ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Text Detection on Images using Region-based Convolutional Neural Network

Hamsa D. Majeed

Department of Information Technology, University of Human Development, Sulaymaniyah, Iraq

ABSTRACT

In this paper, a new text detection algorithm that accurately locates picture text with complex backgrounds in natural images is applied. The approach is based primarily on the region-based convolutional neural network anchor system, which takes into account the unique features of the text area, compares it to other object detection tasks, and turns the text area detection task into an object sensing task. Thus, the proposed text to be observed directly in the neural network's convolutional characteristic map, and it can simultaneously predict the text/non-text score of the proposal and the coordinates of each proposal in the image. Then, we proposed an algorithm for the construction of the text line, to increase the text detection model accuracy and consistency. We found that our text detection operates accurately, even in multiple language detection functions. We also discovered that it meets the 2012 and 2014 International Conference on Document Analysis and Recognition thresholds of 0.86 F-measure and 0.78 F-measure, which clearly shows the consistency of our model. Our approach has been programmed and implemented using Python programming language 3.8.3 for Windows.

Index Terms: Text Detection, Region-based Convolutional Neural Network, Text Images

1. INTRODUCTION

In recent years, enabling computers to read the text in natural images [1]-[3], [4] have been gaining increased attention. It can be used in optical character detection (optical character recognition), photography, and robot navigation. In this research, the two primary activities that concerned us were reading text and understanding text. Our research is based primarily on text identification in natural images, which is far more complex than text detection on a well-maintained text file.

In the previous text, detection works the majority of the works employs bottom-up pipelines, which often include the following steps: Grouping or filtering of characteristics and the configuration of the line text. Some common issues exist in all these processes. First, the effects of character detection using sliding window methods or related component-based approaches largely depend on their efficiency. Mainly low-level characteristics (e.g., based on stroke width transform [5], maximally stable extremal regions [6], or histogram of oriented gradients [7]) are studied in these approaches.

Without background knowledge, it is difficult to define each stroke or character separately. At the same time, it is easy to result in a low recall where ambiguous characters are easily discarded, causing more difficulties for handling them in the following steps. Second, there are several incremental phases with a bottom-up strategy, which makes the method very complex. These difficulties, therefore, reduce the power and efficiency of the program.

Deep learning technology has greatly increased the efficiency of target detection [1], [2], leading to advances

Access this article online

DOI:10.21928/uhdjst.v4n2y2020.pp40-45

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: Hamsa D. Majeed, Department of Information Technology, University of Human Development, Sulaymaniyah, Iraq. E-mail: hamsa.al-rubaie@uhd.edu.iq

Received: 02-05-2020 Accepted: 27-07-2020 Published: 02-08-2020

in text detection. A variety of recent approaches has been used to create pixel predictions of text or non-text in fully convolutional networks (FCNs). In addition, segmentation of text semantic can lead to greater skill in exploiting rich field context data to identify ambiguous content, which leads to less erroneous detections. Two fully coevolutionary networks fell behind the paradigm to make the findings more robust. The second FCN produces word-level or character-level predictions on text region detected by the first FCN. All these steps just lead to a much more complex method. Many techniques are being used to forecast the limits of text in natural images through sliding windows by coevolutionary features, such as the state-of-the-art region-based convolutional neural network (R-CNN) technique where A Region Proposal Network (RPN) is proposed to generate high-quality object proposals directly from convolutional feature maps. And then, these region proposals are fed into a Faster R-CNN model for further classification. In object detection, each object has a well-defined closed boundary, while it is difficult to find one in-text, which makes it more challenging to predict the text line accurately.

The Region Proposal Network presented in [8] is extending in this paper to localize the text lines accurately. We have put into target detection the issue of text line detection. In the meantime, we use the benefits of profound convolution and computer networking systems. In Fig. 1, the results are shown on our network architecture and text proposal identification.

First, we break the function of text identification into a series of fine text proposals. To forecast the proposal position and text and non-text data together, we refine Faster R-CNNs anchor regression method. This can lead to better localization accuracy.

Second, the text line construction algorithm has been proposed to integrate with the fine-scale proposal into a text line area. The proposed method is to join and single process the multiscale and multilingual text.

Third, with using International Conference on Document Analysis and Recognition (ICDAR) 2012 and ICDAR 2014, our model showed reliable and accurate results with text detection.

2. RELATED WORK

The past text on image detections is primarily utilized bottom-up methods [9]-[11] or sliding window methods [12]-[15] to

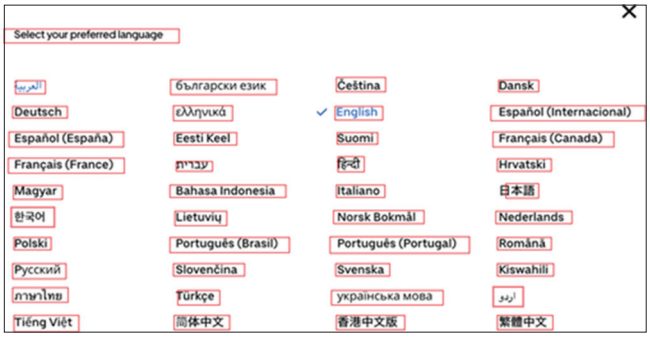

Fig. 1. The output of our model.

detect characters or screen components. The methods used for sliding windows detect text propositions by glancing through an image in a multiscale window and cascading behind the device a classifier to locate text proposals using manually built software or recent CNN features [13], [16], [17]. The methods are based on the related components primarily implement a fast filter to differentiate text from non-text using low-level properties, such as gradient, stroke distance, and color [18].These bottom-up methods do not perform well in character detection and the following steps have produced cumulative false answers. In addition, the bottom-up method is complex and computationally expensive; particularly the sliding windows approach that needs a ranking on several sliding windows).

The efficiency of text detection [1], [2], [4], [5], mostly sliding window system, has recently been advanced by deep learning technologies. They are all have enhanced, primarily using very deep CNN and sharing a coevolutionary mechanism [3], also used to minimize computational costs, to benefit from high-level deep characteristics. Many FCN-based methods were, therefore, provided and findings in text detection tasks were promising. Ren [10] recently introduced a Faster R-CNN object detection method that achieved state-of-the-art object detection tests. They proposed an area proposal network (RPN) that produces highly credible entity proposals directly from the coevolutionary functional maps, fast enough to exchange coevolutionary details. These works are inspiring for our model.

3. METHODOLOGY

There are two modules in our text detection system. In particular, the first module uses a very large, convolutional neural network to create fine-scale proposal regions. The second module is a text line construction that can complete text lines through the text proposal regions given in the

previous module. In the natural picture, the machine will correctly predict the text line. Section A provides us with the concept of faster area CNN and how it fits well in the role of text detection. We propose our text line algorithm in Section B. Fig. 2 shows the block diagram of the proposed method.

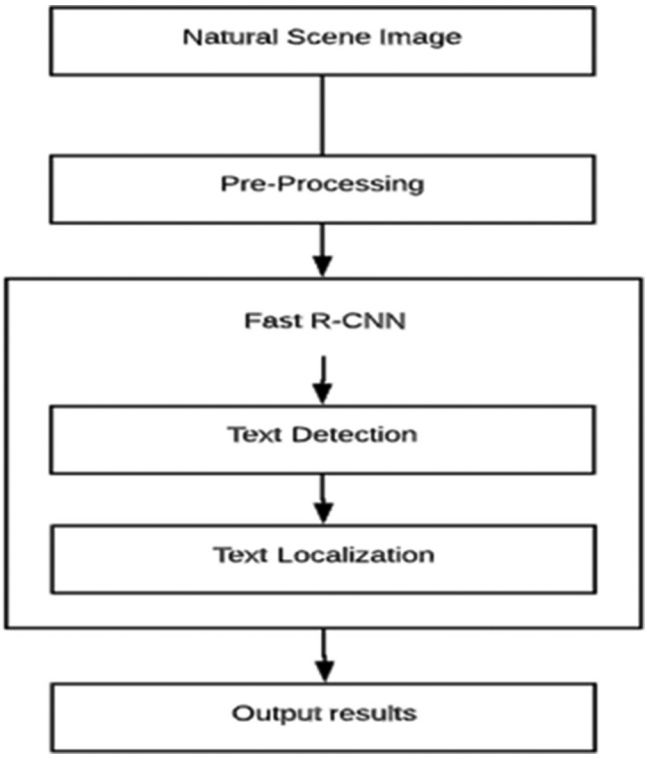

Fig. 2. Method block diagram.

3.1. Fine-scale Proposal Network

Concerning object detection, due to the many cutting-edge tests on object detection benchmark, Faster R-CNN has proved an effective and reliable object detection platform. Its central segment is the area proposal network, which slides a small window in the coevolutionary software, which takes arbitrary formats as input and generates a series of rectangular proposals. Faster R-CNN considers that both the network predicting proposals and classifier networks share a common set of convolutional. RCNN architecture is shown in Fig. 3.

In comparison to general sliding window techniques, (RPN) Fig. 4. applies an effective anchor regression system to identify multiscale items with a single sliding window. This, in turn, reduces to a certain degree the estimated costs for the whole network. The reason that a single-window will forecast mixed objects is primarily that a window maps the multiscaled objects in the original image into multiple anchors with different aspects ratios. We also apply this anchor function in our model in this paper for text detection. The task of text detection is not quite the same as the process of detection of objects. There could be no visible closed boundary in the text field in the picture as can also be used in object detection tasks. It consists of multilevel elements, such as character, text line, stroke, and text area that are not easy to discern.

In all respects, we optimize the anchor function to anticipate components in the text detection process in various stages. We note that a text line can be seen as a series of fine-scale

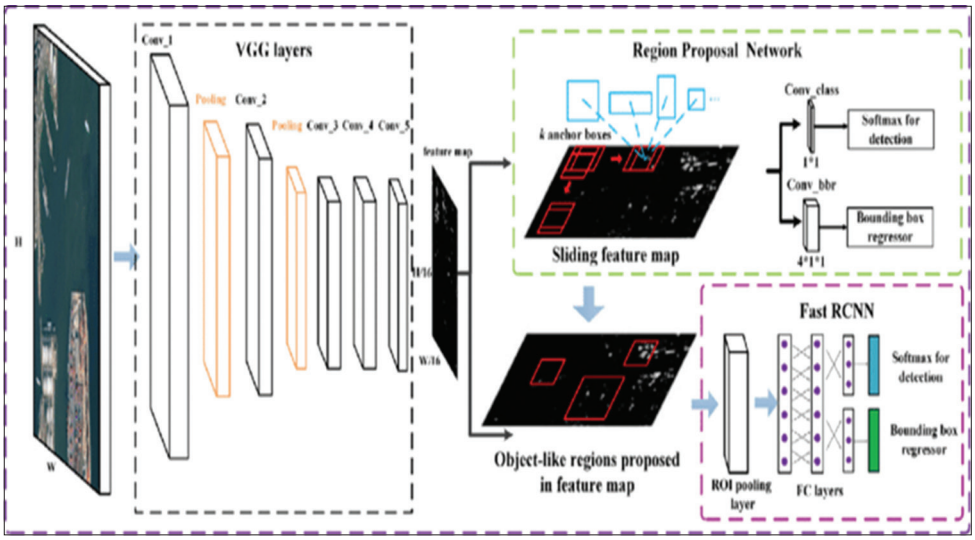

Fig. 3. The architecture of the region-based convolutional neural network.

$$L_{\text{loc}}(t^u, v) = \sum_{i \in \{x,y,w,h\}} \text{smooth}_{L_1}(t_i^u - v_i), \quad (2)$$

in which

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise,} \end{cases} \quad (3)$$

Fig. 4. Loss function of the regressor.

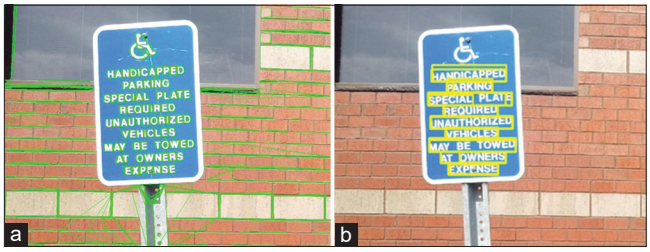

Fig. 5. (a). Results with no non-maximum. (b). Results with non-maximum.

text recommendations that can be used to some degree as an object detection function. We assume that a series of text recommendations from a text line will operate with small scale text detection and different aspect ratios. The anchor function was strengthened as follows: Each text proposal is designated as a 16-pixel fixed-width equivalent to the detector through the conv5 maps, the VGG 16's lasts coevolutionary map that can extract deeper input image features. Moreover, K anchors are established to predict the height of the proposals. (In our experiment, we use K = 10). They are all 16 pixels in width and not the same height to correspond to different scale text regions. Input an image of some size to the VGG 16 model and a Wh HhC functional maps on the conv5 convolution layer (which is the room structure, C is several channels). The transmission detection is as follows: And our model then rolls a 3h3 window through the conv5 function maps, making every window predictable. Moreover, each forecast shall contain K proposals with input picture coordinates and values. All the proposals are compiled and screened, and they do not exceed 0.7 in the next phases.

3.2. Construct the Text Line

Following the development of the text detection network from the Faster R-CNN encoding, a series of fine text proposals is introduced, as shown in Fig. 5a. Our primary goal is to break all these implicit text proposals into various sections of the text field and exclude proposals that do not belong to any text field. It is possible, for some non-text objects with identical composition to text patterns to generate false detection. The non-maximum suppression (NMS) algorithm that was recently widely used in computer vision activities was proposed by Alexander Neubeck and Luc Van Gool. NMS algorithm deletes non-maximum objects that can be called a local maximal search. This is generally used to determine the top scoring range, which normally indicates the tremendous potential for object detection to be the target object. Based on this work, we find the NMS algorithm to be used to eliminate proposals with low scores and create a specific text field for our text detection model. The findings from the results also supported our conjecture. The picture processed after NMS, as shown in Fig. 5b. There are a variety of simple text areas that are composed of a series of adjacent fine text recommendations after each of the above steps are completed. To recreate text lines in input imagery using adjacent text proposals, we suggest an algorithm for text line construction.

The following words are laid down for lines of text:

1. The text line is defined to appear as a quadrilateral in the original image;
2. As the vertical texture region written at the top-to-bottom is not protected in our paper, it is laid down in the text lines that the aspect ratio should not be <2.1;

The next task is to decide where the text region is located. The left and right sides of the quadrilateral text region should be calculated first.

Text line field consists of a series of fine-scale text proposals which are all small, vertical rectangular boxes, the left boundary of the build text line is a left-most boundary of the rectangular text line, and the right border is the right-most rectangular border of all the rectangular boxes that make up the text line. We used the linear regression with the other two parameters to evaluate the commonly used linear regression is often used to predict a sequence of discrete points to evaluate a regression curve that can represent as precisely as possible discrete points.

We use the least squares approach to evaluate a regression line as the top and bottom limits of the text field based on our statement above. Linear regression is used as a tool in the mathematical field specifically to perform curve modification at several discrete points. This allows an understanding of understanding how discrete points are represented as errors as possible.

The text area is completed after the above measure. Moreover, from all text areas built as the true text line blocks

of the original image, we then pick all text areas with a score of at least 0.9.

3.3. Implementation Details

The normal backpropagation and stochastic gradient descent are a qualified method for final text detection. Our samples were obtained from the ICDAR 2014 multilingual text detection competition training results and ICDAR 2012 language localization process competition. We are not larger than 1200 on all training pictures and shorter than 600 on the short side, until the training of the pattern. We have not used the fully ICDAR 2014 training data. We initialize the new fully convolutional layer utilizing the Gaussian distribution of the random 0 means and 0.01 standard variance with the VGG 16 model prepared on ImageNet results. We use 0.9 impulses and weight loss of 0.0005. The analysis limit in 50 K iterations is set to 0.00001. Moreover, our model was implemented in the TensorFlow framework.

4. RESULTS

In the results section, we test our model in both ICDAR 2012 and 2014, with a comparison with the previous researches results in both ICDAR 2012 and 2014 our model had a better result like the following:

4.1. ICDAR 2012 Experiments

The ICDAR 2012 dataset consists of 230 training and 252 sample images in their original. Predominantly, these images are deduced from born-digital images and real-scene photographs. In this article, we check our 2014 version dataset models that were revised in 2014 to optimize the initial version margin. Furthermore, in this case, our model reaches an F-measurement of 0.86, which is a higher than expected result. Our approach has major advantages over the other as Table 1 reveals. Our model supports a more reliable text line identification that enhances efficiency under the ICDAR 2012 and 2014.

The results showed that our proposed method and model had a better recall and precision rate better than previous models.

4.2. ICDAR 2014 Experiments

The ICDAR 2014 dataset has 242 training pictures, and 251 evaluation pictures, mainly of real scene photographs, close to the ICDAR 2012. In Table 2, we equate the performance of our model with other comparative outcomes. The ICDAR 2014 has to be developed for the near-horizontal identification of text, but cannot completely cover the actual text in slanted text images. In this case, our model will perform well, leading to improved results in inclined text detection tasks.

TABLE 1: Experiments results of the ICDAR 2012

Method	Recall	Precision
Yao	0.75	0.73
TextFlow	0.76	0.78
Ching	0.71	0.83
Huang	0.66	0.84
Lai	0.68	0.81
The proposed Method	0.89	0.86

International conference on document analysis and recognition

TABLE 2: Experiments results of the ICDAR 2014

Method	Recall	Precision
Yon	0.61	0.81
FAText	0.73	0.80
SDD	0.66	0.84
Neumen	0.70	0.82
The proposed method	0.79	0.88

International Conference on Document Analysis and Recognition

5. CONCLUSION

In the original picture, we have provided an effective and reliable model for text detection that predicts bounding line boxes. Using the Quicker Region-CNN with our anchor function, we predict a series of fine text proposals. In our experiment, we use an extremely deep network to quantify the image to obtain the deep characteristics of the image to boost the prediction performance. Instead, with the fine-scale texting proposals, we suggested a new structure algorithm for the text rows with the aid of a linear regression method. Such main strategies give the identification of the text line a valuable skill for the identification of text. Both ICDAR 2012 and ICDAR 2014 produced excellent results.

REFERENCES

[1] W. Tao, D. J. Wu, A. Coates and A. Y. Ng. "*End-to-end Text Detection with Convolutional Neural Networks*". Pattern Detection (ICPR), 2012 21st International Conference on IEEE, 2012.

[2] J. Max, A. Vedaldi and A. Zisserman. "Deep features for text spotting". In: *European Conference on Computer Vision*. Springer, Cham, 2014.

[3] N. Lukáš and J. Matas. "*Efficient Scene Text Localization and Detection with Local Character Refinement*". Document Analysis and Detection (ICDAR), 2015 13th International Conference on IEEE, 2015.

[4] M. Rodrigo, N, Thome, M. Cord, J. Fabrizio and B. Marcotegui. "*Snoopertext: A Multiresolution System for Text Detection in*

Complex Visual Scenes". Image Processing (ICIP), 2010 17th IEEE International Conference on IEEE, 2010.

[5] K. Dimosthenis, F. Shafait, S. Uchida, M. Iwamura, L. G. Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan and L. P. de las Heras. "*ICDAR 2013 Robust Reading Competition*". Document Analysis and Detection (ICDAR), 2013 12th International Conference on IEEE, 2013.

[6] H. Weilin, Z. Lin, J. Yang and J. Wang. "*Text Localization in Natural Images Using Stroke Feature Transform and Text Covariance Descriptors*". Computer Vision (ICCV), 2013 IEEE International Conference on IEEE, Sydney, NSW, Australia, 2013.

[7] W. Huang, Y. Qiao, and X. Tang. "*Robust Scene Text Detection with Convolutional Neural Networks Induced MSER Trees*". Vol. 1. European Conference on Computer Vision (ECCV), 2014.

[8] Y. Xu-Cheng, X. Yin, K. Huang and H. W. Hao. "Robust Text Detection in Natural Scene Images". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 970-983, 2014.

[9] E. Boris, E. Ofek and Y. Wexler. "*Detecting Text in Natural Scenes with Stroke width Transform*". Computer Vision and Pattern Detection (CVPR), 2010 IEEE Conference on IEEE, 2010.

[10] T. Zhi, 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*. Springer, Cham, 2016.

[11] T. Shangxuan, Y. Pan, C. Huang, S. Lu, K. Yu and C. L. Tan. "*Text Flow: A Unified Text Detection System in Natural Scene Images*". Proceedings of the IEEE International Conference on Computer Vision, 2015.

[12] Z. Zheng, C. Zhang, W. Shen, C. Yao, W. Liu and X. Bai. "*Multi-oriented Text Detection with Fully Convolutional Networks*". arXiv, 2016.

[13] N. Alexander and L. Van Gool. "*Efficient Non-maximum Suppression*". Vol. 3. Pattern Detection. 18th International Conference on IEEE, 2006.

[14] Y. Cong, X. Bai1, W. Liu, Y. Ma and Z. Tu. "*Detecting Texts of Arbitrary Orientations in Natural Images*". Computer Vision and Pattern Detection (CVPR), 2012 IEEE Conference on IEEE, 2012.

[15] H. Pan, W. Huang, Y. Qiao, C. C. Loy and X. Tang. "*Reading Scene Text in Deep Convolutional Sequences*". Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), 2016.

[16] H. Tong, W. Huang, Y. Qiao and J. Yao. "*Accurate Text Localization in Natural Image with Cascaded Convolutional Text Network*". arXiv, 2016.

[17] L. Minghui, B. Shi, X. Bai, X. Wang and W. Liu. "*TextBoxes: A Fast Text Detector with a Single Deep Neural Network*". Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2017.

[18] R. Shaoqing, K. He, R. Girshick and J. Sun. "Faster R-CNN: Towards real-time object detection with region proposal networks". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 2017.