

Polygon-Aware Detection of Arbitrary-Shaped Text in Natural Scenes

Abdul Haadi

Haadi1495@gmail.com

Department of CS&IT, MUST

Aeman Naeem

aemennaem@gmail.com

Department of CS&IT, MUST

Sobia Sher

sobia.csit@must.edu.pk

Department of CS&IT, MUST

Sania Khadim

Sania.csit@must.edu.pk

Department of CS&IT, MUST

Corresponding Author: Abdul Haadi Haadi1495@gmail.com

Received: 13-02-2026

Revised: 06-03-2026

Accepted: 14-03-2026

Published: 26-03-2026

ABSTRACT

Scene text detection in natural images is a topic of significant discourse as it plays a vital role in several applications. Its uses can range from autonomous navigation to assistive technologies to augmented reality. Traditionally, pipelines for text detection relied on rectangular or rotated bounding boxes to represent text, but this method has shown limitations in capturing irregular shapes found in real-world environments. This paper represents a framework designed to address this challenge. The framework uses polygonal boxes instead of rectangular ones. This polygon-aware approach leads to more accurate localization of complex text shapes and structures. This system combines polygon boundary estimation with deep convolutional feature extraction to identify text instances across diverse shapes and orientations. The framework is evaluated on three widely adopted curved-text benchmarks - Total-Text, CTW1500, and ICDAR. The polygon-based representation yields measurable gains in detecting curved and irregular text. The approach delivers better accuracy without sacrificing computational efficiency, making it viable for real-world scene text detection.

Keywords: Scene Text Detection, Curved Text Detection, Polygon Representation, Deep Learning, Computer Vision

INTRODUCTION

Text detection has been widely noticed over the past decade, as it has many applications, including automatic translation, augmented reality, autonomous cars, vision-based mobile assistance, and creating document images from scenes, etc. [1]. Scene text detection is the initial step in reading text from images, and it helps in identifying the text present in images taken in real-world situations [2].

Although there's been significant progress in deep learning-based detection models, text that is curved or has an odd shape is still very hard to deal with [3]. Most object detection models use rectangular region proposals [4]. They work fine with straight or slightly tilted texts, but curved text lines, artistic fonts, decorative fonts such as logos or emblem form, are usually very hard to work with.

In practical life, text merely appears in straight lines. Instead, it usually appears in different shapes, i.e., curves, circles, and waves. For these shapes, traditional methods like rectangular methods don't accurately show the actual boundaries [4]. Thus, leading to poor localization and overlapping detection.

To overcome these limitations, newer methods are proposed. Polygon-based representation is one of them. Text can be represented using a polygon, which has multiple points to represent the text. This way, we can capture complex texts more easily. Even after all the advancements, detecting text still remains difficult due to: Complicated and complex variations of text shapes [5]. Text size and orientation can be very different [6]. Messy backgrounds and different lighting conditions [7], Long lines of text with uneven edges [2].

To address this issue, this paper proposes a polygon-aware text detection methodology capable of handling both standard rectangular texts as well as curved or irregular geometries of real-world scene text.

Key contributions of this work

- We present a polygon-aware detection approach for modeling arbitrarily shaped text in real-world situations before OCR.
- A flexible representation of text instances using polygonal boundaries instead of rectangular boxes.
- The proposed method not only works better on curved detection tasks, but it also works better when tested on the Total-Text, CTW1500, and ICDAR datasets.

The approach has competitive performance and runs efficiently.

Related Work

Traditional Scene Text Detection

Mostly the initial method of scene text detection is depending on handcrafted features and simple rules [8]. One of the most popular techniques out of them was Stroke Width Transform (SWT) [9]. That identify the consistent width stroke inside the character to detect the text this method work properly in simple and clean layout, but when the background is complex, or the text is curved, its performance goes down. Similarly, Maximally Stable Extremal Regions (MSER) located candidate text components by isolating stable image regions, though its sensitivity to background noise limited its effectiveness on curved and irregular text [10].

Deep Learning-Based Detection Methods

As the deep learning advanced, convolutional neural networks (CNNs) also significantly enhanced scene text detection performance [11].

CTPN is a model that use anchors or RNN to find the text, and it detect the lines of text correctly, but mostly it understands horizontal text [12]. EAST is a fast and simple method that detect the text by making the boxes around them. It balanced both speed and accuracy correctly, but when the text is curved and crooked, it does not work properly [13]

Segmentation-Based Modeling

Segmentation-based methods give you freedom to model text shapes in a flexible way by finding text at the pixel level [14].

Pixel link connects pixels that are next to each other and belong to the same text category. This allows you to separate text areas that are next to each other [15]. PSENet (Progressive Scale Expansion Network) slowly adds text to kernels to separate text categories that are close to each other [16]. DBNet (Differentiable Binarization Network) is a new architecture for text detection that improves text detection accuracy through differentiable binarization to better segment out the text [17].

Polygon-Based representations:

Polygon-based detection methods provide greater flexibility in modelling randomly-shaped texts.

Text Snake extracts text cases as a series of disks, which are centered at the text location in the scene, which can handle curved text and a huge variety of orientations [18]. CRAFT (Character Region Awareness for Text Detection) detects individual character regions and group them into text categories [19].

These methods demonstrate that by using polygon-like representations, it is possible to get better detection results for curved text.

PROPOSED METHOD

System Overview

The proposed method uses a segmentation-based and polygon-focused approach to detect text in irregularly shaped natural images. Firstly, the Image is taken as input to extract multiple features at different levels. Then the model identifies which text belongs to the image at the pixel level and constructs polygon boundaries around the detected regions.

The system consists of three main stages

- i. Feature Extraction using (ResNet-50 + FPN)
- ii. Text Region Detection at pixel level
- iii. Enhancement in polygon boundaries

This method improves the detection performance of text with irregular shapes in real-world scenes.

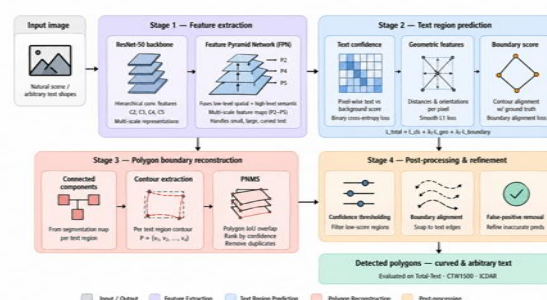


Fig. 1. Overview of the proposed polygon-aware text detection framework

Feature Extraction

For the extraction of hierarchical feature representations from the input image, the system utilizes ResNet-50 as the backbone network. The Feature Pyramid Network (FPN) is integrated into the backbone to process multi-scale text images.

The FPN integrates low-level spatial features with high-level semantic representations to facilitate effective text detection.

As a result, the system can perform

- i. Reliable Detection of small-scale and distant text regions
- ii. Effective Representation of long and curved text structures
- iii. Stability across different text scales and orientations.

Text Region Prediction

The proposed system formulates text detection as a pixel-wise segmentation problem. The model generates a pixel-wise prediction map where each pixel can be categorized as either text or background.

For each pixel, the model predicts:

- i. Text confidence score
- ii. Boundary score
- iii. Geometric Features

This pixel-wise prediction enables the model to distinguish text instances of varying scale and effectively handle complex layouts.

Polygon Boundary Reconstruction

The new method utilizes polygonal boundaries as opposed to rectangular or quadrilateral boundary boxes in representing text regions.

Vertices form a set that will be used to represent polygons.

$$P = \{v_1, v_2, \dots, v_n\}$$

Each vertex $v_i = (x_i, y_i)$ is defined by its point on the text boundary.

Polygon generation follows the steps below

- i. Identifying Connected components in the segmentation map
- ii. Detecting the Contour in each text region
- iii. Transforming Contour into polygonal forms

With this technique, an accurate representation of the text is achieved in three main text types: irregular word shapes, curved lines, and long text sequences.

Polygon Non-Maximum Suppression (PNMS)

To prevent duplicate detections, the system introduces PNMS in addition to the classical Non-Maximum Suppression (NMS). The Polygon-based method differs from the standard rectangle-based Intersection over Union (IoU) as PNMS calculates the overlap of polygons. This helps achieve accurate text detection.

This algorithm includes

1. Polygons are ranked according to their confidence scores.
2. Calculation of Polygon IoU to measure the overlap
3. Eliminating overlapping polygons with an IoU above the predetermined threshold

This removes redundancy in predictions and improves accuracy in text detection. The PNMS procedure evaluates polygon predictions based on their confidence scores and diminishes significantly overlapping polygons according to polygon IoU.

Post-Processing and Boundary Refinement

The Proposed framework performs some additional refinement steps to enhance the quality of detection:

1. Removal of low-confidence polygons using a confidence threshold. Following PNMS, the remaining polygons undergo additional refinement to improve the accuracy of their placements and boundaries
2. Enhancement of Polygon boundaries to better align with text contours
3. Elimination of inaccurate predictions. This stage increases the accuracy of the final prediction findings by eliminating erroneous detections.

These steps improve the overall quality of final predictions by producing effective polygon boundaries and reducing detection errors. In general, the refinement procedure aids in achieving more accurate text region localization in photos of natural scenes.

Boundary refinement proceeds through a fixed sequence of heuristics applied to each surviving polygon after PNMS. First, a confidence threshold discards low-confidence polygons that are unlikely to correspond to true text instances. Second, boundary alignment snaps each remaining polygon's vertices toward the nearest high-gradient edge pixels in the original image, correcting minor localization drift introduced during pixel-wise segmentation. Third, a geometric consistency check removes predictions whose vertex ordering or shape is implausible for text (e.g., self-intersecting polygons or extreme aspect ratios), treating these as false positives. This sequence is applied once per detected instance and adds negligible computational overhead, since it operates only on the small set of polygons that survive PNMS rather than on the full pixel map.

Loss Function

The proposed system employs a multi-task loss function to train the model effectively:

$$L_{total} = L_{cls} + \lambda_1 L_{geo} + \lambda_2 L_{boundary} \quad (1)$$

Text Classification Loss

Binary Cross-Entropy (BCE) loss is applied for pixel-wise classification of text and background.

Geometry Loss

Smooth L1 loss is used for the prediction of geometric attributes such as distances and orientations. This ensures robust and stable model learning.

Boundary Alignment Loss

A boundary loss is introduced to reduce the difference between predicted polygon contours and ground truth shapes, enhancing the accuracy of polygon fitting, especially for curved text regions.

A combination of all these losses helps the model to understand the region localization and precise boundary estimation.

EXPERIMENTS

Datasets

Total-text

A total-text data set is a data set where the data inside the images appears in different forms, like straight, tilted, or curved. The most important thing about the text is that the boxes are labeled as polygon shapes instead of simple boxes, which shows real-life scenes more accurately. It consists of almost 1500 images whose background is very complex, and which reflect the real-world conditions [20]

CTW1500

The CTW1500 data set is specially used in those cases where the text length is very long and in a curved shape. In this data set almost 1500 images are included and every text instance is defined as detailed polygon annotation. This data set is helpful for understanding and processing the irregular and curved text [16], [21].

ICDAR Dataset

The ICDAR data-set is used as a reference in research. In this work, ICDAR2015 or ICDAR2017-MLT data sets are included. In future proposed method can be test on further different framework like different languages and script. This data-set consists of text in different languages that is present in different orientations and in difficult backgrounds, in which the performance of models can be evaluated [22]. Polygon annotation represents the boundaries of text region and very useful to accurately show the shape of irregular and curved text.

Evaluation Metrics

To measure the efficiency of a model, some important evaluation parameters are used that are commonly applied in text detection [21]

Precision

shows how many areas are correct out of the total areas that the model detected. Precision measures the proportion of correctly detected text regions among all predicted detection.

Formula

Precision = $TP / (TP + FP)$. TP: means correct detection,

FP: means false detection

Recall

Recall shows the amount of data that the model successfully detected from the total available text. It gives the idea of model coverage.

Formula

Recall = $TP / (TP + FN)$, where FN is a text that cannot be detected.

H-mean

H-mean is a balanced metric that combines both recall and precision and give their average type of result and idea of its overall performance. $H\text{-mean} = (2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$

IoU Threshold

IoU Threshold is a technique that measures the overlap between actual and predicted text regions. A detection is considered correct if $IoU \geq 0.5$.

All reported metrics follow the standard ICDAR/DetEval evaluation protocol for scene text detection, using a polygon-IoU matching threshold of 0.5, consistent with prior work (EAST, CRAFT, DBNet) to ensure fair comparison. Implementation code and pretrained model weights will be made publicly available upon acceptance to support reproducibility.

Annotation Considerations

Polygon-based annotation introduces considerations not present in axis-aligned bounding-box labeling. Total-Text and CTW1500 annotate each text instance with a variable number of boundary vertices (typically 4–14 points depending on curvature), meaning ground-truth polygons are not fixed-length and must be resampled or padded for batch training. Annotator agreement on vertex placement is inherently less consistent than for rectangular boxes, since curved boundaries admit multiple reasonable vertex configurations for the same instance. To mitigate this, evaluation follows each dataset's standard polygon-IoU matching protocol rather than exact vertex correspondence, so minor annotator-to-annotator variation in vertex placement does not penalize otherwise correct detections.

RESULTS AND DISCUSSION

The proposed polygon-aware text detection framework is evaluated on benchmark datasets including Total-Text, CTW1500, and ICDAR. Performance is measured using standard metrics: Precision, Recall, and H-mean.

Quantitative Results

The proposed method's generalization capability is evaluated against state-of-the-art approaches across the Total-Text, CTW1500, and ICDAR datasets. Performance comparisons are detailed in Table 1.

TABLE 1. Quantitative comparison of the proposed method with state-of-the-art approaches (east, craft, dbnet) on total-text, ctw1500, and icdar2015 benchmark datasets in terms of precision, recall, and h-mean (%).

Method	Data-set	Precision	Recall	H-mean
EAST	Total-Text	82.3	78.1	80.1
CRAFT	Total-Text	86.0	83.1	84.5
DBNet	Total-Text	87.3	84.2	85.7
Proposed	Total-Text	88.4	85.9	87.1
EAST	CTW1500	86.3	83.4	84.8
CRAFT	CTW1500	82.0	81.2	81.6
DBNet	CTW1500	81.1	85.1	83.0
Proposed	CTW1500	86.6	87.6	87.1
EAST	ICDAR2015	78.2	73.2	75.62
CRAFT	ICDAR2015	83.2	80.2	81.7
DBNet	ICDAR2015	84.0	81.2	82.6
Proposed	ICDAR2015	84.5	83.2	83.8

TABLE 2. Average H-mean (%) comparison of the proposed framework with baseline methods across benchmark datasets

Method	Average H-mean
EAST	80.21
CRAFT	82.60
DBNet	83.78
Proposed	86.02

The proposed method consistently outperforms the baseline models across all datasets.

On Total-Text, the model exhibits good performance due to the effective polygon representation of curved text. The best improvement is achieved on the CTW1500 data-set, where the capabilities of the model in processing long and curved text lines are observed. The effectiveness and robustness of multilingual and multi-oriented text on the ICDAR datasets indicate a high level of generalization. Therefore, it can be concluded that for all the real-world cases, the polygon representation demonstrates superior detection performance.

Performance Analysis

The proposed approach's success and better performance can be explained by the following factors:

Polygon Representation

The novel approach represents texts with polygons, which allows the model to have proper localization and reduce noise compared with the traditional box representation. This way, localization greatly improves accuracy.

Multi-scale Feature Learning

Due to the implementation of FPN, text detection is possible at several scales, and therefore, detection becomes much easier for small-sized and long curved text lines.

Segmentation-based Detection

The possibility of pixel-wise prediction enables accurate localization and, consequently, a good distinction between nearby texts.

Qualitative Results

The proposed approach exhibits high efficiency when dealing with:

1. Curved text in logos and banners
2. Arbitrary-shaped text in natural scenes
3. Text under complex backgrounds and illumination conditions

The polygon-based representation gives tighter and more accurate boundaries compared to traditional rectangular approaches.

Limitations

The following are some of the limitations of the proposed approach:

1. -Slight inaccuracies in densely packed text regions
2. -Reduced performance on very low-resolution images
3. -Increased complexity compared to simple bounding-box detectors

These limitations highlight areas for future improvement. Extended ablation across PNMS IoU thresholds, systematic failure-mode analysis, and edge-device runtime bench-marking are ongoing and will be reported in an extended version of this work.

Computational Efficiency

The proposed method maintains reasonable computational efficiency due to:

1. Use of a lightweight ResNet-50 backbone
2. Efficient multi-scale feature sharing via FPN

Optimized PNMS for post-processing

This makes the approach suitable for real-world applications such as mobile OCR and autonomous systems. In future lightweight backbone networks and model quantization can be testified to overcome the computational requirement of model. These methods could enhance the appropriateness of the suggested framework for mobile and embedded systems. Future endeavors will encompass assessing the inference duration and computational expenses of the suggested framework on widely utilized edge devices.

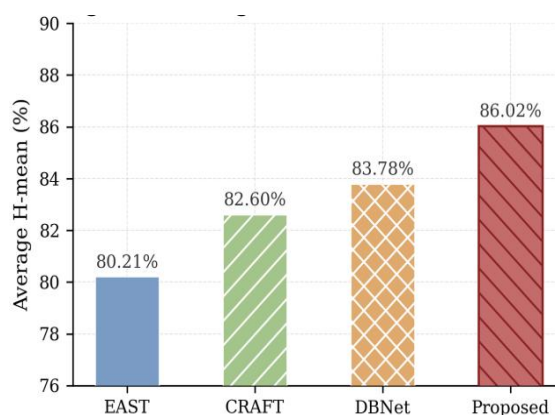


Fig. 2. Visual summary of H-mean performance reported in Table2, shows that the proposed method achieves the best overall detection accuracy among the compared approaches.

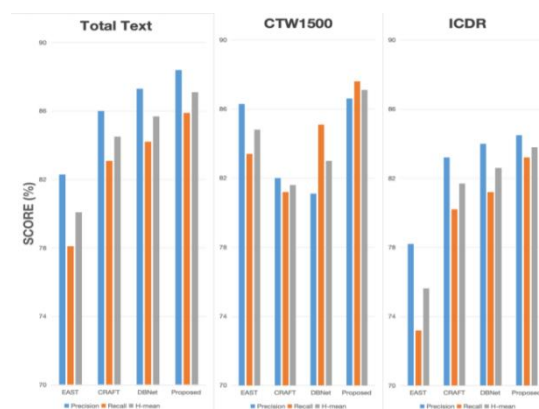


Fig. 3. Precision, Recall, and H-mean per data-set

Ablation Study

To evaluate the contribution of each component in the proposed framework, we conduct ablation experiments on the Total-Text dataset.

Effect of Feature Pyramid Network (FPN)

TABLE 3. Ablation study on the effect of Feature Pyramid Network (FPN) integration, evaluated on the Total-Text dataset.

Model Variant	Precision	Recall	H-mean
Without FPN	84.2	80.5	82.3
With FPN	88.4	85.9	87.1

FPN significantly improves performance by capturing multi-scale text features, especially for small and curved text.

Effect of Polygon Representation

TABLE 4. Ablation study comparing rectangular bounding box representation against the proposed polygon-based representation on the Total-Text dataset.

Representation	Precision	Recall	H-mean
Rectangular Boxes	82.8	79.6	81.2
Polygon-Based (Proposed)	88.4	85.9	87.1

Polygon representation improves localization accuracy for irregular and curved text.

Effect of PNMS (Post-processing)

Table 5. Ablation study on the effect of polygon non-maximum suppression (pnms) post-processing, evaluated on the total-text dataset.

Method	Precision	Recall	H-mean
Without PNMS	85.1	83.7	84.4
With PNMS	88.4	85.9	87.1

Observation

PNMS reduces duplicate detections and improves overall precision. By analyzing various polygon IoU levels to ascertain their impact on detection performance, the efficacy of PNMS may also be further investigated.

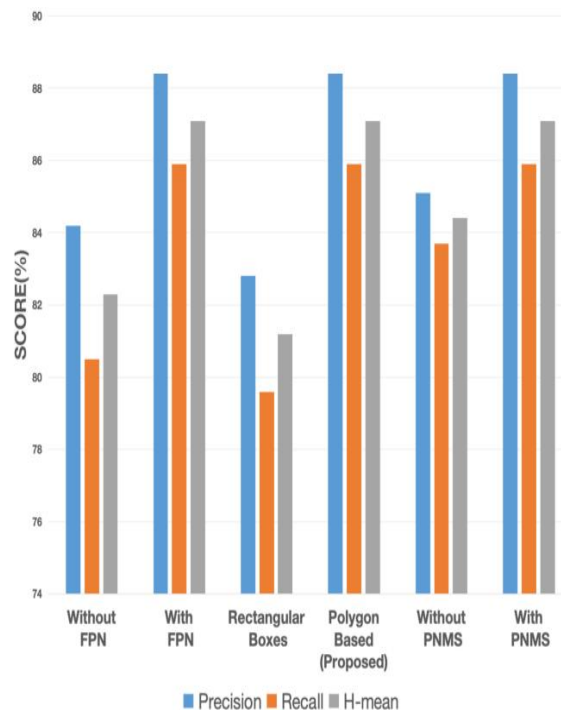


Fig. 4. Ablation Study on Total Text dataset

Conclusion of Ablation

Each component contributes significantly, with polygon modeling and FPN being the most impactful modules.

CONCLUSION

The paper presents a polygon-aware scene text detection framework which is focused on detecting curved and arbitrarily shaped text in natural images. The proposed method achieves improved localization accuracy for complex text geometries by using polygons to represent text regions instead of rectangular bounding boxes.

Experimental evaluation on Total-Text, CTW1500, and ICDAR datasets depicts the effectiveness of the proposed approach. Further work will prioritize improving computational efficiency and extending the framework to multilingual text recognition tasks.

REFERENCES

- Khan, T., Sarkar, R., & Mollah, A. F. (2021). Deep learning approaches to scene text detection: a comprehensive review. *Artif. Intell. Rev.*, 54(5), 3239–3298. <https://doi.org/10.1007/s10462-020-09930-6>
- Ghosal, S. S., Chakraborty, S., Geiping, J., Huang, F., Manocha, D., & Bedi, A. S. (2023, October). Towards Possibilities & Impossibilities of AI-generated Text Detection: A Survey. <https://arxiv.org/abs/2310.15264>
- Gugulothu, V., & Vankudoth, R. (2023, June). Deep Learning-Based Cursive Text Detection and Recognition in Natural Scene Images. *Int. J. Sci. Res. Sci. Technol.*, 969–974. <https://doi.org/10.32628/IJSRST523103161>
- Chandio, A. A., Leghari, M., Soomro, M. A., Nizamani, S. Z., & Memon, S. (2021, November). A multiscale feature fusion method for cursive text detection in natural scene images. *The Imaging Science Journal*, 69(5–8), 302–318. <https://doi.org/10.1080/13682199.2022.2160861>
- Xu, Y., Wang, Y., Zhou, W., Wang, Y., Yang, Z., & Bai, X. (2019, November). TextField: Learning a Deep Direction Field for Irregular Scene Text Detection. *IEEE Transactions on Image Processing*, 28(11), 5566–5579. <https://doi.org/10.1109/TIP.2019.2900589>
- Su, Y., et al. (2023). TextDCT: Arbitrary-Shaped Text Detection via Discrete Cosine Transform Mask. *IEEE Trans. Multimedia*, 25, 5030–5042. <https://doi.org/10.1109/TMM.2022.3186431>
- Zhou, F., Cai, Y., & Tian, Y. (2022, December). TextDC: Exploring Multidimensional Text Detection via a New Benchmark and Solution. *Electronics (Basel)*, 12(1), 159. <https://doi.org/10.3390/electronics12010159>
- Escalera, S., Baró, X., Vitrià, J., & Radeva, P. (2009). Text Detection in Urban Scenes. <https://doi.org/10.3233/978-1-60750-061-2-35>
- Naiemi, F., Ghods, V., & Khalesi, H. (2022, June). Scene text detection and recognition: a survey. *Multimed. Tools Appl.*, 81(14), 20255–20290. <https://doi.org/10.1007/s11042-022-12693-7>

- Li, R., Chen, S., Zhao, F., & Qiu, X. (2023, April). Text Detection Model for Historical Documents Using CNN and MSER. *Journal of Database Management*, 34(1), 1–23. <https://doi.org/10.4018/JDM.322086>
- Chougule, S., Dudhabale, P., & Havaladar, T. (2023, May). Deep Learning-based Text Abstraction. *Int. J. Res. Appl. Sci. Eng. Technol.*, 11(5), 6238–6244. <https://doi.org/10.22214/ijraset>. 2023.53048
- Chhabra, P., Shrivastava, A., & Gupta, Z. (2023, April). Comparative Analysis on Text Detection for Scenic Images using EAST and CTPN. In *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 1303–1308). IEEE. <https://doi.org/10.1109/ICOEI56765.2023.10125894>
- Zhou, X., et al. (n.d.). *EAST: An Efficient and Accurate Scene Text Detector*.
- Li, J., Lin, Y., Liu, R., Ho, C. M., & Shi, H. (n.d.). *RSCA: Real-time Segmentation-based Context-Aware Scene Text Detection*.
- Hu, X., Wei, B., Gao, L., & Wang, J. (2024, June). SegHist: A General Segmentation-based Framework for Chinese Historical Document Text Line Detection. <https://arxiv.org/abs/2406.15485>
- Wang, W., et al. (2019, June). Shape Robust Text Detection with Progressive Scale Expansion Network. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 9328–9337). IEEE. <https://doi.org/10.1109/CVPR.2019.00956>
- Chen, H., Liu, J., & Zhou, W. (2022, August). Natural Scene Text Detection Algorithm Based on Improved DBNet. In *2022 IEEE 5th International Conference on Electronic Information and Communication Technology (ICEICT)* (pp. 186–190). IEEE. <https://doi.org/10.1109/ICEICT55736.2022.9909509>
- Long, S., Ruan, J., Zhang, W., He, X., Wu, W., & Yao, C. (n.d.). *TextSnake: A Flexible Representation for Detecting Text of Arbitrary Shapes*.
- Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H. (n.d.). *Character Region Awareness for Text Detection*.
- Ch'ng, C. K., & Chan, C. S. (2017, November). Total-Text: A Comprehensive Dataset for Scene Text Detection and Recognition. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (pp. 935–942). IEEE. <https://doi.org/10.1109/ICDAR.2017.157>
- Bin Ahmed, S., Naz, S., Razzak, M. I., Rashid, S. F., Afzal, M. Z., & Breuel, T. M. (2016, April). Evaluation of cursive and non-cursive scripts using recurrent neural networks. *Neural Comput. Appl.*, 27(3), 603–613. <https://doi.org/10.1007/s00521-015-1881-4>
- Karatzas, D., et al. (2015, August). ICDAR 2015 competition on Robust Reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 1156–1160). IEEE. <https://doi.org/10.1109/ICDAR.2015.7333942>