WeText: Scene Text Detection under Weak Supervision

Shangxuan Tian[1], Shijian Lu[2], and Chongshou Li[3]

[1]Visual Computing Department, Institute for Infocomm Research
[2] School of Computer Science and Engineering, Nanyang Technological University
[3]Department of Industrial Systems Engineering and Management , National University of Singapore
tianshangxuan@u.nus.edu, Shijian.Lu@ntu.edu.sg, iselc@nus.edu.sg

Abstract

The requiring of large amounts of annotated training data has become a common constraint on various deep learning systems. In this paper, we propose a weakly supervised scene text detection method (WeText) that trains robust and accurate scene text detection models by learning from unannotated or weakly annotated data. With a "light" supervised model trained on a small fully annotated dataset, we explore semi-supervised and weakly supervised learning on a large unannotated dataset and a large weakly annotated dataset, respectively. For the unsupervised learning, the light supervised model is applied to the unannotated dataset to search for more character training samples, which are further combined with the small annotated dataset to retrain a superior character detection model. For the weakly supervised learning, the character searching is guided by high-level annotations of words/text lines that are widely available and also much easier to prepare. In addition, we design an unified scene character detector by adapting regression based deep networks, which greatly relieves the error accumulation issue that widely exists in most traditional approaches. Extensive experiments across different unannotated and weakly annotated datasets show that the scene text detection performance can be clearly boosted under both scenarios, where the weakly supervised learning can achieve the state-of-the-art performance by using only 229 fully annotated scene text images.

1. Introduction

Automatic reading texts in scene images has attracted growing interest in recent years due to its great advantages in image content understanding and contextual information inference. It has been widely used in various tasks such as multilingual image translation (Google translate [31]), vehicle auto-navigation [28], object recognition [15], and assistive smartphone applications for visually impaired people [1]. An indispensable component of an automatic scene text reading system is scene text detection under unconstrained conditions. This is still a very open research challenge due to the tremendous complexity imposed by diverse text fonts and styles, arbitrary text sizes, various geometric distortions, complex image backgrounds, uncontrolled illuminations, *etc*.

Figure 1: Text detection examples of the proposed WeText system. In the top row from left to right are one sample image and detection outputs using the baseline model. Images in the bottom row from left to right are text detections using the proposed semi-supervised and weakly supervised learning approaches, respectively. Blue boxes indicate correct detections while red boxes and green boxes indicate false positives and false negatives, respectively. Detection results have been zoomed in for better visualization.

Two approaches have been explored to address the scene text detection challenge. The first is character based, which first detects character candidates by particular operators such as Stroke Width Transform (SWT) [4], Maximally Sta-

ble Extremal Regions (MSER) [22, 35], sliding windows [29], followed by identifying the real characters with a pre-trained text/non-text classifier. Words or text lines are further determined by grouping the detected characters with heuristic rules [11, 35] or more sophisticated graph models [29, 34]. The second approach is to detect words directly either by generating word proposals [12, 38] or regressing word bounding boxes from default anchor boxes [17]. This approach is simpler and more efficient compared with the character-based approach. On the other hand, it does not work well with multi-oriented text as word proposals tend to detect horizontal texts. In addition, many non-Latin languages such as Chinese do not have a clear word boundary which greatly restricts its applicability.

We take a character based approach due to its flexibility in dealing with multilingual and multi-oriented texts in scenes. However, the character based approach has two major constraints. First, a robust and accurate character detector requires a large amount of annotated character images that are time consuming and costly to prepare. Second, the current character based approach, which first detects characters candidates and then identifies true characters by a text/non-text classifier, is complicated and also accumulates errors.

In this paper, we propose a weakly supervised scene text detection framework (WeText) that is capable of learning a robust and accurate scene text detector with a small amount of annotated character images. The idea is to first train a "light" supervised model by using a small amount of fully annotated character images and then apply the model on a large amount of unannotated or weakly annotated images to search for positive training samples. The searched samples are combined with the small amount of annotated images to re-train a more robust and accurate detector. We investigated two learning strategies including semi-supervised learning that requires no annotations and weakly supervised learning where the character searching is guided by high-level annotations of words or text lines. In addition, we adapt regression based deep networks and design a proposal-free character detector that integrates the character candidate detection and text/non-text classification into a single process to reduce the error accumulation. To the best of our knowledge, this is the first work that uses regression based deep networks for scene character detection, and it is also the first work that fully studies the impact of weakly supervised learning for scene text detection. Example results of the proposed WeText framework are given in Figure 1.

The contributions of our work are twofold. First, we propose a weakly supervised framework that trains a robust and accurate scene text detector by using unannotated or weakly annotated data. The proposed framework aims to address the data annotation constraint faced by many deep learning systems. In particular, it exploits word-level annotations to guide the search for character-level training samples, and the benefits are demonstrated by the great performance gains in scene text detection. Second, we design a proposal-free scene character detector which directly predicts character bounding boxes and text confidence without the complicated candidate detection and classification processes. The integrated detection approach solves the error accumulation issue and greatly improves the accuracy and efficiency for scene text detection. Experiments show that our proposed weakly supervised model can achieve state-of-the-art performance using only 229 images with character annotations.

2. Related Work

Most existing text detection methods can be broadly classified into two categories, namely, character detection based and word detection based. The character detection based methods usually first detect multiple character candidates using various techniques, including sliding windows [3, 13, 29], MSERs [9, 11, 21, 22, 35, 37], as well as some sophistically designed stroke detector [4, 10, 33, 36]. The detected character candidates are filtered by a text/non-text classifier to remove false candidates. Finally, the identified characters are grouped into words/text lines by either heuristic rules [11, 35, 37] or sophisticated clustering/grouping models [23, 29]. Though the initial character candidate detection can achieve very high recall, the current approach involving multiple sequential steps accumulates error which often degrades the final performance greatly. In particular, the intermediate text/non-text classification step requires a large amount of annotated character images which are very time consuming and costly to prepare.

The methods in the second category instead detect words directly [7, 8, 12, 17, 25, 30, 38]. In [12], object region proposals are employed to first detect multiple word candidates which are then filtered by a random forest classifier and the word bounding boxes are finally fine-tuned with Fast R-CNN [6]. An Inception-RPN word proposal network [38] is proposed which employs Faster R-CNN [26] to improve the word proposal accuracy. Gupta *et al*. [7] introduce a Fully-Convolutional Regression Network to jointly achieve text detection and bounding-box regression at multiple image scales. Tian *et al*. [30] propose a Connectionist Text Proposal Network that combines CNN and long short-term memory (LSTM) architecture to detect text lines directly. The most recent TextBoxes approach [17] designs an end-to-end trainable network to output the final word boxes directly, exploiting state-of-the-art (SSD) object detector [18]. Though the word detection approach is simpler, it does not work well with multi-oriented texts due to the constraints on word proposals. In addition, visually defining a word

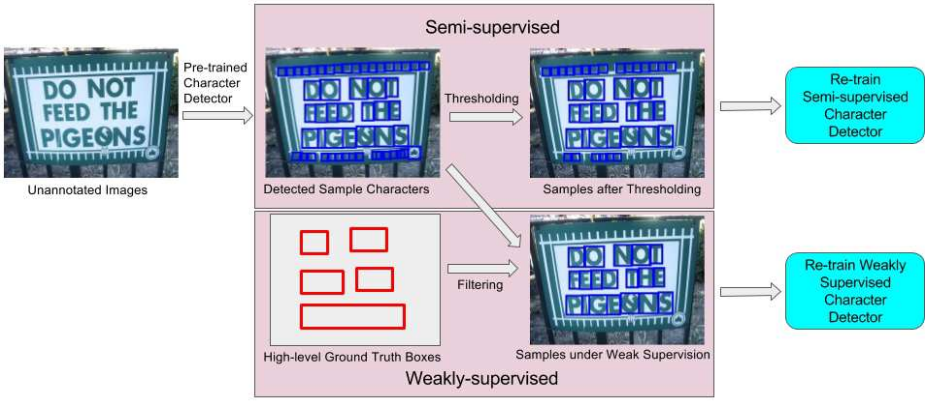

Figure 2: The framework of the proposed WeText system: A "light" supervised model is pre-trained using a small amount of annotated character image set. The light model is then applied to an unannotated dataset to search for more character samples which are combined with the small annotated dataset to train a semi-supervised model. Under certain weak annotations, better character samples can be searched to train a semi-supervised model.

boundary may not be feasible for texts in many non-Latin languages such as Chinese.

Inspired by the idea of weakly supervised learning [14, 24], we propose a weakly supervised scene text detection framework that learns on a small amount of character-level annotated text images, followed by boosting the perfomrance with a much larger amount of weakly annotated images at word/text line level. In the scene text reading domain, similar weakly supervised learning idea has been explored for scene text recognition problem [2, 13]. In [2], a self-supervised training mechanism is designed to augment the training data. In particular, an initial recognition model trained with five million images is applied to search for new training samples where the alignment between images and text is utilized to enhance the quality (based on the assumption that text in real word images also exists verbatim on the web). In [13], similar idea is adopted for automated data mining of Flickr imagery that automatically generates word and character level annotations. The weak correspondence between texts in image titles and texts in scene images is utilized to search for positive training samples.

3. The Proposed Method

3.1. WeText Framework

This section describes the system framework of the proposed weakly supervised scene text detection technique. The system consists of three components including unified scene character detection, semi-supervised and weakly supervised scene text modeling, and graph based text line extraction. The unified scene character detection aims to determine a bounding box together with a confidence score for each character in scene images. The semi-supervised and weakly supervised scene text modeling is achieved by learning from unannotated or weakly annotated scene text images automatically, as illustrated in Figure 2. The graph based text line extraction algorithm [29] is adopted to group characters into text lines.

3.2. Unified Scene Character Detection

We detect characters in scene images by exploiting the recent SSD framework [18] which is designed for generic object detection and has demonstrated superior performance. The adoption of this regression based network aims to address the low efficiency and error accumulation issues of the current scene character detection paradigm where character detection and classification are designed as two separate processes. To the best of our knowledge, this is the first attempt that makes use of the regression based deep networks for scene character detection.

The early layers of the SSD character detector network are based on a standard architecture (VGG-16 [27]) used for image classification. The last two fully-connected layers are converted into convolutional layers with subsampled parameters to speed up the computation. Two auxiliary structures are stacked to the base network to produce character predictions. First, additional convolutional layers are added to the end of the base network, allowing predictions at multiple scales. Unlike Faster R-CNN [26] which uses a single feature layer for prediction, SSD selects multiple feature layers including layers in base network and those additional stacked ones. Second, predictions are computed by applying a set of $3 * 3$ filters to each of the selected feature layers. At each location in the feature layer, we need to predict 6 values for each default anchor, i.e., 4 offsets of the bounding box and 2 scores (text/background).

At the inference stage, Non-Maximum Suppression (NMS) is applied with Jaccard overlap of 0.45 to reduce the detection boxes. As characters tend to appear in groups, the average number of characters in each image is much

Figure 3: Character detection by the baseline model trained using the ICDAR 2013 training images. Thicker bounding boxes indicate higher detection confidence.

larger than general objects in scenes. We therefore keep the top 1000 candidates before NMS and the top 500 detections after NMS per images instead of top 400 and top 200 as used in [18]. Examples of the proposed character detection model are given in Figure 3 (the thickness of the box boundary lines indicates the detection confidence).

3.3. WeText Learning

We investigate two learning strategies to deal with the limited annotation issue which widely exists in many other deep learning systems for object detection/recognition tasks. The first is semi-supervised learning that aims to exploit a large amount of completely unannotated text images. The second is weakly supervised learning where the text images are annotated at the word/text line (instead of character) level. Under both data scenarios, we assume that we have a "light" supervised scene character detection model that is pre-trained by using a small amount of fully annotated scene text images. More details are to be described in the ensuing two subsections.

3.3.1 Semi-Supervised Learning

The scenario of the semi-supervised learning here aims to improve a detection model by learning from a large amount of unannotated data R. Specifically, we have a scene text detection model M that is pre-trained by using a small amount of fully annotated scene character images D, and a large amount of scene text images R that completely has no annotations. The target is to improve M by learning from R with as less manual intervention as possible. It is actually a generic deep learning problem while facing various unannotated "Big Data".

In the WeText system as illustrated in Figure 2, we first run the pre-trained model M on the unannotated dataset R. For each image in R, the model M returns a set of candidate character bounding boxes as well as the corresponding detection score $C = \{(c_1, s_1), (c_2, s_2)..., (c_i, s_i), ...\}$. The positive character samples can be identified by a confidence

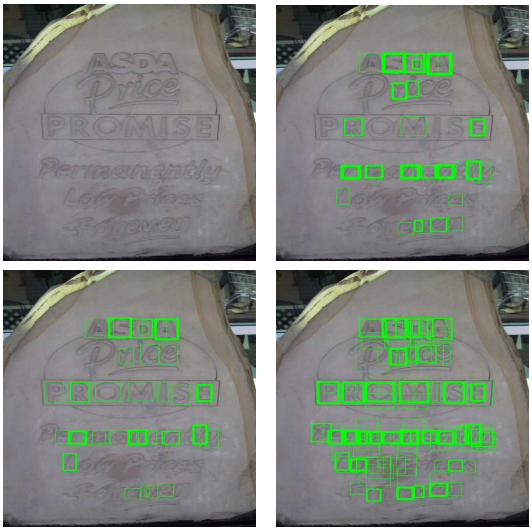

Figure 4: Comparison of different character detectors. Images in the top row from left to right are the input image and output of the baseline detector. Images in the bottom row from left to right are outputs of "COCO-Text_Semi" and "COCO-Text_Weakly" detectors, respectively. The thickness of the box boundary lines indicates the detection confidence.

threshold:

$$P = \{\, c_i \mid s_i > S \;\; and \;\; c_i \in C \,\} \tag{1}$$

where s_i denotes the detection score of the i-th detected character candidate c_i. The notation S is the detection confidence threshold that is used to identify the positive samples. Note that S cannot be too large otherwise the identified sample images lose diversity. At the same time, S cannot be too small otherwise a large amount of non-text samples will be returned. Our experiments show that scene characters can be well searched when S is set to around $[\,0.4,\, 0.6\,]$.

The finally identified positive character sample set P can then be combined with the annotated image set D to train a more robust and accurate scene character detector M'. The top right and bottom left images in Figure 4 show the scene characters that are detected by M and M', respectively. It can be seen that the semi-supervised model M' clearly outperforms the initial model M.

3.3.2 Weakly Supervised Learning

The weakly supervised learning in WeText aims to improve a scene character detection model by learning from large amounts of weakly annotated text images. Different from the semi-supervised learning as described in the last subsection, we have a large dataset R' that has weak annotations at word/text line level as denoted by a set of word/text line bounding boxes $G = \{g_1, g_2, ..., g_j, ...\}$. The target is

to improve M by learning from R' with as less manual intervention as possible. Compared with the semi-supervised learning, the weakly supervised learning has high-level annotations of words/text lines which provide very useful guidance while searching for scene characters in R'.

Similar to the semi-supervised learning as described in the last subsection, the pre-trained model M is first applied to the weakly annotated dataset R', and a candidate character set C is accordingly detected for each image within R'. With the weak annotation G at word/text line level, the positive character sample images are determined as follows:

$$\begin{aligned} P' = \{\, c_i \mid s_i > S' \;\; and \;\; c_i \in C \\ and \;\; I_{x_i} \,/\, W_{c_i} > T_x \\ and \;\; I_{y_i} \,/\, H_{c_i} > T_y \,\} \end{aligned} \tag{2}$$

where W_{c_i} and H_{c_i} denote the width and height of the detected character candidate c_i, I_{x_i} and I_{y_i} denote the maximum horizontal and vertical intersection between c_i and all ground truth bounding boxes in G. S' is a predefined confidence threshold to select positive candidates. It can be set at a much lower value between $[0.2, 0.3]$ due to the constraint provided by the high-level annotations. T_x and T_y are both set at 0.8, based on the observation that a detected character candidate box with more than 80% overlap with the ground truth word/text line boxes are usually texts.

The identified positive sample image set P' can then be combined with the annotated image set D to train a more robust and accurate scene character detector M''. The bottom right image in Figure 4 shows the scene characters detected by M''. It can be seen that the weakly supervised detector M'' outperforms both the initial detector M and the semi-supervised M' clearly.

The better performance of the weakly supervised learning can be explained by two factors. First, more falsely detected character candidates can be removed by leveraging on the annotation bounding boxes at the word/text line level. Second, a lower text confidence threshold S' can be set with the guidance of word/text line bounding boxes which helps to detect more positive character samples greatly. Therefore, the weakly supervised learning can search and retrieve more positive samples of higher quality as compared with the semi-supervised learning.

4. Experiments

Our experiments involve four datasets including the ICDAR 2013 dataset [16], the FORU dataset [38], the COCO-Text dataset [32] and the SWT dataset [4].

4.1. Datasets

ICDAR 2013 [1] consists 229 training image and 233 testing images. Each image also has a segmentation map which helps to extract character boxes. In the experiments, the 229 training images with character-level boxes are used to pre-train a baseline character detector, and the 233 testing images are used for evaluation following the protocol in [16].

FORU [2] is collected from the Flickr website. In our experiments, we use the English2k sub-dataset with 1162 images and 14888 annotated characters. Both character-level and word-level bounding boxes will be used to evaluate the proposed weakly supervised learning.

COCO-Text [3] is derived from the MS COCO dataset with "incidental" texts. In our experiments, we use the training images with at least one legible English text region which leads to 14712 images. Note only word-level bounding boxes are annotated on this dataset.

SWT is introduced in [4] which contains 307 images with word bounding boxes for testing. The dataset is very challenging with cluttered scene images under low contrast and it also contains many small text regions. For evaluation, we use the protocols provided by the dataset creators.

4.2. Implementation Details

Similar to the original SSD [18], we also fine tune from the pre-trained VGG-16 network [27] with initial learning rate 10^{-3}, momentum 0.9, weight decay $5 * 10^{-4}$ and batch size 32 for all the experiments. In addition, all character detection models are trained with input of image scale $512 * 512$ and tested at a **single** image scale $600 * 600$. Further, the parameters in Equation 1 and 2 are empirically set at $S = 0.5$, and $S' = 0.2$ for all experiments. The text confidence threshold for the weakly supervised learning is much lower than that for the semi-supervised learning because word-level bounding boxes in the weakly supervised learning helps to better remove false positives and retrieve true positive samples with lower scores.

The initial "light" character detection model is trained using character annotations within the 229 training images in the ICDAR 2013 dataset. 15k learning iterations are set and the learning rate is reduced to 10^{-4} after 10k iterations. This model will serve as the **Baseline** for both character detection and text line detection as shown in Tables 1. Experiments on the FORU dataset and the COCO-Text dataset target to improve this Baseline model by deriving more positive training samples from the two datasets.

FORU We study three settings on this dataset including **1)** Fully supervised learning where the ground truth character bounding boxes of this dataset are directly combined with the ICDAR training character images to train a better model. This experiment sets an upper bound for the usage of the FORU dataset and experimental result will be used to verify the effectiveness of the semi-supervised and weakly supervised learning; **2)** Semi-supervised learning where no

[1] http://rrc.cvc.uab.es/?ch=2&com=downloads

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

[3] http://vision.cornell.edu/se3/coco-text/

annotation information is used and positive samples are obtained as described in Section 3.3.1; and **3)** Weakly supervised learning where ground truth word bounding boxes are used to guide the sample image searching as described in Section 3.3.2. For all the three settings, we fine-tune from the initial character detector for 3k iterations with learning rate 10^{-3} which is reduced 10^{-4} for another 1k iterations and further reduced to 10^{-5} for the last 1k iterations.

COCO-Text We only study the semi-supervised and weakly supervised settings for this dataset as it does not have character-level bounding boxes. Similar to the FORU dataset, we fine-tune from the character detector trained on the ICDAR 2013 for 10k iterations with learning rate 10^{-3} which is reduced 10^{-4} for another 3k iterations and further reduced to 10^{-5} for the last 2k iterations.

For each setting on the FORU and COCO-Text datasets, the derived positive training samples are combined with the initial ICDAR 2013 training images to re-train a character detection model. Hence we have another five character detection models including **FORU_GT**, **FORU_Weakly**, **FORU_Semi**, **COCO-Text_Semi**, and **COCO-Text_Weakly** as listed in Table 1. Leveraging on the six character detection models (five newly trained plus the Baseline model), we have six corresponding text line detection models after incorporating text line extraction process, including **Baseline_TL**, **FORU_GT_TL**, **FORU_Semi_TL**, **FORU_Weakly_TL**, **COCO-Text_Semi_TL**, and **COCO-Text_Weakly_TL** as listed in Table 2.

4.3. Experimental Results

We evaluate the WeText framework on the ICDAR 2013 testing dataset and the SWT dataset.

4.3.1 Character Detection

We first show the character detection performance on the ICDAR 2013 test dataset, to validate the effectiveness of the proposed semi-supervised and weakly supervised learning from COCO-Text dataset. The PASCAL VOC [5] intersection-over-union (IoU) overlap is used as the evaluation metric (positive detection if IoU≥ 0.5). The recall-precision curve in Figure 5 shows that the semi-supervised model performs clearly better than the baseline model. In addition, the weakly supervised model is superior to both the baseline model and the semi-supervised model. The remarkable performance is largely due to the high precision where the word/text line level ground truth boxes help to filter out lots of false positive samples.

Table 1 shows the precision, recall, and F-score of all character detection models described in the previous subsection, where a confidence threshold 0.05 is used for all detected character candidates (on the ICDAR 2013 testing images). The confidence threshold 0.05 is used for the optimal text line extraction to be described in the next subsection. As Table 1 shows, both semi-supervised and weakly supervised models obviously surpass the baseline model. At the same time, the weakly supervised model clearly outperforms the semi-supervised model due to the availability of the high-level annotations.

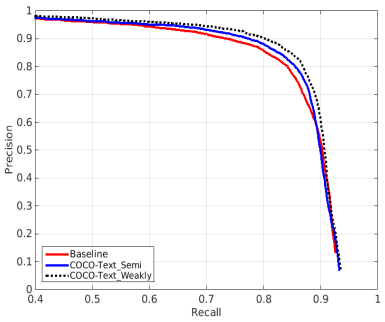

Figure 5: Comparison of character detection performance on ICDAR 2013 test dataset under different learning schemes from COCO-Text dataset. Baseline detector is trained only on ICDAR 2013 training dataset with character boxes. "COCO-Text_Semi" and "COCO-Text_Weakly" detectors are trained without annotation and with text block bounding boxes as described in Section 3.3.1 and Section 3.3.2, respectively.

Table 1: **Character detection** results on ICDAR 2013 dataset (%)

Method	Recall	Precision	F-score
Baseline	84.80	61.44	71.26
FORU_Semi	85.71	63.65	73.05
FORU_Weakly	85.18	67.59	75.37
FORU_GT	85.37	71.83	78.02
COCO-Text_Semi	85.35	66.74	74.91
COCO-Text_Weakly	85.45	72.39	78.38

4.3.2 Text Line Extraction

The detected characters are grouped into text lines using the TextFlow algorithm [29], where we use all detected character candidates that have a detection confidence larger than 0.05. The use of a much smaller confidence threshold (as compared with the S and S' that are used for semi-supervised and weakly supervised training) is because the min-cost flow based text line extraction helps to remove lots of false positive character candidates.

Quantitative Results As shown in Table 2, we achieve state-of-the-art performance on the ICDAR 2013 dataset through the proposed weakly supervised learning strategy. As all our experiments are run at a single scale image, our method outperforms the method in [17] significantly by 6% F-score (86.9% vs 81.0%) when the method in [17] also uses a single scale image as input. In fact, our method still

Table 2: **Text Line detection** results on ICDAR 2013 dataset (%)

Method	Year	Recall	Precision	F-score
Lu *et al*. [19]	2015	69.6	89.2	78.2
Tian *et al*. [29]	2015	75.9	85.2	80.3
Liao *et al*. [17] (single scale)	2017	74.0	88.0	81.0
Zhang *et al*. [37]	2016	78.0	88.0	83.0
Gupta *et al*. [7]	2016	75.5	**92.0**	83.0
Liao *et al*. [17] (multi-scale)	2017	83.0	89.0	86.0
He *et al*. [8]	2016	83.0	90.0	86.0
Baseline_TL	-	80.7	84.2	82.3
FORU_Semi_TL	-	82.0	84.7	83.4
FORU_Weakly_TL	-	82.4	88.6	85.4
FORU_GT_TL	-	82.2	90.9	86.3
COCO-Text_Semi_TL	-	81.8	86.9	84.2
COCO-Text_Weakly_TL	-	**83.1**	91.1	**86.9**

perform better by 1% than [17] where multi-scale testing are adopted. Besides, our baseline model even outperforms the model in Tian et al. [29]. This verifies that the proposed character detector is much more accurate and robust considering the two methods both used similar min-cost flow based text line extraction algorithm.

In addition, models by all three learning schemes, i.e. semi-supervised, weakly supervised, and fully supervised, perform better than the baseline model. In particular, the semi-supervised model improves more than 1% and the fully supervised model achieves the best improvement by 4%. The performance of the weakly supervised model is close to that of the fully supervised model, demonstrating the effectiveness of the proposed weakly learning scheme.

Furthermore, the text line extraction performance is further improved to 84.2% and 86.9%, respectively, for semi-supervised and weakly supervised models when the COCO-Text dataset is used. Similar to the FORU dataset, the "COCO-Text_Weakly_TL" performs better than the "COCO-Text_Semi_TL" for both recall and precision. This verifies that weakly labeled data can effectively helps to remove falsely detected samples and retrieve more difficult positive samples. Additionally, both "COCO-Text_Semi_TL" and "COCO-Text_Weakly_TL" outperform the "FORU_Semi_TL" and "FORU_Weakly_TL", respectively, demonstrating that a larger unannotated or weakly annotated dataset helps to train better semi-supervised and weakly supervised models.

To further verify the proposed framework, we also report results on the SWT dataset [4] in Table 3. All the settings are kept the same as those on ICDAR 2013 dataset except that the test scale is set to $800 * 800$. It can be seen that similar improvement is achieved as on the SWT dataset. The proposed method surpasses the baseline clearly and the learning from a bigger COCO-Text dataset outperforms the learning from a smaller FORU dataset.

Qualitative Results Figure 6 shows text line extraction of several ICDAR 2013 test images that are processed by using the Baseline model, the "FORU_Weakly_LT" model, and the "COCO-Text_Weakly_TL" model, respectively. As Figure 6 shows, the scene text detection performance is clearly improved when more training samples are incorporated in the weakly supervised models. In particular, the recall of the first two sample images is greatly improved. False alarms are successfully removed in the third and forth images. In addition, the "COCO-Text_Weakly_LT" detects one more small word than the "FORU_Weakly_LT", demonstrating the advantage of learning from a much larger dataset. Overall, the proposed weakly supervised learning helps not only detect more positive texts but also remove more false alarms. On the other hand, it could still fail while handling handwriting texts, ultra-low contrast texts, etc. largely due to the limited amount of unannotated or weakly annotated text images. Some of the miss detections are marked by red bounding boxes in Figure 6.

Table 3: **Text Line detection** results on SWT dataset (%)

Method	Recall	Precision	F-score
Epshtein *et al*. [4]	42.0	54.0	47.0
Mao *et al*. [20]	**58.0**	41.0	48.0
Zhang *et al*. [36]	53.0	68.0	**60.0**
Baseline	44.2	69.0	53.9
FORU_Semi	47.5	68.8	56.2
FORU_Weakly	49.3	67.9	57.1
FORU_GT	48.2	**75.7**	58.9
COCO-Text_Semi	48.7	72.9	58.4
COCO-Text_Weakly	49.7	74.9	59.8

4.4. Discussion

We also perform some preliminary study on iterative implementation of the proposed semi-supervised and weakly supervised learning schemes as described in Section 3. Specifically, we repeat the positive sample searching and model re-training process by re-applying the newly trained character detection models back to the unannotated and weakly annotated dataset to search for more sample images for further model re-training. We evaluate the iterative learning idea on the FORU dataset. In the second round, the performance of the newly trained models "FORU_Semi_TL" and "FORU_Weakly_TL" improves from 83.4% to 84.3% and 85.4% to 86.2%, respectively, as compared with the re-trained models after the first round semi-/weakly supervised learning. In particular, the weakly supervised model "FORU_Weakly_TL" after the second round performs nearly as good as the fully supervised model "FORU_GT_LT". We also tested the models after the third round iterative learning but little further improvement is observed. It is probably due to the very close performance to the fully supervised model and further improvements could be achieved when more unannotated or weakly annotated data become available.

Figure 6: Comparison of text detection approach. Images from top to bottom are the text extraction outputs of the "Baseline", "FORU_Weakly" and "COCO-Text_Weakly" character detectors, respectively. Green boxes are outputs of our methods and red boxes are missing detections.

The proposed technique is also fast. For the ICDAR 2013 test dataset, the proposed character detection model takes 0.19s per image and the text line extraction takes about 0.13s per image on Titan X GPU. The total processing time is about 0.32s on average which shows very good potential for various real-time scene text reading tasks.

5. Conclusion

In this paper, we propose a novel weakly supervised learning technique that aims to address the data annotation constraints which exist widely in most deep learning systems. Leveraging on a "light" supervised model that is trained using a small amount of fully annotated images, two learning schemes, namely, semi-supervised learning and weakly supervised learning, are investigated by learning from a large amount of unannotated and weakly annotated images. The proposed technique is evaluated on two publicly available scene text datasets and experiments show that both semi-supervised and weakly supervised models outperform the "light" supervised model clearly. In addition, the weakly supervised model performs almost as well as the fully supervised model.

References

[1] Looktel. http://www.looktel.com/.

[2] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven. Photoocr: Reading text in uncontrolled conditions. In *International Conference on Computer Vision (ICCV)*, pages 785–792, 2013.

[3] X. Chen and A. Yuille. Detecting and reading text in natural scenes. In *Computer Vision and Pattern Recognition (CVPR)*, pages 366–373, 2004.

[4] B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2963–2970, 2010.

[5] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[6] R. Girshick. Fast R-CNN. In *International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[7] A. Gupta, A. Vedaldi, and A. Zisserman. Synthetic data for text localisation in natural images. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2315–2324, 2016.

[8] T. He, W. Huang, Y. Qiao, and J. Yao. Accurate text localization in natural image with cascaded convolutional text network. *arXiv preprint arXiv:1603.09423*, 2016.

[9] T. He, W. Huang, Y. Qiao, and J. Yao. Text-attentional convolutional neural network for scene text detection. *IEEE Transactions on Image Processing*, 25(6):2529–2541, 2016.

[10] W. Huang, Z. Lin, J. Yang, and J. Wang. Text localization in natural images using stroke feature transform and text covariance descriptors. In *International Conference on Computer Vision (ICCV)*, pages 1241–1248, 2013.

[11] W. Huang, Y. Qiao, and X. Tang. Robust scene text detection with convolution neural network induced mser trees. In *European Conference on Computer Vision (ECCV)*, pages 497–511. 2014.

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

[13] M. Jaderberg, A. Vedaldi, and A. Zisserman. Deep features for text spotting. In *European conference on computer vision (ECCV)*, pages 512–528, 2014.

[14] A. Joulin, L. van der Maaten, A. Jabri, and N. Vasilache. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision (ECCV)*, pages 67–84, 2016.

[15] S. Karaoglu, J. C. Van Gemert, and T. Gevers. Object reading: text recognition for object recognition. In *European Conference on Computer Vision (ECCV)*, pages 456–465, 2012.

[16] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, L. P. de las Heras, et al. ICDAR 2013 robust reading competition. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 1484–1493, 2013.

[17] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu. Textboxes: A fast text detector with a single deep neural network. *arXiv preprint arXiv:1611.06779*, 2016.

[18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *European Conference on Computer Vision (ICCV)*, pages 21–37, 2016.

[19] S. Lu, T. Chen, S. Tian, J.-H. Lim, and C.-L. Tan. Scene text extraction based on edges and support vector regression. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 1–11, 2015.

[20] J. Mao, H. Li, W. Zhou, S. Yan, and Q. Tian. Scale based region growing for scene text detection. In *ACM International Conference on Multimedia*, pages 1007–1016, 2013.

[21] L. Neumann and J. Matas. Text localization in real-world images using efficiently pruned exhaustive search. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 687–691, 2011.

[22] L. Neumann and J. Matas. Real-time scene text localization and recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3538–3545, 2012.

[23] Y.-F. Pan, X. Hou, and C.-L. Liu. A hybrid approach to detect and localize texts in natural scene images. *IEEE Transactions on Image Processing*, 20(3):800–813, 2011.

[24] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *International Conference on Computer Vision (ICCV)*, pages 1742–1750, 2015.

[25] A. Polzounov, A. Ablavatski, S. Escalera, S. Lu, and J. Cai. Wordfence: Text detection in natural images with border awareness. *arXiv preprint arXiv:1705.05483*, 2017.

[26] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[28] R. Smith, C. Gu, D.-S. Lee, H. Hu, R. Unnikrishnan, J. Ibarz, S. Arnoud, and S. Lin. End-to-end interpretation of the french street name signs dataset. In *European Conference on Computer Vision (ECCV)*, pages 411–426, 2016.

[29] S. Tian, Y. Pan, C. Huang, S. Lu, K. Yu, and C. Lim Tan. Text flow: A unified text detection system in natural scene images. In *International Conference on Computer Vision (ICCV)*, pages 4651–4659, 2015.

[30] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao. Detecting text in natural image with connectionist text proposal network. In *European Conference on Computer Vision (ECCV)*, pages 56–72, 2016.

[31] L. Ulanoff. Hands on with google translate: A mix of awesome and ok. In *Mashable*, 2015.

[32] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie. COCO-Text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016.

[33] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu. Detecting texts of arbitrary orientations in natural images. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1083–1090, 2012.

[34] X.-C. Yin, W.-Y. Pei, J. Zhang, and H.-W. Hao. Multi-orientation scene text detection with adaptive clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1930–1937, 2015.

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

[36] Z. Zhang, W. Shen, C. Yao, and X. Bai. Symmetry-based text line detection in natural scenes. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2558–2567, 2015.

[37] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai. Multi-oriented text detection with fully convolutional networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4159–4167, 2016.

[38] Z. Zhong, L. Jin, S. Zhang, and Z. Feng. Deeptext: A unified framework for text proposal generation and text detection in natural images. *arXiv preprint arXiv:1605.07314*, 2016.