the staggering variety of text sizes and fonts, potentially poor image quality, low contrast, image distortions, or presence of patterns visually similar to text such as: signs, icons or textures. Most works employ knowledge-based algorithms and heuristics in order to tackle these challenges. Common techniques include: text line extraction [6, 3], character candidate detection [4] or secondary classifiers to remove false positive detections [2].

Recent successes in computer vision are centered on convolutional neural networks (CNNs). Some of the problems being addressed with CNNs include: object-classification in natural images, pixelwise semantic segmentation [7, 8, 9], bounding box detection [10, 11, 12] and text detection in scene images [2, 1, 13, 5, 14, 15].

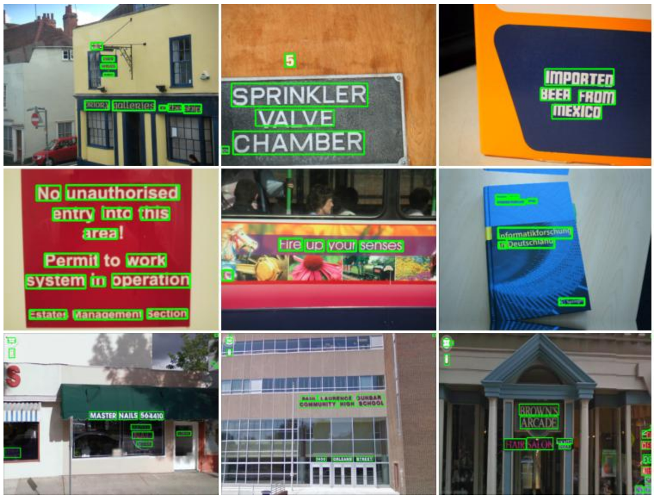

Fig. 1. Word detection bounding box results on ICDAR2011 (top), ICDAR2013 (middle) and SVT (bottom) datasets. Bounding boxes are the output of the proposed method.

A major limitation of CNNs is that networks have trouble taking different scales of images into account. Networks generally use max-pooling layers to reduce the search space for training - this operation reduces resolution and loses spatial information between different features. Yu and Koltun [9] argued that max-pooling does not maintain global scale information and propose dilated convolutions to increase the effective receptive field of convolutional operations. Other works tackled the scale problem with methods such as fully convolutional networks (FCNs) [7] or with atrous convolutions [12, 8]. Another challenge addressed by CNNs is semantic segmentation - where each pixel in the image has to be matched to a specific label. Semantic segmentation has recently been enhanced by dilated convolutions [9], FCNs [7] and probabilistic graphical models [8].

Traditionally text recognition has focused on documents and several optical character recognition (OCR) techniques

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

have been developed for this task. Text recognition in scene imagery however, requires localizing the text first. Generally, text recognition works by first providing a ”candidate bounding box” – or a proposal for a single word or a word-line. This word proposal is then cropped out of the natural image and fed to a word recognition network which then matches words against an internal dictionary.

In the aforementioned scenario, text localization is considered to be the key task, since a well-cropped proposal can be fed to a word recognition system [16]. Before CNNs, popular methods for text localization utilized computer vision techniques with hand-crafted feature descriptors. More recent works have used CNN features. However, all of these approaches have a limitation of feature driven engineering - there are simply too many edge cases to account for. The detectors generate a large amount of non-text false positives, requiring additional filtering techniques. Often, a number of post-processing steps is needed to reach a good performance.

With the prominence of deep learning, CNN based regression of candidate bounding boxes has started becoming utilized for filtering false positive candidates. Bounding box detection has been proposed in the context of object detection by works such as You Only Look Once (YOLO) [10], Faster-RCNN (F-RCNN) [17] and SSD: Single Shot MultiBox Detector [11]. Advances in semantic segmentation [13, 6] have allowed dense prediction to provide input to bounding box regressors. Building on successful implementations of CNNs for semantic segmentation using FCNs for dense prediction [7], several researchers have introduced object localization via FCNs [18].

Early work by Zhang *et al.* [13] used a semantic segmentation model to extract text proposals and refine them by applying hand-crafted heuristics. He *et al.* [6] improved on previous approaches by introducing a cascade of networks. Gupta *et al.* [1] adapted YOLO’s approach [10] for text detection and introduced SynthText - a new synthetic text dataset for training. Analogously, F-RCNN [17] was adapted for text recognition by Zhong *et al.* [5] and Tian *et al.* [3]. The former integrated the F-RCNN framework into a more powerful model. However, a large number of proposals needed to be filtered with a time consuming process. Tian *et al.* [3] fused F-RCNN with a recurrent neural network (RNN), allowing the RNN to consider the proposals as a sequence.

Most current state-of-the-art region of interest (ROI) detectors like F-RCNN [17] use a variation of the following steps: propose bounding boxes, resample pixels of the ROI and then apply a second classifier to filter and improve proposals. In contrast with F-RCNN, our high quality segmentations allow us to extract accurate bounding box proposals directly from the segmentation. The segmentation maps are obtained by inference at different image scales, combining the results with an efficient voting mechanism. Merging the results from different scales helps to eliminate duplicate proposals for the same word and to remove most false positive detections.

Our proposed architecture is inspired by previously mentioned works, but it allows to perform bounding box detection in a single step. Instead of producing a highly non-linear bounding box coordinate prediction as in YOLO [10] and Faster-RCNN [17], our network takes advantage of semantic segmentation to produce a dense pixel labeling map. Afterwards, word proposals are extracted from the given heat map in linear time (see Fig. 1 for examples).

2. WORDFENCE DETECTION NETWORK

Inspired by the success of deep CNNs with residual connections (ResNets), such as the one for semantic segmentation by Chen *et al.* [8], our proposed WordFence Detection Network (WDN) takes advantage of recent deep learning research to produce highly accurate text detection results. The network includes a ResNet-101 (introduced by He *et al.* [19]), followed by a number of dilated convolutions [9] that add full image context to the final classification, before performing a bilinear interpolation on the resulting belief map. Afterwards, connected components are extracted. Each component represents a standalone word on the image which is further processed in the recognition step. Bounding boxes are then extracted from the connected components. Examples are shown in Fig. 2.

2.1. Word Localization as Semantic Segmentation

Object segmentation, has recently been considerably improved with the introduction of the deconvolutional layer [7], dilated convolutions (increasing effective receptive field) [9], *etc.* Several published works [13, 6] have adapted object segmentation for text localization. Segmentation for text localization, despite showing promising results, has had trouble distinguishing individual words from segmented images. Generally, post processing methods and heuristics were applied to refine word localization results, or the task was not addressed at all as in the case of textline approaches.

2.2. ResNet of Exponential Receptive Fields

Recently ResNets have achieved great success in different computer vision tasks [19, 8], even surpassing human performance. Their structure allows ResNets to train very deep neural networks without a vanishing gradient.

In contrast to the semantic segmentation model introduced by Chen *et al.* [8], we do not use parallel replications of ResNet-101 on different scales as it makes the network computationally expensive to train. Instead, we use three parallel convolutional layers of the same kernel size, but different dilation parameters. This way we transform the convolutional features into parallel segmentation maps of different receptive fields. Separate dilated convolutions allow us to enlarge the

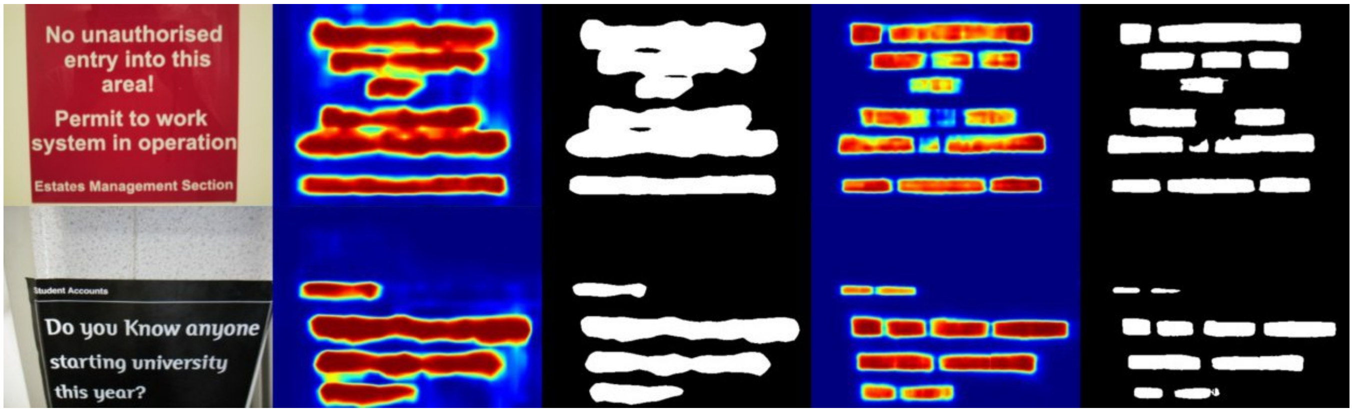
No unauthorised entry into this area!
Permit to work system in operation
Estates Management Section
Student Accounts
Do you Know anyone starting university this year?

Fig. 2. Segmentation comparisons with and without WordFence. First column from the left shows the original images. Second and third columns show the text position belief map and the resulting segmentation, respectively (trained with method outlined by Fisher and Koltun [9]). Last two columns show the belief map and the segmentation from our method. We found that an eight pixel border provides the best separation for most text sizes. Localizing words without WordFence causes individual words to bleed over into each other, which causes difficulty for posterior recognition.

effective receptive field of the CNN. This context information improves the network's understanding of text at different scales. Dilated convolutions do not increase the number of parameters, ensuring that the model remains easy to train. Finally, the obtained parallel segmentation maps are fused together by element wise summation, providing the final segmentation map, which are then used for word extraction.

2.3. Weighted Softmax Loss Function

A common loss function for training semantic segmentation networks is a pixelwise classification softmax loss. Such a function is appropriate for dense pixelwise labelling if there are many classes. For text localization, the pixelwise softmax loss tends to force the network to produce merged segmenations on the borders of words results such as the ones illustrated in Fig. 2. Post processing techniques are required to enhance the segmentation bounding boxes in order to use them for text recognition. In order to overcome this problem, a simple and efficient technique is introduced: instead of a binary text/non-text classification we define the notion of a border for each separate word as a third class. The border acts as a penalization for training. The model is driven to surround each separate word with an artificial barrier, which greatly reduces the ease and computational cost of reading separate words. During inference, individual words are cleanly segmented from each other and can then be extracted using connected components analysis.

The number of text pixels in a text recognition dataset may not be balanced among labels and the vast majority of all pixels are simply background - networks tend to predict background everywhere. To solve this issue, we introduce a weighted normalization. The novel loss function automatically penalizes predictions for pixels which form the majority of a given image and emphasizes pixels which are fewer in number. This makes the loss function well constrained for the task of text segmentation. Weight normalization is applied in two places: loss calculation and loss backpropagation. Normalization factors are calculated on the fly and are inversely proportional to pixel counts of each class. The algorithm of the weighted softmax loss function is shown in Alg. 2.3.

Algorithm 1 Pixelwise Weighted Softmax Loss

```
Require: Predicates after fusion $\mathbf{Pr}$, ground truth labels $\mathbf{L}$
$probs \leftarrow Softmax(\mathbf{Pr})$    ▷ pixel probabilities
$m \leftarrow NumberOfUniqueLabels(\mathbf{L})$
$n_1, n_2, \dots, n_m \leftarrow CountsOfUniqueLabels(\mathbf{L})$    ▷ get counts of each label on a ground truth image
$loss \leftarrow -\sum \frac{1}{n_{gt}} \log(probs_{gt})$    ▷ weighted loss calculation
$Backpropagate(loss, \frac{1}{n_1}, \frac{1}{n_2}, \dots, \frac{1}{n_m})$    ▷ loss backprop with normalization factors
```

3. EXPERIMENTS AND RESULTS

3.1. Datasets

Our model is trained and evaluated on a number of different text detection datasets. The COCO-Text dataset [20] is based on the earlier MS-COCO dataset for object classification [21]. SynthText [1] consists of natural images with synthetic text labels. ICDAR 2011 [22] and ICDAR 2013 [23] are common benchmark datasets from the International Conference of Document Analysis and Recognition. Street View Text dataset (SVT) [24] was harvested from Google Street

Model	**PASCAL VOC IoU = 0.5**								
	ICDAR11			**ICDAR13**			**SVT**		
	Prec.	Rec.	F-score	Prec.	Rec.	F-score	Prec.	Rec.	F-score
Tian *et al.* [3]	0.89	0.79	0.84	0.93	0.83	**0.88**	-	-	-
Gupta *et al.* [1]	0.78	0.63	70.0	0.78	0.63	0.70	0.47	0.45	0.46
Jaderberg *et al.* [2]*	0.89	0.68	77.4	0.89	0.68	0.77	0.59	0.49	0.54
Gupta *et al.* [1]*	**0.94**	0.77	**0.85**	**0.94**	0.76	0.84	**0.65**	0.60	**0.62**
WDN Recognition (ours)	0.64	**0.92**	0.75	0.65	**0.92**	0.76	0.47	**0.63**	0.54

Table 1. State-of-the-art comparison for word detection. Precision, Recall and F-Score are reported. Recall maximization was necessary for obtaining good word detection results. Methods marked with * use a multistage false-positive filtering process to increase precision, the code was not published thus the results are not directly comparable with ours.

View images. We train our model on MS-COCO, finetune on SynthText and evaluate on ICDAR11, ICDAR13 and SVT for comparison with state-of-the-art methods.

3.2. Word Localization Experiments

For the evaluation of our word detection results we use a PASCAL VOC style protocol where a proposal with intersection-over-union (IoU) ≥ 0.5 is considered a positive detection. PASCAL VOC is suitable for detecting individual words as it penalizes areas covering multiple words.

Running the image inference at different scales produces different segmentation maps that need to be processed afterwards. When merging segmentations from different scales, the results will contain many duplicates and false positives, but recall will be high since true positives will likely have been found. We adopt a mechanism for merging segmentation maps of different scales before extracting the bounding boxes, while maintaining a high recall. We use a voting scheme to produce a final segmentation map. We upscale all segmentation maps and find labels that correspond to maximal class probabilities in the segmentation maps. We extract the probability values for the found labels and sum them up on corresponding channels producing the map of summed maximum probabilities from different scales. The final segmentation is obtained by finding labels with maximum probabilities on the combined map giving fewer false positives.

Table 1 shows the performance of WDN on benchmark datasets. On average we improved recall by 15% over the previous multi-scale detection method by Gupta *et al.* [1].

3.3. End-to-end Word Detection and Recognition

Using ideal, single-word proposals recognition accuracy can be as high as 98% [2]. In order to show the effectiveness and quality of proposals we integrate our model with a state-of-the-art recognition model by Shi *et al.* [25]. The recognition model consists of an RNN to recognize words of different length. Our word proposals are cropped out and evaluated with the recognition network. We followed the evaluation protocol outlined by Wang *et al.* [26], where all word proposals that are three characters long or less or those that contain non-alphanumeric characters are ignored. An IoU overlap of 0.5 is required for a positive detection. Results for common recognition dataset are illustrated in Table 2. Our detection network achieves state-of-the-art recall rates - ensuring good candidate words. This combined with the recognition module obtains very accurate results for end-to-end word recognition. The network outperforms results by Jaderberg *et al.* [2] and is on par or better than Gupta *et al.* [1].

Model	**Year**	**ICDAR11**	**ICDAR13**
Neumann *et al.* [27]	2013	0.45	-
Jaderberg *et al.* [2]	2015	0.69	0.76
Gupta *et al.* [1]	2015	**0.84**	0.85
WDN Recognition	2016	**0.84**	**0.86**

Table 2. Evaluation of end-to-end word recognition on ICDAR 2011 and 2013 datasets. Our work is compared against other methods. F-score is reported.

4. CONCLUSION

We have presented the WordFence Detection Network. WDN relies on space between words to accurately split words using purely visual information for a wide variety of fonts, text sizes, scales, orientations and text languages. After segmenting an image proposal bounding boxes are extracted at multiple scales with high detection recall. Lastly, end-to-end word recognition achieves state-of-the-art results with 84 % and 86 % F-Score on ICDAR11 and ICDAR13, respectively. We obtain such high end-to-end scores by leveraging high quality proposals and high recall in the detection stage. Experimental results show that our approach achieves competitive performance on ICDAR11 and ICDAR13 without utilizing any heuristics or knowledge based approaches.[1]

[1] Work was partially supported by the Spanish project TIN2016-74946-P (MINECO/FEDER, UE) and CERCA Programme, Generalitat de Catalunya.

5. REFERENCES

[1] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman, "Synthetic data for text localisation in natural images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2315–2324. 1, 2, 3, 4

[2] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Reading text in the wild with convolutional neural networks," *International Journal of Computer Vision*, vol. 116, no. 1, pp. 1–20, 2016. 1, 4

[3] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao, "Detecting text in natural image with connectionist text proposal network," in *European Conference on Computer Vision*. Springer, 2016, pp. 56–72. 1, 2, 4

[4] Shangxuan Tian, Yifeng Pan, Chang Huang, Shijian Lu, Kai Yu, and Chew Lim Tan, "Text flow: A unified text detection system in natural scene images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4651–4659. 1

[5] Zhuoyao Zhong, Lianwen Jin, Shuye Zhang, and Ziyong Feng, "Deeptext: A unified framework for text proposal generation and text detection in natural images," *arXiv preprint arXiv:1605.07314*, 2016. 1, 2

[6] Tong He, Weilin Huang, Yu Qiao, and Jian Yao, "Accurate text localization in natural image with cascaded convolutional text network," *CoRR*, vol. abs/1603.09423, 2016. 1, 2

[7] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440. 1, 2

[8] Chen Liang-Chieh, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *International Conference on Learning Representations*, 2015. 1, 2

[9] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," in *ICLR*, 2016. 1, 2, 3

[10] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788. 1, 2

[11] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*. Springer, 2016, pp. 21–37. 1, 2

[12] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *CoRR*, vol. abs/1312.6229, 2013. 1

[13] Zheng Zhang, Chengquan Zhang, Wei Shen, Cong Yao, Wenyu Liu, and Xiang Bai, "Multi-oriented text detection with fully convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4159–4167. 1, 2

[14] Tao Wang, David J Wu, Adam Coates, and Andrew Y Ng, "End-to-end text recognition with convolutional neural networks," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 3304–3308. 1

[15] Tong He, Weilin Huang, Yu Qiao, and Jian Yao, "Text-attentional convolutional neural network for scene text detection," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2529–2541, 2016. 1

[16] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," *CoRR*, vol. abs/1406.2227, 2014. 2

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99. 2

[18] Yi Li, Kaiming He, Jian Sun, et al., "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 379–387. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. 2

[20] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie, "Coco-text: Dataset and benchmark for text detection and recognition in natural images," in *arXiv preprint arXiv:1601.07140*, 2016. 3

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755. 3

[22] Asif Shahab, Faisal Shafait, and Andreas Dengel, "Icdar 2011 robust reading competition challenge 2: Reading text in scene images," in *2011 international conference on document analysis and recognition*. IEEE, 2011, pp. 1491–1496. 3

[23] Lluis Gomez and Dimosthenis Karatzas, "Multi-script text extraction from natural scenes," in *2013 12th International Conference on Document Analysis and Recognition*. IEEE, 2013, pp. 467–471. 3

[24] Kai Wang and Serge Belongie, "Word spotting in the wild," in *European Conference on Computer Vision*. Springer, 2010, pp. 591–604. 3

[25] Baoguang Shi, Xiang Bai, and Cong Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016. 4

[26] Kai Wang, Boris Babenko, and Serge Belongie, "End-to-end scene text recognition," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 1457–1464. 4

[27] Lukáš Neumann and Jiří Matas, "Real-time scene text localization and recognition," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3538–3545. 4