Engineering and Technology Journal for Research and Innovation (ETJRI)
ISSN 2581-8678, Volume IV, Issue I, Jan 2022

Detection of Objects and Arbitrary Text in Scene based on Deep Learning

Shilpi Goyal[1], Deepak Motwani[2]
[1]*Research Scholar, Amity University Madhya Pradesh, Gwalior, ASET-CSE, India, shilpi.goyal@s.amity.edu*
[2]*Associate Professor, Amity University Madhya Pradesh, Gwalior, ASET-CSE, India, dmotwani@gwa.amity.edu*

***Abstract*— Images are essential part of our life that increases our basic understanding and to detect information from images lead to give semantic information that could be used for further analysis in various fields such as automated recommendation engines, automatic data entry, give answers to question that is framed from scene. Scene text detection is to detect regions of text from scene with complex background. It consists of images having both object and text that include different orientation, fonts. For better interpretation, images needs to be reconstructed over time. Traditionally, optical character recognition technology (OCR) were used, but faces loss of accuracy with different style of text, orientation, scales, complexity of background details like occlusion. With the recent advent of deep learning with computer vision, many convolution layers are applied with recurrent neural network to undergo more deep to get better results. But it requires lots of GPUs and the process is time consuming so in this, we will propose the blend of traditional OCR and advance deep learning methods to improve results that work well with arbitrary (different alignment) shapes of text as well. In this paper, we will localize objects using R-CNN methods and detect scene text using OCR and different deep learning methods by detect, recognize, spot technique and will evaluate using Normalized Edit Distance metric and case-insensitive word accuracy.**

Keywords*— *Convolutional Neural Networks, Deep Learning, Object Localization, Recurrent Neural Network, Scene Text Detection.

I. INTRODUCTION

Computer vision is the field that deals with images. Images whether it is natural or synthetic capture important information also known as scene illustrated in Fig.1., already explored from last several decades where it is seen digitally and working on different aspects to fulfill different needs of society and require more efforts to explore unseen areas that are even more challenging. Digital documents are mostly preferable when it comes to finding, storing, editing information that could be directly used in some applications or further processed to analyze it. Extracting textual information from images has numerous applications. Some applications are automate service on roads like recognition of text from number plate, automated recommendation engines, automatic data entry, guidance device, converting typed text/ handwritten text to digital text, automatic cheque processing, industrial automation, content based image retrieval, to control driverless vehicle, and many more.

Scene Text Detection (STD) is the process of finding the exact location of text if it is present in scene images and Scene Text Recognition (STR) is the process to recognize the text from that image. Traditionally it was done through OCR illustrated in Fig. 2. where the problem is constrained and it provides good results but, as the problem is unconstrained, it leads to more challenging tasks. The challenges that make the problem unconstrained are inclusion of complex backgrounds, noise, lightening, different fonts, orientation and geometrical distortions in the image.

Earlier systems deal with images that has structured format but the scenario has changed. It has turned around 360 degree angle. The images are usually unstructured where text is positioned at random places, no standard format, font, background, no proper row structure. The text having these characteristics termed as wild text. Extensive use of portable mobile devices changed the overall scenario, what we want to record for future use, we just clicked that information where a need arises to excerpt text from image information. At present, scene text detection and recognition has become a significant aspect of both pattern recognition techniques and computer vision. Our motive is to recognize textual content from scene

* Shilpi Goyal- shilpi.goyal@s.amity.edu, Deepak Motwani- vdmotwani@gwa.amity.edu

Engineering and Technology Journal for Research and Innovation (ETJRI)
ISSN 2581-8678, Volume IV, Issue I, Jan 2022

which has vast application. It uses computer vision

Fig. 1. Text in natural scene images

and deep learning algorithms to deal with the challenges of text extraction.

It can be used in high level technologies also. Driverless vehicles can be controlled by enable its signal by providing text that can be extracted from

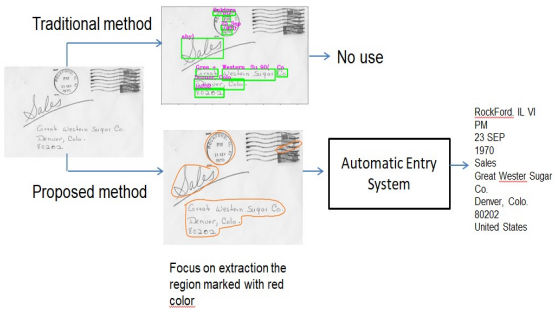

Fig. 2. Difference between traditional and desired method

sign boards (school, hospital). Image occupies more memory than text and more memory leads to more time taking process.

Scene Text Detection (STD) and Scene Text Recognition (STR) are two important areas that are to be focused for scene images. STD involves localization of text. Here, in this paper, we are more focus on text detection.

II. RELATED WORK

Traditionally, OCR methods are used to recognize text from born-digital images where specific fonts and characters are used, but as problem goes unconstrained having many issues to handle except using a pre-determine text. However, the performance of these methods is poor.

With the advent of deep learning based text detection and text recognition methods, the performance is improved with significant amount. Lots of work has been done with the help of Convolutional Neural Network (CNN) [2]. Initial work is based on sliding window method [3], rectangle shaped bounding boxes [4], use of anchor boxes [5], character level segmentation approach [6], word level segmentation approach [6]. Some papers focus first on finding the region of interest and then use different scales to localize the text [7].

Recent papers focus on quadrilateral bounding box [8] as text is arbitrarily oriented. But the problem persists with curve type text. So to handle this, multi-scale shape regression method [9] is used with different scales considering dense boundary box [10]. For highly rotated text, symmetrical axis concept is used by calculating its radius and orientation [11] through fully connected layers. Reference [12] introduced the concept of mask RCNN. The pyramid attention network (PAN) [13] was applied as a backbone for Mask R-CNN to improve the feature representation capacity of R-CNN. The false alarms produced by text-like conditions were effectively suppressed by this PAN. [14] propose segmentation based method and center and border probability. Center direction map is used to segment two close texts by considering close pixels to same text and distant pixels belonging to different text. For this they evaluate text center-border probability (TCBP) and text center-direction (TCD). TCBP can be evaluated where the values are 0 or 1 and values at center affects the values at border and use TCD method to help better learn the probability map. Reference [15] emphasis on detecting text key points and key point links that use control points to exactly determine the location of text and then predict accurately by linking associated land-marks by converting polygon boundary boxes using splines. This method fails if the background contains certain arrangement that is similar to text strokes and it performance also degrades when multiple text lines are extremely close to each other. A centroid-centric vector regression method [16] is proposed by shifting the quadrilateral boundary points to its centroid. It is proposed to eliminate the need of four vectors that are very changed in its length and orientation. An

additional concept, region removal multi-testing approach is used to upturn the performance to lessen the overhead on non-maximal suppression to handle different aspect ratio and size of text. This method can detect the skewed, multi-oriented, and handwritten text.

To detect arbitrary oriented text in the scene image, the main challenges are with the curved text, when the images are cluttered or the shape is in zigzag motion.

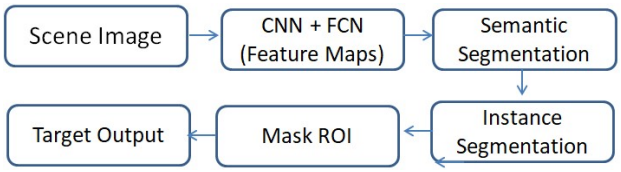

Fig. 3. Proposed pipeline for detection of arbitrary shaped text in scene images.

III. METHODOLOGY

The proposed model pipeline is shown in Fig. 3. It includes convolution neural network and fully connected networks to extract feature maps from scene images. Then apply semantic segmentation to divide the image at pixel level to clearly determine different portions to focus on. After that instance segmentation is applied to focus at each instance of segmented image. Apply mask region of interest to fine tune and apply spline technique to outline the arbitrary shape over text. Finally text is detected from scene. This model works for dense boundary points where texts are very near to each other.

Robust Reading Competitions (RRC) is held every year conducted by International Conference on Document Analysis and Recognition (ICDAR) [17] to highlight different issues of computer vision deal with text images. Number of datasets is provided by RRC, namely ICDAR2015, ICDAR2003, ICDAR2019-ArT, Total-Text [18]. Different benchmark datasets are also available like SVHN, SynthText, SCUT-CTW1500, IIIT-5K word.

The performance of the proposed method will be evaluated on the basis of a Normalized Levenshtein Distance Metric [19]. This calculates how much similarity is there between the detected text and text to be detected.

IV. CONCLUSION AND FUTURE SCOPE

In this paper, we try to propose a novel approach based on segmentation technique to process a scene image to handle arbitrary type of text covering highly curved text instances also. We are very sure that with this approach, we will reduce the processing time as we try to ignore the region proposal method that depends on multi scale anchor boxes and that is very heavy to process. It can be applied on available standard datasets to make comparative study and further used in real time applications like driverless vehicle and automated data entry for industrial use.

REFERENCES

[1] Zhu, Yingying, Cong Yao, and Xiang Bai. "Scene text detection and recognition: Recent advances and future trends." Frontiers of Computer Science 10.1 (2016): 19-36.

[2] Ciresan, Dan Claudiu, et al. "Flexible, high performance convolutional neural networks for image classification." Twenty-second international joint conference on artificial intelligence. 2011.

[3] Yin, Xu-Cheng, et al. "Robust text detection in natural scene images." IEEE transactions on pattern analysis and machine intelligence 36.5 (2013): 970-983.

[4] Cao, Dongping, et al. "Scene Text Detection in Natural Images: A Review." Symmetry 12.12 (2020): 1956.

[5] Arafat, Syed Yasser, and Muhammad Javed Iqbal. "Urdu-text detection and recognition in natural scene images using deep learning." IEEE Access 8 (2020): 96787-96803.

[6] Bluche, Théodore. "Joint line segmentation and transcription for end-to-end handwritten paragraph recognition." arXiv preprint arXiv:1604.08352 (2016).

[7] Long, Shangbang, Xin He, and Cong Yao. "Scene text detection and recognition: The deep learning era." International Journal of Computer Vision 129.1 (2021): 161-184.

[8] Keserwani, Prateek, et al. "Quadbox: Quadrilateral Bounding Box Based Scene Text Detection Using Vector Regression." IEEE Access 9 (2021): 36802-36818.

[9] Xue, Chuhui, Shijian Lu, and Wei Zhang. "Msr: Multi-scale shape regression for scene text detection." arXiv preprint arXiv:1901.02596 (2019).
[10] Hayder, Zeeshan, Xuming He, and Mathieu Salzmann. "Boundary-aware instance segmentation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
[11] Long, Shangbang, et al. "Textsnake: A flexible representation for detecting text of arbitrary shapes." Proceedings of the European conference on computer vision (ECCV). 2018.
[12] He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017.
[13] Huang, Zhida, et al. "Mask R-CNN with pyramid attention network for scene text detection." 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2019.
[14] Zhu, Yixing, and Jun Du. "Textmountain: Accurate scene text detection via instance segmentation." Pattern Recognition 110 (2021): 107336.
[15] Xue, Chuhui, Shijian Lu, and Steven Hoi. "Detection and Rectification of Arbitrary Shaped Scene Texts by using Text Keypoints and Links." arXiv preprint arXiv:2103.00785 (2021).
[16] Keserwani, Prateek, et al. "Quadbox: Quadrilateral Bounding Box Based Scene Text Detection Using Vector Regression." IEEE Access 9 (2021): 36802-36818.
[17] Karatzas, Dimosthenis, et al. "ICDAR 2015 competition on robust reading." 2015 13th International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2015.
[18] Ch'ng, Chee Kheng, and Chee Seng Chan. "Total-text: A comprehensive dataset for scene text detection and recognition." 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Vol. 1. IEEE, 2017.
[19] Marzal, Andres, and Enrique Vidal. "Computation of normalized edit distance and applications." IEEE transactions on pattern analysis and machine intelligence 15.9 (1993): 926-932.