Accurate Scene Text Detection Through Border Semantics Awareness and Bootstrapping

Chuhui Xue, Shijian Lu(✉), and Fangneng Zhan

School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore
xuec0003@e.ntu.edu.sg, {shijian.lu,fnzhan}@ntu.edu.sg

Abstract. This paper presents a scene text detection technique that exploits bootstrapping and text border semantics for accurate localization of texts in scenes. A novel bootstrapping technique is designed which samples multiple 'subsections' of a word or text line and accordingly relieves the constraint of limited training data effectively. At the same time, the repeated sampling of text 'subsections' improves the consistency of the predicted text feature maps which is critical in predicting a single complete instead of multiple broken boxes for long words or text lines. In addition, a semantics-aware text border detection technique is designed which produces four types of text border segments for each scene text. With semantics-aware text borders, scene texts can be localized more accurately by regressing text pixels around the ends of words or text lines instead of all text pixels which often leads to inaccurate localization while dealing with long words or text lines. Extensive experiments demonstrate the effectiveness of the proposed techniques, and superior performance is obtained over several public datasets, e.g. 80.1 f-score for the MSRA-TD500, 67.1 f-score for the ICDAR2017-RCTW, etc.

Keywords: Scene text detection · Data augmentation
Semantics-aware detection · Deep network models

1 Introduction

Scene text detection and recognition has attracted increasing interests in recent years in both computer vision and deep learning research communities due to its wide range of applications in multilingual translation, autonomous driving, etc. As a prerequisite of scene text recognition, detecting text in scenes plays an essential role in the whole chain of scene text understanding processes. Though studied for years, accurate and robust detection of texts in scenes is still a very open research challenge as witnessed by increasing benchmarking competitions in recent years such as ICDAR2015-Incidental [19], ICDAR2017-MLT [30], etc.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11220, pp. 370–387, 2018.
https://doi.org/10.1007/978-3-030-01270-0_22

With the fast development of convolutional neural networks (CNN) in representation learning and object detection, two CNN-based scene text detection approaches have been investigated in recent years which treat words or text lines as generic objects and adapt generic object detection techniques for the scene text detection task. One approach is indirect regression based [7,17,23,25] which employs object detectors such as Faster-RCNN [34] and SSD [24] that first generate proposals or default boxes and then regress to accurate object boxes. These techniques achieve state-of-the-art performance but require multiple proposals of different lengths, angles and shapes. Another approach is direct regression based [11,52] which adapts DenseBox [14] for the scene text detection task. This approach does not require proposals and is capable of detecting words and text lines of different orientations and lengths, but it often suffers from low localization accuracy while dealing with long words or text lines.

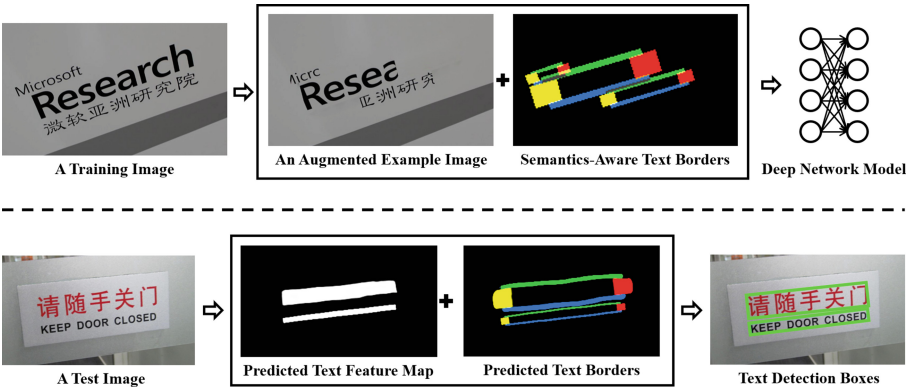


Fig. 1. Overview of proposed scene text detection technique: For each training image, a set of augmented images and semantics-aware text borders are extracted and fed to a multi-channel fully convolutional network to train a scene text detector (as shown above the dotted line). Given a test image, the scene text detector predicts a text feature map and four text borders (highlighted in four colors) for accurate scene text detection (as shown below the dotted line). (Color figure online)

Both direct and indirect regression based approaches are thus facing three common constrains while adapted for the scene text detection task. The first is broken detections where a text line is detected as multiple broken text segments. The reason is that text lines often suffer from more variation as compared with characters or words due to their larger spatial coverage in scenes, e. g. different words within a text line may have different colors, fonts, environmental lighting, etc. The second is inaccurate localization, where the regression fails to produce an accurate text box either by missing certain parts of texts or including certain neighboring background. The inaccurate localization is largely due to the long shape of text lines where the regressing text pixels around the text line center are very far from text line ends where text bounding box vertices are

located. The third is limited training data. A large amount of annotations are required to capture the rich variation within scene texts, but existing datasets often have limited training images, e.g. 300 training images in MSRA-TD500 [44], 229 training images in ICDAR2013 [20], etc.

We design two novel techniques to tackle the three constraints of state-of-the-art scene text detection techniques. First, we design a novel bootstrapping based scene text sampling technique that repeatedly extracts text segments of different lengths from annotated texts as illustrated in Fig. 1. The bootstrapping based sampling helps from two aspects. First, it augments the training data and relieves the data annotation constraint by leveraging existing scene text annotations. Second, the repeated sampling of text segments of various lengths helps to decouple different types of image degradation and reduce the complexity of training data effectively, e. g. scene texts with different lighting within the same text line could be sampled by different text line segments with less variation as illustrated in Fig. 2. The proposed bootstrapping based scene text sampling technique thus helps to improve the consistency of the produced text feature map and performance of regression which are critical for detecting a complete instead of multiple broken boxes for a long word or text line. The idea of repeated sampling has been exploited in training generic object detectors by cropping multiple samples around annotated objects of interest.

Second, we design a novel semantics-aware text border detection technique for accurate localization of texts in scenes. In particular, four text border segments are defined by a pair of long-side borders and a pair of short-side borders which can be extracted based on the text annotation boxes automatically as illustrated in Figs. 1 and 4. By labeling the four text border segments as four types of objects, the trained scene text detector is capable of detecting the four types of text border segments separately as illustrated in Fig. 1 (four colors are for illustration only). The differentiation of the four text border segments helps to improve the text localization accuracy from two aspects. First, the text bounding box can be regressed more accurately by using text pixels lying around the two ends of words or text lines (which can be identified by using the short-side text border segments) that are much closer to the text bounding box vertices as compared with text pixels lying around the middle of text lines. Second, the long-side text border segments can be exploited to separate neighboring text lines especially when they are close to each other.

2 Related Work

Scene Text Detection. Quite a number of scene text detection techniques have been reported in the literature [46,53] and they can be broadly classified into three categories depending on whether they detect characters, words, or text lines directly. The first category takes a bottom-up approach which first detects characters [2,39,47] or text components [35,40] and then groups them into words or text lines. The earlier works detect characters using various hand-crafted features such as stroke width transform (SWT) [4,44], maximally stable extremal

regions (MSERs) [2,15,18,31], boundary [28], FAST keypoints [1], histogram of oriented gradients (HoG) [39], stroke symmetry [49], etc. With the fast development of deep neural networks, CNNs have been widely used to detect characters in scenes, either by adapting generic object detection methods [35,40] or taking a semantic image segmentation approach [9,45,50]. Additionally, different techniques have been developed to connect the detected characters into words or text lines by using TextFlow [39], long short-term memory (LSTM) [50], etc [16,26,45].

The second category treats words as one specific type of objects and detects them directly by adapting various generic object detection techniques. The methods under this category can be further classified into two classes. The first class leverages Faster-RCNN [34], YOLO [33] and SSD [24] and designs text-specific proposals or default boxes for scene text detection [5,7,17,23,25,38]. The second class takes a direct regression approach [11,52] which first detects region of interest (ROI) and then regresses text boxes around the ROI at pixel level.

The third category detects text lines directly by exploiting the full convolution network (FCN) [27] that has been successfully applied for semantic image segmentation. For example, He *et al.* [8] proposed a coarse-to-fine FCN that detects scene texts by extracting text regions and text central lines. In [32,42], FCN is exploited to learn text border maps, where text lines are detected by finding connected components with text labels.

Our proposed technique takes the direct regression approach as in [11,52] that regresses word and text line boxes directly from text pixels. On the other hand, we detect multiple text border segments with specific semantics (instead of a whole text border as in [32,42]) that help to improve the scene text localization accuracy greatly, more details to be described in Sect. 3.2.

Data Augmentation. Data augmentation has been widely adopted in deep network training as a type of regularization for avoiding over-fitting. For various computer vision tasks such as image classification and object detection, it is widely implemented through translating, rotating, cropping and flipping of images or annotated objects of interest for the purpose of creating a larger amount of training data [6,22,37]. Some more sophisticated augmentation schemes have been proposed in recent years, e. g. using masks to hide certain parts of objects to simulate various occlusion instances [51]. Data augmentation has become one routine operation in deep learning due to its effectiveness in training more accurate and more robust deep network models.

Our bootstrapping based scene text sampling falls under the umbrella of data augmentation. It is similar to image cropping but involves innovative designs by catering to text-specific shapes and structures. By decoupling image variations in long words or text lines, it helps to produce more consistent scene text features which is critical in predicting a single complete instead of multiple broken boxes for a word or text line, more details to be described in Sect. 3.1.

3 Methodology

We proposed a novel scene text detection technique that exploits bootstrapping for data augmentation and semantics-aware text border segments for accurate scene text localization. For each training image, the proposed technique extracts a set of bootstrapped training samples and two pairs of text border segments as illustrated in Fig. 1, and feeds them (together with the original scene text annotations) to a multi-channel fully convolutional network to train a scene text detection model. The bootstrapping based sampling improves the consistency of the produced text feature map which greatly helps to predict a single complete instead of multiple broken boxes for long words or text lines. The semantics of the detected text border segments greatly help to regress more accurate localization boxes for words or text lines in scenes as illustrated in Fig. 1.

3.1 Bootstrapping Based Image Augmentation

Fig. 2. Illustration of the bootstrapping based scene text sampling: Given an image with a text line as annotated by the green box, three example text line segments are extracted as highlighted by red boxes where the centers of the sampling windows are taken randomly along the center line of the text line (the shrunk part in yellow color). The rest text regions outside of the sampling windows are filled by inpainting. (Color figure online)

We design a bootstrapping based image augmentation technique that repeatedly samples text line segments for each text annotation box (TAB) as labeled by the green box in the top-left image in Fig. 2. With L denoting the TAB length, the center line of the TAB (as highlighted by the dashed line) is first shrunk by $0.1 * L$ from both TAB ends which gives the yellow line segment as shown in Fig. 2. Multiple points are then taken randomly along the shrunk center line for text segment sampling. The length of each sampled text segment varies from $0.2 * L$ to twice the distance between the sampling point to the closer TAB end. In addition, the rest of the TAB outside the sampled text segment is filled by inpainting [42] as illustrated in Fig. 2. With the sampling process as described

(a) (b) (c) (d) (e)

Fig. 3. The inclusion of augmented images improves the scene text detection: With the inclusion of the augmented images in training, more consistent text feature maps and more complete scene text detections are produced as shown in (d) and (e), as compared with those produced by the baseline model (trained using original training images only) shown in (b) and (c). The coloring in the text feature maps shows the distance information predicted by regressor (blue denotes short distances and red denotes long distance). (Color figure online)

above, the number of the augmented images can be controlled by the number of text segments that are sampled from each text box.

The proposed bootstrapping based scene text image augmentation technique improves the consistency of predicted text feature map as well as the performance of regression greatly as illustrated in Fig. 3. For the sample images in Fig. 3a, Figs. 3b and 3d show the text feature maps that are produced by the baseline model (trained by using the original training images) and the augmented model (trained by further including the augmented sample images), respectively (training details to be described in Sect. 3.3). The coloring in the text feature maps shows the distance to the left-side boundary as predicted by regressor - blue denotes short distance and red denotes long distance. Figures 3c and 3e show the corresponding detection boxes, respectively, where red boxes show false detections and green boxes show correct detections. It can be seen that the inclusion of the augmented images helps to produce more consistent text feature maps as well as smoother geometrical distance maps (for regression of text boxes) which leads to more complete instead of broken scene text detections.

3.2 Semantics-Aware Text Borders

We extract two pairs of semantics-aware text border segments for each scene text annotation as illustrated in Fig. 4. With W and L denoting the width and length of a text annotation box (TAB), a pair of long text border segments in green and blue colors can be extracted along the two long edges of the TAB as

illustrated in Fig. 4, where the segment length is set at L and the segment width is empirically set at $0.2 * W$. In addition, the center line of the two long text border segments overlaps perfectly with the long edges of TAB so that the text border segments capture the transition from text to background or vice versa.

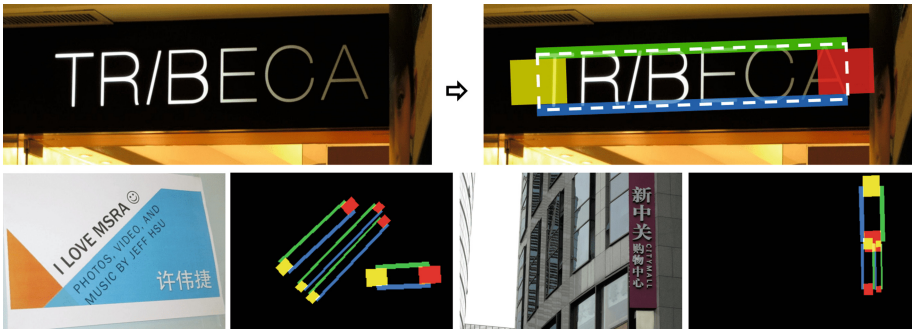
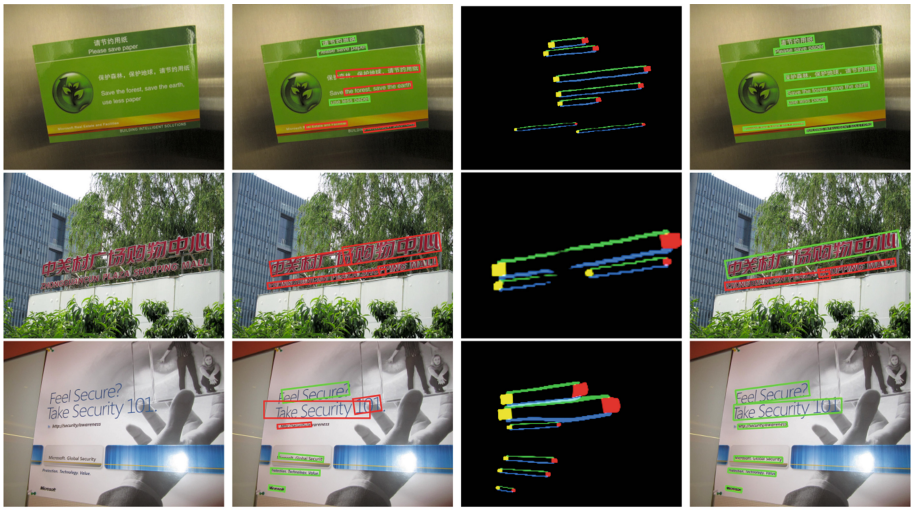


Fig. 4. Semantics-aware text border detection: Four text border segments are automatically extracted for each text annotation box including a pair of short-edge text border segments in yellow and red colors and a pair of long-edge text border segments in green and blue colors. The four types of text border segments are treated as four types of objects and used to train deep network models, and the trained model is capable of detecting the four types of text border segments as illustrated in Fig. 5c. (Color figure online)

A pair of short text border segments can also be extracted based on the TAB as illustrated in Fig. 4. In particular, the dimension along the TAB width is set at $0.8 * W$ which fits perfectly in between the two long text segments. Another dimension along the TAB length is set the same as W with which the trained text border detector can detect a certain amount of text pixels to be used in text bounding box regression. Similarly, the center line of the short text border segments (along the TAB width) overlaps perfectly with the TAB short edge so that the extracted text border segments capture the transition from text to background or vice versa.

The use of the semantics-aware text borders helps to improve the localization accuracy of the trained scene text detection model greatly (training details to be described in Sect. 3.3). With the identified text border semantics as shown in Fig. 5c (four colors are for illustration only), text pixels around the text line ends can be determined by the overlap between the short text border segments and the predicted text feature map. The text bounding box can thus be regressed by using the text pixels lying around the text line ends which often leads to accurate text localization as illustrated in Fig. 5d. The reason is that text pixels around the middle of texts are far from the text box vertices for long words or text lines which can easily introduce regression errors and lead to inaccurate localization as illustrated in Fig. 5b. At the other end, the long text border segments also



(a) (b) (c) (d)

Fig. 5. The use of semantics-aware text borders improves scene text detection: With the identified text border semantics information as illustrated in (c), scene texts can be localized much more accurately as illustrated in (d) as compared with the detections without using the border semantics information as illustrated in (b). Green boxes give correct detections and red boxes give false detections. (Color figure online)

help for better scene text detection performance. In particular, the long text border segments can be exploited to separate text lines when neighboring text lines are close to each other.

3.3 Scene Text Detection

The original scene text annotations, together with the augmented images and the extracted semantics-aware text borders as described in Sects. 3.1 and 3.2, are fed to a multi-channel FCN to train a scene text detection model. The training aims to minimize the following multi-task loss function:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_{loc} * \mathcal{L}_{loc} + \lambda_{brd} * \mathcal{L}_{brd} \tag{1}$$

where $\mathcal{L}_{cls}$, $\mathcal{L}_{loc}$ and $\mathcal{L}_{brd}$ refer to loss of text feature (confidence score of each pixel being a text pixel), regression (distances from each pixel to four sides of text boundaries) and border feature (confidence score of each pixel being a border pixel), respectively. Parameters λ_{loc} and λ_{brd} are weights of the corresponding losses which are empirically set at 1.0 in our system.

For the regression loss $\mathcal{L}_{loc}$, we adopt the IoU loss [48] in training. For the classification losses $\mathcal{L}_{cls}$ and $\mathcal{L}_{brd}$, we use Dices Coefficient [29] which is a widely-used in image segmentation tasks. Given a ground truth region G and a predicted region P, the Dices Coefficient is defined by:

Algorithm 1 Text bounding box detection.

Input: Regressor R, text region map T and text border region map B
Output: A list of text bounding boxes **BB**
Delineate text lines t in T using the long text border regions in B
Determine left-side and right-side regressing text pixels p_l and p_r by overlaps between the delineated t and the two short text border regions in B
Derive two sets of text boxes $\mathbf{BB}_l$ and $\mathbf{BB}_r$ by regressing p_l and p_r
$\mathbf{BB} \leftarrow \Phi$
for each box in $\mathbf{BB}_l$ and $\mathbf{BB}_r$ **do**
  if the two boxes are regressed from text pixels of the same t **then**
    Merge the two boxes and add the merged box to **BB**
  end if
end for
Apply NMS to **BB**

$$\mathcal{L}_{brd} = \frac{2 * |G \cap P|}{|G| + |P|} \tag{2}$$

Given a test image, our trained scene text detector produces three maps including a text feature map, a text border feature map, and a regressor. The text border feature map has four channels which give a pair of short text border segments and a pair of long text border segments as illustrated in Fig. 5c. The regressor also has four channels that predict the distances to the upper, lower, left and right text boundaries, respectively, as illustrated in Figs. 3c and 3e (which shows one channel distance to the left-side text boundary).

Algorithm 1 shows how the text bounding boxes are derived from the outputs of the trained scene text detector. Given a text feature map and a text border feature map, a text region map and four text border region maps are first determined (as the algorithm inputs) by global thresholding where the threshold is simply estimated by the mean of the respective feature map. Overlaps between the text region map and four text border region maps can then be determined. Text lines can thus be delineated by removing the overlaps between the text region map and the two long text border region maps. Further, text bounding box vertices at the left and right text line ends can be predicted by regressing the text pixels that overlap with the left-side and right-side text border region maps, respectively. Finally, the text bounding box is determined by merging the regressed left-side and right-side text box vertices.

4 Experiments

4.1 Datasets and Evaluation Metrics

MSRA-TD500[1] [44] comprise 300 training images and 200 testing images with scene texts printed in either Chinese or English. For each training image, annotations at either word or text line level is provided, where each annotation consists

[1] http://tc11.cvc.uab.es/datasets/MSRA-TD500_1.

Table 1. Recall (R), precision (P) and f-score (F) of different scene text detection methods over the MSRA-TD500 and ICDAR2013 datasets.

MSRA-TD500				ICDAR2013			
Method	R	P	F	Method	R	P	F
Kang *et al.* [18]	62.0	71.0	66.0	He *et al.* [11]	81.0	92.0	86.0
Yin *et al.* [47]	63.0	81.0	71.0	Tian *et al.* [38]	83.1	91.1	86.9
Zhang *et al.* [50]	67.0	83.0	74.0	He *et al.* [7]	86.0	88.0	87.0
He *et al.* [11]	70.0	77.0	74.0	Zhou *et al.* [52]	82.7	92.6	87.7
Yao *et al.* [45]	75.3	76.5	75.9	Jiang *et al.* [17]	82.6	**93.6**	87.7
Zhou *et al.* [52]	67.4	**87.3**	76.1	He *et al.* [10]	83.0	93.0	88.0
Shi *et al.* [35]	70.0	86.0	77.0	Tian *et al.* [40]	87.0	88.0	88.0
Wu *et al.* [42]	**78.0**	77.0	77.0	Hu *et al.* [12]	**87.5**	93.3	**90.3**
Baseline(ResNet)	73.4	70.3	71.8	Baseline(ResNet)	79.3	86.9	83.0
Border(ResNet)	72.0	76.4	74.3	Border(ResNet)	84.5	85.4	84.9
Aug.(ResNet)	71.1	77.7	74.3	Aug.(ResNet)	86.7	83.8	85.2
Aug.+Border(ResNet)	73.3	80.7	76.8	Aug.+Border(ResNet)	86.9	87.8	87.4
Aug.+Border(DenseNet)	77.4	83.0	**80.1**	Aug.+Border(DenseNet)	87.1	91.5	89.2

of a rectangle box and the corresponding box rotation angle. Due to the very small number of training images, 400 training images in the HUST-TR400 [2] [43] are included in training.

ICDAR2013[3][20] consists of 229 training images and 233 testing images with texts in English. The text annotations are at word level, and no rotation angles are provided as most captured scene texts are almost horizontal. We also include training images from ICDAR2015 in training.

ICDAR2017-RCTW[4][36] comprises 8,034 training images and 4,229 testing images with scene texts printed in either Chinese or English. The images are captured from different sources including street views, posters, screen-shot, etc. Multi-oriented words and text lines are annotated using quadrilaterals.

ICDAR2017-MLT[5][30] comprise 7,200 training images, 1,800 validation images and 9,000 testing images with texts printed in 9 languages including Chinese, Japanese, Korean, English, French, Arabic, Italian, German and Indian. Most annotations are at word level while texts in non-Latin languages like Chinese are annotated at text-line level. Similar to ICDAR2017-RCTW, texts in this dataset are also multi-oriented with text annotated using quadrilaterals.

Evaluation Metrics. For MSRA-TD500, we use the evaluation protocol in [41]. For ICDAR2013, ICDAR2017-RCTW and ICDAR2017-MLT, we perform eval-

[2] http://mclab.eic.hust.edu.cn/UpLoadFiles/dataset/HUST-TR400.zip.
[3] http://rrc.cvc.uab.es/?ch=2&com=introduction.
[4] http://www.icdar2017chinese.site:5080/dataset/.
[5] http://rrc.cvc.uab.es/?ch=8.

uations by using the online evaluation systems that are provided by the respective dataset creators. In particular, one-to-many (one rectangle corresponds to many rectangles) and many-to-one (many rectangles correspond to one rectangle) matches are adopted for better evaluation for the ICDAR2013 dataset.

4.2 Implementation Details

The network is optimized by Adam [21] optimizer with starting learning rate of 10^{-4} and batch size of 16. Images are randomly resized with ratio of 0.5, 1, 2, or 3 and cropped into 512×512 without crossing texts before training. 20 augmented images are sampled for each training images by using the proposed data augmentation technique. The whole experiments are conducted on Nvidia DGX-1. All our models are fine-tuned from a pre-trained model using the ImageNet dataset [3]. Two base networks including ResNet [6] and DenseNet [13] are implemented for evaluation. Multi-scale evaluation is implemented by resizing the longer side of test images to 256, 512, 1024, 2048 pixels.

Table 2. Recall (R), precision (P) and f-score (F) of different detection methods over the ICDAR2017-RCTW and ICDAR2017-MLT datasets.

ICDAR2017-RCTW				ICDAR2017-MLT			
Method	R	P	F	Method	R	P	F
gmh [36]	57.8	70.6	63.6	Sensetime OCR [30]	**69.4**	56.9	62.6
NLPR_PAL [36]	57.3	77.2	65.8	SCUT_DLVClab [30]	54.5	**80.3**	65.0
Foo & Bar [36]	**59.5**	74.4	66.1	NLPR_PAL [11]	57.9	76.7	66.0
Baseline(ResNet)	52.2	66.6	58.5	Baseline(ResNet)	60.9	64.5	62.6
Border(ResNet)	58.5	74.2	65.4	Border(ResNet)	60.6	73.9	66.6
Border(DenseNet)	58.8	**78.2**	**67.1**	Border(DenseNet)	62.1	77.7	**69.0**

4.3 Experimental Results

Quantitative Results. Table 1 shows quantitative experimental results on the MSRA-TD500 and ICDAR2013 datasets as well as comparisons with state-of-the-art methods. As Table 1 shows, five models are trained including: (1) 'Baseline (ResNet)' that is trained by using ResNet-50 and the original training images as described in Sect. 4.1, (2) 'Border (ResNet)' that is trained by including text border segments as described in Sect. 3.2, (3) 'Aug. (ResNet)' that is trained by including augmented scene text images as described in Sect. 3.1, (4) 'Aug.+Border (ResNet)' that is trained by including both text border segments and augmented images, and (5) 'Aug.+Border (DenseNet)' that is trained by using DenseNet-121 with the same training data as the 'Aug.+Border (ResNet)'.

As Table 1 shows, the detection models using either semantics-aware text borders or augmented images or the both outperform the baseline model consistently. In addition, the models using both text borders and augmented images outperform the ones using either text borders or augmented images alone. Further, the trained models outperform the state-of-the-art methods clearly when the DenseNet-121 is used, demonstrating the superior performance of the proposed technique. We observe that the performance improvement is mainly from higher precisions for the MSRA-TD500 dataset as compared with higher recalls for the ICDAR2013 dataset. This inconsistency is largely due to the different evaluation methods for the two datasets, i.e. the evaluation of the MSRA-TD500 follows one-to-one (one rectangle corresponds to one rectangles) match whereas the evaluation of the ICDAR2013 follows one-to-many and many-to-one. As studied in [42], the ICDAR2013 online evaluation system usually produces lower precision as compared with the real values. We conjecture that the actual precision by our method should be higher than what is presented in Table 1.

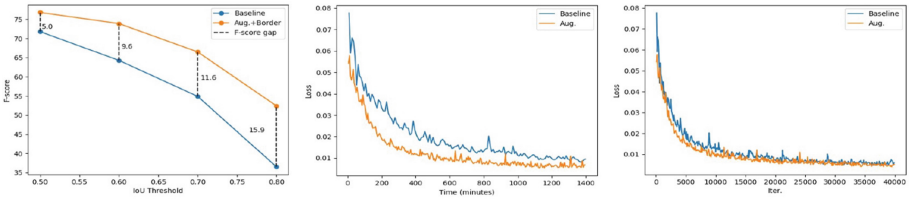
(a) (b) (c) (d)

Fig. 6. Illustration of the proposed scene text detection technique: Successful detections where sample images are picked from the four studied datasets including (a) MSRA-TD500, (b) ICDAR2013, (c) ICDAR2017-RCTW and (d) ICDAR2017-MLT, respectively.

The proposed technique is also evaluated on two more recent large-scale datasets including the ICDAR2017-RCTW and the ICDAR2017-MLT. As the two datasets both have a large amount of training images, we evaluated the proposed semantics-aware text borders only. As Table 2 shows, the use of semantics-aware text borders helps to improve both detection recall and detection precision clearly for the ICDAR2017-RCTW dataset where a mixture of word level

and text-line level annotations is created. For the ICDAR2017-MLT dataset, the improvement is mainly from higher precisions as most annotations in this dataset are at word level. In addition, the proposed techniques outperform state-of-the-art methods (including baselines and winning methods as reported in the ICDAR2017-RCTW and ICDAR2017-MLT benchmarking competition papers [30,36]) for both datasets when DenseNet is used.

Qualitative Results. Figure 6 shows several sample images and the corresponding detections by using the proposed technique, where all sample images are picked from the four studied datasets including several images in Figs. 3 and 5. As Fig. 6 shows, the proposed technique is capable of detecting scene texts that have different characteristics and suffer from different types of degradation. In particular, the inclusion of the bootstrapping based augmentation helps to produce more complete detections though texts may not be localized accurately as illustrated in Fig. 3. On the other hand, the inclusion of the semantics-aware text borders helps to produce more accurate scene text localization though texts may be detected by multiple broken boxes as illustrated in Fig. 5. The combination of bootstrapping based augmentation and semantics-aware text borders overcomes both constraints (broken detection and inaccurate localization) and produces more complete and accurate text detections as illustrated in Fig. 6.



(a) (b) (c)

Fig. 7. The inclusion of the semantics-aware text borders and bootstrapping based augmentation helps to improve the scene text localization accuracy greatly as illustrated in (a), where the f-score gap keeps increasing with the increment of IoU threshold on MSRA-TD500 dataset. Additionally, the inclusion of the bootstrapping based augmentation also leads to fast learning and convergence as illustrated in (b) and (c) on MSRA-TD500 dataset.

Discussion. The proposed technique is capable of producing accurate scene text localization which is critical to the relevant scene text recognition task. This can be observed in Fig. 7a that shows f-scores of the proposed model (the semantics-aware text borders and augmented images are included in training) vs the baseline model (training uses the original images only) when different IoUs (Intersection over Union) are used in evaluation. As Fig. 7a shows, the f-score gap increases from 5.0 to 15.9 steadily when the IoU threshold increases from 0.5 to 0.8, demonstrating more accurate scene text localization by the proposed technique.

Fig. 8. Illustration of the failure cases of proposed scene text detection technique: Sample images are from the four studied datasets, where green boxes are correct outputs of our methods, red boxes are false detections and yellow boxes give the ground-truth missing detections. (Color figure online)

Another interesting observation is that the inclusion of the augmented images often accelerates the training convergence as illustrated in Fig. 7b. For training over 40,000 iterations (batch size of 16) on the MSRA-TD500 dataset, the model using the augmented images takes 36 hours while the baseline model using the original training images only takes 56 hours when both models converge and obtain f-scores of 74.3 and 71.8, respectively, as shown in Table 1. This can be further verified by checking the training loss vs training iteration number as shown in Fig. 7c. Experiments on other datasets show similar convergence pattern as shown in Figs. 7b and 7c. This does not make sense at the first sight as the augmentation increases the number of training images by 20 times (20 augmented images are sampled for each training image). We conjecture that the faster convergence is largely due to the augmented text line segments that are shorter than the original text lines and accordingly decouple different types of image variation which leads to the faster learning and model convergence.

The proposed technique could fail under several typical scenarios as illustrated in Fig. 8. First, it may introduce false positives while handling scene texts of a big size, largely due to NMS errors as shown in the first image. Second, it may produce incorrect broken detections when a text line has large blanks as shown in the second image. This kind of failure often results from annotation inconsistency where some long text line with large blanks is annotated by a single box whereas some is annotated by multiple boxes. Third, it could be confused when vertical texts can also be interpreted as horizontal and vice versa as shown in the third image. Without the text semantic information, it is hard to tell whether it is two vertical text lines or five horizontal words.

5 Conclusions

This paper presents a novel scene text detection technique that makes use of semantics-aware text borders and bootstrapping based text segment augmentation. The use of semantics-aware text borders helps to detect text border segments with different semantics which improves the scene text localization accuracy greatly. The use of augmented text line segments helps to improve the

consistency of predicted feature maps which leads to more complete instead of broken scene text detections. Experiments over four public datasets show the effectiveness of the proposed techniques.

Acknowledgement. This work is funded by the Ministry of Education, Singapore, under the project "A semi-supervised learning approach for accurate and robust detection of texts in scenes" (RG128/17 (S)).

References

1. Busta, M., Neumann, L., Matas, J.: FASText: efficient unconstrained scene text detector. In: IEEE International Conference on Computer Vision (ICCV), vol. 1 (2015)
2. Cho, H., Sung, M., Jun, B.: Canny text detector: fast and robust scene text localization algorithm. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3566–3573 (2016)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 248–255. IEEE (2009)
4. Epshtein, B., Ofek, E., Wexler, Y.: Detecting text in natural scenes with stroke width transform. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2963–2970. IEEE (2010)
5. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. He, P., Huang, W., He, T., Zhu, Q., Qiao, Y., Li, X.: Single shot text detector with regional attention. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
8. He, T., Huang, W., Qiao, Y., Yao, J.: Accurate text localization in natural image with cascaded convolutional text network. arXiv preprint arXiv:1603.09423 (2016)
9. He, T., Huang, W., Qiao, Y., Yao, J.: Text-attentional convolutional neural network for scene text detection. IEEE Trans. Image Process. **25**(6), 2529–2541 (2016)
10. He, T., Tian, Z., Huang, W., Shen, C., Qiao, Y., Sun, C.: An end-to-end textspotter with explicit alignment and attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5020–5029 (2018)
11. He, W., Zhang, X.Y., Yin, F., Liu, C.L.: Deep direct regression for multi-oriented scene text detection. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
12. Hu, H., Zhang, C., Luo, Y., Wang, Y., Han, J., Ding, E.: WordSup: exploiting word annotations for character based text detection. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
13. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, p. 3 (2017)
14. Huang, L., Yang, Y., Deng, Y., Yu, Y.: DenseBox: unifying landmark localization with end to end object detection. arXiv preprint arXiv:1509.04874 (2015)

15. Huang, W., Qiao, Y., Tang, X.: Robust scene text detection with convolution neural network induced MSER trees. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 497–511. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_33
16. Jaderberg, M., Vedaldi, A., Zisserman, A.: Deep features for text spotting. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 512–528. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_34
17. Jiang, Y., et al.: R2CNN: rotational region CNN for orientation robust scene text detection. arXiv preprint arXiv:1706.09579 (2017)
18. Kang, L., Li, Y., Doermann, D.: Orientation robust text line detection in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4034–4041 (2014)
19. Karatzas, D., et al.: ICDAR 2015 competition on robust reading. In: Document Analysis and Recognition (ICDAR), pp. 1156–1160. IEEE (2015)
20. Karatzas, D., et al.: ICDAR 2013 robust reading competition. In: 2013 12th International Conference on Document Analysis and Recognition (ICDAR), pp. 1484–1493. IEEE (2013)
21. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
22. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
23. Liao, M., Shi, B., Bai, X., Wang, X., Liu, W.: TextBoxes: a fast text detector with a single deep neural network. In: AAAI, pp. 4161–4167 (2017)
24. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
25. Liu, Y., Jin, L.: Deep matching prior network: toward tighter multi-oriented text detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
26. Liu, Z., Lin, G., Yang, S., Feng, J., Lin, W., Goh, W.L.: Learning Markov clustering networks for scene text detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
27. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
28. Lu, S., Chen, T., Tian, S., Lim, J.H., Tan, C.L.: Scene text extraction based on edges and support vector regression. Int. J. Doc. Anal. Recognit. (IJDAR) **18**(2), 125–135 (2015)
29. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
30. Nayef, N., et al.: ICDAR 2017 robust reading challenge on multi-lingual scene text detection and script identification-RRC-MLT. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp. 1454–1459. IEEE (2017)
31. Neumann, L., Matas, J.: Real-time scene text localization and recognition. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3538–3545. IEEE (2012)

32. Polzounov, A., Ablavatski, A., Escalera, S., Lu, S., Cai, J.: WordFence: text detection in natural images with border awareness. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 1222–1226. IEEE (2017)
33. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
34. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
35. Shi, B., Bai, X., Belongie, S.: Detecting oriented text in natural images by linking segments. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
36. Shi, B., et al.: ICDAR 2017 competition on reading Chinese text in the wild (RCTW-17). In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 1429–1434. IEEE (2017)
37. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
38. Tian, S., Lu, S., Li, C.: WeText: scene text detection under weak supervision. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
39. Tian, S., Pan, Y., Huang, C., Lu, S., Yu, K., Lim Tan, C.: Text flow: a unified text detection system in natural scene images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4651–4659 (2015)
40. Tian, Z., Huang, W., He, T., He, P., Qiao, Y.: Detecting text in natural image with connectionist text proposal network. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 56–72. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_4
41. Wang, K., Babenko, B., Belongie, S.: End-to-End scene text recognition. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 1457–1464. IEEE (2011)
42. Telea, A.: An image inpainting technique based on the fast marching method. J. Graph. Tools **9**(1), 23–34 (2004)
43. Yao, C., Bai, X., Liu, W.: A unified framework for multioriented text detection and recognition. IEEE Trans. Image Process. **23**(11), 4737–4749 (2014)
44. Yao, C., Bai, X., Liu, W., Ma, Y., Tu, Z.: Detecting texts of arbitrary orientations in natural images. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1083–1090. IEEE (2012)
45. Yao, C., Bai, X., Sang, N., Zhou, X., Zhou, S., Cao, Z.: Scene text detection via holistic, multi-channel prediction. arXiv preprint arXiv:1606.09002 (2016)
46. Ye, Q., Doermann, D.: Text detection and recognition in imagery: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **37**(7), 1480–1500 (2015)
47. Yin, X.C., Pei, W.Y., Zhang, J., Hao, H.W.: Multi-orientation scene text detection with adaptive clustering. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1930–1937 (2015)
48. Yu, J., Jiang, Y., Wang, Z., Cao, Z., Huang, T.: UnitBox: an advanced object detection network. In: Proceedings of the 2016 ACM on Multimedia Conference, pp. 516–520. ACM (2016)
49. Zhang, Z., Shen, W., Yao, C., Bai, X.: Symmetry-based text line detection in natural scenes. In: IEEE Computer Vision and Pattern Recognition (CVPR), vol. 1, p. 3. (2015)

50. Zhang, Z., Zhang, C., Shen, W., Yao, C., Liu, W., Bai, X.: Multi-oriented text detection with fully convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4159–4167 (2016)
51. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. arXiv preprint arXiv:1708.04896 (2017)
52. Zhou, X., et al.: East: an efficient and accurate scene text detector. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
53. Zhu, Y., Yao, C., Bai, X.: Scene text detection and recognition: recent advances and future trends. Front. Comput. Sci. **10**(1), 19–36 (2016)