International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

# Vertical Text Detection and Recognition in Natural Scene Images: A Vertical Text Classifier and Detector with Gated Dual Adaptive Attention Mechanism

**A. S. Venkata Praneel[*1], Dr. T. Srinivasa Rao[2]**

**Submitted**: 12/12/2023 **Revised**: 16/01/2024 **Accepted**: 01/02/2024

**Abstract:** This paper proposes a novel approach to improving vertical text detection and recognition in natural scene images by integrating the Vertical Text Classifier and Detector Module (VTCD), which incorporates the IoU with Inclination (IoUI) Algorithm into the Gated Dual Adaptive Attention Mechanism (GDAAM). GDAAM is a unique framework for successful text recognition in demanding settings. The suggested Vertical Text Classifier and Detector Module integration intends to increase the Gated Dual Adaptive Attention Mechanism accuracy and resilience in dealing with vertical text in complicated visual situations. The Gated Dual Adaptive Attention Mechanism encoder accurately localizes text areas in natural scene images. The Vertical Text Classifier and Detector Module are used after localization to fine-tune the bounding boxes and improve vertical text detection. The Vertical Text Classifier and Detector Module's enhanced data is smoothly integrated into the Gated Dual Adaptive Attention Mechanism decoder, impacting fine-grained attention modelling. The model constantly adjusts its attention weights depending on the revised bounding boxes, enabling exact text identification by selectively focusing on key visual and textual signals. In addition to tackling the issues given by irregular text forms and different orientations. The reasoning module uses VTCD's revised bounding boxes to gather contextual information, while character awareness is improved to handle complicated text layouts and occlusions. The visual-semantic ensemble fusion decoder integrates input from both modalities to provide coherent and contextually consistent text recognition results. Extensive trials on benchmark datasets such as ICDAR 2013, ICDAR 2015, and the VTIG-500 show that the proposed Gated Dual Adaptive Attention Mechanism with Vertical Text Classifier and Detector Module works well. The results indicate higher performance in terms of accuracy and resilience compared to cutting-edge techniques, notably in difficult text recognition tasks. The addition of the Vertical Text Classifier and Detector Module to the Gated Dual Adaptive Attention Mechanism broadens in natural scene images on text recognition, displaying promising results when dealing with vertical text in complex visual conditions.

***Keywords:*** *Vertical Text Classifier and Detector Module, IoU with Inclination Algorithm, Text Detection, Text recognition, GDAAM, Semantic reasoning, Vertical Text, Character awareness.*

## 1. Introduction

Text understanding in images from natural scenes is vital for transmitting information and analysing the surroundings. Independent Scene Text Recognition (STR), for example, enables the visually handicapped to quickly access information and explore freely while getting directions [1]. STR, on the other hand, is a difficult study issue because scene images have complex backgrounds, poor resolutions, numerous sizes and fonts, and ununiform illumination [2]. The presence of text-like elements in scene images is also recognized as text; a vehicle's wheels are identified as the letter 'O' [3]. Furthermore, letters NSIs spread without previous knowledge of their location [4]. Scene text might be horizontal, curved, oriented, or vertical. As a result, multiple researchers, as mentioned in [5], [6], and [7], have been investigating various strategies for Scene Text Detection (STD) and identification for identifying horizontal, curved, and rotational texts. However, only a limited amount of research has been conducted on vertical STD and identification, such as various vertical text kinds. In natural scene images (NSI), identifying vertical text is similarly significant since it can give crucial information for interpreting natural sceneries. STR is one of the many challenging and enduring research challenges in computer vision. [8]. Several academics have worked to enhance STR's ability in natural situations. Furthermore, STR is useful in diversified applications, such as visual navigation systems, image retrieval, and text reading for the visually handicapped. STR is difficult since our surroundings contain a variety of scene text orientations. Fig. 1 shows three types of classification done on Vertical Scene Text: top_to_bottom and vertical texts, which are horizontal-stacked. Scenes might be horizontal, freely aligned, curved, or vertically orientated [9]. Furthermore, images with text may be distinguished between scanned papers and NSIs. The wording of the digitized document remains consistent with homogeneous backdrops, but words in NSI are uneven, typically with complex contexts. Optical Character Recognition (OCR) is a

*1Department of Computer Science and Engineering, GITAM (Deemed-to-be University), Visakhapatnam-530045, AP, India*

*ORCID ID: 0000-0002-9511-0645*

*2 Department of Computer Science and Engineering, GITAM (Deemed-to-be University), Visakhapatnam-530045, AP, India*

*ORCID ID: 0000-0002-6263-2666*

** Corresponding Author Email: praneelsri@gmail.com*

well-known technique for text recognition. It may obtain high recognition rates when the font text is uniform and straightforward, and the backdrop is clear. However, because of the complicated and congested surroundings, low-quality and fuzzy images, variable text orientations, diverse typefaces, and so on, OCR cannot be used to detect text in natural scenarios [10].


(a) (b) (c)

**Fig. 1**: Three varieties of Vertical text like (a) Horizontal Stacked, (b) Top_to_Bottom, (c) Botton_ to_Top

Text recognition methods have difficulties in distinguishing between texts and non-texts since certain items resemble readers. Vehicle wheels, for example, can be identified as the letter 'O'. As a result, numerous academics have suggested diverse ways to text recognition in natural settings. The majority of this work has been focused on modelling horizontal scenes and arbitrarily angled and curved STR in NSI [11]. Although vertically oriented STR is common in our surroundings, it has received little attention. As a result, understanding vertically oriented scene texts (VOST) in natural contexts is critical since they offer information. According to [2], there has been limited work on modelling vertically oriented STR in NSIs.

## 2. Related Work

STD is critical for determining the position of texts in images [4]. While several approaches for recognizing curved, arbitrarily oriented, and horizontally oriented writings have been developed over the years, there was a large void in research on identifying vertically oriented texts in NSIs.

Existing thorough studies show that existing algorithms for text identification rely heavily on basic elements [8]. Methods based on Connected Components and Sliding Windows are the two most common. [12] and [13] used the Connected Component-based technique to extract text areas from images and reduce false positives with trained classifiers.[14], in contrast, used a multi-scale window to identify text sections in all feasible places.

The introduction of DL-based approaches heralded a dramatic change in computer vision for text detection in NSIs. These systems, which use deep learning techniques, outperform classic methods in terms of STD. DL-based algorithms for text detection may be classified into three varieties: character[15], word [16], & text-line-based [17] methods.

[18] and [19] used Extremal Regions to determine the placement of characters in their character-based technique. The word-based technique, developed by [20], suggested the RCNN recognizes words using a class-independent detector. [21] proposed that Faster RCNN, built on RCNN, can recognize lengthy horizontal words in images. Another innovation in this line is [22] Single Shot Detector (SSD), which does numerous detections in one shot. [23] proposed a text-line-based technique for revealing messages in images that use the FCN. In addition to these classification techniques, a shape-based classification for STD has been presented.

Various techniques for STD in images have been presented, each with a distinct orientation. Horizontal text detection algorithms include Stroke Segmentation [24], Boundary Clustering, and String Classification. [25] suggested a technique for identifying horizontal texts and arbitrarily orienting by connecting letters to generate words. To recognize arbitrarily oriented texts, [26] used a U-shaped Fully Convolutional Network. [27] focused on curved texts, employing a Semantic Segmentation-based detector to establish relationships.

There has been very little study on vertically oriented literature. [28] created a model capable of identifying both horizontal and vertical texts which are horizontally stacked. However, it fails to recognize vertically oriented letters from top to bottom and bottom to top in NSIs. Another method, Capture2Text [29], has been developed to detect vertical texts which horizontally stacked. Notably, it requires manual text area selection and only recognizes vertical Japanese letters that are horizontally stacked.

In recent years, the computer vision field has paid particular attention to STR. The key tasks in STR are text detection and recognition. STD finds the position of text in the input NSI, and STR converts the found text sections into machine-readable strings [4]. Word-based methods, Character-based approaches, and sequence_to_sequence algorithms have all been proposed [4] for text recognition. However, these approaches have shortcomings in detecting and identifying diverse text orientations, including orientated, curved, and vertical texts. As a result, deep-learning approaches are commonly used to recognize various text forms. The use of a selective attention network and direction encoding mask enabled the effective recognition of vertical text in NSIs. However, it is only capable of identifying vertical texts which are horizontally stacked and do not support top_to_bottom or bottom_to_top. since the letters in the vertical text, which are horizontally stacked, are interpreted as horizontal letters. Despite advances in detecting horizontally, randomly, and curved letters, there is a research gap in recognizing vertically oriented texts in real-world images. Given their potential importance as a source of valuable information in natural scenes, there is an urgent need to develop a model capable of detecting various types of vertically oriented texts.

## 3. Proposed Method

### 3.1 Framework

In this section, we will delve into the architecture and comprehensive structure of our proposed method, aligning with the well-established encoder–decoder framework commonly found in machine translation [30] and text recognition [31,32]. Our architecture is comprised of three key components:

This component is tasked with extracting 2D features from the

input. It encompasses the MS-RCNN, a Mask Scoring Region Convolutional Neural Network, along with PBTPN and TCN, which collectively contribute to the feature extraction process.

This novel text identification method employs the Mask Scoring Region Convolutional Neural Network (MS-RCNN), which is notable for its ability to properly handle curved text and multi-oriented scene pictures concurrently, displaying outstanding adaptability. Given the varied structure of text in many natural surroundings, adaptability emerges as an important characteristic.

This research makes an important addition by introducing the PBTPN as a unique backbone architecture for the MS-RCNN. This upgrade significantly enhances the model's feature extraction capabilities, enhancing its dependability and accuracy of STD. The proposed strategy's utility is proved by its successful reduction of false alerts produced by text-like backgrounds.

While admitting significant advances in text detection, the study reveals areas that require additional examination and development, which is a frequent feature of every research project. Future research will seek to overcome these limits, demonstrating the dedication to continuous improvement and refinement inherent in scientific investigation. Experiments using STD benchmarks show that the suggested technique performs exceptionally well. Future studies should look into incorporating the Transformation Scaling Extension Algorithm [33] into a comprehensive training plan.

This module introduces an STD technique that can handle multi-oriented and curved text in natural scene photos. The use of PBTPN as a backbone network increases feature extraction while decreasing false alarms. The approach's utility is shown by performance measurements on existing benchmark datasets, and recognizing its limits lays the way for further advances in this field.

### 3.1.2 Vertical Text Classifier and Orientation Module:

The text detection is accomplished using the proposed technique, which is based on deep learning. It introduces the Text Detection Classifier (TDC) module, which can recognize texts in a wide range of orientations, including vertical texts. The module uses the IoU overlap algorithm with inclination [34]. After TDC identifies text portions, the Text Orientation Detector (TOD) determines whether the text is vertically oriented from top_to_bottom or bottom_to_top. Subsequently, the vertical text is translated into a horizontal orientation for further processing. Fig.2 shows the general design of the suggested paradigm.

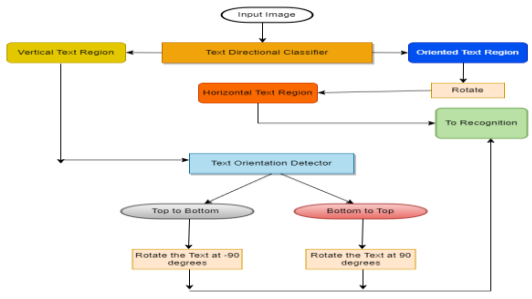


**Fig.2:** Architecture of the Vertical Text Classifier and Orientation Module

#### 3.1.3.1 Text Detection Classifier Module

Text Direction Classifier (TDC), suggested for text detection, is an expanded variant of Region Proposal Networks (RPN) [35]. A dense-proposal-based technique often used in detection applications like face and vehicle recognition employs a neural network to construct an affine transformation matrix for image rectification. TDC uses Rotation RoI (RRoI) pooling layers to identify text in a variety of orientations, including vertical text [35]. This technique takes angle regression into account while detecting text.

During TDC training, the ground truth representation consists of five tuples (x, y, h, w, and Θ) for every text area. The tuples represent the bounding box's center coordinates (x, y), width (w), height (h), and angle (Θ). After rotation, for calculating the ground truth, these tuples are critical in expressing angle discrepancies in rotating-oriented text boxes. TDC, inspired by [35], uses rotation anchors to change text orientation. In Fig.3, the six alternative angles (-π/6, 0, π/6, π/3, π/2, 2π/3) and aspect ratios (1:2, 1:5, 1:8) provide a wide variety of text at the same time keeping scales of 8, 16, and 32. For an Input image, for each anchor, these yield 54 rotation anchors, resulting in *h* x w x 54 feature maps.

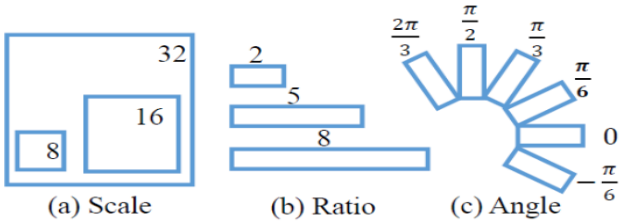


**Fig.3:** Anchors for Rotation

These rotation anchors serve as the foundation for network learning, and the intersection-over-union (IoU) overlap, as depicted in Fig. 4, determines whether a rotation is positive or negative.

In contrast to normal RoI, rotation RoI (RRoI), a fixed feature map combines regions of aspect ratios, varying sizes, and angles. The text area is separated into subregions., all of which have the same alignment. Max-pooling is used to determine whether each sub-region is background or text.

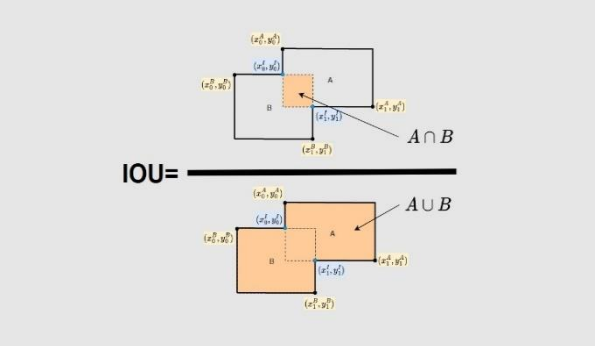


**Fig. 4:** calculations of IoU to find +ve Anchors

IoU with inclination algorithm (IoUI) varies from typical non-maximum suppression (NMS) in that it is intended to accommodate arbitrarily oriented scene text portions. The purpose of this algorithm is to reduce overlapping text sections found using the intersection over union (IOU) method. This method improves performance, especially in detecting VOST inside NSIs.

The IOU is determined by dividing the overlapping area of the ground truth bounding box by the sum of the areas of the detected

bounding boxes [36]. If the IOU value is > 0.7 and the intersection_angle < π/12, the bounding box is regarded as a +ve; otherwise, it is classified as a -vely detected anchor box, as given in Algorithm I.

Algorithm I: IoU with Inclination calculation

Input: calculate the IoU value with inclination

Output: +ve and -ve anchors

1. If IoU value with inclination_angle > 0.7
2. If intersection_angle, Θ < π/12
3. It's a +ve anchor
4. else it's a -ve anchor
5. else it's a -ve anchor

6, End.

After recognizing all +ve anchor boxes, the model includes an RROI pooling layer, which varies from the standard ROI approach. The RROI supports both randomly and VOST. Positive anchor boxes that have been recognized are separated into subregions to aid in the rotation of the discovered areas form axis-aligned text portions.

Vertically oriented area proposals with width (w) and height (h) are separated into Hr x Wr subdivisions, with each roughly measuring h/Hr x w/Wr. Following this subdivision, the program performs maximum pooling on the RROI. Max pooling is used to find the largest or greatest value in a region, allowing the results to be downsampled.

**3.1.3.2 Text Orientation Detector (TOD)**

After identifying a text section, the orientated text's angle is calculated. If the angle is less than π/2 degrees, the text is rotated horizontally to facilitate text identification. Horizontal or vertical text parts may be readily identified. Vertical texts that go from bottom to top and top to bottom, on the other hand, need extra segmentation procedures.

The Text Orientation Detector (TOD) is used in the proposed model to identify whether the observed text sections are vertically orientated from bottom to top or top to bottom. The conceptual basis for TOD is derived from [37], in which the model pulls vertical, horizontal, and placement data from the Vertical Detector Network (VDN), Horizontal Detector Network (HDN) and Clue Detector Network (CDN) for Character Placement to determine text orientation. However, in this model, VDN and CDN are only considered for TOD since the vertical text parts that require orientation determination are either bottom_to_top or top_to_bottom.

In TOD, the VDN is responsible for fundamental feature maps to be encoded to vertical feature vectors. As seen in Fig.5, the VDN, before conducting direct down sampling, rotates the vertical feature map with 5 shared conv blocks. The feature sequence is then further encoded using Bidirectional Long Short-Term Memory (BiLSTM).

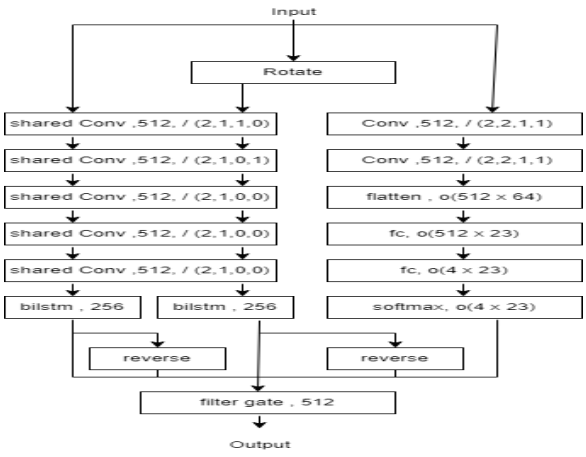


**Fig. 5:** Text Orientation Detector (TOD)

To obtain the inverted feature sequence [37], a reverse procedure is then performed. This reversal has an indirect effect on CDN training, which accelerates the process. Additionally, In Fig.5, the right-side blocks depict the extraction process of hints in the VDN.

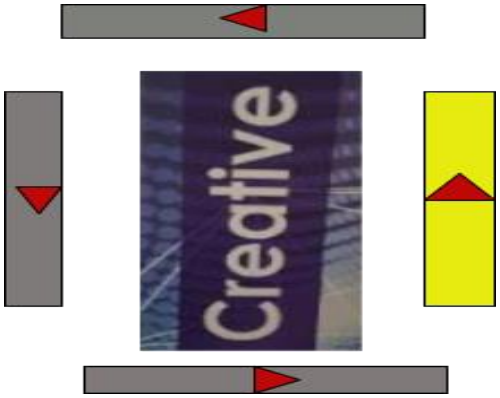


**Fig. 6:** Example showing the direction of the text

As a result, VDN is in charge of creating the feature sequence, while CDN is learning the weights that will guide the development of the final feature sequence. In the last stage, a filter gate is used to remove unnecessary features from the feature sequences acquired from VDN and CDN using a weight-sum operation. Fig.6 depicts the eventual decision of the vertical text direction, with a yellow bar with a brown arrow encircling the shown pieces, indicating bigger weights and providing insight into the text direction.

After establishing the text orientation, the text sections are rotated at - π/2 degrees and π/2 degrees. This orientation makes the text easier to read as horizontal text. GDAAM approach recognises the text regions after being translated into a horizontal orientation.

3.1.3 Relational Attention Mechanism:

Leveraging dot products, this mechanism calculates the similarity or likeness of visual and textual components. It plays a crucial role in capturing relationships between different aspects of the input data.

To seamlessly integrate visual and verbal information, we introduce the groundbreaking Gated Dual Adaptive Attention Mechanism [38]. This innovative module enhances the model's ability to integrate visual patterns with linguistic representations effectively within the encoder–decoder architecture.

We describe a single neural component-based STR system without a preset vocabulary. The decoder comprises a Language Model (LM) designed to process contextual data, which facilitates the connection between character and visual embeddings. The encoder extracts 2D visual patterns using an MS-RCNN, PBTPN and TCN with ResNet 50 as its backbone. A GDAAM is proposed, which integrates linguistic dependencies with visual signals, considerably enhancing recognition accuracy. The recommended architecture is adaptable, with options for both encoder and decoder components. During training, the model is concurrently trained with the teacher-forcing approach, which results in faster convergence if used a pre trained LM.

GDAAM is a novel framework for accurate text recognition in NSIs. It combines encoders such as MS-RCNN, PBTPN and TCN, as well as character awareness, semantic reasoning, and a visual-semantic ensemble [39] fusion decoder. GDAAM's encoder component employs two robust architectures: MS-RCNN, PBTPN, and TCN. Because of its excellent object recognition capabilities, which enable precise text localization inside scene images, we are using MS-RCNN. PBTPN and TCN recognize temporal relationships and contextual information in images, GDAAM merges these encoders to collect entire information from spatial and temporal dimensions, enabling for effective representation of text components.

It combines the GDAAM into its decoder to allow for fine-grained attention modelling. This method allows the model on preferentially emphasize important visual and textual signals while dynamically altering attention weights to the input. GDAAM uses gated processes that successfully combine visual and textual data, improving text recognition accuracy in tough NSIs. Another important part of GDAAM is its semantic reasoning component. The reasoning module combines contextual information, allowing the model to participate in thinking while making sound judgments. GDAAM focuses on important visual and textual cues, using attention processes to increase comprehension and text recognition. Character awareness is stressed by GDAAM while dealing with complicated text layouts, inconsistencies, and occlusions that are frequent in NSIs.

The flexible GDAAM module, which may be adjusted to provide additional text recognition outputs, demonstrates the STR architecture's adaptability, as discussed in [33]. Future tests are intended to validate this capacity. The existing framework may be improved as a meta-algorithm in a variety of ways. Exploring a broader spectrum of visual representation strategies, has the potential to improve recognition accuracy. Furthermore, using a bidirectional Language Model, such as BERT, has the potential to increase language dependency.

### 4.1 Experiment

#### 4.1.1 Datasets

As previously stated, research into the detection of VOST has been restricted, and there is currently a scarcity of datasets particularly targeted for assessing such texts in natural scene photos. In response to this need, the Vertical Text Images Gathered 500 Dataset (VTIG-500) was created for study and assessment. This dataset includes 500 images of VOST, including stacked, top_to_bottom, bottom_to_top.. The dataset is divided into 350 training and 150 test photos. Fig.7 shows photos with vertically aligned scene captions.


**Fig.7:** Vertical Text Image Samples

In addition to the VTIG 500 dataset, the Vertical Text Classifier and Detector with Gated Dual Adaptive Attention Mechanism performance is assessed against benchmark datasets, especially the ICDAR 2013 dataset [40] and the ICDAR 2015 dataset [41]. The ICDAR 2013 dataset has 229 training and 233 testing pictures. The ICDAR 2015 dataset is split into 1,000 training pictures and 500 testing images.

Three quantitative measurements are used to evaluate STD: recall, precision, and f-measure. The precision quantifies the algorithm's confidence by measuring the proportion of accurately recognized text sections in comparison to the ground truth.

#### 4.1.2. Results and Comparisons

The Vertical Text Classifier and Detector module was introduced in the first part of this research endeavour, and the GDAAM was offered in the second. Finally, the study resulted in the invention of the Vertical Text Classifier and Detector with GDAAM, which performs end-to-end text detection and identification for vertically oriented scenes.

Precision, Recall, and F1 Score are essential performance measures used in machine learning to evaluate a classification model's success, particularly in binary classification tasks.

Precision:

$$\text{Precision (P)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

Recall:

$$\text{Recall (R)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

F1-Score:

$$\text{F1 Score(F)} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Rate Of Recognition: is an efficiency metric utilized to assess the accuracy of text recognition models. It is described as the % of correctly recognised words among all the words in a dataset. ROR is calculated as follows:

$$\text{Rate of Recognition} = \frac{\text{No.of Recognised words}}{\text{Total No. Of words}} * 100$$

ROR is often used to assess the efficacy of text recognition models on datasets that contain a large number of words, like STR datasets. It is a more comprehensive metric than character-level accuracy, as it takes into account the entire recognized word rather than just individual characters.

**Table 1:** The comparison of the performance of various methods on the IC13 dataset with respect to P, R, and F for vertical text STD and STR.

| Method | Comparison with existing Methods on IC13 of Vertical STD and STR | | | |
|---|---|---|---|---|
| | Detection | | | Recognition |
| | P | R | F | Rate of Recognition |
| Proposed Method | 0.93 | 0.92 | 0.92 | 0.86 |
| Vertical text Interpreter [34] | 0.91 | 0.88 | 0.87 | 0.84 |
| Wang et al., 2020 [42] | 0.89 | 0.85 | 0.87 | 0.84 |
| Qiao et al., 2020 [43] | 0.92 | 0.88 | 0.90 | 0.84 |
| Lyn et al., 2018 [44] | 0.95 | 0.88 | 0.91 | 0.86 |
| Liao et al., 2018 [8] | 0.74 | 0.86 | 0.80 | 0.85 |
| He et al., 2018[45] | 0.91 | 0.88 | 0.90 | 0.86 |

**Table 2:** The comparison of the performance of various methods on the IC15 dataset with respect to P, R, and F for vertical text STD and STR.

| Method | Comparison with existing Methods on IC15 of Vertical STD and STR | | | |
|---|---|---|---|---|
| | Detection | | | Recognition |
| | P | R | F | Rate of Recognition |
| Proposed Method | 0.91 | 0.87 | 0.89 | 0.65 |
| Vertical text Interpreter [34] | 0.85 | 0.83 | 0.84 | 0.62 |
| Wang et al., 2020 [42] | 0.88 | 0.87 | 0.88 | 0.64 |
| Qiao et al., 2020 [43] | 0.91 | 0.81 | 0.86 | 0.63 |
| Lyn et al., 2018 [44] | 0.86 | 0.81 | 0.86 | 0.62 |
| Liao et al., 2018 [8] | 0.87 | 0.77 | 0.82 | 0.52 |
| He at al., 2018 [45] | 0.83 | 0.84 | 0.83 | 0.63 |

**Table 3:** The comparison of the performance of various methods on the VTIG 500 dataset with respect to P, R, and F for vertical text STD and STR.

| Method | Comparison with existing Methods on VTIG 500 of Vertical STD and STR | | | |
|---|---|---|---|---|
| | Detection | | | Recognition |
| | P | R | F | Rate of Recognition |
| Proposed Method | 0.87 | 0.77 | 0.82 | 0.69 |
| Vertical text Interpreter [34] | 0.74 | 0.86 | 0.80 | 0.62 |
| Capture 2 Text [29] | | | Not applicable | 0.15 |

The Vertical Text Classifier and Detector were trained and validated with the VTIG 500 dataset, that is specifically built for VOST. The findings were documented and shown in Table 2. According to the table, the Vertical Text Classifier and Detector perform well. The model achieves 87% precision (P), 77% recall (R), and 82% f-measure (F) with an accuracy of 69%.

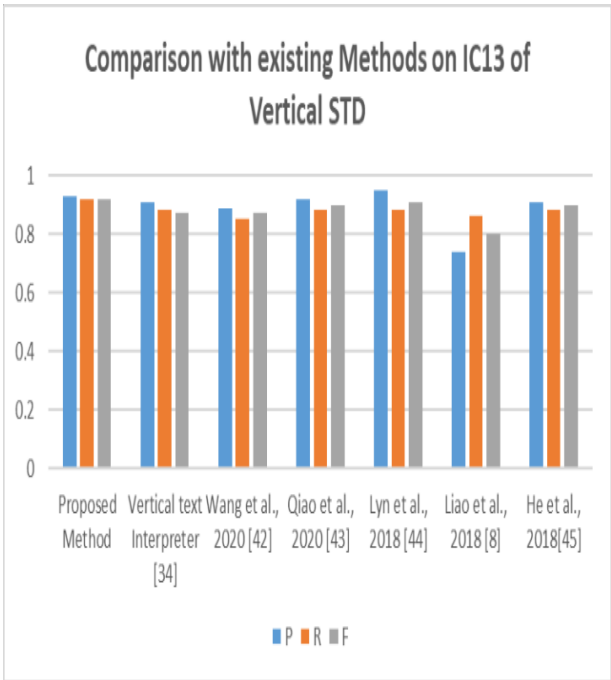


**Fig. 8:** Comparison with existing Methods on IC13 of Vertical STD

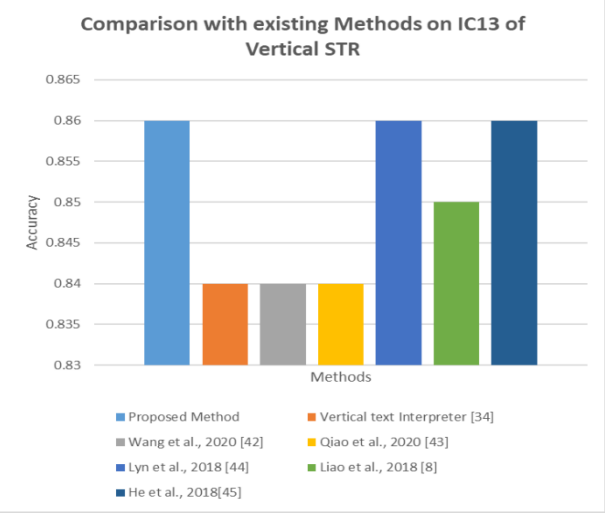


**Fig. 9**: Comparison with existing Methods on IC13 of Vertical STR

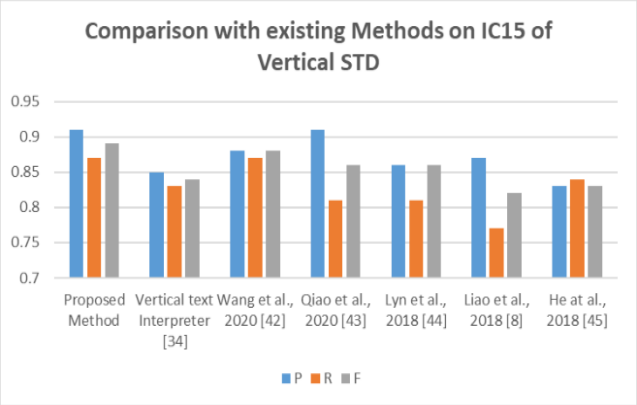


**Fig. 10:** Comparison with existing Methods on IC15 of Vertical STD

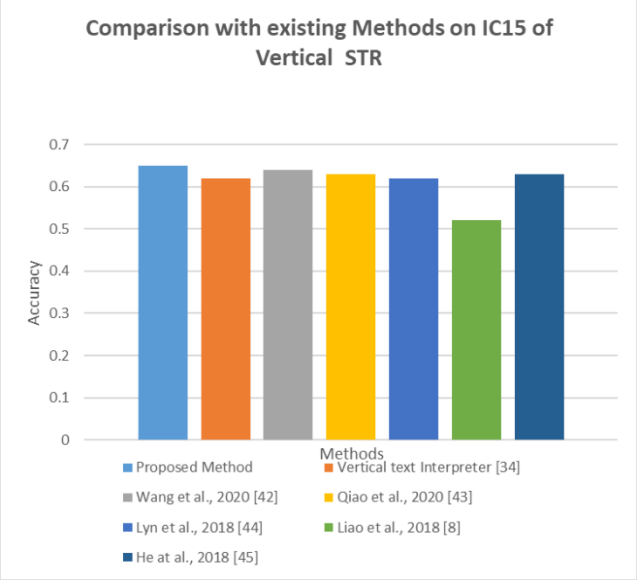


**Fig. 11:** Comparison with existing Methods on IC15 of Vertical STR

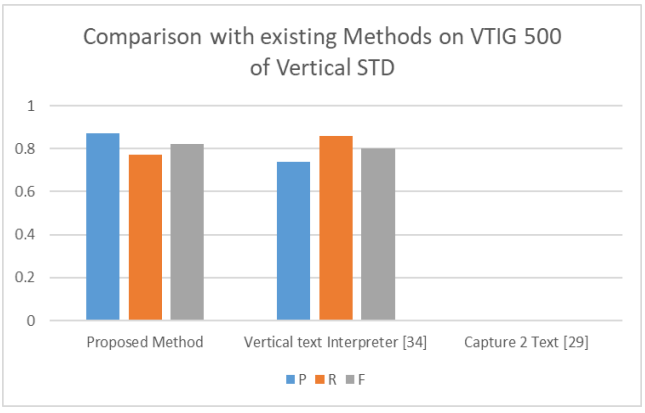


**Fig. 12:** Comparison with existing Methods on VTIG500 of Vertical STD

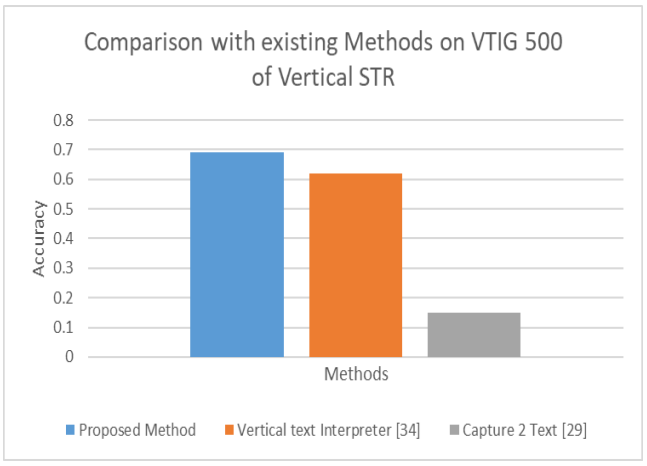


**Fig. 13:** Comparison with existing Methods on VTIG500 of Vertical STR

### 4.1.3: Experimental Results

Our method performance is compared with previous methods, which have the same pipeline of combining detection and recognition of text. We compared our method with the [34],[42],[43],[44][8] and [45] on the recognising Scene text performance. Fig.14 shows the examples of various kinds of Vertically oriented texts successfully recognised.


**Fig. 14:** some examples which are successfully recognised.

### 5,1 Conclusion

Recognizing vertical writing in NSI is necessary for accessing information in natural situations. As a result, this study proposes a model for identifying and recognizing various forms of vertical text in images of natural scenes. The suggested model uses TDC for text detection and GDAAM for vertical STR. TOD is used during the procedure to distinguish different forms of vertical messages. As a result, the suggested DL model can detect texts in a variety of orientations, including vertical scene text. Different forms of vertical text are detected and recognized through training and testing. The generated outcomes are assessed using the metrics suggested. This suggested model is anticipated to identify and recognize rotating texts, horizontal texts, and many sorts of vertical texts in natural settings, such as top_to_bottom, bottom_to_top, and horizontally stacked.

**References**


[1] Ahmed, Abdullah Khalid. "Signage recognition based wayfinding system for the visually impaired." (2015).

[2] M. Liao, B. Shi, and X. Bai, "Textboxes++: A single-shot oriented scene text detector," IEEE Transactions on Image Processing, vol. 27, no. 8, pp. 3676-3690, 2018.

[3] Mishra, Anand. "Understanding Text in Scene Images." International Institute of Information Technology Hyderabad (2016).

[4] Lin, Han, Peng Yang, and Fanlong Zhang. "Review of scene text detection and recognition." Archives of computational methods in engineering 27, no. 2 (2020): 433-454.

[5] Ye, Qixiang, and David Doermann. "Text detection and recognition in imagery: A survey." IEEE transactions on pattern analysis and machine intelligence 37, no. 7 (2014): 1480-1500.

[6] Yuliang, Liu, Jin Lianwen, Zhang Shuaitao, and Zhang Sheng. "Detecting curve text in the wild: New dataset and new solution." arXiv preprint arXiv:1712.02170 (2017).

[7] C. Yao, X. Bai, and W. Liu, "A unified framework for multi-oriented text detection and recognition," IEEE Transactions on Image Processing, vol. 23, no. 11, pp. 4737-4749, 2014.

[8] Chen, Yilin, and Juan Yang. "Research on scene text recognition algorithm basedon improved CRNN." In Proceedings of the 2020 4th International Conference on Digital Signal Processing, pp. 107-111. 2020.

[9] Ling, Ong Yi, Lau Bee Theng, Almon Chai, and Chris McCarthy. "A model for automatic recognition of vertical texts in natural scene images." In 2018 8th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), pp. 170-175. IEEE, 2018.

[10] Tang, Jun, Zhibo Yang, Yongpan Wang, Qi Zheng, Yongchao Xu, and Xiang Bai. "Seglink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping." Pattern recognition 96 (2019): 106954.

[11] Ong, Yi Ling, Bee Theng Lau, Almon Chai, and Chris McCarthy. "A deep learning framework for recognizing vertical texts in natural scene." In 2019 International Conference on Computer and Drone Applications (IConDA), pp. 48-53. IEEE, 2019.

[12] Y. Zhu, C. Yao, and X. Bai, "Scene text detection and recognition: Recent advances and future trends," Frontiers of Computer Science, vol. 10, no. 1, pp. 19-36, 2016.

[13] Huang, Weilin, Zhe Lin, Jianchao Yang, and Jue Wang. "Text localization in natural images using stroke feature transform and text covariance descriptors." In Proceedings of the IEEE international conference on computer vision, pp. 1241-1248. 2013.

[14] He, Tong, Weilin Huang, Yu Qiao, and Jian Yao. "Text-attentional convolutional neural network for scene text detection." IEEE transactions on image processing 25, no. 6 (2016): 2529-2541.

[15] Neumann, Lukáš, and Jiří Matas. "Real-time scene text localization and recognition." In 2012 IEEE conference on computer vision and pattern recognition, pp. 3538-3545. IEEE, 2012.

[16] Jaderberg, Max, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Reading text in the wild with convolutional neural networks." International journal of computer vision 116 (2016): 1-20.

[17] Huang, Weilin, Yu Qiao, and Xiaoou Tang. "Robust scene text detection with convolution neural network induced mser trees." In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13, pp. 497-511. Springer International Publishing, 2014..

[18] Yao, Cong, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. "Detecting texts of arbitrary orientations in natural images." In 2012 IEEE conference on computer vision and pattern recognition, pp. 1083-1090. IEEE, 2012.

[19] Li, Yao, Wenjing Jia, Chunhua Shen, and Anton van den Hengel. "Characterness: An indicator of text in the wild." IEEE transactions on image processing 23, no. 4 (2014): 1666-1677..

[20] Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580-587. 2014.

[21] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems 28 (2015).

[22] Liu, Wei, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. "Ssd: Single shot multibox detector." In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pp. 21-37. Springer International Publishing, 2016.

[23] Gupta, Ankush, Andrea Vedaldi, and Andrew Zisserman. "Synthetic data for text localisation in natural images." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2315-2324. 2016.

[24] Yi, Chucai, and Yingli Tian. "Localizing text in scene images by boundary clustering, stroke segmentation, and string fragment classification." IEEE transactions on image processing 21, no. 9 (2012): 4256-4268.

[25] Tian, Zhi, Weilin Huang, Tong He, Pan He, and Yu Qiao. "Detecting text in natural image with connectionist text proposal network." In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14, pp. 56-72. Springer International Publishing, 2016.

[26] Zhou, Xinyu, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. "East: an efficient and accurate scene text detector." In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 5551-5560. 2017.

[27] Shi, Baoguang, Xiang Bai, and Cong Yao. "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition." IEEE transactions on pattern analysis and machine intelligence 39, no. 11 (2016): 2298-2304.

[28] Choi, Chankyu, Youngmin Yoon, Junsu Lee, and Junseok Kim. "Simultaneous recognition of horizontal and vertical text in natural images." In Computer Vision–ACCV 2018 Workshops: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers 14, pp. 202-212. Springer International Publishing, 2019. [29] Capture2Text. "Capture2Text." http://capture2text.sourceforge.net/ (accessed 15 August, 2018)

[30] Dehghani, Mostafa, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. "Universal transformers." arXiv preprint arXiv:1807.03819 (2018).

[31] Li, Hui, Peng Wang, Chunhua Shen, and Guyu Zhang. "Show, attend and read: A simple and strong baseline for irregular text recognition." In Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, pp. 8610-8617. 2019.

[32] Cheng, Zhanzhan, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. "Focusing attention: Towards accurate text recognition in natural images." In Proceedings of the IEEE international conference on computer vision, pp. 5076-5084. 2017.

[33] A.S.Venkata Praneel, et al. 2023. “Text Detection Using Transformation Scaling Extension Algorithm in Natural Scene Images”. International Journal on Recent and Innovation Trends in Computing and Communication 11 (10):1233-44. https://doi.org/10.17762/ijritcc.v11i10.8664.

[34] Ling, Ong Yi, Lau Bee Theng, Almon Chai Weiyen, and Christopher Mccarthy. "Development of vertical text interpreter for natural scene images." IEEE Access 9 (2021): 144341-144351.

[35] Ma, Jianqi, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. "Arbitrary-oriented scene text detection via rotation proposals." IEEE transactions on multimedia 20, no. 11 (2018): 3111-3122.

[36] Ling Ong, Yi, Bee Theng Lau, Almon Chai, and Chris McCarthy. "Detecting of vertically-oriented texts in images containing natural scenes." In MobiQuitous 2020-17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, pp. 444-450. 2020.

[37] Cheng, Zhanzhan, Yangliu Xu, Fan Bai, Yi Niu, Shiliang Pu, and Shuigeng Zhou. "Aon: Towards arbitrarily-oriented text recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5571-5579. 2018.

[38] Praneel, AS Venkata, and T. Srnivasa Rao. "Gated Dual Adaptive Attention Mechanism with Semantic Reasoning, Character Awareness, and Visual-Semantic Ensemble Fusion Decoder for Text Recognition in Natural Scene Images." International Journal of Intelligent Systems and Applications in Engineering 12, no. 1 (2024): 221-234.

[39] Venkata Praneel, A. S., T. Srinivasa Rao, and M. Ramakrishna Murty. "A survey on accelerating the classifier training using various boosting schemes within cascades of boosted ensembles." In Intelligent Manufacturing and Energy Sustainability: Proceedings of ICIMES 2019, pp. 809-825. Springer Singapore, 2020..

[40] Karatzas, Dimosthenis, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. "ICDAR 2013 robust reading competition." In 2013 12th international conference on document analysis and recognition, pp. 1484-1493. IEEE, 2013.

[41] Karatzas, Dimosthenis, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas et al. "ICDAR 2015 competition on robust reading." In 2015 13th international conference on document analysis and recognition (ICDAR), pp. 1156-1160. IEEE, 2015.

[42] Wang, Hao, Pu Lu, Hui Zhang, Mingkun Yang, Xiang Bai, Yongchao Xu, Mengchao He, Yongpan Wang, and Wenyu Liu. "All you need is boundary: Toward arbitrary-shaped text spotting." In Proceedings of the AAAI conference on artificial intelligence, vol. 34, no. 07, pp. 12160-12167. 2020.

[43] Qiao, Liang, Sanli Tang, Zhanzhan Cheng, Yunlu Xu, Yi Niu, Shiliang Pu, and Fei Wu. "Text perceptron: Towards end-to-end arbitrary-shaped text spotting." In Proceedings of the AAAI conference on artificial intelligence, vol. 34, no. 07, pp. 11899-11907. 2020.

[44] Lyu, Pengyuan, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai. "Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes." In Proceedings of the European conference on computer vision (ECCV), pp. 67-83. 2018.

[45] He, Tong, Zhi Tian, Weilin Huang, Chunhua Shen, Yu Qiao, and Changming Sun. "An end-to-end textspotter with explicit alignment and attention." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5020-5029. 2018.