**IAES International Journal of Artificial Intelligence (IJ-AI)**
Vol. 13, No. 1, March 2024, pp. 839~849
ISSN: 2252-8938, DOI: 10.11591/ijai.v13.i1.pp839-849 

# CRNN model for text detection and classification from natural scenes

**Puneeth Prakash, Sharath Kumar Yeliyur Hanumanthaiah, Somashekhar Bannur Mayigowda**
Department of Information Science and Engineering (ISE), Maharaja Institute of Technology Mysore
(Affiliated to Visvesvaraya Technological University), Belagavi, India

**Article Info**

***Article history:***

Received Sep 13, 2023
Revised Oct 16, 2023
Accepted Nov 7, 2023

***Keywords:***

Scene text detection
Natural scene segmentation
Ensemble learning
PDT2023
ICDAR datasets

**ABSTRACT**

In the emerging field of computer vision, text recognition in natural settings remains a significant challenge due to variables like font, text size, and background complexity. This study introduces a method focusing on the automatic detection and classification of cursive text in multiple languages: English, Hindi, Tamil, and Kannada using a deep convolutional recurrent neural network (CRNN). The architecture combines convolutional neural networks (CNN) and long short-term memory (LSTM) networks for effective spatial and temporal learning. We employed pre-trained CNN models like VGG-16 and ResNet-18 for feature extraction and evaluated their performance. The method outperformed existing techniques, achieving an accuracy of 95.0%, 96.3%, and 96.2% on ICDAR 2015, ICDAR 2017, and a custom dataset (PDT2023), respectively. The findings not only push the boundaries of text detection technology but also offer promising prospects for practical applications.



***Corresponding Author:***

Puneeth Prakash
Department of Information Science and Engineering
Maharaja Institute of Technology Mysore (Affiliated to Visvesvaraya Technological University)
Belagavi, Karnataka 571477, India.
Email: puneeth.phd20@gmail.com.

## 1. INTRODUCTION

Images of natural scenes often include text that can be used for various purposes such as automatic license plate identification, image retrieval, satellite navigation, guiding robots on their way, street sign recognition, and a better understanding of the images themselves [1], [2]. Although natural scene text recognition has come a long way, it is still a complex process because of factors including complicated backdrops, varying text size, color, orientation, low resolution, occlusion, environmental noise, and blur [3].

Many early deep learning scene text identification approaches [4]–[8] use bounding boxes that closely encompass the text it represents. An image classifier checks each region of interest to see if it contains any instances of text. Two-stage approaches and one-stage approaches can be distinguished among these techniques. Anchors are placed in the original image by text detectors that use two stages. The use of bounding boxes has been abandoned in favor of image segmentation, giving rise to various scene text detectors in recent years [9]–[13]. SSC-Net, inspired by [14] approach to segmentation, which links all picture elements in the same instance.

Although deep learning has made strides in text recognition within natural settings [15], the bulk of the research has primarily concentrated on foreign scripts (such as Greek texts) [16], [17]. The text detection and identification domain for cursive languages like Kannada, Hindi and Tamil is still in the early stages. A noticeable research gap exists for cursive languages like Kannada, Hindi, and Tamil, especially in natural

scene text detection and identification. Even as models like convolutional recurrent neural networks (CRNN) show promise in recognizing cursive texts [18]–[20], the challenge is amplified with texts in various natural scenes due to their complexities and variability in backgrounds, fonts, sizes, and colors. Furthermore, much of the existing literature is confined to studying isolated characters and scripts [21]–[23].

In addition, the recognition accuracy is reduced when natural scene images contain multiple non-text elements like leaves, cursive text lines, human agents and other complex environmental conditions, as illustrated in Figures 1, 2 and 3, and the highlight of the proposed model in Figures 4, 5 and 6. Figures 1, 2 and 3 illustrate variations of cursive texts in natural scenes. The yellow bounding boxes as depicted in Figures 4, 5 and 6 show the robustness of the proposed approach to outliers.


Figure 1. Non-textual objects such as green leaves and lighting superimposed on the word “Snipes”


Figure 2. Non-textual objects such as a motorbike, human driver, and complex environment

Figure 3. A challenging and noisy environment with poorly illuminated texts


Figure 4. Detection of texts from natural scene


Figure 5. Natural scene text detection from heterogeneous background

Figure 6. Detection of text from a noisy and poorly illuminated environment

For these reasons, deep learning approaches have recently been developed for natural scene text recognition [24]. It is worth noting that most of these innovations have been geared toward Latin scripts [25]–[27]. However, text identification and recognition in natural scene photographs is still a developing topic for cursive scripts like Kannada, Tamil, and Telugu.

This paper focuses on a novel network for segmenting English and select Indian scripts like Tamil and Kannada, based on the foundational U-Net [28] architecture. It proposes a receptive field expansion through added convolution layers for richer feature extraction. Historically, text detection and recognition have been pivotal in computer vision. Before the deep learning era, scene text detection often encompassed text extraction followed by candidate filtering. This entailed extracting texts based on predefined criteria.

Traditional methods relied on thresholding for document binarization, using global thresholds as filters to distinguish between text and images. Techniques like sliding windows and connecting components were foundational. Sliding-window methods involved classifying various window sizes to detect text, while connected component algorithms, like MSER [29] and SWT [30], grouped pixels. Notably, Tian *et al.* [31] introduced minimum cost flow networks, addressing error accumulation in texts by spotting related components of candidate characters through a cascade-boosting strategy. Other notable approaches include using a 2D Gaussian kernel [32] for region variability and mathematical morphology to segregate the image background. However, traditional methods often faced limitations, particularly struggling with complex backgrounds and inconsistent brightness, leading to inconsistent results.

The proliferation of vast datasets has led to advancements in DL models, including the likes of ResNet50 and VGG19 [33]. Notably, generative adversarial network (GAN) by Goodfellow *et al.* [34] has gained prominence for image enhancement and transformation tasks. Research has shown GANs, including models like CycleGAN [35] and pix2pix-HD [36], to be instrumental in semantic segmentation and high-resolution image translations. Moreover, innovative text detection strategies have emerged, with works like [37], [38] leveraging region proposal networks (RPN) for texts with varied orientations. While scholars like Xu *et al.* [39], and Luc *et al.* [40] have researched image segmentation-based text detection, the full potential of deep neural networks in learning remain a promising research area.

This research aims to address the existing gaps by employing the CRNN model that eliminates the need for segmentation by reformulating the text recognition challenge as a sequential temporal or time-based categorization task. Our findings indicate that deep convolutional neural networks (CNNs) with skip connections are more effective in feature extraction. Incorporating bidirectional recurrent mechanisms allows the model to capture extended contextual data in both forward and reverse directions. This dual-directional contextual understanding is crucial for enhancing prediction accuracy, particularly in the case of cursive writing styles where character shapes often bear similarities. Further, the proposed method is tested on a variety of Indian language scripts, and evaluated alongside state-of-the-art solutions.

Our contributions are in four folds:

a) We proposed a robust approach to handling complex document images;
b) A unique approach to image enhancement that incorporates CLAHE leading to robust segmentation;
c) An efficient neural network for scene text image segmentation and recognition with less computational complexity; and
d) Proposing a robust multi-lingual approach that recognizes Kannada, Tamil, and English texts from images, benchmarked in detail on ICDAR2015, ICDAR2017, and our contributed dataset (PDT2023).

This study is subdivided into four sections: i) Section one focuses on the background, motivation, and survey of relevant literatures; ii) Section 2 is focused on the methodology; iii) The results of the experiments benchmarking with most state-of-the-art are presented in section 3; and iv) Finally, section 4 highlights the conclusion and scope for future study.

## 2. METHOD

This paper introduces a novel semantic segmentation technique [11] that quantifies the distance from the center to the edge of the text, offering a granular perspective on text structure [41] within images. Initially, we provide a comprehensive overview of the dataset selected for this study, detailing its composition and relevance to our research goals. Subsequently, we explore an in-depth discussion of the proposed methodology, highlighting how our approach innovatively addresses the challenges of text segmentation in complex visual data. This foundation sets the stage for the detailed analysis and findings presented in the subsequent sections of the paper.

### 2.1. Dataset description and image enhancement

Three datasets were used for this study, namely ICDAR2015 [11], ICDAR2017 [43], and PDT2023 (a dataset of Kannada, Tamil, and English text images). The ICDAR2015 and 2017 datasets include 1670 camera-captured images and 18,000 images used as a benchmark for the robust character recognition competition on computer vision. The aim was to assist the research community in developing robust algorithms to address the difficulties faced in detecting texts from the wild, and to facilitate the comparison of different approaches.

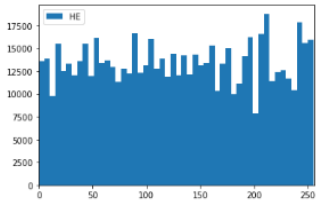
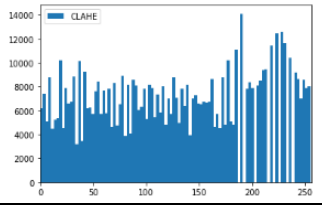
Equally, we present a few cameras captured images to test the robustness of the proposed technique. Throughout the literature, the dataset would be referred to as PDT2023. It consists of 2000 images of varying dimension, captured in and around Mysore region in India. The images were in .jpg format. A discussion on the techniques for data augmentation and preprocessing are subsequently presented. The images were preprocessed and resized to 256×256 pixels to maintain uniformity. The parameters chosen are in Table 1.

Table 1. Parameters for image augmentation on the PDT2023 dataset

| Method | Default | Augmented |
|---|---|---|
| Rotation | - | 300, 450, 600 |
| Rescale | - | 1./255 |
| Zoom range | - | 0.25 |
| x-Shift, y-Shift | None | 0.1 |
| x-Scale, y-Scale | None | 0.1 |
| Adjusted image | Varies | 256 x 256 |

The proposed technique addresses challenges related to scene-text enhancement, detection and classification. Generative adversarial networks (GANs) are utilized due to their proven superiority in generating high-quality samples compared to auto-encoders. This model is designated as PDT-Net, specifically tailored for detection of texts from natural scene images containing select Indian languages. To further improve the image quality, contrast limited adaptive histogram equalization (CLAHE) was applied. It is notable that methods like Binary and Otsu thresholding were deemed unsuitable as they led to excessive noise introduction, making CLAHE the preferred choice for processing outputs. The results of preprocessing and image enhancement are presented in Table 2

Table 2. The results of pre-processing on the PDT 2023

| Original Input Image (1000×750) | Resized image (256×256) | Histogram Equalization (HE) | CLAHE |
|---|---|---|---|
|  |  |  |  |
| | |  |  |

## 2.2. Proposed methodology

The architecture of the proposed framework for recognition of English and select Indian texts from natural scenes is illustrated in Figure 7. The model utilizes VGG-16 and ResNet-18 architectures without fully connected layers for the feature extraction stage. Unlike the architecture proposed in [11], which used seven convolutional layers, the feature extraction models are augmented with skip connections to improve gradient flow. On top of the feature extraction component, a bidirectional long short-term memory (BiLSTM) layer with 256 hidden units is employed to decode feature sequences into per-frame label predictions. Finally, a connectionist temporal classification (CTC) layer is used to map the per-frame label predictions into the final output.

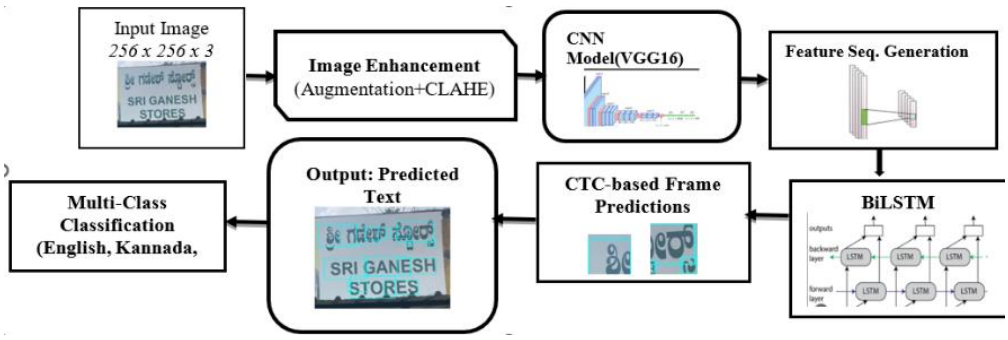



Figure 7. General architecture of the proposed detection (PDT) and recognition of English and select Indian texts from natural scenes

The proposed includes four intermediate steps between the input image and the final result, namely: i) image preprocessing and enhancement using CLAHE, ii) CNN model for feature extraction, iii) feature sequence generator using gated recurrent unit and bidirected LSTM, and iv) frame predictions. The system is trainable with a single loss function and incorporates two networks (CNN and RNN). We draw inspiration from Jaderberg *et al.* [44], fine-tuning some hyperparameters to address the difficulties in Kannada text recognition. Equally, the framework for Kannada character recognition uses the VGG-16 [45], ResNet18 [46], and its various models, with a new proposed network comparable, but with skip connections. First, per-frame label sequence prediction is accomplished by layering a recurrent network with the feature extraction network, such as a BiLSTM, followed by a layer to map the predicted sequences to their final labels. Further, VGG-16 network was employed without linked layers, and a BiLSTM with 256 hidden units, just like in [47], and obtain segmentation accuracy comparable to most state-of-the-art model.

### 2.2.1. Feature extraction with pre-trained convolutional neural networks (CNNs)

The core of the framework is the feature extraction component, which employs multiple deep learning architectures. For effective feature extraction, it is assumed that the images are a sequence of characters. The objective is to identify the most accurate depictions of the patterns in the provided images, preserving vital data at several depths. The feature map was slightly modified to account for the horizontal direction depending on the number of textual instances, and adapted the VGG-16 architecture that integrates shortcut connections. The idea is to better capture both low-level and high-level features from the text images, which are mostly horizontally oriented in the PDT2023 dataset.

a) VGG-16 with baseline model

The primary model for feature extraction in the proposed framework starts with the VGG-16 architecture. Which is adapted by adding an extra block containing two convolutional layers and one max-pooling layer to reduce the feature map's height to 1, while maintaining critical aspects of text sequence representation. Different max-pooling window sizes, specifically 2×2 and 2×1, are utilized in this version of the VGG-16 model.

b) VGG-16 with skip connections

Building upon the standard VGG-16, the improved version (Figure 8) introduces shortcut connections to mitigate the vanishing gradient problem. These connections allow gradients to flow more freely through the network, leading to better feature extraction capabilities. Mathematically, the output feature vector of the enhanced VGG-16 model is represented in (1):

$$O = F(x,\{W_i\} + x_i \quad (1)$$

This model is based on the ensemble of convolution and pooling layers to extract features sequentially.

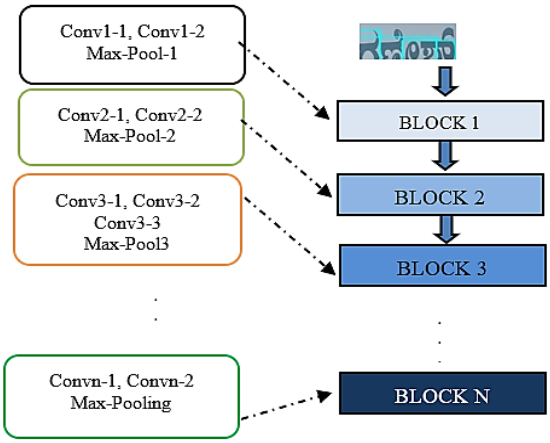



Figure 8. Adapted VGG-16 model with conv blocks

To handle the varnishing gradient problem, VGG-16 employs a skip connection. The output vector (σ) is defined as:

$$\sigma = f(a,B) \quad (2)$$

where $a$ is the input vector; $f$ = a mapping function; $B$ = the weighted values. Thus, in VGG-16, the output feature space is

$$\sigma = f(a, \{Bi\}) + ai \quad (3)$$

where $B_i$ is the weighted parameter of the *ith* CNN model, and $a_i$ = the output feature. where $x_i$ is the output from an earlier convolutional layer.

c) Residual networks (ResNet)

The residual network model employs skip connections that enhance convolutional layer outputs. The original ResNet-18 contains eight residual blocks with two 3×3 kernel layers, while the modified version introduces a 1, 3, 1 kernel layer sequence. Experiments on ResNet architectures explored different shortcut strategies, with post-activation units in ResNet-18 showing superior effectiveness for the given application.

### 2.3. Feature map to feature sequence conversion

Recurrent neural networks (RNNs) utilize hidden layers for sequence generation but face challenges like vanishing and exploding gradients when processing extended text sequences. Long short-term memory (LSTM) [48] networks overcome these issues, offering enhanced memory recall from past inputs. Bidirectional LSTMs (BLSTMs) further refine this by having two hidden layers: one processing input sequences from past to future and another from future to past. The final layer of CNN models transforms outputs into 1D feature maps, which are segmented to produce feature vectors. In mathematical terms, the feature sequences are denoted as $x = \{x_1, x_2, \ldots x_N\}$ where $x_t \epsilon R^{512}$ = the length of the feature sequences.

#### 2.3.1. Per frame predictions

To train a BiLSTM, one must locate where each ground truth text's character is horizontally in a given image. This is because the BiLSTM gives a score at each time step for each horizontal position in the image. As a result of the nature of the text and the overlap, it becomes challenging to separate each character of the ground truth text in an image when using cursive scripts like Kannada and Tamil. The method proposed by [49] was employed, an approach that has been successful in many character recognition tasks. Bounding box generation is an all-important task; to that effect, algorithm 1 was proposed:

Algorithm 1: The strategy employed on PDT-Net for text recognition

```
Load input images: Resize (256 x 256)
Perform data augmentation: Rotation, Scaling & Normalization
For n = 1, k do:
    Apply image enhancement, Img ∈ [1, k]
        Apply CLAHE
        PDT-Net, K do:
            Feature extraction in Fn vector space
            Sequence generation using RNN, and BiLSTM
            Frame predictions: Label (P^i) ← max(Label (P^i_j))
        end do
      Label (P^i) ← Null
endFor
return Predicted Text
```

### 2.4. The training process and experimental setup

PDT2023 dataset: To train PDT-Net using a 3×3 filter with a stride of 1, 240 samples were reserved for training and 60 for validation, corresponding to the 80:20 rule for the training and testing sets, respectively. Normalizing the input characteristics to 0 and 1 speeds up network training and convergence. In order to improve the accuracy of deep neural networks, the training dataset is expanded by incorporating a data augmentation technique that rotates the images at random (at 30, 45, and 60 degrees, respectively).

ICDAR2015, ICDAR 2017 dataset: These datasets were used as a benchmark, which contain 1670 and 18000 images respectively. When creating the training set, only the cropped versions of the images of the words produced by data augmentation were considered. Adaptive momentum (ADAM) with a learning rate set to 10e-5. The network was trained on an NVIDIA 1060 GPU with a memory of 24 GB, which analyses ten input images per batch. The experiments were conducted using Python, leveraging TensorFlow library as the backend. The approach requires 0.5 and 0.4 seconds for both the training and testing phases.

## 3. RESULTS AND DISCUSSION

In this section, we demonstrate the outcomes achieved by applying the proposed techniques. The selection of images for analysis was intentional, capturing a spectrum of lighting conditions and complexities to mirror the diverse challenges encountered in real-world scenarios. This approach ensures that our results are theoretically sound and practically applicable, providing a robust validation of the techniques' effectiveness across various environments.

- Accuracy (Acc) measures the number of correct predictions to the sum of predictions. It is defined as (4).

$$Acc = \frac{True_P + True_N}{True_P + True_N + False_P + False_N} \quad (4)$$

- Precision addresses the question of the proportion of identifications that was correct. The criterion is expressed as (5):

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

where TP = true positive.

- Sensitivity, also known as recall, accounts for the actual positives identified correctly. It is defined mathematically as (6):

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

where FN = false positive.

### 3.1. Results of using PDT-Net across benchmark datasets

The outcome of employing PDT-Net on established benchmark datasets is presented. The selected images within these datasets encompass a wide array of illumination conditions and present numerous complexities to closely simulate actual environmental conditions. This careful curation ensures that the effectiveness of PDT-Net is thoroughly evaluated, showcasing its adaptability and robustness in handling diverse and challenging scenarios encountered in practical applications. In the experiments as presented in Table 3, different sets of images were tested to evaluate the model's performance, and subsequently evaluated on each of these dataset as presented in Table 4.

The ICDAR2015 and ICDAR2017 datasets showcased the model's adept text detection and recognition capabilities, as evidenced by annotated outputs. The custom PDT2023 dataset initially highlighted baseline performance, but after refinement, the model demonstrated increased accuracy and linguistic versatility, recognizing English, Kannada, and Tamil texts. This underscores its potential adaptability for diverse real-world linguistic scenarios.

On the ICDAR 2017, an accuracy of 98% highlights that the model performed better than when trained on other datasets. However, a precision of 98% on the ICDAR2015 proved that ICDAR2015 was better since it contained fewer images than its successor. An accuracy of 79.2% on the newly proposed dataset (PDT2023) shows the learnability of the model.

### 3.2. The discussion on results obtained using VGG16 and ResNet18

In this sub-section, we explored a detailed analysis of the classification performance achieved through the utilization of pre-trained models such as VGG16 and ResNet18. The comparative results, meticulously tabulated in Tables 5 through 7, clearly depict each model's efficacy in our study's context. This examination not only highlights the strengths and limitations of the employed models but also sets the groundwork for further discussion on the implications of these results for the field of image classification.

### 3.3. Comparative analysis

To determine the efficiency of our proposed method, we conducted a comprehensive comparative analysis, pitting our approach against established methods presented in studies [7], [39], [49], [4], and [50]. This comparison aimed to benchmark the performance of the method in terms of key metrics: accuracy (Acc), precision, and recall. These metrics were computed based on the formula provided in (5), (6), and (7). By analyzing these metrics across different methodologies, we can offer an understanding of how our method stands relative to the state-of-the-art, highlighting its strengths and potential areas for improvement. This comparative approach ensures a robust evaluation, providing readers with a clear perspective on the advancements our research brings to the field.

Table 3. Segmentation and recognition results of the three different datasets

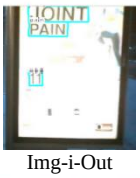
| Dataset | Input image(s) | | Segmented output with Bbox predictions | |
|---|---|---|---|---|
| ICDAR2015 |  Img-i |  Img-ii |  Img-i-Out |  Img-ii-Out |
| ICDAR2017 |  Img-iii |  Img-iv |  Img-iii-Out |  Img-iv-Out |
| PDT2023 |  Img-a |  Img-b |  Img-a-Out |  Img-b-Out |
| |  Img-c |  Img-d |  Img-c-Out |  Img-d-Out |
| |  Img-e |  Img-f |  Img-e-Out |  Img-f-Out |

Table 4. Performance metrics on three different datasets

| Dataset | Accuracy (%) | Precision (%) | Recall (%) |
|---|---|---|---|
| ICDAR2015 | 95.0 | 98.0 | 89.0 |
| ICDAR2017 | 96.3 | 97.0 | 92.0 |
| PDT2023 | 79.2 | 85.0 | 75.0 |

Note: Bold values represent better metrics across datasets

Table 5. Text classification metrics across different pre-trained models

| S.No | CNN Model | RNN Structure | # of Units | Precision (%) | Recall (%) |
|---|---|---|---|---|---|
| 1 | VGG-16 | BiLSTM | 256 | 93.70 | 92.60 |
| 2 | VGG-16 | “ | 512 | 95.30 | 92.70 |
| 3 | ResNet18 | “ | 512 | 95.20 | 96.60 |

Table 6. Text recognition accuracy across different number of RNN models with the same number of units

| S.No | CNN Model | RNN Structure | # of Units | Accuracy (%) | Precision (%) |
|---|---|---|---|---|---|
| 1 | VGG-16 | BiLSTM | 512 | 94.70 | 95.40 |
| 2 | VGG-16 | BiGRU | 512 | 93.30 | 93.70 |
| 3 | ResNet18 | BiLSTM | 512 | 90.20 | 94.60 |
| 4 | ResNet34 | BiGRU | 512 | 93.40 | 90.20 |
| 5 | ResNet50 | BiLSTM | 512 | 96.30 | 96.80 |

Table 7. Comparison with relevant methods

| Author | Methods | Acc | Precision | Recall |
|---|---|---|---|---|
| [7] | EAST | NR | 83.3 | 78.3 |
| [39] | R2CNN | NR | 85.6 | **79.7** |
| [49] | DocUNet | 0.41 | NR | NR |
| [4] | RRPN | NR | **90.0** | 72.0 |
| [50] | AdaptiveBinarization | NR | NR | 17.53 |
| **Ours** | PDT-Net | **98.2** | 85.0 | 75.0 |

Where NR = Not Reported
Note: Bold values represent better metrics across datasets

From Table 5, VGG-16 seems to outperform ResNet18 in terms of precision when the number of RNN units was 256. However, when the number of units was increased to 512, ResNet18 significantly surpasses VGG-16 in recall. Number of units: Increasing the number of BiLSTM units from 256 to 512 leads to a higher precision in VGG-16 but only a marginal increase in recall. This could mean that adding more units improves the model's ability to identify true positives but does not significantly improve its ability to capture all the positives (Recall). In Table 6, the experimental results on the recognition accuracy using the same number of units were reported. BiLSTM generally performs better than BiGRU when the number of units is held constant at 512. This is indicative of BiLSTM's effectiveness in capturing long-term dependencies for this specific task.

In terms of accuracy and precision, ResNet50 with BiLSTM and 512 units has the highest accuracy and precision among the models, suggesting that for a complex task like text recognition, deeper networks might be more effective. From Table 7, when compared with other relevant models, the overall performance of the PDT-Net (proposed) stands out with an accuracy of 98.2%, significantly outperforming other state-of-the-art methods. However, its precision and recall are not the highest, suggesting that while the model is highly accurate, there may be room for improvement in its ability to correctly identify true positives (precision) and its ability to identify all positives (recall).

The method by Pratikakis *et al.* [50] has a very low recall, indicating that it misses a large number of true positive cases. These tables collectively offer a robust evaluation of your PDT-Net's performance in contrast to existing methods and varying configurations, providing both a validation of the approach and insights for future optimization.

## 4. CONCLUSION

The research presented offers a comprehensive examination of text recognition methodologies, focusing on varying architectures, configurations, and evaluation metrics. Our proposed PDT-Net model demonstrated superior performance, achieving an accuracy rate of 98.2%, thereby outpacing existing state-of-the-art methods. This result is indicative of the efficacy of the combined CNN-RNN architecture, which leverages the strengths of convolutional layers for feature extraction and recurrent layers for sequence modeling.

However, it is worth noting that while PDT-Net excels in terms of overall accuracy, there are areas where it could be further optimized. Specifically, its precision and recall metrics, while respectable, did not reach the upper echelons observed in some other methods. This suggests that while the model is generally reliable, it may still miss some true positives or falsely identify negatives, indicating room for refinement.

Furthermore, the comparative analysis revealed the merits and limitations of different configurations and RNN units. For instance, increasing the number of BiLSTM units from 256 to 512 led to a noticeable rise in precision for the VGG-16 model but only a marginal improvement in recall. This finding is valuable for future studies aiming to balance these metrics effectively. Finally, real-world testing and applications, as well as efforts to make the model more interpretable, could pave the way for its integration into various systems that demand high-efficiency text recognition.

## REFERENCES

[1] L. Yang, D. Ergu, Y. Cai, F. Liu, and B. Ma, "A review of natural scene text detection methods," *Procedia Computer Science*, vol. 199, pp. 1458–1465, 2022, doi: 10.1016/j.procs.2022.01.185.
[2] A. A. Chandio, M. D. Asikuzzaman, M. R. Pickering, and M. Leghari, "Cursive text recognition in natural scene images using deep convolutional recurrent neural network," *IEEE Access*, vol. 10, pp. 10062–10078, 2022, doi: 10.1109/access.2022.3144844.
[3] W. Feng, F. Yin, X.-Y. Zhang, and C.-L. Liu, "Semantic-aware video text detection," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021. doi: 10.1109/cvpr46437.2021.00174.
[4] J. Ma *et al.*, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3111–3122, 2018, doi: 10.1109/tmm.2018.2818020.
[5] P. Lyu, C. Yao, W. Wu, S. Yan, and X. Bai, "Multi-oriented scene text detection via corner localization and region

segmentation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018. doi: 10.1109/cvpr.2018.00788.
[6] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "deep direct regression for multi-oriented scene text detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2017, pp. 745–753. doi: 10.1109/ICCV.2017.87.
[7] X. Zhou *et al.*, "EAST: An efficient and accurate scene text detector," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. doi: 10.1109/cvpr.2017.283.
[8] H. Hu, C. Zhang, Y. Luo, Y. Wang, J. Han, and E. Ding, "WordSup: Exploiting word annotations for character based text detection," *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017. doi: 10.1109/iccv.2017.529.
[9] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2019, pp. 9357–9366. doi: 10.1109/CVPR.2019.00959.
[10] D. Deng, H. Liu, X. Li, and D. Cai, "PixelLink: Detecting scene text via instance segmentation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018, doi: 10.1609/aaai.v32i1.12269.
[11] W. Wang *et al.*, "Shape robust text detection with progressive scale expansion network," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019. doi: 10.1109/cvpr.2019.00956.
[12] C. Xue, S. Lu, and F. Zhan, "Accurate scene text detection through border semantics awareness and bootstrapping," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 370–387. doi: 10.1007/978-3-030-01270-0_22.
[13] Y. Tang and X. Wu, "Scene text detection using superpixel-based stroke feature transform and deep learning based region classification," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2276–2288, 2018, doi: 10.1109/tmm.2018.2802644.
[14] V. I. Agughasi and M. Srinivasiah, "Semi-supervised labelling of chest x-ray images using unsupervised clustering for ground-truth generation," *Applied Engineering and Technology*, vol. 2, no. 3, pp. 188–202, 2023, doi: 10.31763/aet.v2i3.1143.
[15] X. Li, "A deep learning-based text detection and recognition approach for natural scenes," *Journal of Circuits, Systems and Computers*, vol. 32, no. 05, 2022, doi: 10.1142/s0218126623500731.
[16] I. Marthot-Santaniello, M. T. Vu, O. Serbaeva, and M. Beurton-Aimar, "Stylistic similarities in greek papyri based on letter shapes: a deep learning approach," *Document Analysis and Recognition – ICDAR 2023 Workshops*. Springer Nature Switzerland, pp. 307–323, 2023. doi: 10.1007/978-3-031-41498-5_22.
[17] M. Ghosh, H. Mukherjee, S. M. Obaidullah, X.-Z. Gao, and K. Roy, "Scene text understanding: recapitulating the past decade," *Artificial Intelligence Review*, vol. 56, no. 12, pp. 15301–15373, 2023, doi: 10.1007/s10462-023-10530-3.
[18] A. Yadav, S. Singh, M. Siddique, N. Mehta, and A. Kotangale, "OCR using CRNN: A deep learning approach for text recognition," *2023 4th International Conference for Emerging Technology (INCET)*. IEEE, 2023. doi: 10.1109/incet57972.2023.10170436.
[19] R. Najam and S. Faizullah, "Analysis of recent deep learning techniques for arabic handwritten-text OCR and post-OCR correction," *Applied Sciences*, vol. 13, no. 13, p. 7568, 2023, doi: 10.3390/app13137568.
[20] P. Chhabra, A. Shrivastava, and Z. Gupta, "Comparative analysis on text detection for scenic images using EAST and CTPN," *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, 2023. doi: 10.1109/icoei56765.2023.10125894.
[21] A. Rahman, A. Ghosh, and C. Arora, "UTRNet: High-resolution urdu text recognition in printed documents," in *Lecture Notes in Computer Science*, Springer Nature Switzerland, 2023, pp. 305–324. doi: 10.1007/978-3-031-41734-4_19.
[22] S. Kaur, S. Bawa, and R. Kumar, "Heuristic-based text segmentation of bilingual handwritten documents for Gurumukhi-Latin scripts," *Multimedia Tools and Applications*, 2023, doi: 10.1007/s11042-023-15335-8.
[23] E. F. Bilgin Tasdemir, "Printed Ottoman text recognition using synthetic data and data augmentation," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 26, no. 3, pp. 273–287, 2023, doi: 10.1007/s10032-023-00436-9.
[24] S. Long, X. He, and C. Yao, "Scene text detection and recognition: The deep learning era," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 161–184, 2021, doi: 10.1007/s11263-020-01369-0.
[25] T. Khan, R. Sarkar, and A. F. Mollah, "Deep learning approaches to scene text detection: a comprehensive review," *Artificial Intelligence Review*, vol. 54, no. 5, pp. 3239–3298, 2021, doi: 10.1007/s10462-020-09930-6.
[26] E. Hassan and L. V L., "Scene text detection using attention with depthwise separable convolutions," *Applied Sciences*, vol. 12, no. 13, p. 6425, 2022, doi: 10.3390/app12136425.
[27] X. Liu, G. Meng, and C. Pan, "Scene text detection and recognition with advances in deep learning: a survey," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 22, no. 2, pp. 143–162, 2019, doi: 10.1007/s10032-019-00320-5.
[28] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.
[29] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004, doi: 10.1016/j.imavis.2004.02.006.
[30] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010. doi: 10.1109/cvpr.2010.5540041.
[31] S. Tian, Y. Pan, C. Huang, S. Lu, K. Yu, and C. L. Tan, "Text Flow: A unified text detection system in natural scene images," *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015. doi: 10.1109/iccv.2015.528.
[32] W. Xiong, X. Jia, J. Xu, Z. Xiong, M. Liu, and J. Wang, "Historical document image binarization using background estimation and energy minimization," *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018. doi: 10.1109/icpr.2018.8546099.
[33] A. Victor Ikechukwu, S. Murali, R. Deepu, and R. C. Shivamurthy, "ResNet-50 vs VGG-19 vs training from scratch: A comparative analysis of the segmentation and classification of Pneumonia from chest X-ray images," *Global Transitions Proceedings*, vol. 2, no. 2, pp. 375–381, 2021, doi: 10.1016/j.gltp.2021.08.027.
[34] I. J. Goodfellow *et al.*, "Generative adversarial networks," Jun. 2014, [Online]. Available: http://arxiv.org/abs/1406.2661
[35] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2017, pp. 2242–2251. doi: 10.1109/ICCV.2017.244.
[36] A. V. Ikechukwu, M. S, and H. B, "COPDNet: An explainable ResNet50 model for the diagnosis of COPD from CXR images," *2023 IEEE 4th Annual Flagship India Council International Subsections Conference (INDISCON)*. IEEE, 2023. doi: 10.1109/indiscon58499.2023.10270604.
[37] Y. Jiang *et al.*, "R2CNN: Rotational region CNN for orientation robust scene text detection," Jun. 2017, doi: .1706.09579.
[38] "i-Net: a deep CNN model for white blood cancer segmentation and classification," *International Journal of Advanced Technology and Engineering Exploration*, vol. 9, no. 95, 2022, doi: 10.19101/ijatee.2021.875564.

[39] Y. Xu, Y. Wang, W. Zhou, Y. Wang, Z. Yang, and X. Bai, "TextField: Learning a deep direction field for irregular scene text detection," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5566–5579, 2019, doi: 10.1109/tip.2019.2900589.
[40] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," 2016, [Online]. Available: http://arxiv.org/abs/1611.08408
[41] Z. Liu, G. Lin, S. Yang, F. Liu, W. Lin, and W. L. Goh, "Towards robust curve text detection with conditional spatial expansion," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019. doi: 10.1109/cvpr.2019.00744.
[42] D. Karatzas *et al.*, "ICDAR 2015 competition on Robust Reading," *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2015. doi: 10.1109/icdar.2015.7333942.
[43] N. Nayef *et al.*, "ICDAR2017 robust reading challenge on multi-lingual scene text detection and script identification - RRC-MLT," *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2017. doi: 10.1109/icdar.2017.237.
[44] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *International Journal of Computer Vision*, vol. 116, no. 1, pp. 1–20, Dec. 2016, doi: 10.1007/s11263-015-0823-z.
[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, Sep. 2015, [Online]. Available: http://arxiv.org/abs/1409.1556
[46] A. Victor Ikechukwu and M. S, "CX-Net: an efficient ensemble semantic deep neural network for ROI identification from chest-x-ray images for COPD diagnosis," *Machine Learning: Science and Technology*, vol. 4, no. 2, p. 25021, 2023, doi: 10.1088/2632-2153/acd2a5.
[47] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994, doi: 10.1109/72.279181.
[48] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
[49] K. Ma, Z. Shu, X. Bai, J. Wang, and D. Samaras, "DocUNet: Document image unwarping via a stacked U-Net," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018. doi: 10.1109/cvpr.2018.00494.
[50] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos, "ICDAR2017 competition on document image binarization (DIBCO 2017)," *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2017. doi: 10.1109/icdar.2017.228.

## BIOGRAPHIES OF AUTHORS

**Puneeth Prakash** is an Assistant Professor at Maharaja Institute of Technology Mysore. He has 8 years of Teaching and Research experience. He has done bachelor's degree in Information Science and a master in computer science from VTU Belagavi. His key area of interest is image processing, machine learning, and computer vision. He mainly works on Scene text-related images and has published papers in national and international Journals. He is keen on multiprogramming paradigm implementation. He can be contacted at email: puneeth.phd20@gmail.com.

**Sharath Kumar Yeliyur Hanumanthaiah** is a Professor and Head of the Department of Information Science & Engineering, MITM, Mysore. His areas of interest are image processing, pattern recognition, and information retrieval. He has around 12 years of experience in teaching. He completed a B.E. in Computer Science & Engineering from VTU and an M.Tech. in Computer Cognition Technology from the University of Mysore. Further, completed Ph.D. from the University of Mysore. Published 50 research articles in reputed conferences and journals. He served as the BoE and BoS of the University of Mysore from 2016 to 2020. He can be contacted at email: sharathyhk@gmail.com.

**Somashekhar Bannur Mayigowda** is an Assistant Professor at Maharaja Institute of Technology Mysore. He has 13 years of teaching and research experience. He has done bachelor's degree in information science, and a master's in computer science from VTU Belagavi, His Key area of interest is image processing. He mainly works on multi-modeling biometrics and has published papers in national and international journals. He has taken up many projects from image processing and IoT, which are utilized in the real world. He can be contacted at email: somumtech@gmail.com.