Proceedings of the 2nd International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/19/20230532

Multiple improvement strategies for high-density text detection based on DBNet

Jingyang Men[1]

[1] Department of Mathematics, New York University, 251 Mercer Street, New York, NY 10012-1110, United States

jm7828@nyu.edu

Abstract. High-Density text detection is actually challenging task, which deserves more attention on its performance improvement. Based on a typical existing text detection model called DBNet, this paper introduces a new approach in improving the performance of it in text-rich and readability interfered scenarios, which are mostly characterized by dozens of lines of text using different font sizes and text background harder for detecting and reading, the most significant scenario is text detection on the passport information page. The approach proposed in this study includes several steps, namely a baseline setting, a method of dataset enrichment and new loss function aiming at raising text detection for short texts and single characters, such methods has achieved enhanced improvement on text detection especially on short texts: an overall precision reaching 0.7854 and recall reaching 0.8758. Compared with the base model, the experimental results demonstrated the effectiveness of the proposed methods in this study.

Keywords: text detection; ID images; text rich materials.

1. Introduction

Text detection has already become a widely applied technology in the modern world in areas such as image understanding, traffic plate detection, image searching and grouping, vehicle auto-driving, etc [1]. The key component and aim of text detection models are to localize a bounding region of text instances in a given image, of which the text instances can be in various fonts, shapes, orientations and scales. Currently, the majority of existing detection models are targeting at low-density text scenarios, such as billboards, notices, receipt and common life scenes shown in Figure 1. These scenarios are designed to convey information to audiences as clearly as possible, hence they are typically written in large fonts, sparse character space and high contrast between background and text. These characteristics make it easier for models to detect text location and boundaries.

DOI: 10.54254/2753-8818/19/20230532

Figure 1. Example of Low-Density Text Scenarios [2].

However, compared to the low-density text scenarios, high-density text scenarios which are more challenging to detect also require a lot of attention. In contrast, what this paper targets at is high-density text scenarios with reduced readabilities, characterized by compact texts (usually over 20 lines of texts in different locations of the page), combined text size and different fonts, complex background (usually with curved lines, repeated pictured or text water-marks), different reflexive layers that make the picture of such material coming with the loss on the information of text boundaries, best examples are info-page of ID or passport and text location in fake-proof labels. In this paper, the targeting material is the passport info page as shown in Figure 2.

Figure 2. High Text Density Scenario [3].

The rest of the paper is organized as follows. Section 2 describes the recent related works in terms of text information detection from the passport. The employed base deep learning model and some detailed procedures of methods are described in Section 3 and Section 4. Section 5 presents the experimental results and provides the corresponding discussion for them. Finally, the conclusion and future work are summarized in Section 6.

DOI: 10.54254/2753-8818/19/20230532

2. Related work

2.1. Template-based methods

The majority of currently existing solutions on recognizing text information from the passport, such as Passport OCR service offered by Huawei Cloud [4], are focusing on Machine Readable Zone (MRZ) code, a two-line region located at the bottom of the passport info page standardized by ICAO [5]. However, an emerging problem is that fake passport only needs to make counterfeit parts of MRZ as the progress of automation during passport checking, thus, algorithms for text detection and recognition from other regions of the page must be developed.

Another solution is offered by Nanonet [6], which obtains information from different passports by using corresponding pre-set templates after recognizing basic information from the MRZ, so that they can retrieve key-value paired information from the info page. The issue with this technique is that applying templates to an image necessitates a complete, distortion-free image, which is not possible when the captured image is skewed or only partially formed.

A third example to this text detection is offered by Google Cloud Service [7], a specific service targeting at dense document text scenarios. However, their definition to dense document is different from what specified by this paper. The example given by Google is a receipt that has many lines of text on it, yet still, the contrast between text and background is obvious, space between lines is also significant to contribute to readability.

2.2. Deep learning-based methods

Another general approach is to use deep learning models to detect text on passport images directly, yet the performance was not satisfactory. The majority are segmentation-based methods, which typically involve using post-processing algorithms on pixel-level predictions on the original image and produce boundaries of text boxes. After collecting passport samples from Council of European Union and labeled corresponding text locations and deployed the model offered in DBNet repository and trained using ICDAR dataset. The model handles the text location task on the passport info page with relatively low precision as shown in Figure 3.

Figure 3. Baseline performance on the test image.

From this baseline, this paper is aiming to solve several problems, as indicated by arrows on the image:

1. MRZ missing from detection. As shown in Figure 3, general DBNet is unable to detect the entire line of MRZ code, especially on non-alphabetic characters.

2. Oversizing boundary box. The passport info page has compact line spaces between different text sizes, that allows a single large text box to include more non-text spaces and overlap other text locations.

3. Missing detection from relatively sparse places. The majority of the text on papers is typically positioned close to the image, with other information scattered over the image's right side. In addition, shorter texts and single characters are generally harder to detect.

DOI: 10.54254/2753-8818/19/20230532

4. Text under reflexive layers. Most passport has various reflexive layer to prove counterfeits, yet they are most distracting and covering information especially when the machine can only take picture from a single direction (where manual processing can change direction of viewing thus eliminating the affection from reflexive layers).

3. Base model: DBNET

In this section, the ground baseline this paper improves upon will be introduced. The base model is the DBNet model published in 2020 [8], which introduced a differentiable binarization upon previous models such as PSENet [9] or SAE [10].

Standard binarization is usually described as follows, where a given threshold t is preset:

$$B_{i,j} = \begin{cases} 1 & if\ P_{i,j} \geq t \\ 0 & otherwise \end{cases} \tag{1}$$

DBNet proposed an approximate step function to make the training process entirely differentiable, the formula is shown below and an illustration of curves of two binarization is shown in Figure 4.

$$\hat{B}_{i,j} = \frac{1}{1+e^{-k(P_{i,j}-T_{i,j})}} \tag{2}$$

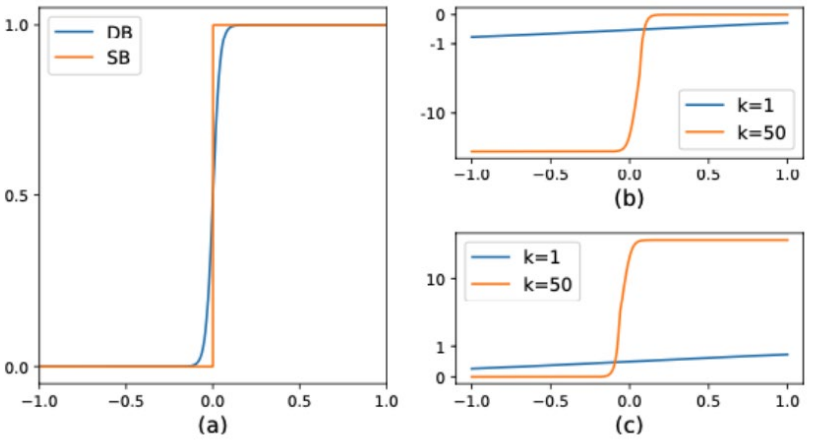

Figure 4. (a) Numerical comparison of Standard Binarization (SB) and differentiable binarization (DB). (b) Derivative of l_+ (c) Derivative of l_-. [8].

4. Methodology

4.1. Labelling Standards

The criteria followed for labeling passport photographs downloaded from the online must first be demonstrated. There are five rules for labeling the images:

- Huge tilted watermarks are not labeled, such has the 'SPECIMEN' on each image, since they will not appear in real situations.
- Hand written signatures are not labeled, as they are usually too abstract to detect.
- Each of the two lines of MRZ are labeled as an entire line
- Any two lines that apart away more than a single line space are labeled separately.
- Smaller and individual letters are labeled in a single box.

After labeling, there are 176 images for testing and 63 images for training. Due to the limited number of training samples (compared with over 1000 images in ICDAR), an expansion of dataset is necessary for training.

DOI: 10.54254/2753-8818/19/20230532

4.2. Expanding Dataset

In order to expand more training samples (32 times in this study), the expansion combines horizontal and vertical flipping, adding Gaussian noise, rgb channel splitting, and merging with various combinations. Figure 5 shows several samples after processing.

Figure 5. (a) Image after adding noise. (b) Image after RGB regrouping.

4.3. Single and Short Text Detection

To capture short and single texts, a new loss function is proposed and applied into the model. In the original DBNet model, the loss is computed combining loss on probability map, threshold map and binarization map, correspondingly named as L_s, L_b, L_t, and the total loss is computed as:

$$L = L_s + \alpha \times L_b + \beta \times L_t \tag{3}$$

where L_s uses Dice Loss with Online Hard Example Mining (OHEM) to overcome an unbalanced ratio between positive regions and negative regions. In the hard mined sample set the ratio of positives and negatives is 1:3. L_b is formulated as a Dice loss based on weighted with width of text boxes. If we let

$$L = \omega_i = \frac{32}{w_i} \tag{4}$$

where w_i is the width of a text box, then the new proposed loss on binarization map is:

$$L_b = \sum_i \omega_i \cdot (1 - \frac{2|X_i \cap Y_i|}{|X_i| + |Y_i|}) \tag{5}$$

where X_i is the corresponding prediction on i^{th} text box and Y_i is the corresponding ground truth of the i th text box. For the loss of threshold, a weighted parameter is also added:

$$L_t = \sum_i \omega_i \cdot (|X_i - Y_i|) \tag{6}$$

DOI: 10.54254/2753-8818/19/20230532

The new Loss function aims at giving shorter texts a higher weight in order to detect them as separate text boxes.

In addition, the original model has a certain ratio of shrinkage on masks in the prediction map, the new model cancelled horizontal mask shrinkage in order to reserve single and short texts, effects of this procedure can be seen in Figure 6.

Figure 6. Masks after cancelling horizontal shrinkage.

5. Experimental results and discussion

In this section, results are shown based on 3 steps, the baseline, which is the original DBNet model after training; the model trained with expanded datasets; and the model with the new loss function trained with expanded datasets. The first results are text labelling using trained models after the second and third steps, as shown in Figure 7. The problems after the second step training are arrowed on the image: 1. large boxes have included texts that should be separated, 2. overlapping of boxes in compact text areas and 3. oversized tilted boxes that cover too much negative space. After step3, single characters are detected into single boxes and most boxes are tight around text.

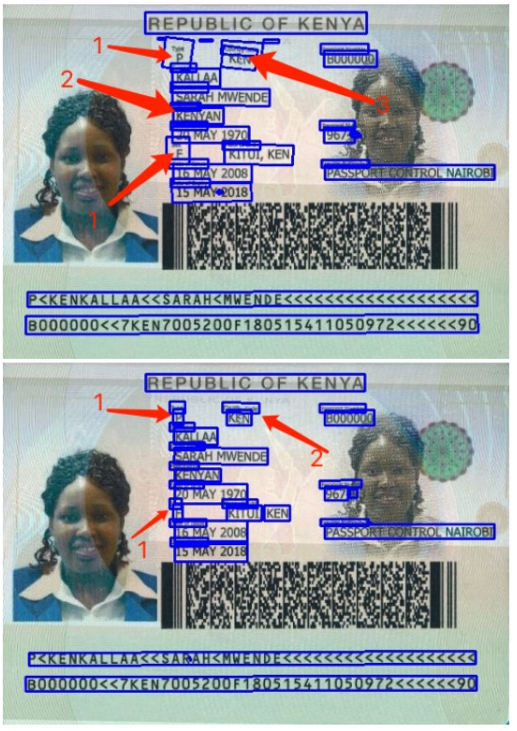

Figure 7. Predictions after second and third step training.

DOI: 10.54254/2753-8818/19/20230532

Table 1. The precision and recall of the model after each step.

	Performance	
	Recall	Precision
Baseline	39.35%	22.86%
Step2	84.53%	70.37%
Step3	87.58%	78.54%

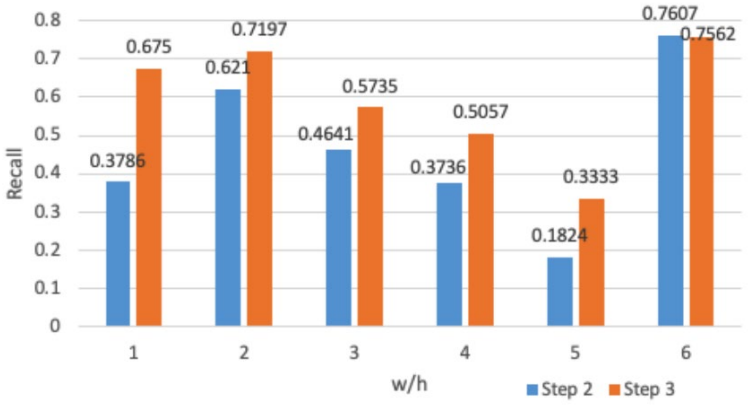

Figure 8. Recalls based on the width-height Ratio.

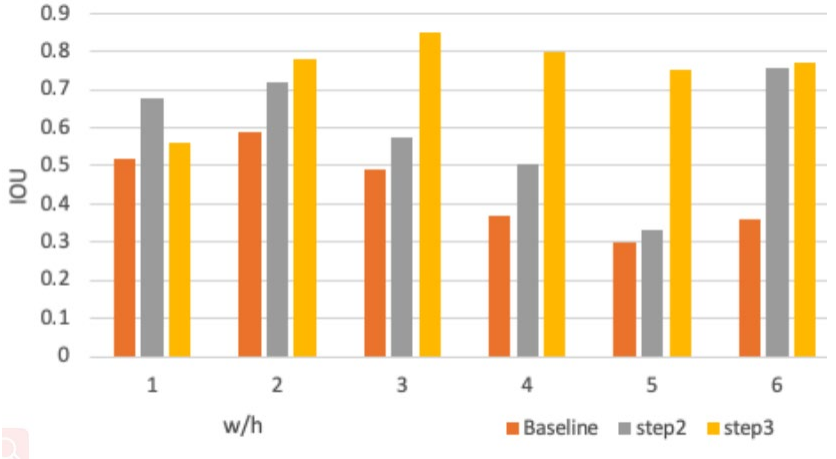

Figure 9. IOU Loss based on the width-height Ratio.

The statistical accuracy is illustrated in Table 1, Figure 8 and Figure 9, where the first one showed an uplifting in both recall and precision in general. In Figure 8 and Figure 9, a more detailed recall and IOU based on width-height ratio is illustrated. The major improvement is focused on short texts: there is a major improvement in w/h ratio between 1 and 5, and a general advancement for w/h ratio under 20, which can further prove the effectiveness of the new loss function.

DOI: 10.54254/2753-8818/19/20230532

6. Conclusion

In this paper, solutions targeting at information dense texts materials are raised and experimented, implementation has shown results of improved effect in both precision and recall. Moreover, shorter texts and single characters which are usually neglected by the base model acquired an enhanced detection recall through applying a new error calculation formula. The improved methods raised in this paper is based on DBNet framework but can be extended to any currently existing general text detection framework. In the future, a comparison based on different base framework should be experimented. In addition, following functions such as text reading, logic linking, and dynamic template application should be added based on the improvements from this paper.

References

[1] Ye Q and Doermann D 2014 Text detection and recognition in imagery: A survey IEEE transactions on pattern analysis and machine intelligence 37(7) 1480-1500

[2] Karatzas D et al. 2015 ICDAR 2015 competition on robust reading 2015 13th international conference on document analysis and recognition (ICDAR) IEEE 2015

[3] C. of European Union. Public register of authentic travel and identity documents online. 2022. https://www.consilium.europa.eu/prado/en/prado-latest-authentic.html

[4] HUAWEI TECHNOLOGIES CO. 2022 Optical character recognition: Api reference. https://support.huaweicloud.com/intl/en-us/api-ocr/api-ocr.pdf

[5] I. C. A. Organization 2021 Machine Readable Travel Documents The Authority of the Secretary General 8 ed., 2021

[6] Kurama V How to easily ocr passport and id cards. 2020. https://nanonets.com/blog/ocr-for-passports-and-id-cards/

[7] G. C. V. Support. Dense document text detection tutorial. 2022. https://cloud.google.com/vision/docs/fulltext-annotations

[8] Minghui L et al. 2022 Real-time scene text detection with differentiable binarization and adaptive scale fusion IEEE Transactions on Pattern Analysis and Machine Intelligence

[9] Wenhai W et al. 2019 Shape robust text detection with progressive scale expansion network Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

[10] Zhuotao T et al. 2019 Learning shape-aware embedding for scene text detection Proceedings of the IEEE/CVF conference on computer vision and pattern recognition