Int J Syst Assur Eng Manag (December 2023) 14(6):2289–2300
https://doi.org/10.1007/s13198-023-02070-4

ORIGINAL ARTICLE

Rfpssih: reducing false positive text detection sequels in scenery images using hybrid technique

Avaneesh Kumar Yadav[1] · Animesh Sharma[2] · Vikas Yadav[1] · Neha Kalia[3]

Received: 27 April 2022 / Revised: 29 June 2023 / Accepted: 19 July 2023 / Published online: 12 September 2023

Abstract Text detection from scenic photographs with text is a difficult issue that has recently attracted a lot of attention. There are two main elements in scenery photographs (1) Recognizing text in photographs and (2) Character recognition. The model's entire accuracy depends on the output of this phase, finding the text in the photos is the most crucial aspect. An approach consisting of two phases has been proposed in this article. (1) Text recognition and (2) Text checker. Text detection is accomplished using the Maximally Stable Extremal Regions (MSER) feature detector. The output of the MSER feature detector is subjected to various filters in order to exclude components, i.e., unlikely to contain text. The second phase uses a machine learning methodology to classify the text and non-text on phase-1 final output. It has been discovered that the proposed method nearly removes all false-positive results on the MSER method's final output.

✉ Vikas Yadav
1999vikasy@gmail.com

Avaneesh Kumar Yadav
avaneesh0741@gmail.com

Animesh Sharma
er.animeshsharma@gmail.com

Neha Kalia
nehakalia111@gmail.com

1 Department of Computer Science and Engineering, Motilal Nehru National Institute of Technology Allahabad, Prayagraj 211004, India

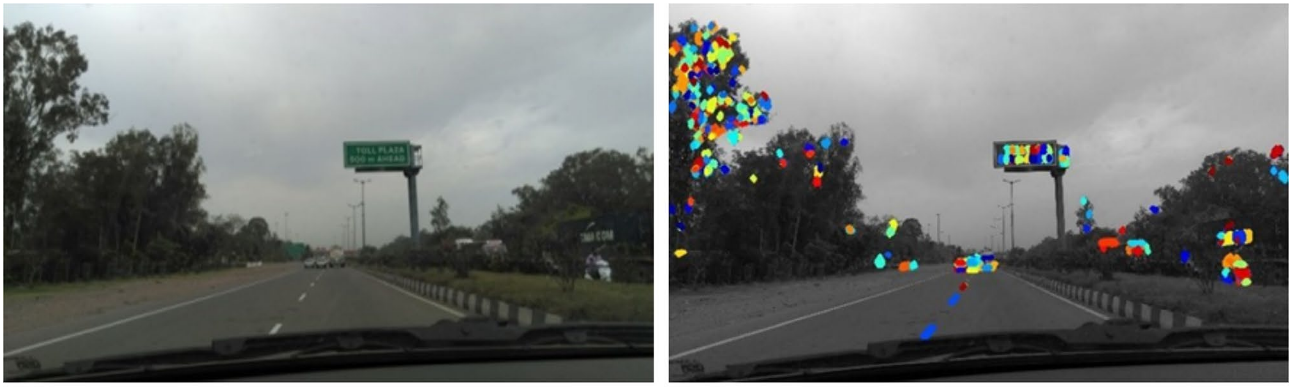
2 Department of Computer Science and Engineering, Thapar University, Patiala 147004, India

3 Department of Computer Science , Hindu Girls College, Sonipat, Haryana 131001, India

Keywords Text detection · Scenery photos · Artificial neural network · MSER

1 Introduction

Extracting text from scenery images is a difficult task that has many practical applications, such as assisting blind people in finding images based on text in them and car navigation systems that automatically read street signboards and navigate the car or send alert messages to the driver based on the sign and text on the street board. Traditional OCR systems are designed for scanned documents; therefore, one cannot directly apply the scenery image in the input of the OCR system without segmenting the text region (Imam et al. 2022). In a traditional OCR system, one has to correctly isolate the characters from the background pixels and then recognize them. In the scenery image, isolating the characters from the background pixel is difficult because of the random background color, noise, lightning effect, etc. Also, the page layout for the traditional OCR system is well structured, but this is not the case for the scenery images (Yang et al. 2019) because there are only a few texts and lots of variable structures with different geometry and appearances. This paper's hybrid approach consists of two phases - phase-1 is based on the connected component approach, and phase 2 is based on the machine learning approach. In phase-1 MSER feature detector (Chen et al. 2011; He et al. 2016; Gupta and Jalal 2019; Ch'ng et al. 2020; Rashtehroudi et al. 2023) is applied to detect the text, and from the image shown in Fig. 1, an impression has emerged that the MSER feature detector works well in scenery images. It works well because the text has a consistent color and strong contrast, resulting in a consistent intensity profile. However, the MSER feature



(a) (b)

Fig. 1 **a** Original images, **b** MSER regions

detector has a problem with many non-text regions being recognized alongside the text.

Suppose the output of the MSER feature detector is processed in the further stages of the OCR (Imam et al. 2022; Gupta and Jalal 2019; Tong et al. 2022; Rajeswari and Aradhana 2021). In that case, the overall accuracy will be reduced by a large amount because it contains a large number of false-positive regions. A collection of data (Rashtehroudi et al. 2023; Hao et al. 2016) contains more than 90 pictures of scenes with various amounts of architecture such as geometry, blur (haze), color, and appearances. On all of dataset-1's sample photos, the MSER feature detector is used, and from its output, an investigation is done to determine which types of regions give false-positive results. According to the investigation, background elements like building windows, rooftops, tree branches, and background components including tree leaves frequently provide false-positive results (Tong et al. 2022).

The research mentioned above produces a new dataset (Gupta and Jalal 2019; Soni et al. 2020) that includes 100 photos of text-free regions and 100 images from non-text areas. Give this dataset the identification as dataset-1. Figure 2 shows several random photos from dataset-1.

The dataset should be referred to as dataset 2. Some random photos of dataset-2 are shown in Fig. 3.

The goal of this study is to reduce the number of false-positive regions. More filters are applied to the candidate text region (output of MSER feature detector). Then (1) elimination of non-text regions in the output of the MSER feature detector is done based on simple geometric properties. (2) Another filter is applied to eliminate some more non-text regions in the output of step 1 based on stroke width variation. (3) The output of step 2 is finally fed into the ANN classifier (ANN classifier is trained by dataset-2 so that it has the ability to classify regions as text or non-text) in our test dataset, which contains more than 25 photos and achieves state-of-the-art results. We'll refer to the test dataset as dataset-3. Table 1 frequently refers to terms used in this article.

1.1 Paper organization

The rest of the section follows as the related study of word extraction in scenery images is presented in Sect. 2. Section 3 is a brief overview of the proposed methodology. The experimental findings of the suggested algorithm are described in Sect. 4, and the results are compared to the

Fig. 2 Sample images of the dataset-1

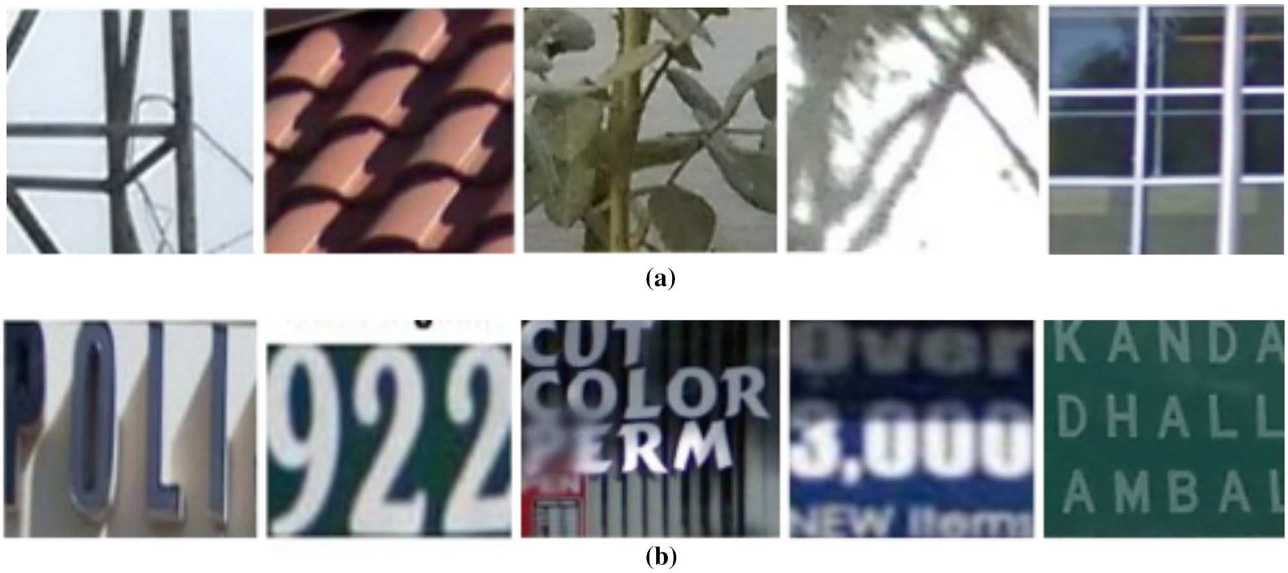


Fig. 3 Sample images of the dataset-2 containing **a** Non-text regions **b** Text regions

Table 1 Frequently used abbreviation

Abbrivation	Full name	Abbrivation	Full name
P(1)	Positive (1)	SVT	Street view text
TP	True positive	ANN	Artificial Neural Network
NN	Neural network	CRF	Conditional random field
N(0)	Negative (0)	CTW	Chinese text-dataset Wild
TN	True negative	CNN	Convolutional Neural Network
FN	False negative	OCR	Optical character recognition
ML	Machine learning	MSER	Maximally stable extremal regions
PV	Predicted value	MOSTL	Multi-oriented scene text localization
AV	Actual value	ICDAR	International Conference on Document Analysis and Recognition
FP	False positive		

results of different algorithms. Section 5 is discussed the conclusion and future scope.

2 Related work

Finding text from scenery photographs and video frames takes a lot of effort. The extraction of linear characteristics is a topic that has been studied in several disciplines. A comprehensive review of several text detection methods may be found in the literature (Liang et al. 2005; Panchal et al. 2022). In general, the approaches for locating text in photographs can be classified into two types: (a) texture-based methods (b) region-based techniques. Texture-based approaches (Naiemi et al. 2021; Yang et al. 2022) scan images at various sizes, classifying pixel neighborhoods based on various text properties. Using conditional random fields (Gllavata et al. 2004), A cluster of filters was used by Gllavata et al. (2004) to study the appearance scenery in different blocks and the joint texture variations in adjacent regions. The disadvantage of these methods is that Before categorizing the image, they employed a non-content based image region to divide it into equal-sized segments. The non-content based picture divider can break down text characters into pieces, but it doesn't satisfy the texture constraints. The drawbacks of this texture-based technique are that it is computationally intensive. This is because photos must be scanned at various scales. Naiemi et al. (2021) discussed the MOSTL techniques. The proposed approach included an enhanced ReLU-layer and an enhanced inception-layer. The design idea is first utilized to extract basic visual information. After that, a further layer was added to enhance feature extraction. The text identification process has benefited from the addition of the i.inception layers and i.ReLU. The output from the i.inception layers and i.ReLU was supplied with an additional layer that allows MOSTL to recognize multi-oriented texts, including vertical and curved features. He et al. (2016) recommended applying scene text recognition to the just-developed CE-MSER detector. A picture from CNN's competent text classifier was used to categorize text features. When employing this techniques, only for the horizontally texts are discovered (Rashtehroudi et al. 2023). Gupta and Jalal (2019) presented a novel approach for locating prominent text in a natural environment. They combined the benefits of the Grabcut segmentation method and the capabilities of the MSER detector. The key components of the natural landscape are identified using the Zhang model. Li et al. (2000) discussed the two types of texture-based approaches such as block-based texture and cell-based texture methods. The text features are retrieved for a specific area, and the text is detected using a classifier. Baran et al. (2018) introduced a new approach based on threshold finding (Neumann and Matas 2011) that took into account the vertical location, height, and energy of each character with two neighboring characters prior to the collection of stages. This approach seeks to achieve its goals by removing non-characters, decreasing the number of unwanted results, and creating an initial setting for the rest of the characters. Another arbitrary-shape scenery text benchmark called Total-Text (Ch'ng et al. 2020) has almost 1,256 training and nearest 300-test images. Every text instance has a word-level polygon with an adjustable number of important points annotating it. Ye et al. (2005) introduced a character component technique based on deep learning that focuses on locating each character in the scene image. At the character level, the approach is used for both fundamental image ground truths and synthetic picture character level labeling. The affinity estimation between characters is also incorporated into the algorithm. Different benchmarked datasets were employed for evaluation, including ICDAR-2013, 2015, and 2017, MSRA-TD500, Total Text, and CTW-1500, which attained 95.1, 86.8, 73.9, 82.7, 83.4, and 83.2 F_1-measures, respectively. Weinman et al. (2004) introduced a method for detecting text in photos of scenes. MSERs are used to identify potential text patches in complex backdrop imagery. A CRF-based model was also utilized to distinguish between text and non-text portions of the image. The missing text of obtained MSERs is recovered using the text's context information. The characters on a line were then retrieved using the clustering technique. The canny edge detector is then used to divide the grouped line into words. To improve the system, the false-positive text is deleted with binary and grey images. Because of its resilience and speed, a random forest shape-specific classifier is employed to gaining a secure text area. The ICDAR-2005, ICDAR-2011, ICDAR-2013, and SVT datasets were utilized to evaluate performance and yielded F_1-measures of 0.75, 0.77, 0.76, and 0.41, respectively. Slanted text detection can help enhance the model. Furthermore, these approaches are incapable of detecting slanted text. The regions are used in the other set of text detection algorithms (Wang et al. 2018). The pixels in these algorithms indicate particular features. Approximately constant color, for example, is clustered together. Epshtein et al. (2010) recently developed a content based segmentation known as the stroke-width transform to extract text characters with consistent stroke-widths. The width of every picture pixel is determined to capture the value of the stroke and thus verify its usage in the work of scene text detection in scenery photographs (He et al. 2016; Naiemi et al. 2021). This technique is intriguing since it can recognize texts of various scales at the same time and is not limited to horizontal texts. MSER is a region-based approach for detecting features. However, the problem with this method is that it generates much too many false positives. This problem can be solved using the given method.

3 Proposed methodology

Hybrid approaches are sometimes preferred for better performance. consequently the various techniques are used in the resultant output of the MSER feature identified for better performance. Figure 4 shows the flow diagram for extracting the text from the scenic photograph.

There are two phases in the flow diagram, and the MSER feature detector is used in the first phase. Filters are used in the second step, such as the removal of non-text areas based on fundamental geometric parameters. To screen out areas that are unlikely to contain text, non-text sections can be removed based on variations in stroke width. After that, the text regions are segmented and sent into the ANN classifier, which functions as a text verifier. The input photos can be identified as text or non-text; an ANN classifier was used. Given a segmented region of 30×30 pixels, the presence or absence of textual areas in the viewing area can be determined by the ANN classifier. Because the ANN classifier has two outputs (text and non-text), it is utilized as a binary classifier. The dataset-2 is used to train the ANN classifier. Figure 5(a) depicts our neural network's learning architecture. Due to learning complicated nonlinear elements is one of the ANN classifier's fields of study, it is used. The three layers employed are the input layer, the concealed layer, and the output layer. Due to the image's size of 30 by 30, 900 input layer units were utilized (but we do not include the additional bias unit; then it always yields plus one). We used the variables x and y to represent the training data, and we randomly initialized the weights for the ANN variables W(1) and W(2). In a separate mat file, we saved the dataset as well as the ANN classifier's weight. The settings are sized for a second-layer artificial neural network with 60 units and two output units (falling under the two categories of text and

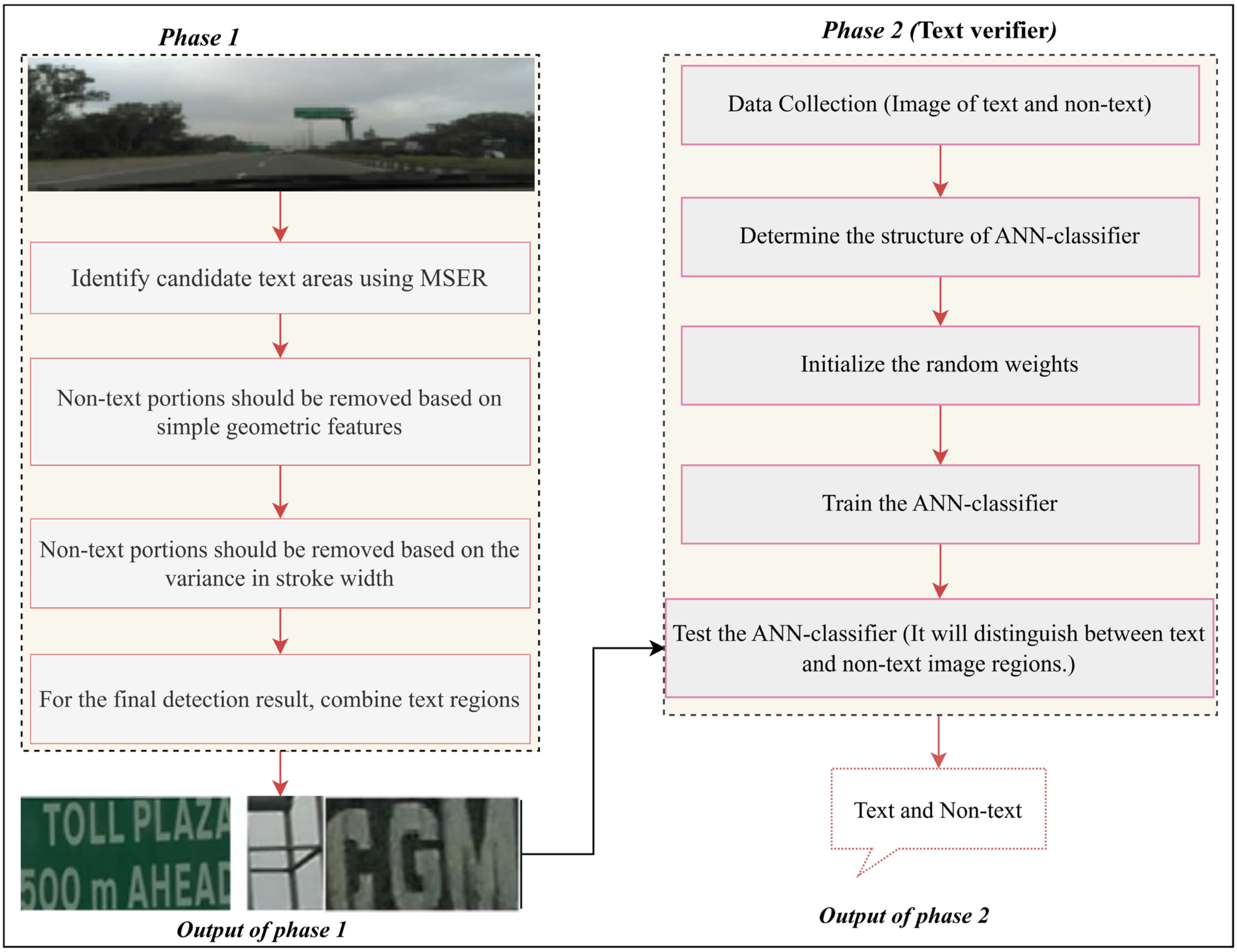


Fig. 4 Flowchart of the proposed method

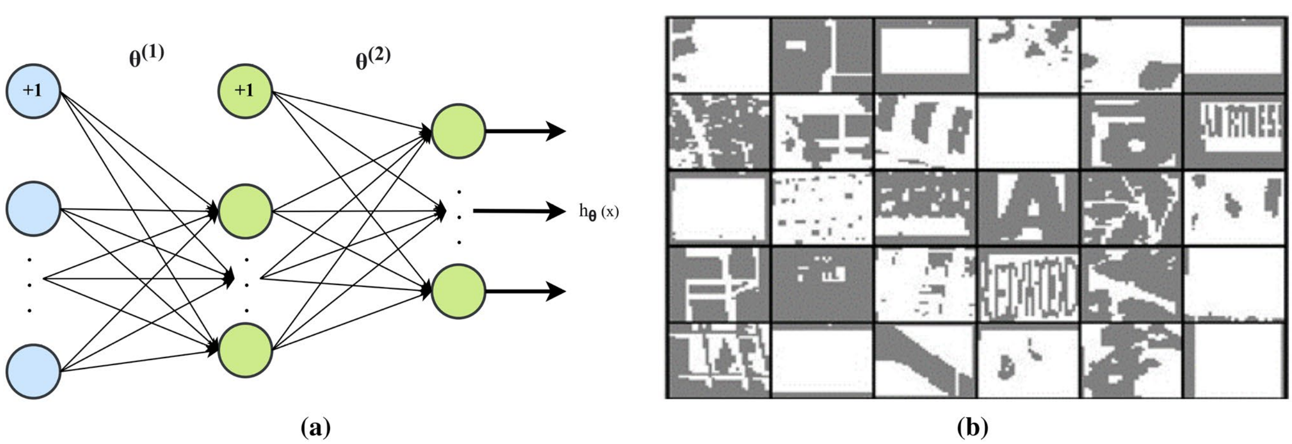


Fig. 5 **a** Learning architecture of the ANN classifier and **b** 30 random grayscale images of dataset-2

non-text). Figure 5(b) depicts dataset-2's 30 random grayscale photos.

The following list outlines the approaches used in order to identify the text section in scenery images:

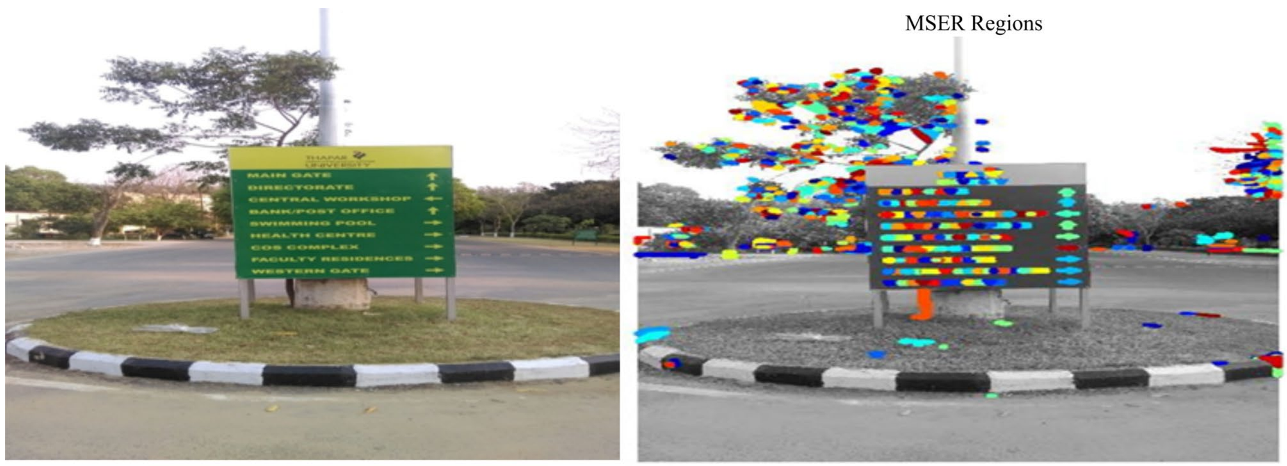
Step 1: It has been possible to locate probable text sections by using the MSER feature detector. Figure 6 displays the original image and the specified region after applying the MSER features detectors.

Step 2: Non-text regions are removed based on basic geometric features.

- Different geometric features, such as Euler number, extent, aspect ratio, eccentricity, and solidity, have been utilized to distinguish between text and non-text regions based on a simple threshold value.
- Fig. 7(a) shows how non-text portions can be removed using simple geometric features.

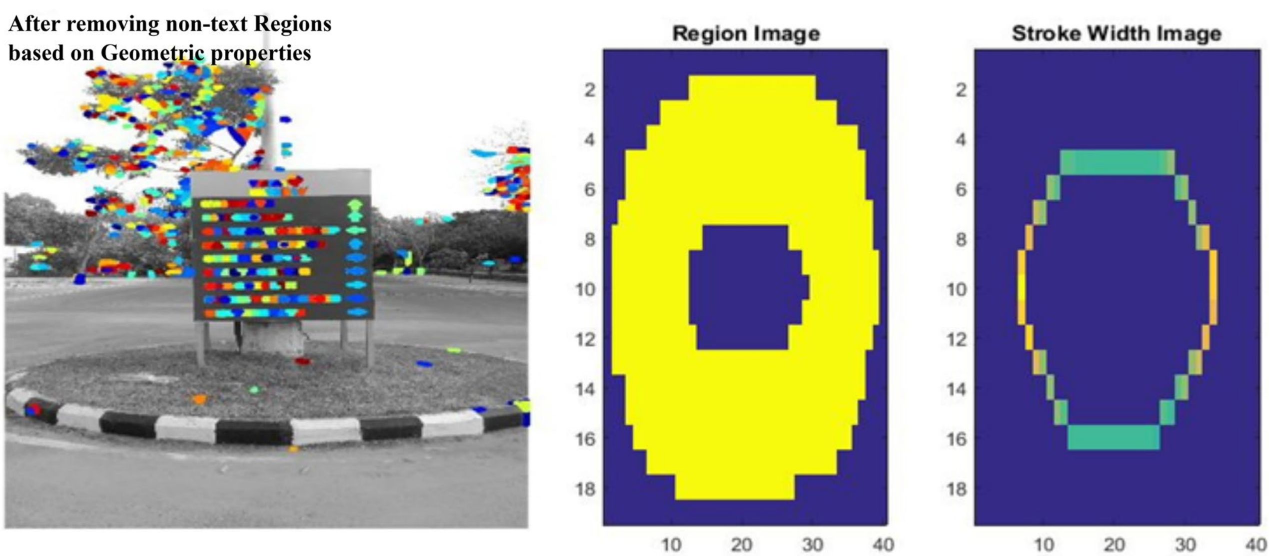
Step 3: Based on the stroke width variation, non text parts are removed.

- The width of the curves and lines that make up a character determines the variation in stroke width. The stroke width of the text parts varies only slightly, whereas the non-text sections have a wide range of stroke widths. Figure 7(b) shows how the stroke width fluctuates sig-



(a) **(b)**

Fig. 6 **a** Actual image **b** the result of using the MSER feature detector



(a) **(b)**

Fig. 7 **a** Image after geometric properties-based removal of the non-text section **b** Example of uniform stroke variation in the text

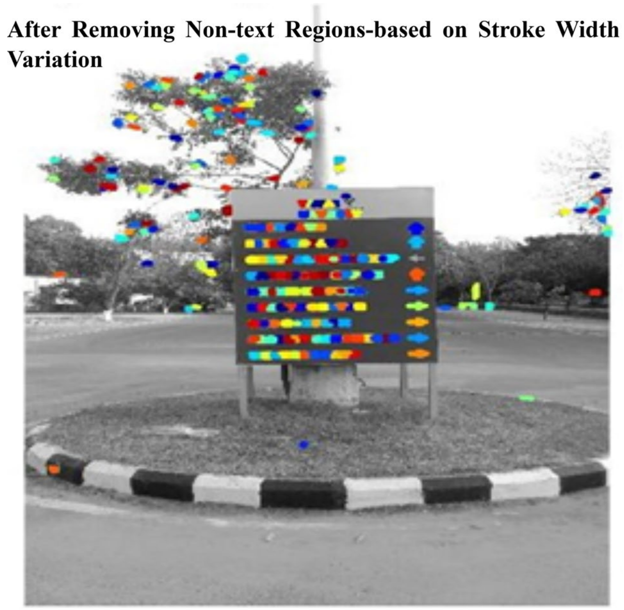


Fig. 8 After deleting non-text based on stroke width fluctuation, the image looks like this

nificantly over different locations because the widths of the curves and lines are frequently equal.

- Fig. 8 depicts the impact of eliminating non text portions through the variance in stroke-width after removing the non-text portions:

Step 4: Combining characters into words or lines of text.

- Given that step three's outcome comprises unique characters, The practice of blending text sections into words or lines by using nearby text has been uncovered. After completing step 4, Fig. 9 shows the results in terms of bounding boxes.
- A bounding box is deleted if its pairwise overlap ratio is less than two after finding the pairwise overlap ratio.

Step 5: The output of step 4 is verified as non text and text using the ANN classifier, which also serves as the verifier. Following are the many steps to implement an ANN classifier: Step 5: The ANN classifier serves as the verifier to check the output of step 4's text areas, including non text and text. The multiple steps required to create the ANN classifier are as follows:

1. Create the training data collection and network architecture for the neural network.
2. The issue of both overfitting and underfitting has been solved using a regularized cost function. Equation (1) is applied to the cost function to calculate it, as shown below:

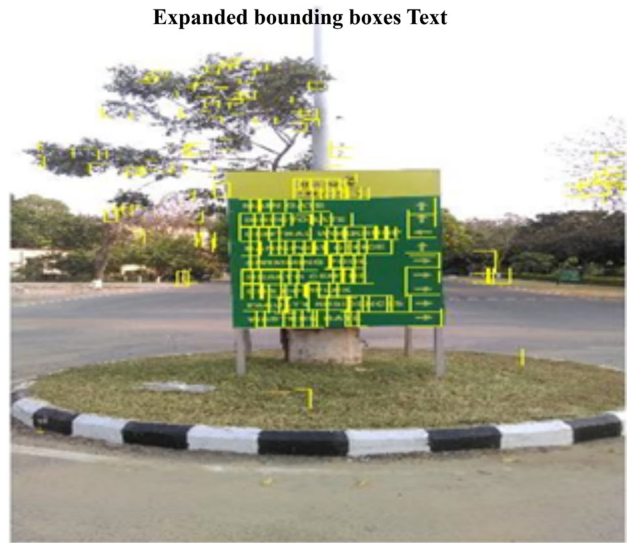


Fig. 9 Individual text reflected by the bounding box

$$J(\theta) = [A] + [B] \tag{1}$$

Calculate the $J(\theta)$ with the help of Eqs. (2) and (3).

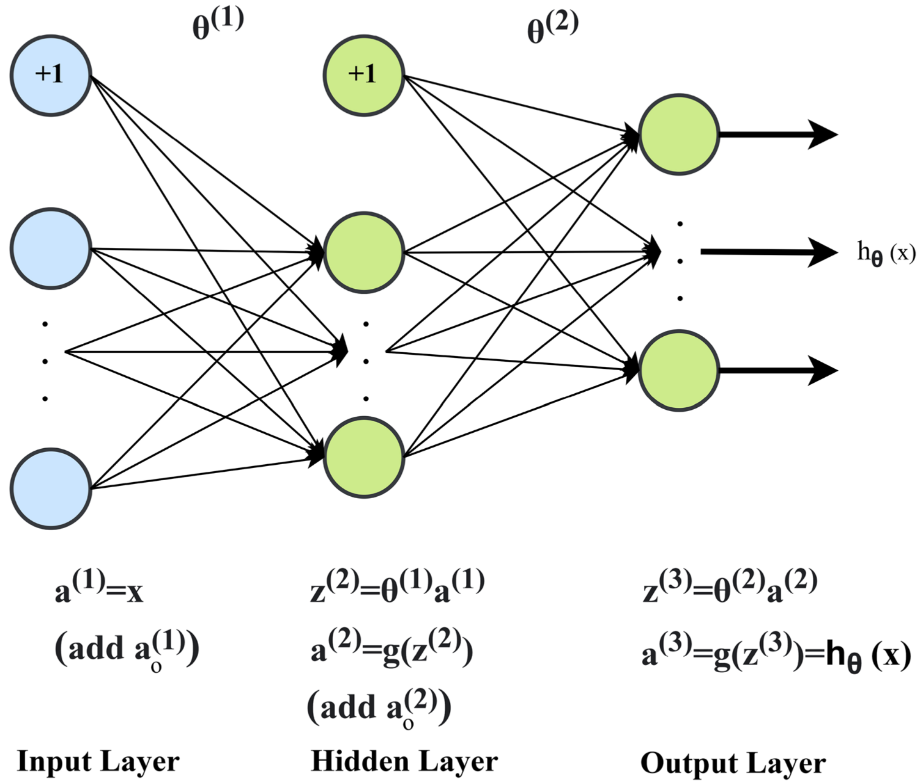
$$[A] = \frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{K} \left[-y_k{}^{(i)} \log \left(\left(h_\Theta \left(x^{(i)} \right) \right)_k \right) - \left(1 - y_k{}^{(i)} \right) \log \left(1 - \left(h_\Theta \left(x^{(i)} \right) \right)_k \right) \right] \tag{2}$$

$$[B] = \frac{\lambda}{2m} \left[\sum_{j=1}^{60} \sum_{k=1}^{900} \left(\Theta_{j,k}{}^{(1)} \right)^2 + \sum_{j=1}^{60} \sum_{k=1}^{2} \left(\Theta_{j,k}{}^{(2)} \right)^2 \right] \tag{3}$$

where $J(\Theta)$= Regularized cost function, j= number of units in second layer, $\left(h_\Theta \left(x^{(i)} \right) \right)_k$= output value of K^{th} output unit, K= number of output layer, $x^{(i)}$= i^{th} training example, m = number of pixels applied as the input in layer 1, λ = Regularized parameter, $y_k{}^{(i)}$= value of output k in i^{th} training example.

3. The ANN classifier's weights have been initialized at random and are close to zero.
4. The hypothesis is then obtained by using forward propagation. The various steps are given below: The output value of the hypothesis $h_\Theta(x)$, given a training example $(x^{(t)}, y^{(t)})$, is computed using the "forward pass," which computes all activations across the whole network. Where the number of layers is l, $a^{(l)}$=activation of layer l, it's set to the t^{th} training example x(t) input layer values (a(1)) Then, using Eq. (4), evaluate the activations $((a^{(2)}, z^{(2)}), (a^{(3)}, z^{(3)}))$ for layers 2 and 3 in a feedforward pass (Fig. 10). To guarantee the vectors of activations applicable in the layers of a(1) and a(2) that are included in the bias unit, another term (a+1) must be included.

Fig. 10 Forward Propagation



In layer 1, activation is performed through input. Mathematically different steps are:

$$\left.\begin{array}{l} a^{(1)} = x \\ z^{(2)} = a^{(1)} \times \Theta^{(1)} \\ sigmoid(z) = g(z) = \dfrac{1}{1+e^{-z}} \\ a^{(2)} = g\left(z^{(2)}\right) \\ z^{(3)} = a^{(2)} \times \Theta^{(2)} \\ a^{(3)} = g\left(z^{(3)}\right) \end{array}\right\} \tag{4}$$

5. Equation (1) is used to determine the cost function during the code-development process. In order to determine the cost function, the code is implemented using Equation (1).
6. The back propagation code has been implemented to compute partial derivatives. Below are the different steps to do this task:

   (a) Layer 3 is the output layer, which employs Eq. (5) to calculate the error for each output unit k.

   $$\delta_k^{(3)} = (a_k^{(3)} - y_k) \tag{5}$$

   Where $a_k^{(3)}$= activation unit k in layer-3, $y_k \in 1, 2$, If y_k=1 belongs to class "k", or if y_k= 2 belongs in different class, it is marked as an example of current training.

   (b) Middle layer is known as the hidden-layer, i.e., hidden-layer l = 2; Eq. (6) determines the error.

   $$\delta^{(2)} = \left(\Theta^{(2)}\right)^T \delta^{(3)}. * g\prime(z^{(2)}) \tag{6}$$

   (c) The gradient is accumulated using the following Eq. (7).

   $$\Delta^{(l)} = \Delta^{(l)} + \delta^{(l+1)}\left(a^{(l)}\right)^T \tag{7}$$

   (d) Divide the collected gradients by $\frac{1}{m}$ to get the (unregularized) gradient for the NN cost function, as shown in Eq. (8).

   $$\frac{\partial}{\partial \Theta_{ij}^{(l)}} J(\Theta) = D_{ij}^{(l)} = \frac{1}{m} \Delta_{ij}^{(l)} \tag{8}$$

7. The numerical estimation of the cost function gradient and the partial derivatives are then compared using gra-

dient checking. The parameters are checked for gradients, which can be done by "unrolling" the parameters $\Theta^{(1)}, \Theta^{(2)}$ into a long vector Θ. Instead of thinking of the cost function as $J(\Theta)$, one can think of it as $J(\Theta)$ and use the gradient checking approach below. Function $f_i(\Theta)$ that purports to compute $\frac{\partial}{\partial\theta_i}J(\Theta)$; it's a good idea to double-check that f_i is returning correct derivative values.

$$\left.\begin{array}{l} \Theta^{(i+)} = \Theta + \begin{bmatrix} 0 \\ 0 \\ . \\ . \\ \varepsilon \\ . \\ . \\ . \\ 0 \end{bmatrix} \\ \Theta^{(i-)} = \Theta - \begin{bmatrix} 0 \\ 0 \\ . \\ . \\ \varepsilon \\ . \\ . \\ . \\ 0 \end{bmatrix} \end{array}\right\} \tag{9}$$

After finding the $\Theta^{(i+)}$ and $\Theta^{(i-)}$ through Eq. (9). According to, $\Theta^{(i+)}$ is same as "Θ", to except that its i^{th}-element is increased by ε. Similarly, $\Theta^{(i-)}$ is the corresponding-vector, but with the i^{th} member reduced by ε. By checking the Eq. (10) for each i, one can now quantitatively verify the accuracy of $f_i(\Theta)$'s.

$$f_i(\Theta) \approx \frac{J\left(\Theta^{(i+)}\right) - J\left(\Theta^{(i-)}\right)}{2\varepsilon} \tag{10}$$

8. Backpropagation and advanced optimization techniques were employed to minimize the cost function.

Testing was done on dataset-3, which comprises more than 25-photos. After developing the ANN-classifier using the preceding techniques, it was discovered that the text verifier practically removes all false-positive findings.

4 Experiment result and analysis

The experimental results for each methodology are determined in terms of recall, precision, and F_1-score value. These terms are definded individually with mathematical formula that is used in image technology. "True Positive (TP) is taken as text area correctly identified as text area and False Positive (FP) as non-text area incorrectly identified as a text area." "True Negative (TN) as non-text area identified as non-text area and False Negative (FN) as text area incorrectly identified as non-text area" (Yadav et al. 2021, 2023). All these three term are calculated based on confusion matrix, i.e, shown Table 2. Confusion matirix are four possible predicted and actual value combination in the Table 2.

Precision The precision is the percentage of the predicted text appearing in the photograph out of all the information in the photograph. Its mathematical formulas are shown in Eq. (11) (Yadav et al. 2021, 2023).

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

Recall The percentage of accurately identified text out of all text that is true text is known as recall. Its mathematical formulas are shown in Eq. (12) (Yadav et al. 2021, 2023).

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

F_1-score This is calculated through precision and recall. It follows the harmonic process, and its mathematical formula is shown in Eq. (13) (Yadav et al. 2021, 2023).

$$F_1 - score = \frac{1}{\frac{\alpha}{precision} + \frac{1-\alpha}{Recall}} \tag{13}$$

The standard F_1-measure is often used to integrate precision and recall. The relative weight can be adjusted using the variable. α has been adjusted to 0.5 to give recall and precision equal weight. Therefore, imbalanced weight produced a greater recall in the $\alpha = 0.75$ setting (recall value = 0.75, precision value = 0.25). The precision and recall for phase-1 are 0.45 and 0.84, respectively. The precision rises to 0.87 after incorporating the ANN classifier (phase-2), but the recall remains unchanged at 0.84 since the text region that was not retrieved in phase-1 cannot be retrieved in phase-2. The Table 3 contains a various methods along with reults. The $ICDAR_1$ dataset findings are obtained for the comparative results. The particular dataset in the table references dataset-3.

Table 2 Representation of confusion matrix

Confusion Matrix		AV	
		Positive (1)	Negative (0)
PV	Positive (1)	TP	FP
	Negative (0)	FN	TN

Table 3 Experimental result based on proposed work

Methods	Dataset	Recall	Precision	F_1-score
Todoran (TD)	$ICDAR_1$	0.18	0.19	0.18
Jisoo Kim (JK)	$ICDAR_1$	0.28	0.22	0.22
Qiang Zhu (QZ)	$ICDAR_1$	0.40	0.33	0.33
Hinnerk Becker (HB)	$ICDAR_1$	0.62	0.67	0.58
Nobuo Ezaki (NE)	$ICDAR_1$	0.36	0.18	0.22
Epshtein (EPN)	$ICDAR_1$	0.61	0.72	0.66
Gupta et al. (GP)	$ICDAR_1$	0.77	0.87	0.82
He at al. (HE)	$ICDAR_1$	0.80	0.82	0.81
Ashida (ASH)	$ICDAR_1$	0.46	0.55	0.50
Wolf (WF)	$ICDAR_1$	0.44	0.30	0.35
Alex Chen (AC)	$ICDAR_1$	0.60	0.60	0.58
HWDavid (HWD)	$ICDAR_1$	0.46	0.44	0.45
Proposed (Our*)	*Special*	0.80	0.89	0.84

Table 3 shows experimental values of different state-of-the-art methods and their comparison of recall, precision, and F_1-score, shown in graphical ways as Figs. 11, 12 and 13. The proposed algorithm has better performance than other state-of-the-art methods which is described in this section. The next section discuss the conclusion and future scope of this article.

5 Conclusion and future scope

The latest research on text detection methods is developed using a hybrid methodology with two components. The connected component method (MSER) with numerous filters makes up the first phase, while the ML approach makes up the second (ANN-classifier). When the proposed approach is used on scenery photographs with text, it is discovered that phase-1 yields a precision and recall rate is 0.45 and 0.84, respectively. The precision and recall rate improve by 0.87 and 0.84 once the ANN classifier is integrated (text verifier-phase-2). Any approach whose output produces a significant amount of false positive results can be merged with phase-2 (text verifier) to enhance the system's performance. It is shown that the best results are obtained by identifying the type of structure in landscape photographs that produce false positive results and then using that sort of structure to train the ANN-classifier. The text verifier uses minimal computational resources than the sliding window approach since it only needs to execute one scan in the text-containing sections before classifying the image as text or non-text. When compared to other methodologies, our proposed approach shows the highest recall and precision values. The ANN classifier makes the highest results even if it is trained on a dataset of just 200-photos. Increasing the number of training set photographs on phase-2 will try for more performance improvement in the future.

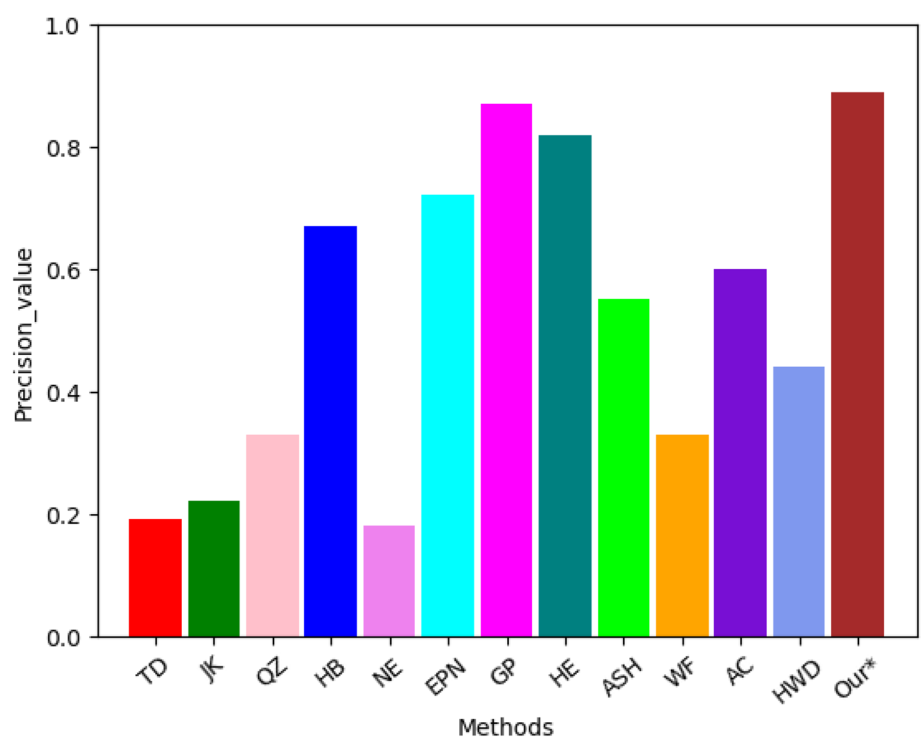


Fig. 11 Comparison of various methods for experimental precision metric effects based on scenery images text detection

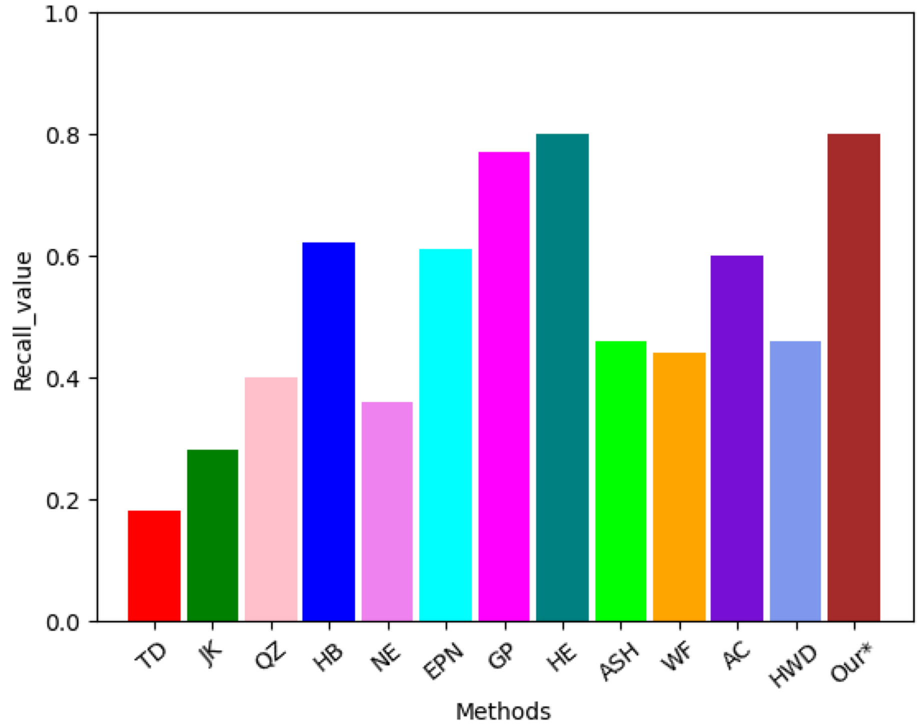


Fig. 12 Comparison of various methods for experimental recall metric effects based on scenery images text detection

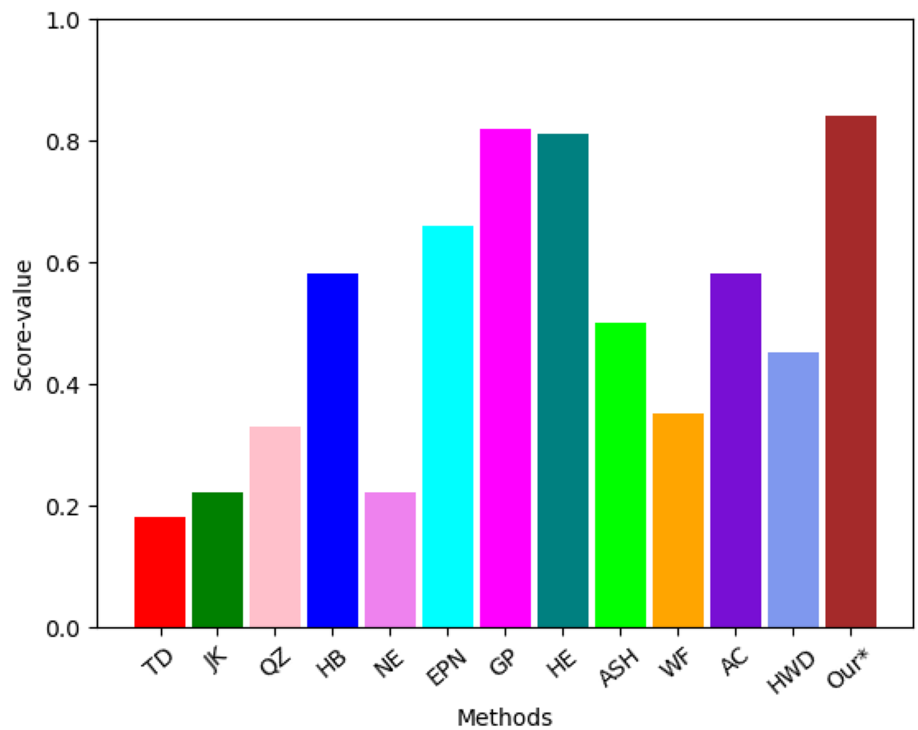


Fig. 13 Comparison of various methods for experimental F-score metric effects based on scenery images text detection

Acknowledgements We would like to thank to reviewers.

Funding No funding.

Declarations

Conflict of interest None.

Informed Consent There is no plagiarized.

Human and animals participants None.

References

Baran R, Partila P, Wilk R (2018) Automated text detection and character recognition in natural scenes based on local image features and contour processing techniques, in: Intelligent human systems integration: proceedings of the 1st international conference on intelligent human systems integration (IHSI 2018): Integrating people and intelligent systems, January 7-9, 2018, Dubai, United Arab Emirates, Springer, pp. 42–48

Chen H, Tsai SS, Schroth G, Chen DM, Grzeszczuk R, Girod B, Robust text detection in natural images with edge-enhanced maximally stable extremal regions, in, (2011) 18th IEEE international conference on image processing. IEEE 2011:2609–2612

Ch'ng C-K, Chan CS, Liu C-L (2020) Total-text: toward orientation robustness in scene text detection. Int J Doc Anal Recognit (IJDAR) 23(1):31–52

Epshtein B, Ofek E, Wexler Y, Detecting text in natural scenes with stroke width transform, in, (2010) IEEE computer society conference on computer vision and pattern recognition. IEEE 2010:2963–2970

Gllavata J, Ewerth R, Freisleben B (2004) Text detection in images based on unsupervised classification of high-frequency wavelet coefficients, in: Proceedings of the 17th international conference on pattern recognition, 2004. ICPR 2004., Vol. 1, IEEE, pp. 425–428

Gupta N, Jalal AS (2019) A robust model for salient text detection in natural scene images using mser feature detector and grabcut. Multimed Tools Appl 78:10821–10835

Hao M, Shi W, Zhang H, Wang Q, Deng K (2016) A scale-driven change detection method incorporating uncertainty analysis for remote sensing images. Remote Sens 8(9):745

He T, Huang W, Qiao Y, Yao J (2016) Text-attentional convolutional neural network for scene text detection. IEEE Trans Image Process 25(6):2529–2541

Li H, Doermann D, Kia O (2000) Automatic text detection and tracking in digital video. IEEE Trans Image Process 9(1):147–156

Liang J, Doermann D, Li H (2005) Camera-based analysis of text and documents: a survey. IJDAR 7:84–104

Imam NH, Vassilakis VG, Kolovos D (2022) Ocr post-correction for detecting adversarial text images. J Inform Secur Appl 66:103170

Naiemi F, Ghods V, Khalesi H (2021) Mostl: an accurate multi-oriented scene text localization. Circuits Syst Signal Process 40:4452–4473

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

Neumann L, Matas J (2011) A method for text localization and recognition in real-world images, in: Computer Vision–ACCV 2010: 10th Asian conference on computer vision, Queenstown, New Zealand, November 8-12, 2010, Revised Selected Papers, Part III 10, Springer, pp. 770–783

Panchal BY, Chauhan G, Panchal SR, Chaudhari UM (2022) An investigation on feature and text extraction from images using image recognition in android. Mater Today Proc 51:798–802

Rajeswari R, Aradhana B (2021) Character recognition in scene images using mser and cnn, in: Cognition and recognition: 8th international conference, ICCR 2021, Mandya, India, December 30–31, 2021, Revised Selected Papers, Springer,2023, pp. 99–107

Rashtehroudi AR, Akoushideh A, Shahbahrami A (2023) Pestd: a large-scale persian-english scene text dataset. Multimed Tools Appl. https://doi.org/10.1007/s11042-023-15062-0

Soni R, Kumar B, Chand S (2020) Text region extraction from scene images using agf and mser. Int J Image Graphics 20(02):2050009

Tong G, Dong M, Sun X, Song Y (2022) Natural scene text detection and recognition based on saturation-incorporated multi-channel mser. Knowl-Based Syst 250:109040

Wang X, Liu S, Du P, Liang H, Xia J, Li Y (2018) Object-based change detection in urban areas from high spatial resolution images based on multiple features and ensemble learning. Remote Sens 10(2):276

Weinman J, Hanson A, McCallum A (2004) Sign detection in natural images with conditional random fields, in: Proceedings of the 2004 14th IEEE signal processing society workshop machine learning for signal processing, IEEE, 2004, pp. 549–558

Yadav AK, Maurya AK, Yadav RS, et al (2021) Extractive text summarization using recent approaches: a survey, Ingénierie des Systèmes d'Information 26(1)

Yadav AK, Yadav RS, Maurya AK, et al (2023) State-of-the-art approach to extractive text summarization: a comprehensive review, Multimed Tools Appl 1–63

Yang X, Wang H, Xie D, Deng C, Tao D (2022) Object-agnostic transformers for video referring segmentation. IEEE Trans Image Process 31:2839–2849

Yang Z, Dong J, Liu P, Yang Y, Yan S (2019) Very long natural scenery image prediction by outpainting, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 10561–10570

Ye Q, Huang Q, Gao W, Zhao D (2005) Fast and robust text detection in images and video frames. Image Vis Comput 23(6):565–576

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.