Grenze International Journal of Engineering and Technology, June Issue

CapText-Unet: Efficient Scene Text Detection using UNET

Mahesha M[1], V. N. Manjunath Aradhya[2], Basavaraju H T[3] and Siddesha S[4]
[1]Dept. of Computer Applications, JSS Science and Technology University, Mysuru, India
Email: mahesha_m@sjce.ac.in
[2-4]Dept. of Computer Applications, VVCE, Mysuru, India

***Abstract*—Scene text detection in natural images remains a challenging task due to variations in text appearance, orientation, and background complexity. This paper proposes a novel approach, CapText-Unet, which leverages the U-Net architecture to detect text regions in scene images. Our method adapts U-Net to predict heatmaps representing text regions, using bounding box annotations to generate pseudo-segmentation labels. The proposed approach achieves state-of-the-art results on the MSRA-TD500 dataset, demonstrating its effectiveness in detecting text regions with high accuracy. The CapText-Unet (CaptureText-UNet) architecture is simple, efficient, and requires only bounding box annotations, making it a promising solution for scene text detection applications.**

***Keywords*—Scene Text Detection, CNN, UNET, Heatmap.**

I. INTRODUCTION

Text appearance in the outside atmosphere has many significances and challenges, studies on scene text detection and solution to detect the text in scene images has plenty of challenges due various constraints such as low image clarity, dull Color, different orientation, occlusion so on. Deep learning approaches also provided various techniques to solve computer vision and scene text detection problems in particular Gujjeti et al [1]. Feature Pyramid Network (FPN) is used extract the global and low-level features and semantic information discussed in Wu, Y et.al [2]. Many works related to scene text with complex scenarios like detecting text in traffic sign addressed by pioneers such as Greenhalgh and Mirmehdi [3]. The detection of real time and wide scale range text in scene images using scale region proposal network (SRPN) at faster rate is proposed by He et al [4]. Due to advent of internet and mobile is increasing, more and more applications of extraction of text scene images also increasing so it has opened up wide research need in line with computer vision and pattern recognition domain. Various international conferences like International Conference on Document Analysis and Recognition (ICDAR), the International Conference on Computer Vision, Computer Vision and Pattern Recognition, the European Conference on Computer Vision, and the AAAI Conference on Artificial Intelligence, have done remarkable improvement in the field of scene text detection and recognition (STDR), considered as separate specialized section to gather attention of research community. In our proposed algorithm we have used the UNet architecture to extract the textual components from scene images. UNet has a contracting and tiered alternating orientation resolution, the tiered contracts requisite to lower resolution is feature only and not translatable, and unfolds the resolution through feature capturing with no understanding of site. Skip connections are made from the equivalent sections in the encoder to the decoder in Order To enhance segmentation accuracy.

Grenze ID: 01.GIJET.11.2.762

Two segmentation models are used in the system, namely Fully Convolutional Network (FCN) and UNet. Analysis of features based on area other than single pixels has been implemented by applying FCN while line and text segmentation has been executed through analysis of images. UNets constructor is more advanced and presents an encoder-decoder structure with skip connections radially that helps cut out extra garner projectile around and point of what is concerned. UNet generates outputs with high precision using a symmetric encoder-decoder architecture. The encoder compresses the input image to extract features, while the decoder reconstructs the segmented text regions. The model also uses skip connections between corresponding encoder and decoder layers to retain fine details, making it especially effective in segmenting intricate or overlapping text. We have trained the U-Net model using the train_dataset and val_dataset over 20 to 40 epochs and saved the trained model after the training process for further to evaluate the efficacy of the proposed technique. And yielded very good detection and recognition results. Significant feature of our approach is that its yielded better efficacy even with the smaller number of trained samples. We have experimented our CapText-Unet on MSRA-TD500 dataset and achieved the very good results, serves as simple technique to solve scene text detection applications in the natural scene images and videos.

II. Literature Survey

Natural scene text detection has emerged as a pivotal task in computer vision, finding applications in diverse domains such as optical character recognition, visually impaired navigation, and numerous other fields [1]. The continuous advancement of deep learning technology has opened up new avenues for the evolution of natural scene text detection. In recent research, two predominant approaches have been prominent: segmentation-based and regression-based [2]. The core concept of regression-based natural scene text detection methods is to enhance target detection by treating text as the target and regressing to extract its crucial information, including position, size, and orientation [3]. STDR even though focuses locating and detection of text that appears in real time natural photographs, it does extract the critical high level semantic data from scene images [5-7]. As results it has proved its importance and increasing its applications in all industries limited Computer Vision, Knowledge management, photo content management, seen reading and understanding, content augmentation, scene analysis and fraud deterrence. Further, STDR not only extract any particular content but helps to extract from a character to knowledge extraction of full pages from images and videos. This automation of detection and recognition helped in revolutionary applications like autonomous driving systems and driver assisting systems to interpret the sign boards and route direction boards. Robot assisted mass production factory lanes to read boards on palates, lane numbers and product details. Multimedia and language translation assist applications to help on road safety and security systems in automobile sectors which largely depends on technology with passenger safety and total control [8,9]. For example, scene text images of industry applications where detecting text on path lanes production environment helps robot to stacks the palates. Using deep learning technology to address the scene text detection challenges, scene text detection techniques and various methods available, open areas researcher can focus so on discussed in [10]. Number plate recognition in real time images using deep learning and used dictionary search method for text recognition and un attended challenges discussed [11]. Therefore, choosing an appropriate method for detecting and recognizing text successfully is a hard task when several methods or models have been proposed in the literature [5–7]. Rotating Candidate Region Generation Network (RRPN) is introduced to tackle text detection challenges in natural scenes, particularly for multi-directional and irregular text instances that Faster R-CNN [12] struggles to handle. RRPN leverages rotating rectangular candidate boxes to generate tilted text candidate regions. Similarly, extended Faster R-CNN by devising an Adaptive Region Proposal Network (RPN) to produce more accurate text candidate regions[13]. They added an auxiliary branch for detecting text contours within these candidate regions to mitigate false detections. This approach substantially elevates detection accuracy and is adaptable to detecting text with arbitrary shapes. While regression-based methods have yielded commendable results in text detection, they often face challenges in obtaining smooth text contour curves, particularly for curved text. Consequently, researchers have introduced a natural scene text detection approach based on image segmentation. This method begins with pixel-level classification, determining whether each pixel belongs to a text target. It produces a probability map of the text region and subsequently derives the text segmentation region's enclosing curve through post-processing [4]. The Progressive Scale Expansion Network (PSENet) for text segmentation to address the intricacies of distinguishing between neighbouring text instances in segmentation-based text algorithms [14-15]. PSENet leverages multi-level prediction to predict the central region of text at various scales. Subsequently, it employs a progressive scale expansion 3 algorithm to produce the segmentation results of text instances. In response to the complexities of post-processing in PSENet and its relatively low efficiency, the Pixel Aggregation Network (PAN) built upon

PSENet. PAN uses pixel clustering to merge nearby pixels along the predicted text kernel's central region[16]. This approach maintains high accuracy while significantly enhancing prediction speed. Addressing the time-consuming post-processing in segmentation-based methods due to threshold binarization. Various approaches being experimented significantly improves detection accuracy in natural scene text detection [17-18]. More machine learning approaches and simple CNN based algorithms for scene text detection be found in [19-22]. In summary, segmentation-based text detection methods offer notable advantages in flexibility and robustness, and they continue to make strides in enhancing detection efficiency.

III. Proposed Methodology

Here we propose the simple technique for scene text detection using U-Net Architecture. To perform scene text detection without explicit masking, U-Net can be adapted to predict heatmaps representing text regions. Instead of relying on detailed pixel-level masks, bounding box annotations are used to generate pseudo-segmentation labels, where text regions are represented by filled rectangles or Gaussian blobs. These heatmaps highlight text areas, simplifying the task. During training, the U-Net model learns to map input scene images to these heatmaps, enabling the detection of text regions. After training, postprocessing techniques, such as thresholding and connected components analysis, are applied to the predicted heatmaps to extract bounding boxes for the detected text regions. This approach is effective for scene text detection, requiring only bounding box annotations and leveraging U-Net's capabilities for segmentation-like tasks. The U-Net model, when adapted for heatmap-based tasks such as scene text detection, operates by transforming input images into output heatmaps that highlight regions of interest (in this case, text areas). The encoder portion of U-Net extracts hierarchical spatial features from the input image through successive convolutional and pooling layers, capturing both local and global contextual information. These features are then processed in the bottleneck, which acts as a latent representation summarizing the most salient information. The decoder portion of U-Net takes this representation and reconstructs spatial details using transposed convolutions (upsampling) while concatenating corresponding feature maps from the encoder via skip connections. These skip connections ensure that fine-grained details from the early layers are preserved and reintroduced into the decoding process, allowing precise localization of text regions. The final output layer generates a single-channel heatmap where each pixel value represents the likelihood of it belonging to a text region. High-intensity areas in the heatmap correspond to detected text regions, while low-intensity areas represent the background. During postprocessing, thresholds are applied to the heatmap to segment text regions, which can then be refined and converted into bounding boxes or polygons for practical use. This process makes U-Net an effective tool for generating dense, pixel-level predictions, even for tasks like text detection where explicit masks may not be available. UNet has a contracting and tiered alternating orientation resolution, the tiered contracts requisite to lower resolution is feature only and not translatable, and unfolds the resolution through feature capturing with no understanding of site. Skip connections are made from the equivalent sections in the encoder to the decoder in order to enhance segmentation accuracy, Two segmentation models are used in the system, namely Fully Convolutional Network (FCN) and UNet. Analysis of features based on area other than single pixels has been implemented by applying FCN while line and text segmentation has been executed through analysis of images. UNets constructor is more advanced and presents an encoder-decoder structure with skip connections radially that helps cut out extra garner projectile around and point of what is concerned. Over block diagram of proposed CapText-Unet architecture shown in diagram Figure 1.

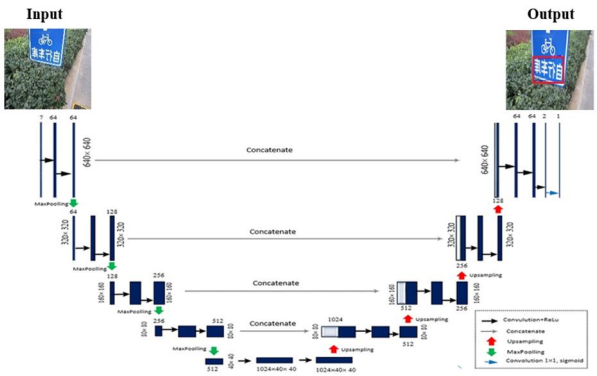

Figure 1: Block diagram of the proposed CapText-UNet Model

IV. Experimental Setup and Discussion

The proposed approach adapts U-Net to predict heatmaps representing text regions, using bounding box annotations to generate pseudo-segmentation labels. The U-Net model is trained to map input scene images to these heatmaps, enabling the detection of text regions.

Experimental Setup uses MSRA Dataset: A dataset of scene images with bounding box annotations for text regions. CapText U-Net architecture with an encoder-decoder structure, skip connections, and a final output layer generating a single-channel heatmap.

The CapText U-Net model is trained using a binary cross-entropy loss function and Adam optimizer. As part of the Postprocessing, the thresholding and connected components analysis are applied to the predicted heatmaps to extract bounding boxes for detected text regions. The experimentation yielded promising results, with the U-Net model achieving high detection accuracy. The model detected text regions with high accuracy, demonstrating its effectiveness in scene text detection. The model showed robustness to variations in text size, orientation, and color, as well as to cluttered backgrounds. Cap-Text-U-Net architecture enabled efficient processing of input images, making it suitable for real-time applications.

A. Experimental Analysis

The experimentation highlights the effectiveness of U-Net architecture in scene text detection. The use of pseudo-segmentation labels and heatmap-based prediction enables the model to detect text regions accurately, without requiring explicit pixel-level masks.

The results demonstrate the robustness of the U-Net model to variations in text and background, making it a reliable choice for scene text detection tasks. This experimentation analysis demonstrates the effectiveness of U-Net architecture in scene text detection.

The proposed approach leverages the strengths of U-Net in segmentation-like tasks, achieving high detection accuracy and robustness to variations. The results highlight the potential of U-Net in real-time scene text detection applications. is a simpler approach than heavy deep learning models such as TextBox, EAST, RRPN, TextSpotter, CRAFT.

The performance metrics such as precession (P) (Eq 1), recall (R) (Eq 2), and the F-measure (F) (Eq 3) to assess the efficiency of the proposed technique, which detects true text regions and extracts text sequences from natural scene images.

Three characteristics are utilized to determine the effectiveness of the text detection approach: TDR (True Detected Text Regions), APTR (Actually Text Regions Present) in the image, and FIR (False Identified Text Regions), which incorrectly identify non-text regions as text in the image.

B. Experimental Results

The proposed Unet model effectively detects The Table 1. Shows the results of Precision (P) and Recall (R) of overall images and their instances as well as individual classes for different train-validation ratios.

$$\text{Precision (P)} = \frac{TDR}{TDR+FIR} \quad (1)$$

$$\text{Recall (R)} = \frac{TDR}{APTR} \quad (2)$$

$$\text{F} - \text{Measure (F)} = \frac{2*p*r}{P+R} \quad (3)$$

MSRA -TD500 dataset contains varieties of images with different scenarios, different environment and many complexities involved.

These include Multiple scripts, multiple orientation, text embossing, lighting effects, blurred, night vision, low resolution, low contrast, text overlapping, images with incomplete texts, shaded and dull colored.

TABLE I: Comparative Analysis with Different Approaches

Methods	Recall	Precision	F-measure
MRF[21]	71.37	64.92	67.99
Min Liang et.al, Scale guided regression (SRM), Ref in [24]	85.56	88.89	87.19
Guanyu Deng, et al RFRN, [25]	84.9	76.3	80.4
Tao Wang et al, SH cascade RCNN [26]	84.06	86.01	85.02
DBnet++ [27]	73.26	73.29	72.77
Proposed Methodology (CapText-Unet)	82.80	93.84	88.10

Figure 4. Illustration of Possible Outcome of the CapText-Unet approach

V. Conclusion

Here we propose technique using UNet model to extract the text components from scene images. The model is effective such that it is trained with minimal set of training and validation images. After generation of training to generate the weights then images validated by the trained mode. Initially e have used UNet model for training of images with standard weights, after training the base model with labelled images the model with best performance weights generated, and best model generated is tested for actual scene text images represents performance efficacy of the model. Many methods including advanced deep-learning models and transformers have been developed in the past to address these challenges. Most of the method involve labelling and training of large number of datasets, which is tedious process. Our proposed algorithm combining double edge method, Unet and post processing techniques for efficient scene text detection, which produces good results with few numbers of samples. We have trained and tested on various ratios like 20 to 40 epochs and image size of 540*540. Significant feature of our algorithm is that its yielded better efficacy even with the smaller number of trained samples, this helps in lessening the burdon of tedious and time-consuming training process. Our method achieves state-of-the-art results on MSRATD500 dataset, serves handy for text detection/recognition in the natural scene and videos.

References

[1] S. Gujjeti and S. Radhika, ‘‘Analysis of various approaches for scene text detection and recognition,’’ J. Data Acquisition Process., vol. 38, no. 3, p. 1735, 2023.
[2] Y. Wu, L. Zhang, H. Li, Y. Zhang, and S. Wan, ‘‘Feature fusion pyramid network for end-to-end scene text detection,’’ ACM Trans. Asian Low Resource Language Inf. Process., pp. 1–6, Jan. 2023.
[3] Jack Greenhalgh and Majid Mirmehdi “recognizing text-based traffic signs” IEEE transactions on intelligent transportation systems, vol.16, no.3,june2015. DOI:10:109/TITS.2014.2363167
[4] Wenhao He, Xu-Yao Zhang, Fei Yin, Zhenbo Luo, Jean-Marc Ogier, Cheng-Lin Liu,“Realtime multi-scale scene text detection with scale-based region proposal network”,Pattern Recognition, Volume 98,2020,
[5] Sheng F, Chen Z, Xu B (2019) Nrtr: A no-recurrence sequence-to-sequence model for scene text recognition. In: ICDAR, pp 781–786.
[6] Bagi R, Dutta T, Gupta H P.Cluttered TextSpotter: An end-to-end trainable light-weight scene text spotter for cluttered environment. IEEE Access 8:111433–111447,2020.
[7] Munjal RS, Prabhu AD, Arora N, Moharana S, Ramena G Stride: Scene text recognition in-device. arXiv preprint arXiv:2105.07795,2021.
[8] Li HP, Doermann D, Kia O Automatic text detection and tracking in digital video. IEEE Trans Image Processing 9: 147–156. 2000.
[9] Zhong Y, Karu K, Jain AK Locating text in complex color images. Pattern Recognition 28: 1523–1535.1995.

[10] Long, S., He, X. & Yao, C. Scene Text Detection and Recognition: The Deep Learning Era. Int J Comput Vis 129, 161–184 (2021). https://doi.org/10.1007/s11263-020-01369-0
[11] S. Gujjeti, S. M and G. V, "A Systematic Investigation on Different Scene Text Detection and Recognition Method using Deep Learning Techniques," 2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bangalore, India, 2024, pp. 1-5, doi: 10.1109/IITCEE59897.2024.10467856.
[12] S. Gujjeti and S. Radhika, ''Analysis of various approaches for scene text detection and recognition,'' J. Data Acquisition Process., vol. 38, no. 3, p. 1735, 2023.
[13] Y. Wu, L. Zhang, H. Li, Y. Zhang, and S. Wan, ''Feature fusion pyramid network for end-to-end scene text detection,'' ACM Trans. Asian Low Resource Language Inf. Process., pp. 1–6, Jan. 2023.
[14] Sheng F, Chen Z, Xu B (2019) Nrtr: A no-recurrence sequence-to-sequence model for scene text recognition. In: ICDAR, pp 781–786.
[15] Bagi R, Dutta T, Gupta H P.Cluttered TextSpotter: An end-to-end trainable light-weight scene text spotter for cluttered environment. IEEE Access 8:111433–111447,2020.
[16] Wang, W., Xie, E., Song, X., Zang, Y., Wang, W., Lu, T., Yu, G., Shen, C. (2020). Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), pp.8440-8449.
[17] Li HP, Doermann D, Kia O Automatic text detection and tracking in digital video. IEEE Trans Image Processing 9: 147–156. 2000.
[18] Zhong Y, Karu K, Jain AK Locating text in complex color images. Pattern Recognition 28: 1523–1535.1995.
[19] Basavaraju, H.T., Aradhya, V.M., Guru, D.S., Harish, B.S. (2018). LoG and structural based arbitrary oriented multi lingual text detection in images/video. International JournalofNaturalComputingResearch,7(3):1-https://doi.org/10.4018/IJNCR.2018070101
[20] Basavaraju, H.T., Aradhya, V.N.M., Guru, D.S. (2019). Text detection through hidden Markov random field and EM-algorithm. In Information Systems Design and Intelligent Applications. Advances in Intelligent Systems and Computing, vol. 862, pp. 19-29. Springer, Singapore. https://doi.org/10.1007/978-981-13-3329-3_3
[21] Mahesha, M., Manjunath Aradhya, V.N., Basavaraju, H.T., Siddesha, S. (2024). MRFScene: Multi-lingual Multi-oriented Scene Text Detection Using Markov Random Fields. In: Tiwari, R., Saraswat, M., Pavone, M. (eds) Proceedings of International Conference on Computational Intelligence. ICCI 2023. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-97-3526-6_34
[22] Mahadevappa, M., Aradhya, V.N.M., Basavaraju, H.T., Shivarudraswamy, S. (2024). CNN-DEdge: Multilingual scene text detection and extraction. Mathematical Modelling of Engineering Problems, Vol. 11, No. 11, pp. 3152-3160. https://doi.org/10.18280/mmep.111125