FAST: Facilitative and Accurate Scene TextProposal through FCN Guided Pruning

Dena Bazazian[a,*], Raúl Gómez[a,b], Anguelos Nicolaou[a], Lluís Gómez[a], Dimosthenis Karatzas[a], Andrew D. Bagdanov[a,c]

[a] *Computer Vision Centre (CVC), Universitat Autònoma de Barcelona (UAB), Barcelona, 08193, Spain*
[b] *Eurecat - Catalunya Technology Center, unit of Multimedia Technologies, Barcelona, 08018, Spain*
[c] *Media Integration and Communication Center (MICC), University of Florence, Florence, 50134, Italy*

ABSTRACT

Class-specific text proposal algorithms can efficiently reduce the search space for possible text object locations in an image. In this paper we combine the Text Proposals algorithm with Fully Convolutional Networks to efficiently reduce the number of proposals while maintaining the same recall level and thus gaining a significant speed up. Our experiments demonstrate that such text proposal approaches yield significantly higher recall rates than state-of-the-art text localization techniques, while also producing better-quality localizations. Our results on the ICDAR 2015 Robust Reading Competition (Challenge 4) and the COCO-text datasets show that, when combined with strong word classifiers, this recall margin leads to state-of-the-art results in end-to-end scene text recognition.

1. Introduction

Text understanding in unconstrained scenarios, such as text in scene images or videos, is attracting significant attention from the scientific community. Localizing and recognizing text in complex images is a challenging task, especially in scenarios involving text that can appear in any perspective or orientation, or in multi-script contexts as evidenced by the latest results of the ICDAR Robust Reading Competition series (Karatzas et al., 2015). Object proposal techniques have emerged as an efficient approach to reducing the search space of possible object locations in an image by generating candidate class-independent object locations and extents (Uijlings et al., 2013). Such generic object proposal methods are typically designed to detect single-body objects, and are not appropriate for text detection which aims to detect groups of disjoint, atomic objects (characters or text strokes). Text Proposals (Gomez-Bigorda and Karatzas, 2016) was recently suggested as an alternative, class-specific object proposal method that takes into account the specific characteristics of text.

Text Proposals combined with a strong word classifier such as (Jaderberg et al., 2014) produces state-of-the-art results in end-to-end scene text recognition (Gomez-Bigorda and Karatzas, 2016). The downside is the necessity to produce and process the

*Corresponding author: Tel.: +34-93-581-18-28; fax: +34-93-581-16-70;
e-mail: dbazazian@cvc.uab.es (Dena Bazazian)

multitude of bounding boxes produced by the algorithm. State-of-the-art object detection methods such as YOLOv2 (Redmon and Farhadi, 2016) and text-specific flavors of them such as TextBoxes (Liao et al., 2017), CTPN (Tian et al., 2016) or Gupta et al (Gupta et al., 2016) are better-suited alternatives when fast localization is sought, but have limits in terms of their capacity to cover all the text regions in the image. If high recall is sought, Text Proposals remains the best alternative available on the most challenging scenarios such as Incidental Text dataset. In contrast to standard object detection methods, the Text Proposals algorithm produces a list of probable text locations following a strategy based on an over-segmentation and the efficient construction of meaningful groupings of regions. The result is very accurate text localization. Yet, like all object proposals approaches, it is optimized for high recall and produces a considerable number of false positives. Furthermore, Text proposal techniques are capable of generating proposals at multiple scales. Therefore, compared to text detection approaches (which are often limited by prior assumptions about text size and spacing), text proposal approaches are able to obtain better coverage of space and scale and to maximize text recall.

In this paper we leverage the power of Fully Convolutional Networks (FCNs) to improve the efficiency of Text Proposals and to substantially reduce the number of proposals produced without affecting the detection capacity of the algorithm. The robustness of FCNs stems from the use of convolutional layers instead of fully-connected layers as in conventional deep networks, which results in spatial information being preserved in the final discriminative layer of the network. The output of such a network is a coarse heatmap of locations of interest which must be further processed to produce proper localization results. This paper extends our previous work in (Bazazian et al., 2016), introducing improvements in the original pipeline, experimentation in end-to-end scenarios, and further comparison and evaluations on two widely used datasets. In (Bazazian et al., 2016) we employed an FCN trained for text detection as a means to suppress a number of the proposals at a post-processing stage. In this article we introduce an alternative pipeline and demonstrate how such an FCN can be integrated with the Text Proposals method and used at the beginning of the pipeline as an early-pruning mechanism that actively limits the number of proposals produced. This early-pruning mechanism results in a significant speed increase with minimal loss in performance. In addition, we present a thorough comparison with state-of-the-art text and object localization methods YOLOv2 (Redmon and Farhadi, 2016), CTPN (Tian et al., 2016) and TextBoxes (Liao et al., 2017), demonstrating the limitations and benefits of using a proposal approach instead of a localization algorithm. Specifically, we show that our improved Text Proposals method yields consistently higher recall, on the order of 20% higher, compared to state-of-the-art text localization methods, while also producing higher-quality localizations. By integrating the improved Text Proposals method in an end-to-end pipeline we demonstrate, in agreement with (Gomez-Bigorda and Karatzas, 2016), that this recall margin leads to state-of-the-art end-to-end text recognition results.

We call our framework FAST, since it simplifies proposals (**F**acilitative) while it maintains high recall (**A**ccurate) in **S**cene **T**ext images, and overall it is a FAST framework for text proposals.

The remainder of the article is organized as follows. In the next section we discuss related work from the literature, and in 3 we describe our approach and architecture. We report on a range of experimental results in section 4, and draw conclusions in section 5.

2. Related Work

Text detection and recognition in scene images are gaining increasing attention from the computer vision community. Despite the immense effort that has been devoted to improving the performance of text detection in unconstrained environments, it is still quite challenging due to the diversity of text appearance and geometry in the scene and the highly complicated backgrounds against which it can appear.

Text detection and localization algorithms are divided in two main approaches. First, connected component-based methods perform image segmentation and then classify and group the detected components into text candidates (Kang et al., 2014; Li et al., 2014; Yao et al., 2012; Yin et al., 2015, 2014). Following this idea, but in the other way around, is (Gomez and Karatzas, 2013) in which interesting components are detected via grouping. Second, sliding-window methods perform region classification at multiple scales and follow with standard post-processing to derive precise text locations (Jaderberg et al., 2014; Tian et al., 2015; Wang et al., 2011; Zhang et al., 2015; Wang et al., 2012a).

Most top-ranking approaches (Koo and Kim, 2013; Yin et al., 2015, 2014; Wang et al., 2012a; Neumann and Matas, 2013) in the latest editions of the ICDAR Robust Reading Competition (Karatzas et al., 2015) are connected component-based methods based on variants of the Maximally Stable Extremal Regions (MSERs) algorithm of (Matas et al., 2004). The MSER algorithm is particularly suited to text detection as it efficiently leverages local contrast patterns. Different approaches in this category include the Stroke Width Transform (SWT) (Epshtein et al., 2010) and recent variants like the Stroke Feature Transform (SFT) (Huang et al., 2013). Moreover, in (Yao et al., 2014) candidates (characters and lines) are generated by employing SWT and clustering in order to detect multi-oriented text regions and recognize characters. Connected component-based approaches permit early fusion with the language model at the time of filtering components and forming words. A recent example proposes a model which combines bottom-up cues from individual character detection and top-down cues from a lexicon (Mishra et al., 2016).

CNN classifiers are increasingly incorporated in text detection approaches. (Wang et al., 2012b) employed CNN models for text/non-text classification. (He et al., 2015) and (Huang et al., 2014) exploit a CNN model to filter out non-character MSER components, while CNN-based region classification has been repeatedly employed for sliding-window approaches (Jaderberg et al., 2014; Wang et al., 2012a). (Tian et al., 2015) proposed TextFlow for character classification, which simplifies multiple post-processing steps through the use of a minimum cost flow network (Wang et al., 2012a). These approaches are all based on hand-crafted features for detecting character candidates, and they apply a pre-trained CNN model to assign a score to each candidate. On

a different line, (Zhang et al., 2015) trained two classifiers that work at character level and text region level, respectively.

Despite the immense success of CNN models for tasks such as character classification or word-spotting, once text regions are localized the problem of text localization still poses significant challenges. To this end, the use of generic object proposal techniques for scene text understanding has been exploited by (Jaderberg et al., 2016). Their end-to-end pipeline combines the object proposal algorithm EdgeBoxes (Zitnick and Dollar, 2014) and a trained aggregate channel features detector (Dollar et al., 2014) with a powerful deep Convolutional Neural Network for holistic word recognition. Their method uses a CNN-based bounding box regression module on top of region proposals in order to improve their quality before word recognition is performed.

Employing generic object proposals is not optimal when text is to be detected, as demonstrated in (Gomez and Karatzas, 2015). The author of (Gomez-Bigorda and Karatzas, 2016) propose instead a text-specific object proposal method that is based on generating a hierarchy of word hypotheses based on a region grouping algorithm. Contrary to generic object proposal algorithms, this approach considers the specific characteristics of text regions that are fundamentally different from the typical notion of single-body object as used for object detection. Replacing the overly complicated pipeline used for text localization in (Jaderberg et al., 2016) with the single-step localization offered by (Gomez-Bigorda and Karatzas, 2016) is currently the state-of-the-art for end-to-end methods in the demanding Challenge 4 (incidental text) of the Robust Reading Competition.

On the other hand, FCNs (Fully Convolutional Networks) (Long and Darrell, 2015) have recently attracted considerable attention from the robust reading community (Zhang et al., 2016b,a; He et al., 2016; Gupta et al., 2016). FCN-based methods replace fully-connected layers with convolutional layers which allows them to preserve coarse spatial information which is essential for text localization tasks. (Zhang et al., 2016b) integrated semantic labeling by FCN with MSER to provide a natural solution for handling text at arbitrary orientations. Addressing the same problem of multi-oriented text detection, a direct regression based method was proposed in (He et al., 2017). In a parallel work (Zhang et al., 2016a) designed a character proposal network based on an FCN which simultaneously predicts characterness scores and refines the corresponding locations. The characterness score is used for proposal ranking. (He et al., 2016) proposed a Cascaded Convolutional Text Network (CCTN) that combines two custom convolutional networks for coarse-to-fine text localization. The CCTN detects text regions roughly from a low-resolution image, and then accurately localizes text lines in each enlarged region. In (Shi et al., 2016), a fully convolutional network fused with a a bidirectional LSTM is used to build a robust text recognition framework.

Beyond the specific problem of text localization, CNN-based object detection methods like Fast R-CNN (Ren et al., 2015), YOLO (Redmon et al., 2015; Redmon and Farhadi, 2016) and SSD (Liu et al., 2016) have reached state-of-the-art performance on generic object detection benchmarks. These methods overcome the need for an external object proposals algorithm with a CNN that simultaneously predicts object bounds and objectness scores at each position in a single forward pass. Proposal approaches

have achieved excellent results on general object detection (Bell et al., 2016; Ren et al., 2015) and text detection (Tian et al., 2016; Liao et al., 2017) tasks. An advantage of region proposal approaches for object and text detection is that they reduce the search space and false positives by focusing the attention on promising areas of the image. Inspired by Fully-Convolutional Networks (Long and Darrell, 2015) and the YOLO detector of (Redmon et al., 2015), (Gupta et al., 2016) propose a text localization network as an extreme variant of Hough voting. TextBoxes (Liao et al., 2017) re-purposes the SSD detector for word-wise text localization. (Tian et al., 2016) follows the idea of Region Proposal Networks (RPNs) (Ren et al., 2015) and proposes a Connectionist Text Proposal Network which improves the accuracy for text localization tasks and also is compatible with multiple scales, aspects, and languages. Still, all these CNN models are designed to perform only text localization, and a second recognition stage is needed for the end-to-end task. We note here that, since these networks are trained to maximize only the text localization performance, their recall rates tend to be inferior than using an external text proposals algorithm. And this recall difference may lead to lower end-to-end text recognition results when using such CNN-based localizers.

In this paper we take into account the respective advantages of FCN models and Text Proposals in order to propose an improved text proposal algorithm. We employ text probability scores obtained through FCN over individual proposals for ranking, which efficiently suppresses non-text proposals. Contrary to (Zhang et al., 2016b) who propose a coarse-to-fine method to obtain bounding boxes from the FCN heatmap with the purpose of maximizing localization accuracy, we employ the FCN heatmap to produce text proposals in order to maximize recall by producing the necessary number of proposals in an efficient way.

We choose to use the FCN result to re-rank text proposals obtained by (Gomez-Bigorda and Karatzas, 2016), and thereby improving the performance of the text proposal algorithm. Furthermore, (Liao et al., 2017) presented an end-to-end text detector which has a text box layer which defines bounding boxes according to a range of different aspect ratios. In this case all the proposed bounding boxes are axis-oriented, while in the technique of (Gomez-Bigorda and Karatzas, 2016) proposed boxes are defined according to the grouping of textness features regardless of axis-orientation. The work we describe here extends the recently text proposal technique described in (Bazazian et al., 2016). In this paper we propose a more efficient use of the FCN probabilities maps, to guide the proposing process into generating less hypotheses instead of suppressing hypotheses in a post-processing step.

3. The Proposed Method

The proposed text proposals pipeline, shown in Figure 1, comprises three elements. First, a **pixel-wise text prediction** stage, makes use of a Fully Convolutional Network to estimate the pixel-level text probability of the image (see Section 3.1). This information is then used to guide the **text proposal stage**, based on (Gomez-Bigorda and Karatzas, 2016), during which the image is decomposed into regions and text location hypotheses are created (see Section 3.2). A number of alternative approaches are proposed to leverage the FCN predictions in order to re-rank and filter the text hypotheses produced in a **hypothesis ranking and**

filtering stage that can be introduced either as a late-suppression mechanism or a more efficient early-pruning mechanism (see Section 3.3). The remainder of this section discusses these three components in detail.

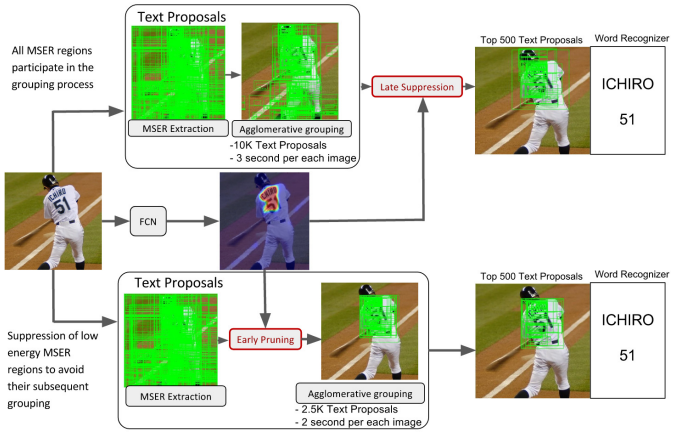

Fig. 1. Comparison of the late suppression and early pruning strategies. The Late Suppression strategy (top) allows the text proposals algorithm to generate all hypotheses before filtering, while the Early Pruning strategy is tightly integrated with the Text Proposals algorithm guiding it to only generate relevant hypotheses.

3.1. FCN-based Text Prediction

Fully Convolutional Networks take input of arbitrary size and produce correspondingly-sized output, allowing to make pixel-level predictions (Long and Darrell, 2015; Shelhamer et al., 2016). Every layer in an FCN computes a local operation on relative spatial coordinates. Since there is no fully-connected layer, it is possible to use FCNs on variable dimension images and produce an output of the corresponding input dimension as well as preserving coarse spatial information of the image which is essential for the text detection tasks.

Here we train an FCN to perform per-pixel text prediction and estimate a *text heatmap* for the input image. The architecture of our network is based on (Long and Darrell, 2015) and is shown in Figure 2. To accomplish it, we transform the pre-trained VGG network (Simonyan and Zisserman, 2015) into a fully convolutional form following (Long and Darrell, 2015). The original FCN network was designed for the semantic segmentation of images into the twenty classes of the PASCAL VOC dataset. In our case, we customize the network to our purpose of performing text/no-text segmentation. We apply softmax normalization on the FCN output in order to employ it as text probability for the subsequent hypotheses ranking and filtering steps. In Figure 3 we show an

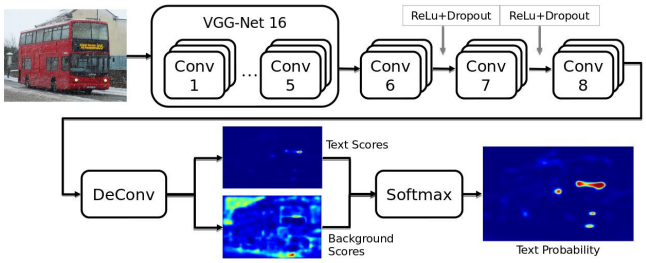

Fig. 2. Architecture of the Fully Convolutional Network used in our pipeline for text prediction.

Fig. 3. Sample image with the corresponding ground truth mask, and the FCN produced text-heatmap.

example image from our training set, along with the ground truth text annotations and text heatmap output by our FCN. Details on the training process are given in section 4.2.

3.2. Text Proposals

In our framework we use the Text Proposals algorithm of (Gomez-Bigorda and Karatzas, 2016) where text detection is posed as a search within a hierarchy produced by an agglomerative similarity clustering process over individual regions. The method is divided in three main steps: (1) an initial over-segmentation of the input image from which we obtain a set of regions; (2) the creation of text hypotheses through several bottom-up agglomeration processes; and (3) a ranking strategy that prioritizes the best text proposals.

Region decomposition is based on the Maximally Stable Extremal Regions (MSER) algorithm which produces a set of regions $\mathcal{R}_c$. This step aims to detect the atomic parts that will give rise to subsequent text groups. Initially each region $r \in \mathcal{R}_c$ starts in its own cluster and then the closest pair of clusters (A, B) is merged iteratively, using the single linkage criterion ($\min\{\, \mathrm{d}(r_a, r_b) : r_a \in A,\ r_b \in B \,\}$), until all regions are clustered together ($C \equiv \mathcal{R}_c$) via the agglomeration processes. The grouping process builds an hierarchy of groupings of the initial set of MSER regions based on a distance definition $\mathrm{d}(r_a, r_b)$ that combines region similarity features and region location. We use the five low computational cost features and diversification strategies as proposed in the original paper (Gomez-Bigorda and Karatzas, 2016). The nodes of the produced similarity hierarchy represent the text hypotheses or text proposals, which are ranked based on a weak classifier that uses the coefficients of variation of the individual region features, as well as simple group-level features that can be computed efficiently in an incremental way. For further information on the specifics of the Text Proposal algorithm, please see (Gomez-Bigorda and Karatzas, 2016)

3.3. Hypotheses Ranking

The Text Proposals algorithm promotes text hypotheses that represent well structured groups of regions. This makes it very efficient in detecting areas in the image corresponding to text, but at the same time can produce a large number of false positives. The text-prediction FCN on the other hand provides only coarse information about the locations of interest in the image, but it is much more robust in the presence of distracting patterns in the image. We analyze here different strategies for fusing the two sources of information in order to drastically reduce the number of false positive text hypotheses produced, and consequently speed up significantly the Text Proposals algorithm.

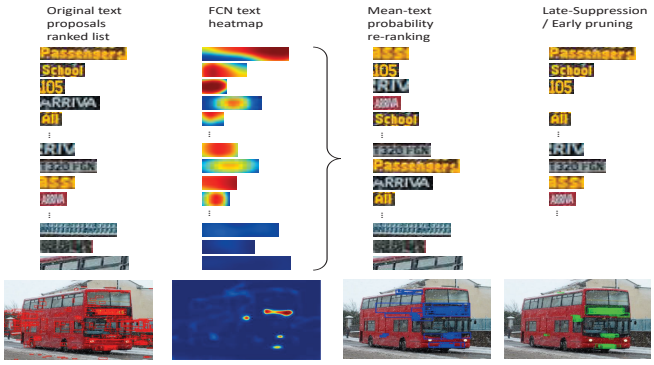

Fig. 4. A list of text hypotheses generated by the Text Proposals algorithm and their corresponding regions from the FCN produced text-heatmap are shown on the left. Re-ranking using the mean text probability from the FCN heatmap filters false positives but promotes small patches within text regions over the full regions. Late-suppression or early-pruning filter out false positives without affecting the original ranking.

We have considered the following strategies in order to re-rank and filter the text hypotheses, using the text heatmaps produced by the FCN. The **baseline** strategy ranks text hypotheses using the score provided by the Text Proposals algorithm of (Gomez-Bigorda and Karatzas, 2016) as explained in section 3.2 and does not perform any re-ranking or filtering of the hypotheses using the FCN text heatmap.

A direct way to re-rank the hypotheses produced by the Text Proposals algorithm is to use the **mean text probability** obtained from the FCN text heatmap by averaging over the proposed hypothesis bounding box. As FCN heatmaps provide only coarse information about text localization, directly using this information to rank hypotheses can be tricky as smaller regions, corresponding to the inner parts of a text block, might appear to be better candidates than the full text block bounding box that covers the extents of the text including lower probability pixels at the borders. This behavior is demonstrated in 4. Indeed, using the mean text probability over text hypotheses as a ranking mechanism has the adverse effect that small patches within text regions are ranked better than full regions corresponding to words or lines, leading to worse performance than the baseline strategy. In order to filter false positive hypotheses without affecting the original ranking of the Text Proposals algorithm, which promotes well-structured text blocks, we introduce a **Late Suppression** strategy. This strategy suppresses the text hypotheses which have low mean text probability, in order to filter false positives, as shown in Figures 1 and 4. Consequently, the rank of the remaining regions is established according to the score assigned by the Text Proposals. The smaller regions corresponding to the inner parts of text blocks are not filtered, however they are generally ranked low based on their grouping quality score as they only cover a small subset of the text parts. The suppression strategy allows us to discard a significant number of false positives, resulting in superior detection with a smaller number of proposals while preserving the high recall rate.

The suppression strategy efficiently reduces the number of text hypotheses that need to be passed to a subsequent recognition process, but nevertheless, does not impede the Text Proposals algorithm to generate the full set of text hypotheses. As an alternative, the **Early Pruning** strategy employs the FCN text heatmap to filter the original MSER regions in a pre-processing step, actively

limiting the number of hypotheses being produced. The resulting set of hypotheses are practically the same as in the late suppression strategy, but produced in a fraction of the time. By filtering MSER regions before the agglomerative clustering process begins, we avoid the generation of a series of text hypotheses all-together, resulting in a significant speed increase compared to the late-suppression strategy with minimal loss in performance. A comparative study of the above strategies is offered in the next section.

4. Experimental Results

In this section we describe experiments which we have performed to evaluate the proposed method. First, we detail the datasets used and the experimental setup, before we compare the proposed ranking strategies to the baseline method and other state-of-the-arts strategies.

4.1. Datasets

We employed two widely used, challenging benchmark datasets, namely the ICDAR-Challenge 4 (Karatzas et al., 2015) and the COCO-Text (Veit et al., 2016), for training and evaluating our method.

- **ICDAR-Challenge4:** ICDAR-Challenge 4 (Karatzas et al., 2015) focuses on incidental scene text, referring to scene text that appears in the scene without the user having taken any specific prior action to cause its appearance or to improve its positioning or quality in the frame. Incidental scene text represents a wide range of applications linked to wearable cameras or massive urban captures where the acquisition process is difficult or undesirable to control. This challenge has 1000 images with publicly available ground truth annotations and a private test set of 500 images that can be used for evaluation of specific tasks through submitting results online to the Robust Reading Competition portal. We evaluate end-to-end results using the online evaluation functionality. For in-house experimentation we randomly selected and set aside 200 images of the public portion of the dataset to use for testing and used the rest for training. The ground truth for ICDAR-Challenge 4 is defined as tight (not axis oriented) four-point polygons. Since all standard methodologies, including the Text Proposals algorithm produce axis oriented bounding boxes, we use the equivalent smallest enclosing axis-oriented bounding boxes as ground truth for in-house experimentation.

- **COCO-Text:** The COCO-Text dataset (Veit et al., 2016) comprises a subset of images from the MS COCO dataset, which contains images of complex everyday scenes. The COCO-Text dataset contains non-text images, legible text images and illegible text images. A training set and a validation set are defined, from which we consider only those images containing at least one instance of legible text. Notice that some of that images may also contain illegible text instances. That is a total of 22,184 images for training and 7,026 images for validation. We evaluate on the validation set, since the official test set of COCO-Text is withheld. Ground truth is defined as axis-oriented bounding boxes.

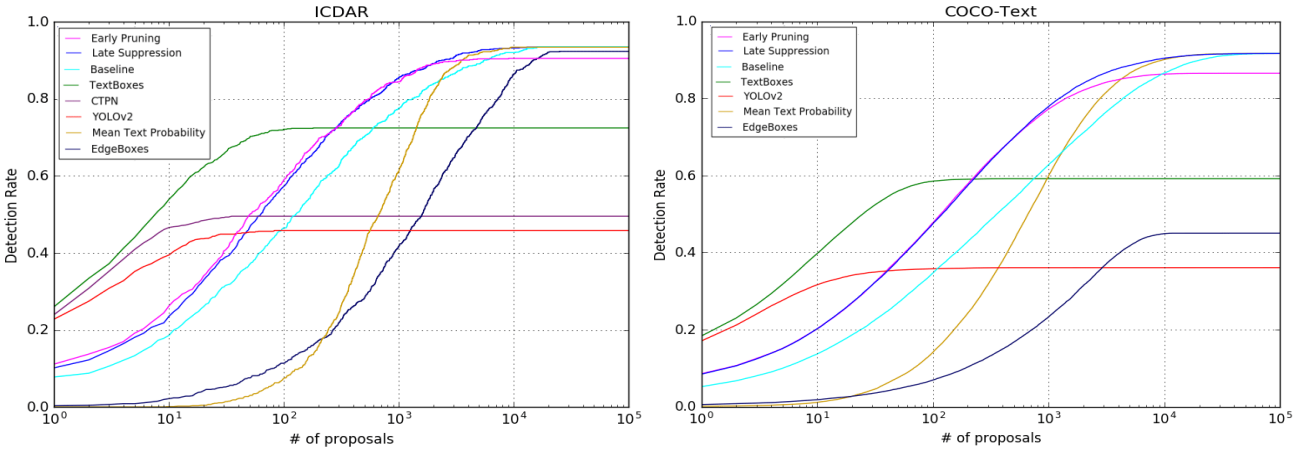

Fig. 5. Detection rates at $0.5IoU$ **over the ICDAR-Challenge 4 (Left) and the COCO-Text (Right) datasets.**

4.2. The FCN Training Process

In this work we trained an FCN network for the task of text detection, by fine-tuning a VGG16 network pre-trained on Imagenet (Simonyan and Zisserman, 2015). We repeated the fine-tuning process based on both the ICDAR-Challenge 4 training set and the COCO-Text training set, and observed minor differences on the quality of the heatmaps produced and the results of the re-ranking. For the rest of the discussion, we have used the FCN network trained on COCO-Text. The FCN network was trained by Stochastic Gradient Descent with mini-batches of 20 images, a momentum of 0.9, L2 weight-decay of 5^{-4}, and a fixed learning rate of 10^{-10}. We apply dropout after the Conv6 and Conv7 layers with a rate of 0.5 as shown in Figure 2.

4.3. Comparison of Proposal Ranking Strategies

In this section we report results obtained using the different hypotheses ranking and filtering strategies to fuse the text-prediction information afforded by the FCN with the Text Proposals algorithm. First, we analyze the alternative strategies proposed **Mean Text Probability**, **Late Suppression** and **Early Pruning** strategies, and compare against the **Baseline** Text Proposals method. Subsequently, we compare against other state-of-the-art approaches, including **Edge Boxes** as a representative general object proposal method, as well as **YOLOv2** and **CTPN** and **TextBoxes** as representative object and text localization approaches correspondingly. We have considered all the proposals of each one of the aforementioned techniques. Then, based on each ranking list strategy, we have computed the detection rate (recall) at each number of proposals. The quantitative results are shown in Figure 5 and Table 1. The last column of Table 1 refers to all the proposals generated by each technique (thus where each achieves its highest recall). Note that we did not perform any experiments on COCO-Text for CTPN strategy, since the provided models by authors (Tian et al., 2016) and parameters are adjusted to the ICDAR dataset.

4.3.1. Mean text probability re-ranking

Figure 6 shows the top-N proposals for the Baseline Text Proposals method versus the ones produced by the Mean Text Probability strategy. The Mean Text Probability strategy uses the mean text probability of each bounding box, as calculated by averaging

Table 1. Detection rates of different strategies considering the top 10, 100 and 1000 proposals on the ICDAR-Challenge4 and COCO-Text datasets.

	ICDAR				COCO-Text			
#of Proposals	10	10^2	10^3	All	10	10^2	10^3	All
Early Pruning	0.26	0.58	0.84	0.90	0.20	0.48	0.77	0.87
Late Suppression	0.23	0.57	**0.85**	**0.94**	0.20	0.47	**0.78**	**0.92**
Baseline	0.18	0.46	0.77	**0.94**	0.13	0.34	0.63	**0.92**
TextBoxes	**0.55**	**0.71**	0.72	0.72	**0.39**	**0.58**	0.59	0.59
CTPN	0.48	0.51	0.51	0.51	-	-	-	-
YOLOv2	0.39	0.45	0.46	0.46	0.30	0.44	0.45	0.45
Mean Text Probability	0.00	0.07	0.61	**0.94**	0.01	0.15	0.60	**0.92**
EdgeBoxes	0.03	0.11	0.42	0.92	0.02	0.06	0.23	0.44

(a)Top 10 proposals (b)Top 100 proposals (c)Top 500 proposals

Fig. 6. The top-N hypotheses generated by the baseline Text Proposals method (top row) and after re-ranking using the mean text probability (bottom row).

the FCN text-prediction heatmap over the bounding box, as the ranking criterion to re-rank the hypotheses produced by the Text Proposals method. As a result, it prioritizes hypotheses corresponding to text regions over false positives. However, as mentioned before, it also promotes hypotheses corresponding to inner parts of text blocks as relevant regions, resulting in an overall worse performance than the baseline method as it shows in Figure 5 and Table 1.

4.3.2. Late Suppression

The late suppression strategy acts as a post-processing filtering mechanism, that suppresses hypotheses with low mean text probability, instead of re-ranking the proposals. By doing so, it preserves the ranking established by the Text Proposals method for the remaining hypotheses. Figure 7 shows the effect of using different suppression thresholds in the ICDAR-Challenge 4 dataset. A threshold of 0.14 yields the best results in our experiments. This low threshold might be counter-intuitive, given that the FCN seems capable to produce good quality heatmaps. Figure 8 shows the mean text probability of all ground truth regions of ICDAR-Challenge 4 and reveals that an almost uniform distribution. This demonstrates that Mean Text Probability is not a useful metric to identify regions containing words, but that it is a useful metric to discard proposals that do not contain text at all. A key reason for this is the fact that text appears in a variety of orientations in the scenes and axis oriented bounding boxes inadvertently include a significant amount of background pixels for non-horizontal (and non-vertical) angles. Similar statistics can be observed

in COCO-Text.

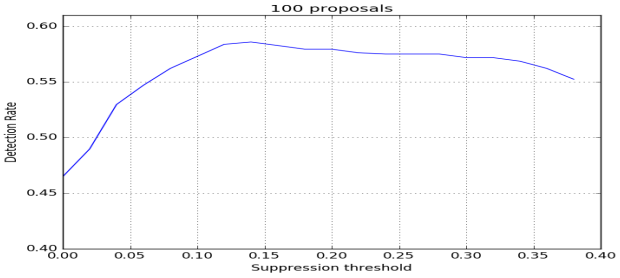

Fig. 7. Detection rate at $0.5IoU$ **for different thresholds for the Suppression strategy.**

4.3.3. Early Pruning

While suppression acts as a post-processing filter on the full list of hypotheses produced by the Text Proposal algorithm, Early Pruning acts as a pre-processing step, filtering early MSER regions that might give rise to irrelevant hypotheses. As a result, the Text Proposals algorithm is guided to produce only a subset of the original hypotheses. This results to a significant speed improvement (see section 4.5). This improvement is particularly notable in those images with large non-text textured areas. On the counterside, filtering regions early might hinder the generation of a group that could eventually lead to a relevant hypothesis. This implies a trade-off between reaching the maximum possible detection rate (such as as Late Suppression or the Baseline method) and gain in computational time.

4.3.4. Comparison with Different Proposal Techniques

We subsequently compare the proposed strategies to relevant state of the art methods. We compare to Edge Boxes (Zitnick and Dollar, 2014) as a top performing generic selective search method. We used the default configuration, taking into account all the output boxes (proposals). We also compare to FCN-based state of the art localization methods YOLOv2 (Redmon and Farhadi, 2016), CTPN (Tian et al., 2016) and TextBoxes (Liao et al., 2017). While the former is a generic object localization approach, the latter is specifically designed for text localization. In order to do a fair comparison in terms of their capacity to detect text, for TextBoxes ,CTPN and YOLOv2 we consider all output boxes (proposals). For YOLOv2 we fine-tuned the model to COCO-Text for $45 * 10^4$ iterations without modifying the training parameters which are: learning rate of 10^{-5}, weight decay of 5^{-4}, and momentum of 0.9. For CTPN we have employed the model for word-level detection provided by the original authors. For TextBoxes we used the default configuration and trained model. Regarding Figure 5 and Table 1, generic object proposal algorithms such as

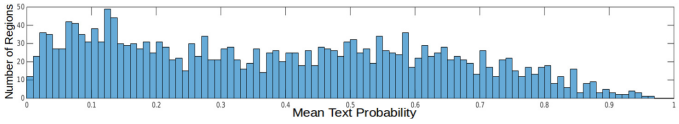

Fig. 8. Mean Text Probability of all ground truth regions of the ICDAR-Challenge 4 dataset.

Edge Boxes (Zitnick and Dollar, 2014), designed to detect single-body objects perform poorly when it comes to text detection. On the other hand, localization methods such as YOLOv2, CTPN and TextBoxes yield the best performance for a small number of bounding boxes, but quickly get saturated. The Text Proposals with early-pruning and late-suppression strategies presented surpass the performance of YOLOv2 after about 60 proposals and the performance of TextBoxes after about 500 proposals. Eventually, the proposed strategies are able to eventually detect about 45% more text than YOLOv2 fine-tuned for text, and about 20% – 25% more text than TextBoxes. For the rest of the analysis we focus on the Text Proposals with Early Pruning alternative.

4.4. Qualitative Analysis

Figure 9 shows all hypotheses generated with IoU over 0.5. Small text seems to pose problems for YOLOv2, CTPN and TextBoxes. In contrast Text Proposals with Early Pruning is able to generate hypotheses across a larger set of scales, and although it demonstrates a higher degree of redundancy, object proposals are well focused on text parts. Following this experiment, we have

Fig. 9. All true hypotheses generated (matching ground truth regions with IoU > 0.5). Text Proposals with Early Pruning hypotheses are shown in blue, TextBoxes in yellow, YOLOv2 in green and CTPN in purple. Small text seems to pose problems for YOLOv2, CTPN and TextBoxes, but not for Text Proposals with Early Pruning.

singled out the top-1 (highest IoU) hypothesis generated for each ground-truth region by each of the algorithms. It can be observed that rotated text poses difficulties to TextBoxes, while YOLOv2 yields strong performance in cases of low-contrast text as illustrated in Figure 10.

4.5. Time Comparison

Generating the FCN heatmaps introduces a small overhead in the process, which nevertheless results in significant time gains over a complete end-to-end pipeline. The time consumed to generate the FCN heatmaps is 0.15 sec per image, while the application of a state of the art word recognizer, such as (Jaderberg et al., 2016), takes about 0.4 ms per bounding box. Table 2 shows empirical times of an end-to-end pipeline. We consider all proposals generated by the different proposed strategies compared with the baseline

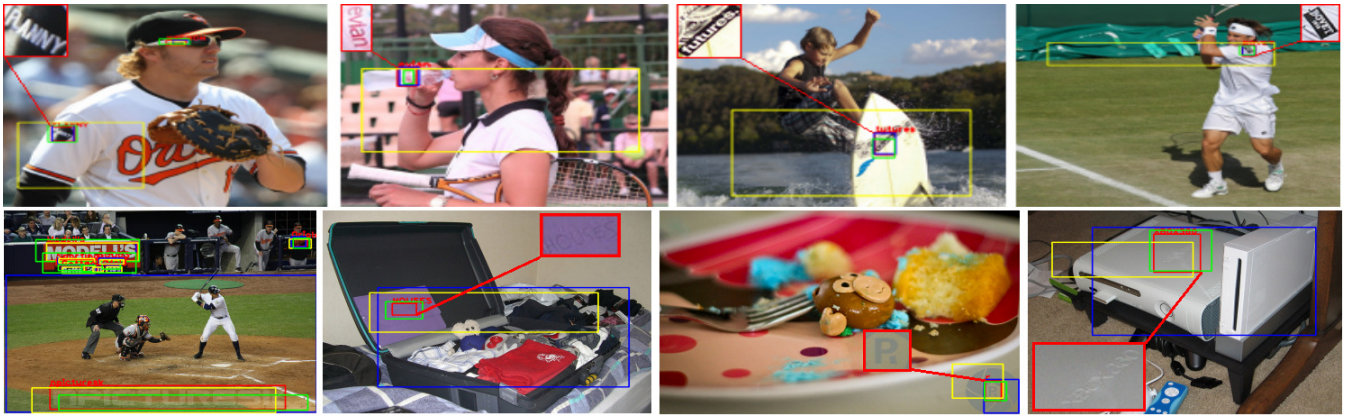

Fig. 10. Top row: examples demonstrating the difficulties of TextBoxes on oriented text. Bottom row: examples demonstrating the ability of YOLOv2 to detect text in low contrast conditions. Bounding boxes of Early Pruning hypotheses are shown in blue, TextBoxes in yellow, YOLOv2 in green, and ground-truth in red.

Table 2. Execution time per image (seconds).

	FCN	Text Proposals	Word Recognizer	Total
Baseline	0	4.11	25.69	29.80
Suppression	0.15	4.11	**5.4**	9.66
Early Pruning	0.15	**1.33**	5.58	**7.06**

Text Proposals on the ICDAR-Challenge 4 dataset. The baseline Text Proposals generates 64k proposals. The Late Suppression strategy reduces this to 13.5k proposals, while Early Pruning generates 14k proposals. The expected gain is of the order of 4 times All experiments were conducted on a system with a GPU GeForce GTX TITAN and an Intel® Core™2 Quad CPU Q9300@ 2.50GHz processor.

4.6. Quality of produced hypotheses

The standard practice of using an $IoU > 0.5$ criterion when evaluating detection performance is not very informative. Especially in the case of text, low IoU scores correspond to missing parts of a word, hindering its correct recognition. The objective of proposal generator algorithms is to produce at least one good localization for each object in the scene, and trade off redundancy of bounding boxes for the chance of detecting all objects well. In this sense, we are interested in analyzing the quality of the hypotheses produced in addition to the detection rate. Figure 11 shows the plot of IoUs for the top-1 hypotheses generated for each of the ground truth regions by each of the algorithms: Text Proposals with Early Pruning, YOLOv2, CTPN and TextBoxes. It can be easily seen that Text Proposals with Early Pruning produces significantly better localizations for a far larger number of ground truth boxes than the other algorithms. Of course this is done at the cost of producing many more redundant bounding boxes per ground truth region than TextBoxes, CTPN or YOLOv2. What remains to be seen is whether this qualitative difference is actually beneficial for an end-to-end pipeline.

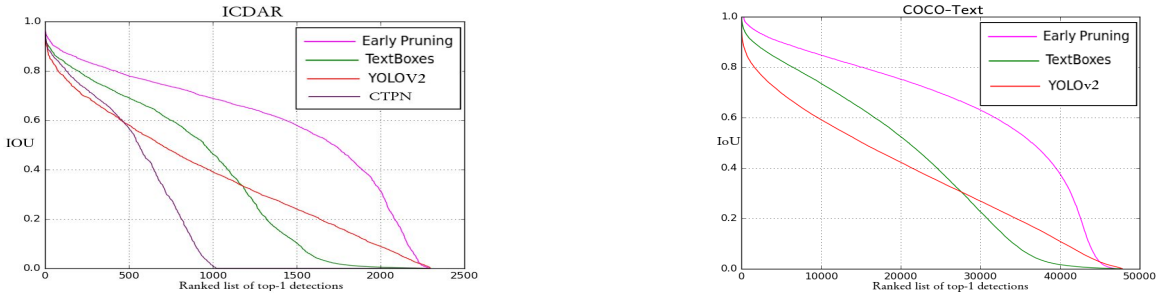

Fig. 11. Ranked list of the top-1 hypotheses for each ground truth region. The quality of the hypotheses produced by our approach is significantly better than state of the art text localization techniques.

4.7. End-to-end Results

In order to demonstrate that the large recall margin afforded by using a proposal generator approach can lead to better end-to-end results, we have built an end-to-end scene text recognition pipeline by combining the proposed Text Proposals with Early Pruning strategy and a state-of-the-art holistic word recognizer: the CNN word classifier of (Jaderberg et al., 2016). The evaluation framework is the standard for end-to-end text recognition datasets (Karatzas et al., 2015; Wang and Belongie, 2010). We obtain evaluation results using the ICDAR Robust Reading Competition on-line evaluation functionality. The evaluation framework considers as correct matches bounding boxes that overlap with a ground truth bounding box by more than 50% and the provided word transcription is correct (ignoring case). Based on this rule a single F-score measure is calculated. Table 3 shows the obtained end-to-end word spotting F-scores on ICDAR-Challenge 4 dataset and compares them with the state-of-the-art. The top block of Table 3 lists published methods with results reported in the ICDAR Robust Reading Competition portal. As we can see, the proposal generator approach yields consistently better results in all scenarios. Also, Early Pruning is 4x faster in execution speed. The middle block of table 3 gives results obtained by combining YOLOv2 (fine-tuned as explained in section 4.3.4), CTPN, and TextBoxes with the same word recognizer as the Text Proposals. We created an end-to-end pipeline by adding three filtering steps to the recognized word regions. First we removed any region with a word recognition confidence of less than 40%. In the following step, contextualized vocabularies (Karatzas et al., 2015) were used to remove any region whose word is out of vocabulary. Finally the remaining words were passed through a standard non-maximum suppression (NMS) with an IoU threshold of 30% as in (Jaderberg et al., 2016). The proposed method performs consistently better than both alternative pipelines constructed on top of the state of the art localizers.

5. Conclusions

In this work we propose a fusion of a proposal generator technique with Fully Convolutional Networks to efficiently reduce the number of proposals while maintaining the same text recall level and thus gaining a significant speed up. In particular, we train an FCN network for text prediction, and propose different strategies for employing this information to re-rank and/or filter

Table 3. Comparison of end-to-end recognition and word spotting F-scores on the ICDAR-Challenge4 dataset.

	End-to-End results		Word spotting Results		speed-up
	Strong	Weak	Strong	Weak	
Deep2Text-MO (Yin et al., 2014), (Jaderberg et al., 2014)	16.77	16.77	17.58	17.58	
Beam Search CUNI	22.14	19.80	23.37	21.07	
NJU Text	32.63	-	34.10	-	
Stradvision-1	33.21	-	34.65	-	
TextSpotter (Neumann and Matas, 2012)	35.06	19.91	37.00	20.93	
Stradvision-2	43.70	-	45.87	-	
Megvii-Image++ (Yao et al., 2015)	46.74	40.00	49.95	42.71	
TextBoxes (Liao et al., 2017)+DictNet (Jaderberg et al., 2014)	44.69	44.03	46.84	46.21	
CTPN (Tian et al., 2016)+DictNet (Jaderberg et al., 2014)	28.78	28.33	30.42	29.97	
YOLOv2 (Redmon and Farhadi, 2016)+DictNet (Jaderberg et al., 2014)	16.29	16.14	17.18	17.06	
TP (Gomez-Bigorda and Karatzas, 2016)+DictNet (Jaderberg et al., 2014)	53.30	**49.61**	56.00	**52.26**	×1
TP (Gomez-Bigorda and Karatzas, 2016)+**EarlyPruning**+DictNet (Jaderberg et al., 2014)	**54.17**	45.98	**57.08**	48.57	×4

the hypotheses generated by the Text Proposals algorithm. We demonstrate that this approach yields significantly higher recall rates than state-of-the-art text FCN based localization techniques, while also producing better-quality localizations. End-to-end performance shows that this recall margin, and the better quality hypotheses lead to state-of-the-art results in scene text reading systems.

Acknowledgments

This work was supported by the CERCA Programme of the Generalitat de Catalunya, the research project TIN2014-52072-P and the Eurecat Catalan Technology Center.

References

Bazazian, D., Gomez, R., Nicolaou, A., Gomez, L., Karatzas, D., Bagdanov, A., 2016. Improving text proposals for scene images with fully convolutional networks. Int Conf on Pattern Recognition, (DLPR workshop), arxiv:1702.05089 .

Bell, S., Zitnick, C., Bala, K., Girshick, R., 2016. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. IEEE Conf on Computer Vision and Pattern Recognition , 2874 – 2883.

Dollar, P., Appel, R., Belongie, S., Perona, P., 2014. Fast feature pyramids for object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence,36 (8) , 1532–1545.

Epshtein, B., Ofek, E., Wexler, Y., 2010. Detecting text in natural scenes with stroke width transform. IEEE Computer Vision and Pattern Recognition , 2963–2970.

Gomez, L., Karatzas, D., 2013. Multi-script text extraction from natural scenes. Int Conf on Document Analysis and Recognition , 467–471.

Gomez, L., Karatzas, D., 2015. Object proposals for text extraction in the wild. Int Conf on Document Analysis and Recognition , 206–210.

Gomez-Bigorda, L., Karatzas, D., 2016. Textproposals:a text-specific selective search algorithm for word spotting in the wild. Preprint submitted to Pattern Recognition, arxiv:1604.02619 .

Gupta, A., Vedaldi, A., Zisserman, A., 2016. Synthetic data for text localisation in natural images. IEEE Computer vision and pattern recognition , 2315–2324.

He, T., Huang, W., Qiao, Y., Yao, J., 2015. Text-attentional convolutional neural networks for scene text detection. arxiv:1510.03283 .

He, T., Huang Yu Qiao, W., Yao, J., 2016. Accurate text localization in natural image with cascaded convolutional text network. arxiv:1603.09423 .

He, W., Zhang, X., Yin, F., Liu, C., 2017. Deep direct regression for multi-oriented scene text detection. arxiv:1703.08289 .

Huang, W., Lin, Z., Yang, J., Wang, J., 2013. Text localization in natural images using stroke feature transform and text covariance descriptors. IEEE Int Conf on Computer Vision , 1241–1248.

Huang, W., Qiao, Y., Tang, X., 2014. Robust scene text detection with convolutional neural networks induced mser trees. Euro Conf on Computer Vision , 497–511.

Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A., 2016. Reading text in the wild with convolutional neural networks. International Journal of Computer Vision, 116(1) , 1–20.

Jaderberg, M., Vedaldi, A., Zisserman, A., 2014. Deep features for text spotting. Euro Conf on Computer Vision , 512–528.

Kang, L., Li, Y., Doermann, D., 2014. Orientation robust text line detection in natural images. IEEE Computer Vision and Pattern Recognition , 4034–4041.

Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, Lukas Chandrasekhar, V., Ramaseshan, Lu, S., Shafait, F., Uchida, S., Valveny, E., 2015. Icdar 2015 robust reading competition. IEEE ICDAR 2015 robust reading competition , 1156–1160.

Koo, H., Kim, D., 2013. Scene text detection via connected component clustering and nontext filtering. IEEE Transactions on Image Processing,22(6) , 2296–2305.

Li, Y., Jia, Shen, C., van den Hengel, A., 2014. Characterness: An indicator of text in the wild. IEEE Trans. Image Processing , 1666–1677.

Liao, M., Shi, B., Bai, X., Wang, X., Liu, W., 2017. Textboxes: A fast text detector with a single deep neural network. Association for the Advancement of Artificial Intelligence , 4161–4167.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., C.Berg, A., 2016. Ssd: Single shot multibox detector. Euro Conf on Computer Vision , 21–37.

Long, J Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. IEEE Computer Vision and Pattern Recognition , 3431–3440.

Matas, J., Chum, O., Urban, M., Pajdla, T., 2004. Robust wide- baseline stereo from maximally stable extremal regions. Image and vision computing , 761–767.

Mishra, A., Alahari, K., Jawahar, C.V., 2016. Enhancing energy minimization framework for scene text recognition with top-down cues. Computer Vision and Image Understanding(145) , 30–42.

Neumann, L., Matas, J., 2012. Real-time scene text localization and recognition. Computer Vision and Pattern Recognition , 3538–3545.

Neumann, L., Matas, J., 2013. On combining multiple segmentations in scene text recognition. Int Conf on Document Analysis and Recognition , 523–527.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2015. You only look once: Unified, real-time object detection. IEEE Computer vision and pattern recognition , 779–788.

Redmon, J., Farhadi, A., 2016. Yolo9000: Better, faster, stronger. arxiv:1612.08242 .

Ren, S., He, K., Girshick, R., Jian Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Conf on Neural Information Processing Systems , 91–99.

Shelhamer, E., Long, J., Darrell, T., 2016. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Shi, B., Bai, X., Yao, C., 2016. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans Pattern Anal Mach Intell. arxiv:1507.05717 .

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. Int Conf on Learning Representations .

Tian, S., Pan, Y., Huang, C., Lu, S., Yu, K., Tan, C., 2015. Text flow: A unified text detection system in natural scene images. IEEE Int Conf on Computer Vision , 4651–4659.

Tian, Z., Huang, W., He, T., He, P., Qiao, Y., 2016. Detecting text in natural image with connectionist text proposal network. Euro conf on Computer vision , 56–72.

Uijlings, J., Van De Sande, K., Gevers, T., Smeulders, A., 2013. Selective search for object recognition. Int journal of computer vision 104, 154–171.

Veit, A., Matera, A., Neumann, L., Matas, J., Belongie, S., 2016. Coco-text: Dataset and benchmark for text detection and recognition in natural images. arxiv:1601.07140 .

Wang, K., Babenko, B., Belongie, S., 2011. End-to-end scene text recognition. IEEE Int Conf on Computer Vision , 1457–1464.

Wang, K., Belongie, S., 2010. Word spotting in the wild. Euro conf on Computer vision , 591–604.

Wang, Q., Yin, F., Liu, C., 2012a. Handwritten chinese text recognition by integrating multiple contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence,34(8) , 1469–1481.

Wang, T., Wu, D., Coates, A., Andrew, N., 2012b. End-to-end text recognition with convolutional neural networks. Int Conf on Pattern Recognition , 3304–3308.

Yao, C., Bai, X., Liu, W., 2014. A unified framework for multioriented text detection and recognition. IEEE Trans. on Image Processing, 23(11) , 4737 – 4749.

Yao, C., Bai, X., Liu, W., Ma, Y., Z, T., 2012. Detecting texts of arbitrary orientations in natural images. IEEE Computer Vision and Pattern Recognition , 1083–1090.

Yao, C., Wu, J., Zhou, x., Zhang, C., Zhou, S., Cao, Z., Yin, Q., 2015. Incidental scene text understanding: Recent progresses on icdar 2015 robust reading competition challenge 4. arxiv:1511.09207 .

Yin, X., Pei, W., Zhang, J., Hao, H., 2015. Multi-orientation scene text detection with adaptive clustering. IEEE Trans. Pattern Analysis and Machine Intelligence , 1930–1937.

Yin, X., Yin, X., Huang, K., Hao, H., 2014. Robust text detection in natural scene images. IEEE Trans. Pattern Analysis and Machine Intelligence, 36(5) , 970–983.

Zhang, S., ZLin, M., Chen, T., Jin, L., Lin, L., 2016a. Character proposal network for robust text extraction. Int Conf on Acoustics, Speech, and Signal Processing , 2633–2637.

Zhang, Z., Shen, W., Yao, C., Bai, X., 2015. Symmetry-based text line detection in natural scenes. IEEE Computer Vision and Pattern Recognition , 2558–2567.

Zhang, Z., Zhang, C., Shen, W., Yao, C., Liu, W., Bai, X., 2016b. Multi-oriented text detection with fully convolutional networks. IEEE Computer vision and pattern recognition , 4159–4167.

Zitnick, L., Dollar, P., 2014. Edge boxes: Locating object proposals from edges. Euro Conf on Computer Vision , 391–405.