SIFT-ResNet Synergy for Accurate Scene Word Detection in Complex Scenarios

Riadh Harizi[1][a], Rim Walha[1,2][b] and Fadoua Drira[1][c]
[1] *REGIM-Lab, ENIS, University of Sfax, Tunisia*
[2] *Higher Institute of Computer Science and Multimedia of Sfax, University of Sfax, Tunisia*

Keywords: Scene Text Detection, Deep Learning, SIFT Keypoints, Bounding Box Regressor.

Abstract: Scene text detection is of growing importance due to its various applications. Deep learning-based systems have proven effective in detecting horizontal text in natural scene images. However, they encounter difficulties when confronted with oriented and curved text. To tackle this issue, our study introduces a hybrid scene text detector that combines selective search with SIFT-based keypoint density analysis and a deep learning training architecture framework. More precisely, we investigated SIFT keypoints to identify important areas in an image for precise word localization. Then, we fine-tuned these areas with a deep learning-powered bounding box regressor. This combination ensured accurate word boundary alignment and enhancing word detection efficiency. We evaluated our method on benchmark datasets, including ICDAR2013, ICDAR2015, and SVT, comparing it with established state-of-the-art scene text detectors. The results underscore the strong performance of our scene text detector when dealing with complex scenarios.

1 INTRODUCTION

Scene text detection is of growing importance due to its various applications in a wide range of fields. It is often described as the process of localizing the specific regions of text within images captured from natural scenes. Indeed, scene text detection can involve multiple candidate regions, including text, word, and character levels, which are then candidates for further processing. At each level of candidate regions, there are distinct advantages and disadvantages based on their utility for specific applications. On the one hand, text-level candidate regions, which treat the entire scene text as a single candidate region, are particularly useful for recognizing text blocks or headings within images. Their simplicity makes them the preferred choice when text recognition is unnecessary, as it doesn't provide fine-grained information about individual words or characters, which can be a drawback when detailed text analysis is required. Word-level candidate regions, on the other hand, focus on providing finer details by localizing individual words within a scene. These regions are highly valuable for applications requiring word-level analysis (Harizi et al., 2022b). But, they may face challenges when dealing with densely packed words or when character-level recognition becomes necessary. Character-level candidate regions offer the ultimate solution for character-level analysis but demand more complex processing compared to word or text-level regions. These regions can be particularly challenging for handwritten or cursive text. In summary, the choice of the candidate level depends on the application's requirements, striking a balance between the need for detail and the complexity of processing.

Several text detection algorithms have been employed to tackle accurately the text localization task. Inspired by the rapid advancements in deep learning and the availability of annotated datasets, numerous text detection techniques have emerged. Deep learning-based systems have proven effective in detecting horizontally aligned text in natural scene images. However, they encounter difficulties when confronted with more challenging and complex scenarios, specifically those involving oriented or curved text. Therefore, many investigations continue to explore innovative architectures and solutions to advance the state of the art in text detection. Their primary focus is on addressing some notable limitations of existing techniques, such as improving the effectiveness of handling images containing text

[a] https://orcid.org/0000-0003-4096-8959
[b] https://orcid.org/0000-0002-0483-6329
[c] https://orcid.org/0000-0001-6706-4218

Harizi, R., Walha, R. and Drira, F.
SIFT-ResNet Synergy for Accurate Scene Word Detection in Complex Scenarios.
DOI: 10.5220/0012426200003636

In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 3*, pages 980-987
ISBN: 978-989-758-680-4; ISSN: 2184-433X

with varying angles or vertical orientations. Indeed, main text detection methods adapted to text of varying shapes and orientations can be categorized into two primary groups: region-based and texture-based (Naiemi et al., 2021). In this study, we will primarily concentrate on process-driven categories of region-based methods, which can be defined by three main groups: pixel-based, model-based and hybrid text detectors approaches. This choice is guided by the nature of our proposition. Pixel-based text detector approaches aim to detect and precisely locate text regions by conducting an in-depth examination of the individual pixels. These approaches utilize a variety of image processing techniques and pixel-level characteristics, including but not limited to color, brightness, texture, pixel connectivity, keypoint density and corners. These approaches are suitable for detecting text in scenarios where text regions are well-defined and exhibit distinct pixel-level features. However, in real-world scenarios, this is often not the case. Therefore, most approaches dealing with this challenging task predominantly fall into two categories: model-based and hybrid-based text detector approaches. On the one hand, model-based text detector approaches rely on pre-trained models or specific neural network architectures to detect text. These approaches utilize features learned from training data to recognize and localize text regions in images. On the other hand, hybrid approaches combine the strengths of both pixel-based and model-based approaches for text detection.

In this context, we introduce an hybrid approach to address the challenges of text detection in unstructured, real-world scenarios. This approach combines the strengths of Convolutional Neural Networks (CNN) and Scale-Invariant Feature Transform (SIFT) techniques to enhance word detection accuracy and robustness. While ResNet excels in feature extraction, SIFT is well-known for its resilience to scale and orientation variations. Our method utilizes word-level candidate regions as they strike a balance between the simplicity of text-level detection and the detail of character-level detection. Word-level candidate regions are considered the optimal choice when priorities include readability, reduced complexity, and efficient processing.

The remainder of this study is structured as follows. Section 2 provides an overview of well-known model-based and hybrid scene text detection approaches in unstructured, real-world scenarios. Section 3 details our proposed SIFT-ResNet hybrid text detection method. Section 4 presents the experimental study to highlight the effectiveness of our approach. Section 5 concludes the study and highlights open issues for future research.

2 RELATED WORKS

In this section, we present an in-depth exploration of related work in the field of model-based versus hybrid-based text detector approaches, with a particular emphasis on deep learning-driven methods. Our focus on this area stems from the growing importance of text detection in various applications, where the choice between model-based and hybrid approaches plays a pivotal role in achieving accurate and efficient results, particularly in addressing the challenges encountered in complex real-world scenarios.

Regarding model-based text detector approaches, common models used include CNNs and other deep learning architectures. These approaches are suitable for detecting text in scenarios where text can vary in terms of orientation, font, and placement (Zhou et al., 2017; Liao et al., 2018a; Long and Yao, 2020; Zhu et al., 2021; Yu et al., 2023). For example, EAST (Efficient and Accurate Scene Text Detector) employs a CNN to predict text regions and their corresponding quadrilateral bounding boxes in a single forward pass (Zhou et al., 2017). Another model-based text detection approach is TextBoxes, which predicts both text regions and their corresponding bounding boxes by incorporating multiple aspect ratios and orientations in the output layer of the network (Liao et al., 2017). Additionally, YOLO (You Only Look Once), originally designed for object detection, can be adapted for text detection tasks by training it on text-specific datasets, making it applicable for text detection in diverse scenarios (Redmon and Farhadi, 2017). Stroke Width Transform (SWT) is another model-based text detector. It operates in a single pass through the image, identifying and grouping regions with similar stroke widths to identify potential text regions (Piriyothinkul et al., 2019) (Epshtein et al., 2010). In (Mallek et al., 2017), the authors explored the integration of a sparse prior into a model-based scene text detection approach. Specifically, the features of the convolutional PCANet network are enhanced by sparse coding principle (Walha et al., 2013), representing each feature map through interconnected dictionaries and hence facilitating the transition from one resolution level to a suitable lower-resolution level. Liao et al. proposed a unified end-to-end network which operates in a single pass through the network to detect and recognize text in images (Liao et al., 2018b). The introduction of the RoIRotate operator is a part of their single-stage architecture, which aims to handle oriented text and gain axis-aligned feature maps efficiently. Another study developed an instance segmentation-based method that employed a deep neural network to simultaneously predict text regions and their interconnecting relationships (Deng et al., 2018).

Concerning hybrid text detector approaches, they may use pixel-level analysis to identify potential text regions and then utilize models for further refinement and classification. Indeed, the classification process is designed to distinguish text from non-text areas, while the refinement process is focused on enhancing the precision of text region detection. This fine-tuning may entail adjustments to the position, size, or shape of the bounding boxes to achieve more accurate text region localization. For instance, Faster R-CNN can be used to initially generate region proposals that are likely to contain text and then refine these proposals for accurate text localization (Ren et al., 2015). Moreover, Mask R-CNN extends Faster R-CNN by incorporating a segmentation mask branch (He et al., 2017). Another example is CTPN (Connectionist Text Proposal Network). It generates text proposals in the first stage and then refines them using a recurrent neural network (RNN) in the second stage (Tian et al., 2016). Zihao et al. presented in (Liu et al., 2018b) an approach that involves linking individual text components to create complete text lines. In (Long et al., 2018), the authors proposed a method representing curved text with straight lines in a two-stage process: first identifying potential text regions, and then refining and classifying these regions for improved accuracy in text detection. In their multi-oriented scene text detection approach (Dai et al., 2018), the authors used a region proposal network for text detection and segmentation, followed by non-maximum suppression to handle overlapping instances. Furthermore, hybrid methods may include a rectification stage to address geometric distortions in text, like perspective distortion or skew. This step improves text legibility and streamlines the recognition process. An example of this method is ASTER (Attentional Scene Text Recognizer) (Shi et al., 2019).

In summary, hybrid text detection efficiency stems from using advanced deep learning frameworks for scene text localization and integrating textual features pixel-wise. This integration of textual features can lead to real-time solutions. Motivated by these considerations, we introduce our hybrid text detector framework in the following section.

3 PROPOSED SCENE WORD DETECTOR

In this section, we present the proposed deep learning-based scene text detection method. It relies on an hybrid-based detection approach that harnesses the advantages of both convolutional-based deep networks and key-points based techniques in order to accurately localize multi-oriented words involved in a given real-world scene image. An overview of the proposed method is depicted in Figure 1. As exhibited in this figure, our proposition consists of three main stages: multiscale SIFT-based RoIs detection, RoIs filtering and grouping, and word bounding box regression. Details concerning each stage of our proposition are provided in the following.

3.1 Multiscale SIFT for RoI Detection

Real-world scene text differs visually from document text. Rather than employing a preliminary segmentation or handling the entire input image content and investigating its overlapping regions, we suggest a more focused approach which conducts a precise selective search. This is achieved by detecting keypoints and exploring a multi-scale spatial grids applied to the input scene image, facilitating thus the identification of pertinent local regions. Specifically, our detection process initiates with the localization of SIFT keypoints, serving as a means to guide the selection of candidate regions that are likely to encompass text areas, thereby eliminating the need for exhaustive processing of all image regions. Following this, a refinement process partitions the image into cells of varying sizes using multi-scale grids. Each cell corresponds to a local patch area having a dimension of $n \times n$ pixels, with n being selected from the set $8, 12, 16, 32$. This secondary step aids in systematically identifying regions of interest (RoIs). The method focuses on pertinent local areas, inspecting multi-scale grids within the image, especially those with SIFT keypoints. Bounding boxes for these chosen cells, referred to as SIFT-RoIs, are created by computing Euclidean distances between SIFT keypoints and cell centroids.

3.2 RoIs Filtering and Grouping

Throughout the training phase, our main emphasis lies in assessing the precision of the chosen SIFT-RoIs bounding boxes. In this regard, we utilize the Intersection over Union (IoU) evaluation measure which represents a commonly employed measure in object detection applications. In fact, this measure is a valuable tool for evaluating the alignment between the predicted bounding boxes, specifically those associated with SIFT-RoIs, and the ground-truth bounding boxes. These latter offer precise annotations that define the true positions of text patterns within the training images. Specifically, the IoU measure quantifies

Table 1: Overview of recent methods in deep learning-based scene text detection.

Reference	Model	Category		Backbone Network	Candidate Region	Text Shape		
		MD	HD			MOT	CT	H
(Zhang et al., 2015)	STLD		✓	CNN	C,T			✓
(Mallek et al., 2017)	DLSP	✓		PCANet	W,C			✓
(Zhou et al., 2017)	EAST	✓		FCN	W,T	✓		✓
(Liao et al., 2017)	TextBoxes	✓		VGG-16+CRNN	W			✓
(Liao et al., 2018a)	TextBoxes++	✓		SSD+CRNN	W	✓		✓
(Deng et al., 2018)	PixelLink		✓	FCN	W	✓		✓
(Baek et al., 2019)	CRAFT	✓		VGG-16	C,W	✓	✓	✓
(Tian et al., 2016)	CTPN		✓	VGG-16	T,W			✓
(Wang et al., 2019a)	PSENET		✓	ResNet-18	T	✓	✓	✓
(Liu et al., 2022)	ABCNet		✓	ResNet-18	T	✓	✓	✓
(Long et al., 2018)	TextSnake		✓	U-Net	W	✓	✓	✓
(Liao et al., 2021)	Mask TextSpotter		✓	FPN+CNN	T	✓	✓	✓
(Harizi et al., 2022a)	CNN		✓	CNN	W,C			✓
(Busta et al., 2017)	Deep TextSpotter	✓		CNN	W			✓
(Liu et al., 2018a)	FOTS		✓	ResNet-50	W	✓		✓
(He et al., 2018)	TextSpotter		✓	PVAnet	T,W,C	✓		✓
(Shi et al., 2019)	ASTER		✓	SSD+LSTM	W	✓		✓
(Xing et al., 2019)	charNet	✓		ResNet-50	C,W	✓	✓	✓
(Wang et al., 2019b)	PAN		✓	ResNet-18	T,W	✓	✓	✓
(Long and Yao, 2020)	UnrealText	✓		FCN	W	✓		✓
(Liao et al., 2020)	SynthText3D	✓		ResNet-50	W	✓		✓
(Zhu et al., 2021)	FCENet		✓	ResNet-50+FPN	T,W	✓	✓	✓
(Yu et al., 2023)	TCM	✓		ResNet-50	T,W	✓	✓	✓
Proposed method	SIFT-ResNet		✓	ResNet-19	W	✓	✓	✓

Note: W : Word-based, T : Text-based, C: Character-based, MD : Model-based text Detector approach, HD : Hybrid-based text Detector approach, MOT: Multi-Oriented Text, CT: Curved Text, : HT: Horizontal Text.

the proportion of the overlapping area between the predicted and actual bounding boxes in relation to the combined area of both boxes. Especially, it is obtained for the j-th ground-truth (G_j) and i-th detection bounding box (D_i) as follows:

$$\mathrm{IOU} = \frac{\mathrm{Area}(G_j \cap D_i)}{\mathrm{Area}(G_j \cup D_i)} \qquad (1)$$

This assessment crucially refines our scene text detection model's precision, providing valuable insights into its proficiency in identifying text regions within images. The IoU values guide the selection of SIFT-RoIs, and from these, a Bag of Word Patterns (BoW) model is constructed to determine the presence of text patterns in specific image regions. After BoW-based filtering, we move to a region grouping stage to enhance coverage of dense text patterns. In this phase, we randomly merge nearest filtered SIFT-RoIs, usually selecting the two or three closest centroids. This step aims to reduce redundancy, consolidate overlapping regions, and potentially improve precision in the subsequent word bounding box detection stage.

3.3 Word Bounding Box Regression

As depicted in Figures 1 and 3, the grouped SIFT-RoIs bounding boxes cover text regions comprehensively but lack precision in localizing individual words. To address this, we introduce a Word Bounding Box Regressor (WBBR), inspired by object detectors like YOLO (Redmon and Farhadi, 2017) and Faster R-CNN (Ren et al., 2015). WBBR enhances word region detection for accurate delineation of word bounding boxes. Our approach adapts bounding box regression for precise localization of arbitrarily oriented words in real-world scene images. Specifically, we analyze SIFT-based RoIs, serving as proposals for WBBR, crucial for accurately localizing each word region

The proposed WBBR model in this study is based on the fully-convolutional structure of ResNet-19 (He et al., 2016), keeping convolution and pooling layers. However, we modify it by replacing the final fully-connected layers with four dense layers, as shown in Figure 2. These layers progressively reduce neurons, with the last layer having four neurons, serving as the

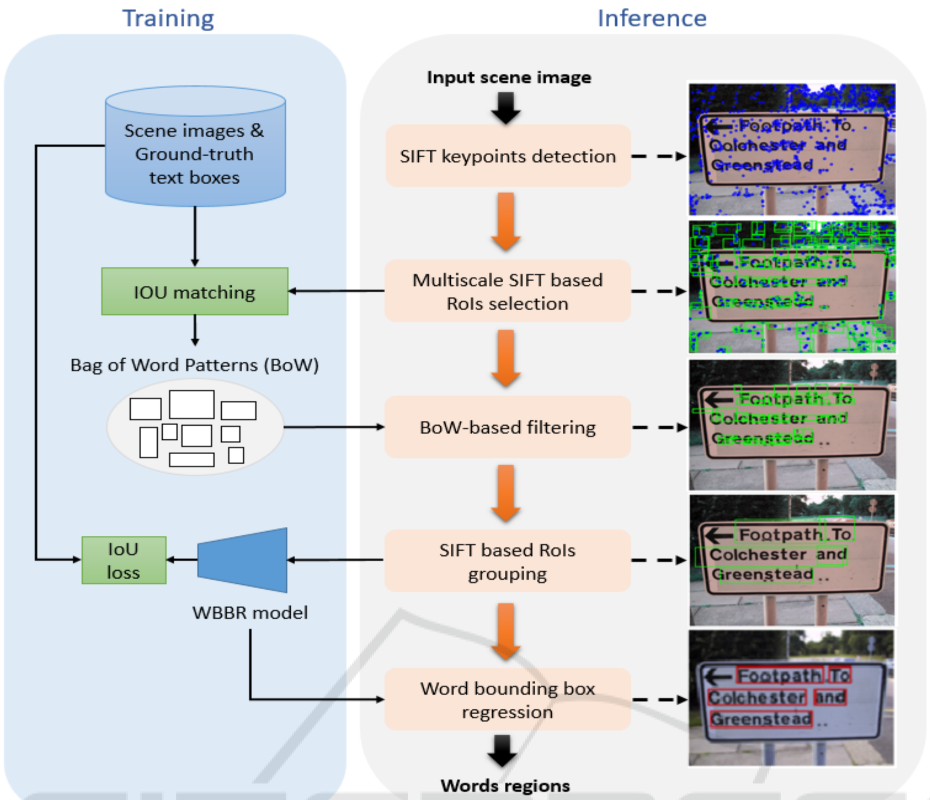

Figure 1: Illustration of the proposed scene text detection method.

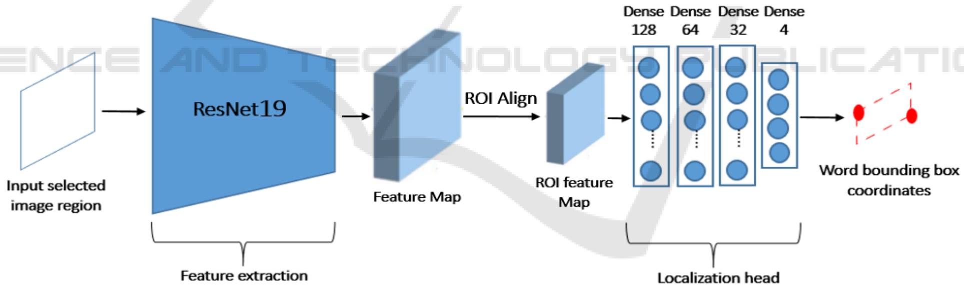

Figure 2: Illustration of the word bounding box regression architecture.

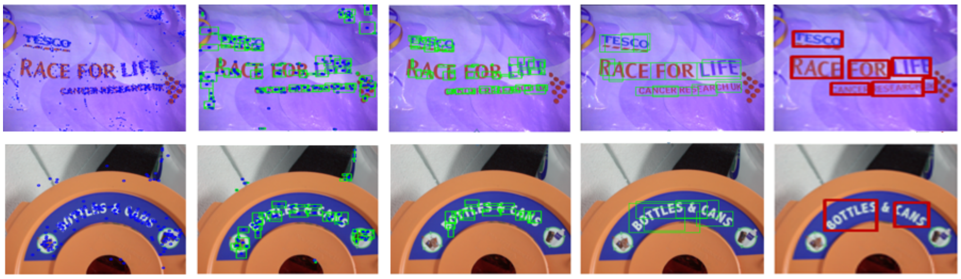

Figure 3: Illustration of intermediate results generated through the proposed scene words detection process.

detection head for predicting word region positions. Refer to Figures 1 and 3 for a visual overview of the proposed word detection process and results on various scene image samples.

Table 2: Features of the datasets used in the study.

Dataset	**Language**	**Images**		**Text shape**			**Text Features**	**Annotation level**	
		Train	Test	H	CT	MO		Char	Word
ICDAR2013 (Karatzas et al., 2013)	EN	229	233	✓	–	–	Large	–	✓
ICDAR2015 (Antonacopoulos et al., 2015)	EN	1000	500	✓	✓	✓	Small, Blur	–	✓
SVT (Wang and Belongie, 2010)	EN	100	250	✓	✓	✓	Low-resolution	–	✓

Note: ML - Multi-Lingual, EN - English, H - Horizontal, CT - Curved Text, MO - Multi-Oriented.

Table 3: Quantitative comparison among some of the recent scene text detection methods evaluated on the ICDAR 2013, ICDAR 2015, and SVT datasets. The best values are in bold.

		ICDAR2013			**ICDAR2015**			**SVT**		
Reference	**Model**	***R***	***P***	***F***	***R***	***P***	***F***	***R***	***P***	***F***
(Zhang et al., 2015)	STLD	0.74	0.88	-	-	-	-	-	-	-
(Liao et al., 2017)	TextBoxes	0.83	0.89	0.86	-	-	-	-	-	-
(Liao et al., 2018a)	TextBoxes++	0.84	0.91	0.88	0.78	0.87	0.82	-	-	-
(Zhou et al., 2017)	EAST	-	-	-	0.78	0.83	0.81	-	-	-
(Liu et al., 2018a)	FOTS	-	-	0.87	0.82	0.89	0.85	-	-	-
(He et al., 2018)	TextSpotter	0.87	0.88	0.88	0.83	0.84	0.83	-	-	-
(Shi et al., 2019)	ASTER	-	-	-	0.69	0.86	0.76	-	-	-
(Tian et al., 2016)	CTPN	0.83	0.93	0.88	0.52	0.74	0.61	0.65	0.68	0.66
(Deng et al., 2018)	PixelLink	0.88	0.89	0.88	0.82	0.86	0.84	-	-	-
(Baek et al., 2019)	CRAFT	**0.93**	**0.97**	**0.95**	0.84	0.89	0.86	-	-	-
(Wang et al., 2019a)	PSENET	-	-	-	0.84	0.86	0.85	-	-	-
(Metzenthin et al., 2022)	WSRL	0.70	0.84	0.77	-	-	-	-	-	-
(Wang et al., 2019b)	PAN	-	-	-	0.82	0.84	0.83	-	-	-
(Long and Yao, 2020)	UnrealText	0.74	0.88	0.81	0.81	0.86	0.83	-	-	-
(Liao et al., 2020)	SynthText3D	0.76	0.71	0.73	0.80	0.87	0.83	-	-	-
(Zhu et al., 2021)	FCENet	-	-	-	0.83	**0.90**	0.86	-	-	-
(Harizi et al., 2022a)	CNN	0.92	0.94	0.93	0.74	0.79	0.76	0.72	0.78	0.75
(Yu et al., 2023)	TCM	-	-	0.79	-	-	**0.87**	-	-	-
Proposed method	SIFT-ResNet	**0.93**	**0.97**	**0.95**	**0.85**	**0.90**	**0.87**	**0.74**	**0.79**	**0.76**

Note: R - Recall, P - Precision, F - F-score.

4 EXPERIMENTS AND EVALUATION

4.1 Datasets

This work utilizes three popular datasets: ICDAR 2013 (Karatzas et al., 2013), ICDAR 2015 (Antonacopoulos et al., 2015), and Street View Text (SVT) (Wang and Belongie, 2010). Table 2 summarizes key features and details of the datasets. We note that the proposed WBBR architecture undergoes training using these datasets. During the training phase, we divide them into three sets: 80% for training images, 10% for validation, and 10% for testing.

4.2 Evaluation Metrics

The evaluation in this study uses recall, precision, and F-score metrics. Recall (R) measures the proportion of true positives to the total positives in ground truth annotations, while precision (P) is the ratio of true positives to the total detected text examples. They are defined as follows: $R = \frac{TP}{TP+FN}$, $P = \frac{TP}{TP+FP}$. Here, TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives. The F-score (F) is computed by combining recall and precision values through the formula: $F = \frac{2\times R\times P}{R+P}$.

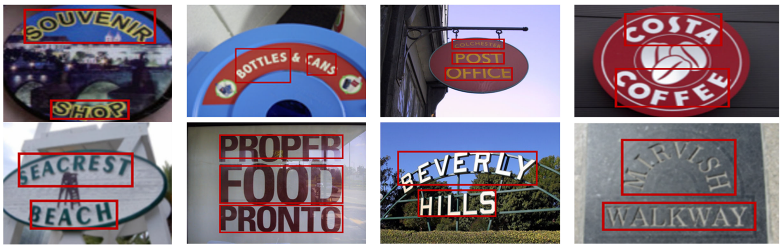

Figure 4: Some visual results generated by the proposed scene text detection method which succeeds to localize curved and oriented scene words with different orientations, curvatures, styles, sizes, illuminations, and spatial resolutions.

4.3 Performance Evaluation Study

Table 3 depicts the results of our scene text detection method compared to other state-of-the-art methods on ICDAR 2013, ICDAR 2015, and SVT datasets. Results, evaluated by precision, recall, and F-score, show superior performance of our method across all datasets. Specifically, on the ICDAR 2013 dataset, our method outperforms the WSRL text detection method (Metzenthin et al., 2022), increasing the F-score from 0.77 to 0.95, with improvements exceeding 13% in precision and 23% in recall. Compared to the efficient CRAFT detector (Baek et al., 2019), our method performs similarly on the ICDAR 2013 dataset but outperforms CRAFT on the more challenging ICDAR 2015 dataset. The improvements in precision, recall, and F-measure, particularly on ICDAR 2015 and SVT datasets, highlight the robustness of our approach in detecting challenging scene text, including small-scale and multi-oriented examples that challenge human perception.

Figure 4 displays results from our SIFT-ResNet scene text detector. The outputs demonstrate successful localization in complex scenarios with varying orientations, curvatures, styles, sizes, illuminations, and spatial resolutions. Notably, the method accurately localizes highly-curved words, as shown in Figures 4. In general, evaluating scene text localization considers both accuracy and efficiency. Our research combines traditional (SIFT) and modern (ResNet-based bounding box regression) techniques, resulting in a highly accurate scene word detector. The detector demonstrates notable improvements in precision, recall, and F-score measures. The proposed method excels in performance due to the effective use of multi-scale SIFT keypoints for character pattern extraction and precise localization with a selective search-based word bounding box regressor in a deep learning framework. By combining the local feature capturing strength of SIFT keypoints with the semantic understanding of deep neural networks, our approach achieves a more precise scene word detector. This collaboration highlights the synergy between traditional computer vision and modern deep learning techniques. The efficiency of our text detection method is notably boosted by the significant contribution of the SIFT technique. It streamlines the process by identifying key regions with a high likelihood of containing text, allowing the detector to focus on these areas rather than the entire image. This concentration improves the overall speed of our text detector. Finally, the achieved scene text detection results could enhance the functionality and performance of diverse applications like text super-resolution and recognition (Walha et al., 2015).

5 CONCLUSION

In summary, our study concentrated on detecting text in real-world scene images. We introduced a hybrid text detection method that combines SIFT-based keypoints localization, BoW-based character patterns filtering, and ResNet-19 based word bounding box regression. Experimental results affirmed the method's efficiency, particularly in handling multi-oriented and curved scene texts. Performance evaluations were conducted on three challenging datasets, comparing favorably with various state-of-the-art text detection methods. As a future work, we aim to extend this research to address the multi-script text detection and recognition tasks.

REFERENCES

Antonacopoulos, A., Clausner, C., Papadopoulos, C., and Pletschacher, S. (2015). ICDAR2015 competition on recognition of documents with complex layouts. In *ICDAR 2015*, pages 1151–1155.

Baek, Y., Lee, B., Han, D., Yun, S., and Lee, H. (2019). Character region awareness for text detection. In *CVPR 2019*, pages 9365–9374.

Busta, M., Neumann, L., and Matas, J. (2017). Deep textspotter: An end-to-end trainable scene text localization and recognition framework. *ICCV 2017*, pages 2223–2231.

Dai, Y., Huang, Z., Gao, Y., Xu, Y., Chen, K., Guo, J., and Qiu, W. (2018). Fused text segmentation networks for multi-oriented scene text detection. In *ICPR 2018*, pages 3604–3609.

Deng, D., Liu, H., Li, X., and Cai, D. (2018). Pixellink: Detecting scene text via instance segmentation. In *AAAI 2018*, pages 6773–6780.

Epshtein, B., Ofek, E., and Wexler, Y. (2010). Detecting text in natural scenes with stroke width transform. In *CVPR 2010*, pages 2963–2970.

Harizi, R., Walha, R., and Drira, F. (2022a). Deep-learning based end-to-end system for text reading in the wild. *Multim. Tools Appl.*, 81(17):24691–24719.

Harizi, R., Walha, R., Drira, F., and Zaied, M. (2022b). Convolutional neural network with joint stepwise character/word modeling based system for scene text recognition. *Multim. Tools Appl.*, 81(3):3091–3106.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. B. (2017). Mask R-CNN. In *ICCV 2017*, pages 2980–2988.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *CVPR 2016*, pages 770–778.

He, T., Tian, Z., Huang, W., Shen, C., Qiao, Y., and Sun, C. (2018). An end-to-end textspotter with explicit alignment and attention. *CVPR 2018*, pages 5020–5029.

Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L. G., Mestre, S. R., Mas, J., Mota, D., Almazán, J., and de las Heras, L. (2013). ICDAR 2013 robust reading competition. In *ICDAR*, pages 1484–1493.

Liao, M., Lyu, P., He, M., Yao, C., Wu, W., and Bai, X. (2021). Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(2):532–548.

Liao, M., Shi, B., and Bai, X. (2018a). Textboxes++: A single-shot oriented scene text detector. *IEEE Transactions on Image Processing*, 27:3676–3690.

Liao, M., Shi, B., Bai, X., Wang, X., and Liu, W. (2017). Textboxes: A fast text detector with a single deep neural network. In *AAAI 2017*, pages 4161–4167.

Liao, M., Song, B., Long, S., He, M., Yao, C., and Bai, X. (2020). Synthtext3d: synthesizing scene text images from 3d virtual worlds. *Sci. China Inf. Sci.*, 63(2):120105.

Liao, M., Zhu, Z., Shi, B., Xia, G., and Bai, X. (2018b). Rotation-sensitive regression for oriented scene text detection. In *CVPR 2018*, pages 5909–5918.

Liu, X., Liang, D., Yan, S., Chen, D., Qiao, Y., and Yan, J. (2018a). Fots: Fast oriented text spotting with a unified network. *CVPR 2018*, pages 5676–5685.

Liu, Y., Shen, C., Jin, L., He, T., Chen, P., Liu, C., and Chen, H. (2022). Abcnet v2: Adaptive bezier-curve network for real-time end-to-end text spotting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(11):8048–8064.

Liu, Z., Shen, Q., and Wang, C. (2018b). Text detection in natural scene images with text line construction. In *ICICSP 2018*, pages 59–63.

Long, S., Ruan, J., Zhang, W., He, X., Wu, W., and Yao, C. (2018). Textsnake: A flexible representation for detecting text of arbitrary shapes. In *ECCV 2018, Part II*, pages 19–35.

Long, S. and Yao, C. (2020). Unrealtext: Synthesizing realistic scene text images from the unreal world. *CoRR*, abs/2003.10608.

Mallek, A., Drira, F., Walha, R., Alimi, A. M., and Lebourgeois, F. (2017). Deep learning with sparse prior - application to text detection in the wild. In *VISIGRAPP - Volume 5: VISAPP 2017*, pages 243–250.

Metzenthin, E., Bartz, C., and Meinel, C. (2022). Weakly supervised scene text detection using deep reinforcement learning. *CoRR*, abs/2201.04866.

Naiemi, F., Ghods, V., and Khalesi, H. (2021). A novel pipeline framework for multi oriented scene text image detection and recognition. *Expert Syst. Appl.*, 170:114549.

Piriyothinkul, B., Pasupa, K., and Sugimoto, M. (2019). Detecting text in manga using stroke width transform. In *KST 2019*, pages 142–147.

Redmon, J. and Farhadi, A. (2017). YOLO9000: better, faster, stronger. In *CVPR 2017*, pages 6517–6525.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28.

Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., and Bai, X. (2019). Aster: An attentional scene text recognizer with flexible rectification. *PAMI*, 41:2035–2048.

Tian, Z., Huang, W., He, T., He, P., and Qiao, Y. (2016). Detecting text in natural image with connectionist text proposal network. In *ECCV, Part VIII*, pages 56–72.

Walha, R., Drira, F., Lebourgeois, F., Garcia, C., and Alimi, A. M. (2013). Single textual image super-resolution using multiple learned dictionaries based sparse coding. In *ICIAP 2013, Part II*, volume 8157, pages 439–448.

Walha, R., Drira, F., Lebourgeois, F., Garcia, C., and Alimi, A. M. (2015). Joint denoising and magnification of noisy low-resolution textual images. In *ICDAR 2015*, pages 871–875.

Wang, K. and Belongie, S. (2010). Word spotting in the wild. In *ECCV 2010*, pages 591–604.

Wang, W., Xie, E., Li, X., Hou, W., Lu, T., Yu, G., and Shao, S. (2019a). Shape robust text detection with progressive scale expansion network. In *CVPR 2019*, pages 9336–9345.

Wang, W., Xie, E., Song, X., Zang, Y., Wang, W., Lu, T., Yu, G., and Shen, C. (2019b). Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In *ICCV 2019*, pages 8439–8448.

Xing, L., Tian, Z., Huang, W., and Scott, M. (2019). Convolutional character networks. In *ICCV*, pages 9125–9135.

Yu, W., Liu, Y., Hua, W., Jiang, D., Ren, B., and Bai, X. (2023). Turning a CLIP model into a scene text detector. In *CVPR 2023*, pages 6978–6988.

Zhang, Z., Shen, W., Yao, C., and Bai, X. (2015). Symmetry-based text line detection in natural scenes. In *CVPR 2015*, pages 2558–2567.

Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., and Liang, J. (2017). East: An efficient and accurate scene text detector. In *CVPR 2017*, pages 2642–2651.

Zhu, Y., Chen, J., Liang, L., Kuang, Z., Jin, L., and Zhang, W. (2021). Fourier contour embedding for arbitrary-shaped text detection. In *CVPR 2021*, pages 3123–3131.