Journal of Applied Engineering and Technological Science
Vol 6(1) 2024: 103-114

SCENE TEXT DETECTION AND RECOGNITION USING MAXIMALLY STABLE EXTREMAL REGION

Golda Jeyasheeli. P[1*], Athinarayanan. B[2], Manish. T[3], Mohamad Umar. M[4]
Department of Computer Science Engineering, Mepco Schlenk Engineering College, Sivakasi, Tamilnadu, India[1,2,3,4]
pgolda@mepcoeng.ac.in[1*], athi.balu26@gmail.com[2], manishcs1011@gmail.com[3], umarsafiq192@gmail.com[4]

Received: 21 August 2024, Revised: 01 November 2024, Accepted: 02 November 2024
**Corresponding Author*

ABSTRACT
In recent years, scene text detection and recognition have become important research areas in computer vision and machine learning. Traditional text detection and recognition methods may struggle with detecting and recognizing text in images with low resolution, complex backgrounds, and varying font sizes. The proposed methodology addresses these challenges by combining multiple algorithms and using deep learning techniques. In this paper, we propose a method for scene text detection based on Maximally Stable Extremal Regions (MSER) combined with Stroke Width Transform (SWT) and recognition using Convolutional Recurrent Neural Networks (CRNN). Our method consists of two stages: text detection and text recognition. To detect text, we use MSER and SWT to extract candidate text regions from the input and then, we eradicate non-text regions using image to image translation. Finally, to recognize text, CRNN is used to recognize the text present in the detected regions. Our CRNN architecture consists of convolutional and recurrent layers, which enable us to capture both spatial and temporal features of the text. The methodology is evaluated on various benchmark datasets and has obtained good results with accuracy of 96% when compared to existing methods.
Keywords: *MSER, SWT, Scene Text Detection, Text Recognition, Deep Learning, CRNN.*

1. Introduction

With the proliferation of digital information and the increasing demand for automated text processing, text detection and recognition techniques have become essential in various applications, ranging from document analysis and optical character recognition (OCR) to image and video captioning, augmented reality, and autonomous vehicles. Text detection involves identifying the presence and location of text regions within images or videos, while text recognition focuses on converting the detected text into machine- readable text that can be further processed and analyzed. These tasks pose significant challenges due to variations in text fonts, sizes, orientations, lighting conditions, and backgrounds, as well as the presence of noise, distortion, and occlusion.

The problem of text detection and recognition in images is to accurately locate and recognize text within an image taken in a natural scene, such as a street or a building. This can include identifying text in various font styles, and in the presence of various levels of image noise, blur, and distortion. Text could be multi-oriented or made up of different colors. The goal is to efficiently bound the text region and recognizing them efficiently with a high degree of accuracy.

The two main objectives of the proposed work is to design a text detection model to localize the text region using image processing and Deep Learning Techniques and to minimize the non-text region in the detected text region.

Hence, due to the enormous diversity of natural scene texts, researchers are presented with a difficulty to overcome. This enormous problem can be rolled over by the use of a reliable system, capable of recognizing the textual information from the images.

In our proposed model, we use Maximally Stable Extremal Region (MSER) and Stroke Width Transform (SWT) to produce text proposals. Generative Adversarial Network (GAN) (Goodfellow et al., 2014), is used for refining the proposal and for this process we use Pix2Pix framework to do image to image translation. And finally, to recognize textual information available in the detected region we use a Convolutional Recurrent Neural Network (CRNN), in

which CNN and RNN are mainly utilized to classify the text and recognize it in an efficient manner.

The proposed methodology of detecting text using MSER & SWT and recognizing text using CRNN differs from existing methodologies in a few fundamental ways. Firstly, it utilizes both MSER and SWT algorithms for text detection, which allows for more accurate and robust detection of text in complex scenes, and we also perform image to image translation to remove the detected region where there is no textual data. Additionally, the use of the CRNN model for text recognition allows for better recognition accuracy, even in cases where the text is distorted or skewed. This methodology also has the advantage of being able to handle multiple languages and fonts, making it more versatile and applicable in a wider range of scenarios. Overall, this approach represents a significant improvement over existing methodologies to detect & recognize text from the images (Karatzas et al., 2015) and (Karatzas et al., 2013).

2. Literature Review

Scene text identification and recognition has been an important subject of interest in the field of research in recent years. The processing of a scene's image requires a great deal of complexity, which is the cause of this. Scene text identification may generally be tackled using one of two major methods: (a) handcrafted feature-based, or (b) deep learning-based. According to (Matas et al., 2004), components with identical intensities of grey level, or MSER, are assumed to be balanced. Nevertheless, MSER struggles with blurry images, and it is difficult to fine-tune the parameters that are in charge of just capturing the text region. (Dutta et al., 2019), Developed a method to solve this problem and has given us a distinct method in which text components are recognized based on the stability and density of pixel data and grey levels are binned or aggregated to ensure foreground uniformity. In their SWT approach, (Epshtein, Ofek, & Wexler, 2010), used clever edge detection and then applied a subsequent transform on the edges to calculate the likely width and distance of strokes for every output pixel. This strategy merely requires less computing work because it generates fewer candidate components. This identified area may be used to identify the text. Though we combine SWT and MSER, there is still a significant amount of non-text region. (Mukhopadhyay et al., 2019), have addressed this issue by proposing to use a Support Vector Machine (SVM) trained with the help of feature vectors taken from the text region and bearing in mind the text/non-text classification as a single-class problem to handle this problem in.

The spontaneous development of deep learning techniques and Artificial Intelligence in recent years has given researchers working on Scene text identification a new path to investigate. (Yao et al., 2016), produced successful results because it enabled the creation of detailed prediction maps that included information about the characters and relationships between them that were present in the text sections. By utilizing an FCNN built on the foundation of the Holistically Nested Edge Detection approach, they were able to make it happen. With this method, it is feasible to identify curved or countless type of differently oriented text images. By utilizing the extractor mode of the semantic segmentation detector, they were able to accomplish this. The aforementioned issue has been addressed in a variety of methods. (He et al., 2017), provided one more approach by down sampling the segmentation to differentiate between areas that contain written or printed text and those that do not, one can employ a distinction based on textual and non-textual regions. The coordinates needed to create a bounding box will be determined using direct regression.

(Tian et al., 2024), presented a novel Dynamic Receptive Field Adaption Framework that incorporates Memory Attention and Dynamic Feature Adaptive modules to dynamically adjust the receptive field based on character proportions, thereby enhancing recognition accuracy in distorted images. This research demonstrated the effectiveness of these techniques in improving scene text recognition performance across various benchmark datasets. (Yan et al., 2023), employed a learned multi-scale representation approach to address the challenges of recognizing texts at various scales. The study also incorporated hierarchical feature fusion and dynamic log-polar transformers to improve context awareness for different text sizes. The study by (Kai et al., 2024), incorporated a Hierarchical Awareness (HA) module for spatial feature extraction, a Feature Enhancement (FE) module for improving recognition accuracy, and a HAFE framework

for in-depth processing of multi-dimensional features and demonstrated its superiority over state-of-the-art methods in terms of accuracy and efficiency in text recognition tasks. (Sun et al., 2024) and (Yu et al., 2024), integrated scene text detection and recognition through a cross-cooperation guided dynamic points generation approach. It employed a deformable transformer to enhance the interaction between point queries and aggregates multi-scale text features for improved semantic representation. This evaluated the impact of various configurations on performance.

(Das et al., 2024), presented a novel text recognition system that outperforms state of the art methods in challenging scenarios, particularly with occluded text. The proposed approach utilized Maximally Stable Extremal Regions (MSER) for text component detection and incorporates Soft Sets for robust threshold determination. (Wu et al., 2023), presented text detection methods categorized into regression and segmentation-based approaches. Regression methods detected text instances as common objects, while segmentation methods progressively obtain pixel-level masks for oriented and curved texts. The proposed End-PolarT method utilized polar representation for text instances, enhancing detection capabilities in complex scenes (Zhang & Kasturi, 2014).

The traditional text detection methodologies use MSER to identify the regions where text is located but the non-text regions also get detected but our proposed methodology uses MSER combined with SWT to identify the regions where text is located and by using SWT we are significantly reducing the detected text region by reducing the amount of region misclassified as text region (Khalid et al., 2024).

The use of MSER and SWT algorithms for detecting text allows for better detection of text in challenging backgrounds, while the CRNN model for text recognition allows for better recognition accuracy even in cases where the text is distorted or skewed. Overall, the proposed methodology represents a significant improvement over traditional methodologies in this field, addressing the shortcomings of the existing approaches and delivering better performance in challenging scenarios.

3. Research Methods

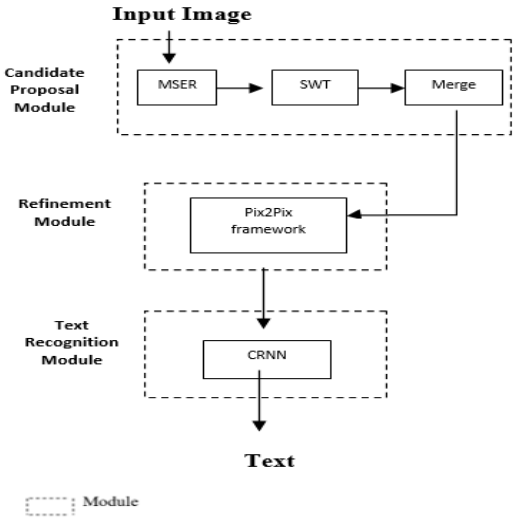


Fig. 1. The proposed work with 3 modules: (i). Candidate Proposal Module; (ii). Refinement Module; (iii). Text Recognition Module

Our proposed model presents a holistic and comprehensive system that addresses the challenges of scene text detection, localization, and recognition. Leveraging a synergistic combination of traditional text detection techniques and deep learning methods, our model delivers a robust and efficient solution (Fang et al., 2022), (Liu et al., 2019), (Xu et al., 2024) and (Wu et al., 2022). The entire system is composed of three distinct sub-levels, each contributing to the overall pipeline: the candidate proposal module, which generates initial text region proposals; the refinement module, which further refines and localizes the text regions with enhanced accuracy; and the text recognition module, which performs accurate and reliable

recognition of the text within the localized regions. Together, these sub-levels form a cohesive system that integrates traditional and deep learning approaches for effective scene text detection, localization & recognition. A visual representation of the modules is shown using a block diagram in Fig. 1.

3.1. Candidate Proposal Generation

This module utilizes MSER (Maximally Stable Extremal Regions) algorithm (Panda et al., 2020), to generate a bounding box that encloses a text region in a single word. The main approach here lies on the concept that regions with minimal changes in the intensity of their grey level are considered stable within a localized neighbourhood area. This idea is employed to identify and bound such regions into a box as part of the candidate proposal module. The MSER function is initialized with a minimum region area of 0.02%, maximum region area of 10% of the test image, and a delta threshold value of 2. These lenient thresholds ensure that text components are not missed from the bounded region, but at the same time, a lot of non-text regions may also be bounded as text regions. The analysis of Chakraborty et al. and Panda et al. supports this approach.

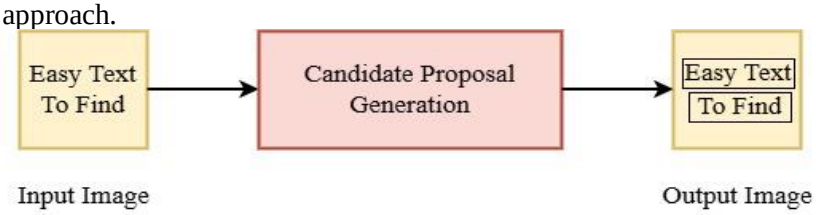


Fig. 2. Working of Candidate Proposal Generation module.

The approach we use to accurately locate and refine text regions involves incorporating the SWT method. The stroke width of a candidate connected component (CC) is determined by measuring the perpendicular interval between successive pixels along its boundary or by considering its thickness. Since text CCs typically have a constant stroke width, we use a threshold of 0.02 to remove the non-text components. However, some characters may not be included due to minimal spacing between them, so we dilate the components and form complete words which is known as merge operation. The candidate proposal is then generated by aggregating all of these terms inside a bounding box. The subsequent module processes this proposal further. Fig. 2, visually depicts the steps involved in this module for a better understanding.

3.2. Proposal Refinement

The Proposal Refinement module refines the bounding box coordinates given by the proposal module. As the complexity of scene text in images increases, the performance of the candidate proposal generation module may deteriorate, revealing the limitations of the proposal generation module. In general, the module may produce an alarming number of incorrect proposals, either by generating proposals for non-text regions or by over or underestimating the proposals. Therefore, we employ a post-processing module to reduce the number of incorrect proposals obtained from the first module. In the Proposal Refinement module, we improve the results of the first module by identifying the bounding boxes that exclusively contain text elements and remove the proposals that doesn't contain textual information.

To improve the accuracy of text region proposals, we utilize an image-to-image translation framework in our model. This method transforms an input image into an output image in which black pixels represent text area and white pixels represent non-text area. This translation is accomplished through a Generative Adversarial Network (GAN) framework, which is trained using input-output data pairs. This novel approach builds on the groundbreaking work of (Goodfellow et al. 2014), who introduced GANs and expanded the possibilities for deep learning research.

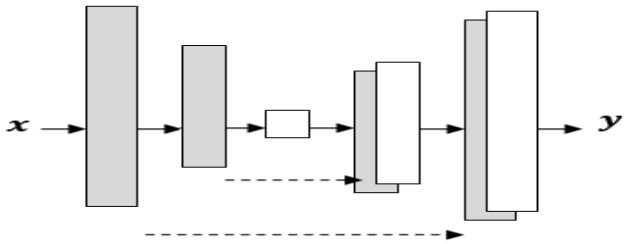


Fig. 3. Architecture of U-net generator model.

The image-to-image translation framework in this module utilizes Pix2Pix, which was introduced by (Isola et al., 2017). Pix2Pix is a conditional Generative Adversarial Network (cGAN) architecture (Mirza et al., 2014). In cGAN, an input image is tuned to generate an output image, and the framework is trained using a loss function to learn the relationship between input and output images. The loss function is a collage of L1 normalization and cGAN losses. The Pix2Pix framework contains a generator G and a discriminator D. The generator produces output images based on tuned input images, while the discriminator predicts whether an image produced G is original or fake. The aim of G is to deceive D by producing output images that are accepted as authentic, resulting in a Nash Equilibrium as G and D play a min-max game. The generator follows a U-Net design, which includes skip connections between mirrored layers of the encoder and decoder stacks. This design is used to preserve low-level properties and avoid bottleneck issues of the Encoder-Decoder network. The generator comprises eight encoder levels and eight decoder layers, each consisting of [Conv / DeConv-BatchNorm-ReLU] units, as illustrated in Fig. 3. The discriminator is a classification model that utilizes a Convolutional Neural Network (CNN) architecture with five layers.

The Pix2Pix component of our model utilizes the mapping of text in images to improve the precision of the bounding box boundaries and remove non-text areas detected by the previous module. Since non-text areas are represented by white pixels, the module can eliminate them by outputting a white image (Geng, 2024) and (Zhang et al., 2015). To obtain refined results, the proposals from the previous module are clipped & used as input to the refinement module.

Our model employs a doubling strategy that takes into account both the bounding box proposals and the contextual information of the image in order to ensure complete coverage of the text regions. Afterwards, this improved image is subjected to the refining module (Ye & Doermann, 2014) and (Mu et al., 2021). The contours of the binary images created by this module are dilated in order to estimate the bounding boxes around the text sections. Improved bounding box region is finally cropped so that the Text recognition module can use them as inputs.

3.3. Text Recognition

The Text Recognition module utilizes the results obtained from the Refinement module to identify the text present in the image. With the aid of the Refinement module, text recognition can be performed more accurately. To accomplish this, we employ a Convolutional Recurrent Neural Network (CRNN), which is a deep learning approach (Shi et al., 2016).

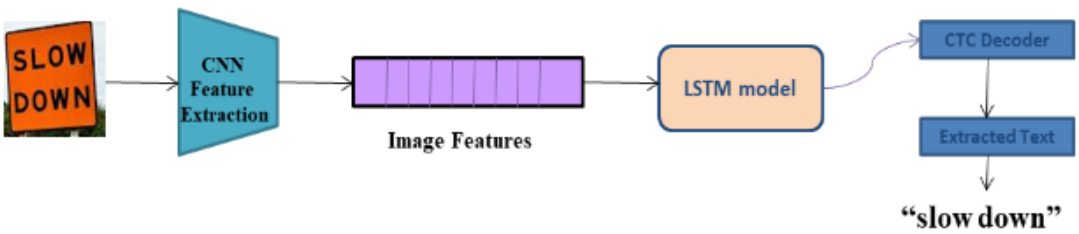


Fig. 4. Working of Text Recognition module.

By combining the benefits of CNN and RNN architectures, the CRNN is a deep learning model that excels in identifying text in images. Hence, it can effectively extract information from the image and precisely identify the text that is present.

The CRNN model starts by using CNN to obtain features from the image. CNN is a type of neural network optimized for grid-like data like images, which can automatically learn relevant features through convolutional and pooling layers that capture local patterns and spatial information. This process enables the CRNN model to create meaningful representations of the text regions in the image, which are then passed to the RNN component for further processing. The convolutional layers in CRNN extract image features such as edges, corners, and textures to improve text recognition accuracy (Bagi et al., 2021), (Yin et al., 2015) and (Zhou et al., 2011). Once CNN has extracted the features, the RNN is used to recognize the text in the image by processing the sequential data. The RNN uses a memory cell to store information from previous time steps and make predictions based on that information. To train the model, CRNN uses a CTC loss function, an algorithm that allows the model to make predictions based on the sequence of characters, rather than their position in the image (Tong et al., 2022). This enhances the model's resilience to image variations like skew or distortion, making it more robust. It is explained well in Fig. 4.

The trained CRNN model uses CNN to first extract features from the picture in order to recognize text in images, and then RNN to process the extracted features in order to identify the text.

4. Results and Discussions

The proposed method is tested, and the performance is evaluated using scene text dataset. The result of each module is discussed in detail below and the proposed methodology has given notable improvement while testing.

4.1. Candidate Proposal Generation

The approach creates candidate proposals in this stage, which are bounding box suggestions that might have text areas. The bounding boxes are initially created using MSER, after which they are filtered using SWT depending on stroke width. Finally, using a joining method, related bounding boxes are combined as separate proposals (Epshtein, Ofek, & Wexler, 2010) and (Gomez & Karatzas, 2014).

As mentioned earlier, the candidate proposal generation module is employed to locate text regions in an image. The results of each process within this module are shown in Fig. 5. The proposed method has been evaluated using images from the ICDAR-2013 Robust Reading competition dataset (Karatzas et al., 2015) and (Karatzas et al., 2013).

4.2. Proposal Refinement Module

This module is responsible to reduce the coordinates of the bounding box obtained from the module 1. Based on the candidate suggestions, which is accomplished through an image-to-image translation task. Every proposal is given a precise text and non-text region distinction by the refinement module.

The module converts any image into a bi-color image, where the text regions are represented in one color and the non-text regions in another color. This image-to-image translation enables the creation of a training dataset based on the resulting bi-color images.

Fig. 5. Shows examples of the results of the production of bounding box proposals after each sub-step.

The ground truth annotations help us to create the bounding boxes for the text regions. To create a new bounding box, a margin between 5 and 100 pixels is added to each side. The area of focus extracted from the original image is subsequently trimmed using this updated bounding box. The clipped area acts as input for the Pix2Pix framework. Text regions are marked using black pixels, while non-text regions are marked by white pixels in the output image, which is of the same size as the input image. The resulting photos are then downsized to 208 160 pixels from the input images. A collection of the textual areas is generated using this approach, resulting in a dataset of text regions. Fig. 6, provides a visual explanation of this generation process. The refinement module is entrusted with both the refinement of the text region's borders and to detect the misclassified proposals from the proposal generation module. White pixels represent non-text regions, so images with no text should produce an output image which is completely white in colour. The MS-COCO dataset contains images of non-text things that are taken into consideration for this. After clipping off the non-text objects, the photos are resized to 208 x 160 pixels. Dataset used for training the refinement module consists of 10,000 images of text regions and 2,000 images of non-text regions, maintaining a ratio of 5:1 for text to non-text images (Islam et al., 2016) and (Koo & Kim, 2013). Fig. 7, and Fig. 8, show sample input-output paired images for the text region and non-text region, respectively.

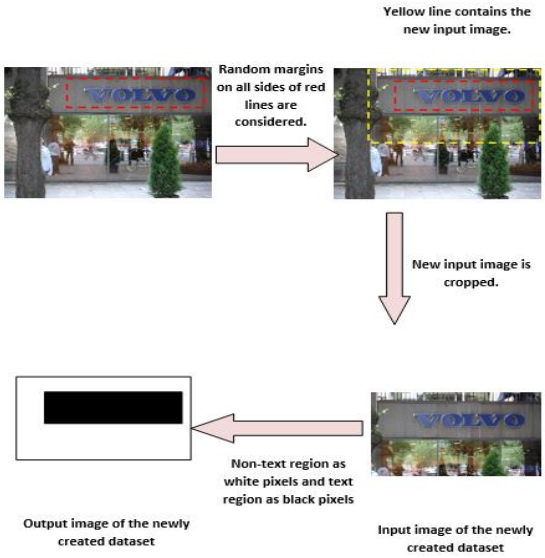


Fig. 6. Examples of images that demonstrate the dataset preparation procedures used to find the text region in the refining module are shown.

Fig. 7. Displays sample input and output images from the dataset, showcasing images that contain texts and their corresponding output images after text region detection.

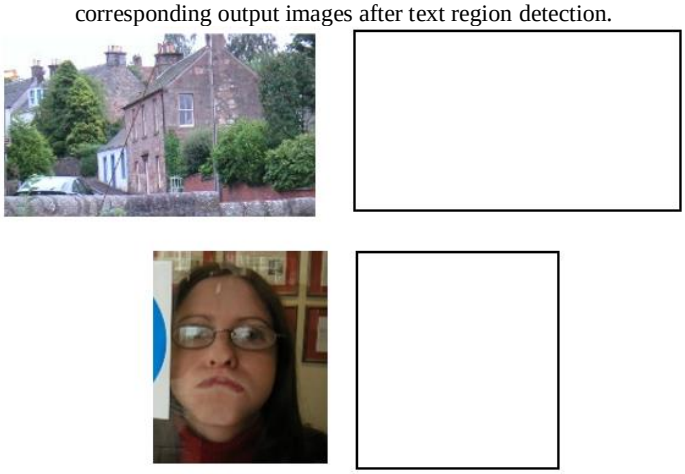
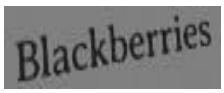


Fig. 8. Sample input and output images from the dataset, highlighting images that do not contain any texts and their corresponding output images after text region detection.

4.3. Text Recognition Module.

To evaluate the performance of our text recognition module, we trained a CRNN model on the MJSynthtext dataset, which contains more than 8 million synthetic text images. The CRNN model consists of a stack of convolutional and recurrent layers, followed by a fully connected layer that outputs the recognized text. The model is trained using the CTC loss function, which is a standard loss function for sequence-to-sequence tasks. We have trained this model separately using CRNN and this model is used in the recognition part which provides a good result. The results of the CRNN are shown below in Table 1.

Table 1 - Sample input images which are given to CRNN and the ground truth texts (Actual Text) and Predicted Texts are shown.

Input Image	Actual Text	Predicted Text
	"STRONGBOX"	"STRONGBOX"
	"RAREFACTION"	"RAREFACTION"
	"BLACKBERRIES"	"BLACKBERRIES"
	"ORATORICALLY"	"OPATORICALLY"
	"CURLICUED"	"CURLICUED"

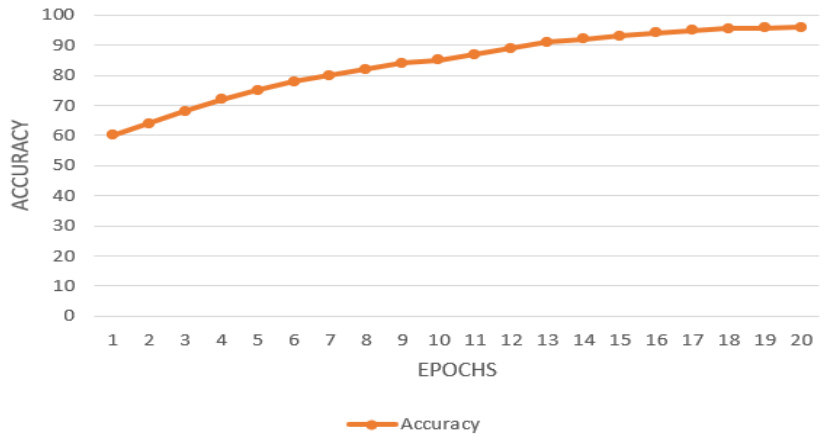


Fig. 9. CRNN accuracy for 20 epochs.

Table 2 - Performance comparison between CRNN and the proposed methodology with ICDAR 2013, 2015 and KAIST datasets.

Dataset	Methodology	Precision	Recall	F1 score
ICDAR 2013	CRNN	88.8%	83%	90%
ICDAR 2013	Proposed	90.8%	84.2%	87.3%
ICDAR 2015	CRNN	90%	85.1%	82%
ICDAR 2015	Proposed	83%	83.1%	82.7%
KAIST	CRNN	87.4%	84.8%	88%
KAIST	Proposed	90.2%	87.2%	88.9%

While training the CRNN model with MJSynthText dataset, we have achieved an accuracy of 96%, which is quite a good measure. The Accuracy graph obtained for 20 epochs is given in Fig. 9. The performance of the CRNN text recognition and the proposed methodology has been compared in the following table. The performance had been measured with various datasets. It is tabulated in a well detailed manner in Table 2.

5. Conclusion

An end-to-end system is proposed for the detection and recognition of scene texts in images. Using a blend of MSER and SWT with dynamically configurable parameters, texts are found in images. The text regions are then refined, and the non-text components are eliminated, using a GAN-based refinement module and finally a CRNN based model is employed to predict the text sequence thus by recognizing the text in the scene images. The different modules supplement each other in such a manner that the whole system satisfactorily achieves the goal of detecting maximum scene texts along with least non-text components and accurately recognizing the detected scene texts. Further analysis is needed to improve the performance of the proposed system and tests need to be conducted to observe the robustness of the system against increasing complexity in images.

References

Bagi, R., Dutta, T., Nigam, N., Verma, D., & Gupta, H. P. (2021). Met-MLTS: leveraging smartphones for end-to-end spotting of multilingual oriented scene texts and traffic signs in adverse meteorological conditions. *IEEE Transactions on Intelligent Transportation Systems*, *23*(8), 12801-12810. https://doi.org/10.1109/TITS.2021.3117793

Cheng, P., Cai, Y., & Wang, W. (2019). A direct regression scene text detector with position-sensitive segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(11), 4171-4181. https://doi.org/10.1109/TCSVT.2019.2947475

Das, A., Palaiahnakote, S., Banerjee, A., Antonacopoulos, A., & Pal, U. (2024). Soft Set-based MSER End-to-End System for Occluded Scene Text Detection, Recognition and Prediction. *Knowledge-Based Systems*, 112593. https://doi.org/10.1016/j.knosys.2024.112593

Dutta, I. N., Chakraborty, N., Mollah, A. F., Basu, S., & Sarkar, R. (2019). Multi-lingual text localization from camera captured images based on foreground homogenity analysis. In *Recent Developments in Machine Learning and Data Analytics: IC3 2018* (pp. 149-158). Springer Singapore. https://doi.org/10.1007/978-981-13-1280-9_15

Epshtein, B., Ofek, E., & Wexler, Y. (2010, June). Detecting text in natural scenes with stroke width transform. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 2963-2970). IEEE. https://doi.org/10.1109/CVPR.2010.5540041

Fang, S., Mao, Z., Xie, H., Wang, Y., Yan, C., & Zhang, Y. (2022). Abinet++: Autonomous, bidirectional and iterative language modeling for scene text spotting. *IEEE transactions on pattern analysis and machine intelligence*, *45*(6), 7123-7141. https://doi.org/10.1109/TPAMI.2022.3223908

Geng, T. (2024). Transforming Scene Text Detection and Recognition: A Multi-Scale End-to-End Approach With Transformer Framework. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3375497

Gomez, L., & Karatzas, D. (2014, August). MSER-based real-time text detection and tracking. In *2014 22nd International Conference on Pattern Recognition* (pp. 3110-3115). IEEE. https://doi.org/10.1109/ICPR.2014.536

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*. https://doi.org/10.48550/arXiv.1406.2661

He, W., Zhang, X. Y., Yin, F., & Liu, C. L. (2017). Deep direct regression for multi-oriented scene text detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 745-753). https://doi.org/10.1109/ICCV.2017.87

Islam, M. R., Mondal, C., Azam, M. K., & Islam, A. S. M. J. (2016, May). Text detection and recognition using enhanced MSER detection and a novel OCR technique. In *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)* (pp. 15-20). IEEE. https://doi.org/10.1109/ICIEV.2016.7760054

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125-1134). https://doi.org/10.48550/arXiv.1611.07004

Kai, H. E., Jinlong, T. A. N. G., Zikang, L. I. U., & Ziqi, Y. A. N. G. (2024). HAFE: A Hierarchical Awareness and Feature Enhancement Network for Scene Text Recognition. *Knowledge-Based Systems*, *284*, 111178. https://doi.org/10.1016/j.knosys.2023.111178

Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., ... & Valveny, E. (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

Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L. G., Mestre, S. R., ... & De Las Heras, L. P. (2013, August). ICDAR 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition* (pp. 1484-1493). IEEE. https://doi.org/10.1109/ICDAR.2013.221

Khalid, S., Shah, J. H., Sharif, M., Dahan, F., Saleem, R., & Masood, A. (2024). A Robust Intelligent System for Text-Based Traffic Signs Detection and Recognition in Challenging Weather Conditions. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3401044

Koo, H. I., & Kim, D. H. (2013). Scene text detection via connected component clustering and nontext filtering. *IEEE transactions on image processing*, *22*(6), 2296-2305. https://doi.org/10.1109/TIP.2013.2249082

Liu, Y., Jin, L., & Fang, C. (2019). Arbitrarily shaped scene text detection with a mask tightness text detector. *IEEE Transactions on Image Processing*, *29*, 2918-2930. https://doi.org/10.1109/TIP.2019.2954218

Matas, J., Chum, O., Urban, M., & Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, *22*(10), 761-767. https://doi.org/10.1109/TIP.2019.2954218

Mirza, M. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*. https://doi.org/10.48550/arXiv.1411.1784

Mu, D., Sun, W., Xu, G., & Li, W. (2021). Random blur data augmentation for scene text recognition. *IEEE Access*, *9*, 136636-136646. https://doi.org/10.1109/ACCESS.2021.3117035

Mukhopadhyay, A., Kumar, S., Chowdhury, S. R., Chakraborty, N., Mollah, A. F., Basu, S., & Sarkar, R. (2019). Multi-lingual scene text detection using one-class classifier. *International Journal of Computer Vision and Image Processing (IJCVIP)*, *9*(2), 48-65. https://doi.org/10.4018/IJCVIP.2019040104

Panda, S., Ash, S., Chakraborty, N., Mollah, A. F., Basu, S., & Sarkar, R. (2020). Parameter tuning in MSER for text localization in multi-lingual camera-captured scene text images. In *Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2019* (pp. 999-1009). Springer Singapore. https://doi.org/10.1007/978-981-13-9042-5_86

Shi, B., Bai, X., & Yao, C. (2016). 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*(11), 2298-2304. https://doi.org/10.1109/TPAMI.2016.2646371

Sun, W., Wang, Q., Hou, Z., Chen, X., Yan, Q., & Zhang, Y. (2024). DPGS: Cross-cooperation guided dynamic points generation for scene text spotting. *Knowledge-Based Systems*, *302*, 112399. https://doi.org/10.1016/j.knosys.2024.112399

Tian, S., Zhu, K. X., Qin, H. B., & Yang, C. (2024). Dynamic receptive field adaptation for scene text recognition. *Pattern Recognition Letters*, *178*, 55-61. https://doi.org/10.1016/j.patrec.2023.12.005

Tong, G., Dong, M., Sun, X., & Song, Y. (2022). Natural scene text detection and recognition based on saturation-incorporated multi-channel MSER. *Knowledge-Based Systems*, *250*, 109040. https://doi.org/10.1016/j.knosys.2022.109040

Wu, L., Xu, Y., Hou, J., Chen, C. P., & Liu, C. L. (2022). A two-level rectification attention network for scene text recognition. *IEEE Transactions on Multimedia*, *25*, 2404-2414. https://doi.org/10.1109/TMM.2022.3146779

Wu, Y., Kong, Q., Qian, C., Nappi, M., & Wan, S. (2023). End-PolarT: Polar Representation for End-to-End Scene Text Detection. *Big Data Research*, *34*, 100410. https://doi.org/10.1016/j.bdr.2023.100410

Xu, Y., Liang, Z., Liang, Y., Li, X., Pan, W., You, J., ... & Scotti, F. (2024). Data-Driven Container Marking Detection and Recognition System with an Open Large-Scale Scene Text Dataset. *IEEE Transactions on Emerging Topics in Computational Intelligence*. https://doi.org/10.1109/TETCI.2024.3377680

Yan, X., Fang, Z., & Jin, Y. (2023). An adaptive n-gram transformer for multi-scale scene text recognition. *Knowledge-Based Systems*, *280*, 110964. https://doi.org/10.1016/j.knosys.2023.110964

Yao, C., Bai, X., Sang, N., Zhou, X., Zhou, S., & Cao, Z. (2016). Scene text detection via holistic, multi-channel prediction. *arXiv preprint arXiv:1606.09002*. https://doi.org/10.48550/arXiv.1606.09002

Ye, Q., & Doermann, D. (2014). Text detection and recognition in imagery: A survey. *IEEE transactions on pattern analysis and machine intelligence*, *37*(7), 1480-1500. https://doi.org/10.1109/TPAMI.2014.2366765

Yin, X. C., Pei, W. Y., Zhang, J., & Hao, H. W. (2015). Multi-orientation scene text detection with adaptive clustering. *IEEE transactions on pattern analysis and machine intelligence*, *37*(9), 1930-1937. https://doi.org/10.1109/TPAMI.2014.2388210

Yu, W., Liu, Y., Zhu, X., Cao, H., Sun, X., & Bai, X. (2024). Turning a clip model into a scene text spotter. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2024.3379828

Zhang, J., & Kasturi, R. (2014). A novel text detection system based on character and link energies. *IEEE Transactions on Image Processing*, *23*(9), 4187-4198. https://doi.org/10.1109/TIP.2014.2341935

Zhang, Z., Shen, W., Yao, C., & Bai, X. (2015). Symmetry-based text line detection in natural scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2558-2567). https://doi.org/10.1109/CVPR.2015.7298871

Zhou, G., Liu, Y., Tian, Z., & Su, Y. (2011, September). A new hybrid method to detect text in natural scene. In *2011 18th IEEE International Conference on Image Processing* (pp. 2605-2608). IEEE. https://doi.org/10.1109/ICIP.2011.6116199