IET Image Processing

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

A New Genetic Algorithm-Based Network for Text Localization in Degraded Social Media Images

Shivakumara Palaiahnakote[1] | Chandrahas Pavan Kumar[2] | Pranjal Aggarwal[2] | Shubham Sharma[2] | Pasupuleti Chandana[2] | Mahadveppa Basavanna[3] | Umapada Pal[4]

[1]School of Science, Engineering and Environment, University of Salford, Manchester, UK | [2]Department of Computer Science and Engineering, Indian Institute of Information Technology, Dharwad, India | [3]Department of Studies in Computer Science, University of Davanagere, India, Davangere, Karnataka, India | [4]Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India

Correspondence: Shivakumara Palaiahnakote (s.palaiahnakote@salford.ac.uk)

Received: 28 October 2024 | **Revised:** 21 December 2024 | **Accepted:** 11 February 2025

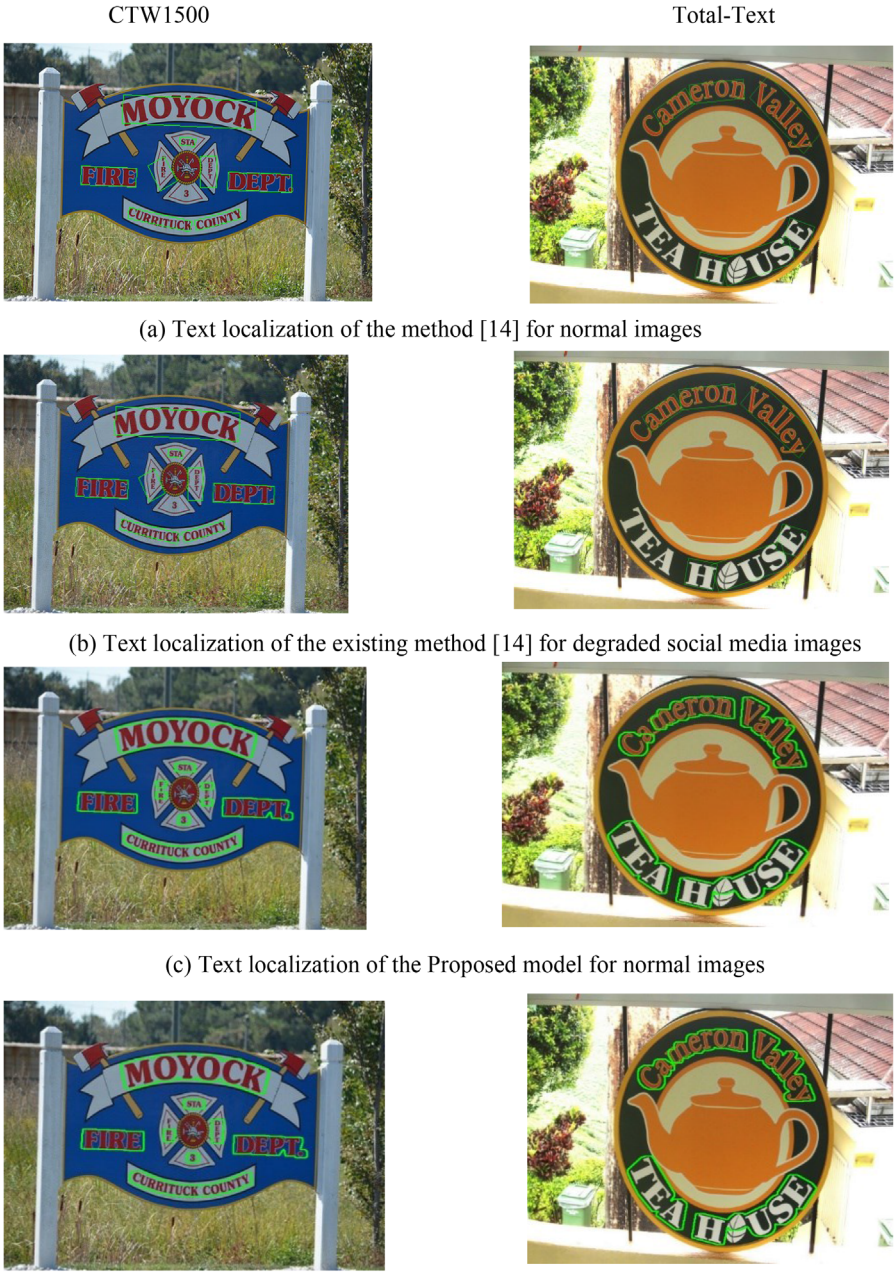
Funding: The authors received no specific funding for this work.

ABSTRACT

This paper presents a novel model for understanding social image content through text localization. For text localization, we explore maximally stable extremal regions (MSER) for detecting components that work by clustering pixels with similar properties. The output of component detection includes several non-text components due to the degradations of social media images. To select the best components among many, we explore the genetic algorithm by convolving different kernels with components, which results in a feature matrix that is further fed to EfficientNet for choosing actual text components. Therefore, the proposed model is called genetic algorithm based network for text localization in degraded social media images (TLDSMI). For evaluating text localization, we consider the images of the standard dataset of natural scenes by uploading and downloading from different social media platforms, namely, WhatsApp, Telegram, and Instagram. The effectiveness of our method is shown by testing on original and degraded standard datasets. For example, for the degraded images of different complexities including degradations caused by social media platforms, the proposed method performs well in almost all situations. In addition, the proposed model achieves the best F1-Score, 0.76, 0.77, 0.70, and 0.78 for the degraded images of CUTE, ICDAR 2013, Total-Text, and CTW1500, respectively, compared to the state-of-the-art methods.

1 | Introduction

Monitoring uploaded content on social media is very important for several real-time applications such as person identification, person behavior identification (big five factors), person mind reading, and person searching to identify the crime [1–3]. This is because most of the time, social media has been used to share personal views, express emotions, conveying messages, which may include messages related to anti-social activities such as anger, aggression, hate speech, etc. Therefore, watching content uploaded on social media can help us to prevent unwanted anti-social activities and suicide cases. In this regard, motivated by the work on text spotting in vehicle tires [4], metal surfaces [5], and shipping containers [6, 7], we propose text localization in degraded social media images in this work to analyze and measure the uploaded content on the web. We believe text in degraded social media images has the same effect as vehicle tires, metal surfaces, and shipping containers. When we upload and download images from different social media platforms multiple times, such as WhatsApp, Telegram, and Instagram, it severely affects image content. Therefore, the images are called degraded social media images. This is due to the adaptation of different compression and storage mechanisms by different social media platforms. In addition, using different devices/machines



(a) Text localization of the method [14] for normal images

(b) Text localization of the existing method [14] for degraded social media images

(c) Text localization of the Proposed model for normal images

(d) Text localization of the Proposed model for the degraded social media images

FIGURE 1 | Challenges of text localization in degraded social media images.

with different configurations for uploading and downloading operations adds more complexity to the localization problems of text. Thus, text localization in degraded social media images is significantly more difficult than text localization in normal images captured by the camera. The methods have been developed for text detection in scene images and videos in the past [8–13]. However, the main focus of the existing methods is not degraded social media images and therefore, the existing methods may not be effective in achieving the best results.

From the illustration in Figure 1, it is clear that the state-of-the-art method [14], which proposes differential binarization and adaptive scale fusion for scene text localization fixes improper bounding boxes for the text for normal and degraded social media images as can be seen in Figure 1a,b. This shows that the existing models are not robust to degradations created by uploading and downloading on social media platforms. On the other hand, as seen in Figure 1c,d, the proposed model performs well for both normal and degraded social media images. This is the contribution achieved by combining maximally stable extremal regions (MSER) [15], genetic algorithm [16] and EfficientNetB0 [17].

It is observed that MSER has the special property of detecting character objects as components irrespective of degradations [15]. Due to unpredictable degradations in the case of social media images, the number of component detections varies without missing text components [15]. Motivated by this observation, we

explore MSER for detecting components in the input images of both normal and degraded social media images. Since the output of MSER includes text and non-text components, it is necessary to choose components that represent text. We believe text components are strong components compared to non-text components in terms of characteristics of text properties. Therefore, we propose a genetic algorithm (GA) to choose text components as strong components for text localization in this work [16]. GA performs cross-over and mutation based on characteristics of text property, which results in a feature matrix. Inspired by the strengths of EfficientNet, which is capable of achieving high accuracy and better efficiency over conventional convolutional neural networks, we explore the same for classifying text and non-text components [17]. The proposed work uses text components for merging to fix the bounding box for the words in the images.

Consequently, the following are the main contributions of the suggested work. (i) As far as we are aware, this is the first attempt at localizing localizing text in degraded social media images. The suggested model, in contrast to existing approaches, is effective for both normal and degraded social media images. (ii) Exploring MSER for detecting components irrespective of degradations and poor quality is new for text localization in natural scene images. (iii) Similarly, exploring the combination of GA and EfficientNet for choosing text components as strong components is a novel idea for the text detection community. (iv) Overall, the way the proposed work integrates the merits of MSER, GA, and EfficientNet is a unique contribution compared to state-of-the-art methods.

The rest of the paper is organized as follows. In Section 2, the associated techniques for text detection in images of natural scenes are examined. MSER for component detection, GA for feature extraction, and EfficientNet for text component classification are presented in Section 3. Section 4 presents extensive tests to demonstrate the viability of the suggested approach. In Section 5, the conclusions are summarized.

2 | Related Work

The literature points out that text localization in degraded social media images is relevant to text detection in images of natural scenes. As a result, this section provides a review of the current application-specific procedures.

There are methods for addressing the challenges of text detection or localization in natural scene images. Du et al. [18] aim to develop an OCR for both scene text images and document images through text detection based on the differential binarization concept. To adapt a differential binarization model for different types of images, the approach modified the backbone and light head modules. Long et al. [19] noted that methods developed so far work well for multi-oriented text but not arbitrarily shaped text. The approach explored a concept called TextSnake which is more flexible to represent text instances of different orientations. Zhang et al. [20] observed that the existing methods for arbitrary shaped scene text detection do not provide satisfactory results. To improve text detection performance for arbitrarily shaped text, the authors introduced deep relational reasoning graph networks. Liao et al. [14] used differential binarization and an adaptive scale fusion approach for real-time text detection in scene images. Zhu et al. [21] developed Fourier contour embedding for arbitrarily shaped text detection in scene images. This method combines Fourier transformation with deep learning to achieve better results. Xing et al. [22] used a learnable embedding network for detecting arbitrarily shaped scene text. The approach constructs boundary segmentation branches and then predicts coarse boundary masks for each text instance. Rong et al. [23] proposed a recurrent dense text localization network. The main focus of the approach is that it addresses the issue of crowded or dense text in natural scene images. Dai et al. [24] developed a model for accurate scene text detection using scale-aware data augmentation and shape similarity. The objective of the technique is to increase the number of relevant samples such that the methods can acquire the required knowledge to address the challenges of arbitrarily shaped text and introduce global shape similarity to improve the performance of text detection. Zhong et al. [25] pointed out that most existing methods work based on bounding box prediction. This idea may not be effective for separating adjacent instances of arbitrarily shaped and curved text. Therefore, the approach uses the idea of describing the progressive variety of text regions by proposing 2D progressive kernels. In addition, to improve the performance of text detection, post-processing is used.

Although the aforementioned models addressed several text identification challenges, their primary drawback is that a model must be designed in stages. As a result, these models are ineffective, especially for text lines with curved or other arbitrary arbitrary shapes. Recently, transformer-based models have been introduced for text detection/spotting in natural scene images [11–13]. These models are better than conventional deep learning models. This is because the transformer-based models avoid preprocessing, and intermediate stages, such as proposal anchors, non-maximum suppression, sliding window process, and heuristic design. It is noted that if the methods involve many stages designed by different modules, these are prone to errors. In addition, the methods may not be stable and consistent for different situations. On the other hand, end-to-end transformer-based models are accurate, and efficient and achieve the best results compared to conventional deep learning approaches.

However, the majority of the models covered above focus on essentially the same scene text detection applications. The difficulties and complexity are thereby constrained although few techniques concentrate on certain uses. For text concealing in photographs of natural scenes, Keserwani et al. [26] presented a conditional generative adversarial network. This is done to protect sensitive information and conceal text that relates to private text. A model for setting tight bounding boxes for arbitrarily oriented text and even objects was created by Dai et al. [27]. The model investigates an anchor-free corner evaluation strategy to do this. The goal of Akallouch et al. [28] is to find Arabic-Latin scene text in traffic panels for highways. They concentrated on including multilingual and multifunctional datasets when designing models. For localizing text in 3D video, Nandanwar et al. [29] established the wavefront concept. For better outcomes, the model incorporates an adaptive B-spline polygon curve network and generalized gradient vector flow. Model-based episodic learning was suggested by Chowdhury et al. [30] for text detection in sports photos. The method's main goal is to identify

textile areas so that text may be correctly identified. Although the model aims at different situations and considers different challenges for text detection, the scope of the model is the same as scene text detection and localization. In addition, the foreground (text) is almost the same for all the applications. However, some models focus on totally different applications where one can expect different inputs and changes in both foreground (text) and background. For instance, text detection in vehicle tires [4], metal surfaces [5], and shipping containers [6, 7]. These images pose different challenges compared to normal text detection because text nature and complexities vary according to background and applications. Therefore, the above methods may not work for these situations. However, these models are limited to specific applications. In other words, these approaches may not work well for normal scene images.

In summary, none of the models consider degraded social media images for text detection. It is noted from the methods of different categories that the methods that work well for normal scene images may not be effective for the images of special applications (vehicle tire, metal surface, and shipping container) and vice versa. Besides, it is also noted that none of the existing models, including the methods of special applications, consider degraded social media images for text localization. Therefore, one can conclude that there are no methods for achieving better results for both normal scene images as well the images of special applications (degraded social media images in this work). Hence, we aim to develop a novel genetic algorithm network for text localization in degraded social media images (TLDSMI) in this work.

3 | Proposed Model

The main goal of the proposed study is to create a model that can withstand the difficulties of text localization in both typical scene photographs and degraded social media images, as was stated in the previous section. The degraded social media images are obtained by uploading and downloading the images of standard natural scene text detection datasets on different social media platforms, namely, WhatsApp, Telegram, and Instagram, feeding the images of normal scene images. Due to repeated uploading and downloading on these platforms, the quality of images is adversely affected (distortion, degradation, loss of quality). In addition, devices with different capacities and configurations used for uploading and downloading make the problem much more complex. Therefore, achieving better results for normal scenes and degraded social media images is an open challenge.

Motivated by special properties of maximally stable extremal regions (MSER) which detect components irrespective of degradation effects, we explore the same for detecting components in the input images. As the level of degradation changes, the number of components detected changes. As a result, component detection includes text and non-text components. In this situation, we believe that text components are strong candidates compared to non-text components. Therefore, we introduce genetic algorithm (GA) for choosing text components as strong candidates. For this, different kernels are convolved with these components to extract features at a component level. Inspired by the ability of EfficientNet which can achieve high accuracy and better efficiency, the feature matrix is fed to EfficientNet for detecting text components.

The text components are merged using iterative dilation and based on nearest neighbor criteria to fix the bounding box for any oriented text lines. To reduce false positives, the proposed work uses a classifier or OCR as a postprocessing step to improve text localization performance. The block diagram of the proposed work can be seen in Figure 2.

3.1 | MSER for Components Detection

Maximally stable extremal regions (MSER) is a method of blob detection that was proposed by [15]. This algorithm extracts several co-variant regions from an image, called MSER. It is a stable connected component of some grey-level sets of the image. In this algorithm, all pixels below a specified threshold are considered 'black,' while those above or equal to it are considered 'white.' If a sequence of threshold result images is formed from a source image, with each image corresponding to an increasing threshold '*t*', the initial image will be white, followed by 'black' patches due to local intensity minima, which will grow larger. When one of these dark regions has the same (or nearly the same) size as the preceding image, a maximally stable extremal zone is found. These "black" areas will ultimately blend turning the entire image dark. The set of all extremal regions in the sequence is composed of all connected components. The area of the region segmented by the threshold t is represented by Q_t and Δ represents the variation of t. When Q_t in Equation (1) is a local minimum, Q_t belongs to an MSER, indicating that the region is stable when the threshold t varies. More MSER components are produced when the value of Δ is lower.

$$q(t) = \frac{Q_{t+\Delta} - Q_{t-\Delta}}{Q_t}, t \in [0, 255] \quad (1)$$

The effect of MSER on component detection can be seen in Figure 3, where one can note that MSER detects almost all text components for normal scene images as shown in (a), and degraded social media images as shown in (b). Importantly, it is observed from Figure 3a,b that the step does not miss any text components although the content of the images degrades due to social media platforms. However, the number of components increases for the degraded social media image compared to a normal scene image. Therefore, the output of MSER includes non-text components as well. In summary, one can infer that the MSER detects components without missing text information irrespective of degradation effects. Inspired by this unique characteristic of MSER, we propose to use this in conjunction with genetic algorithms to enhance the selection of text components. These MSER components are used as an input to the GP Trees generated for Genetic Algorithms

3.2 | Genetic Algorithm and EfficientNet for Text Components Detection

A form of search algorithm called a genetic algorithm draws its inspiration from natural evolutionary processes. By imitating

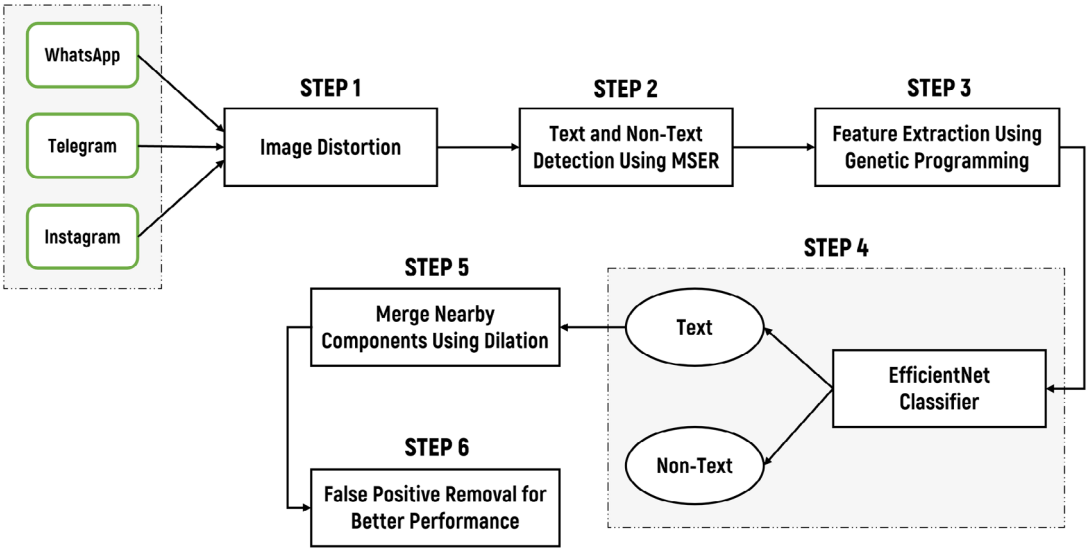
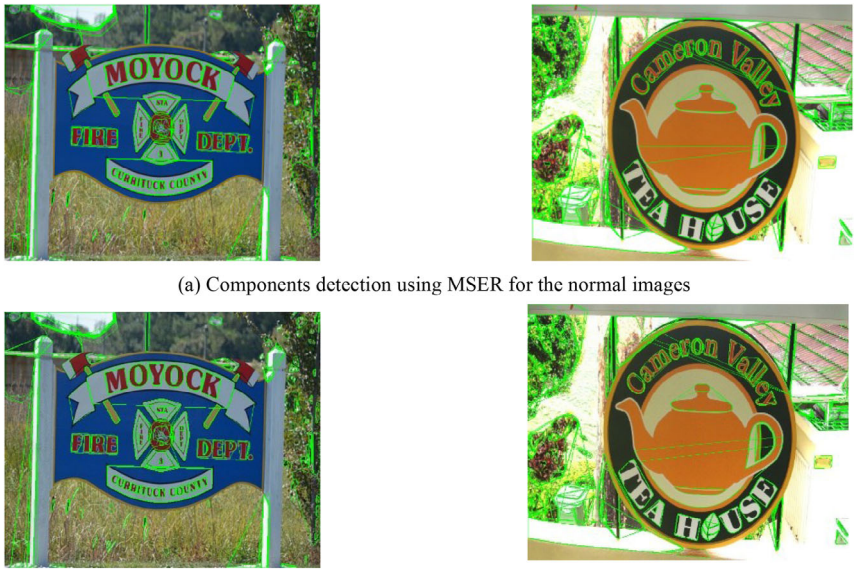


FIGURE 2 | Proposed framework for text localization in degraded social media images.



(a) Components detection using MSER for the normal images

(b) Components detection using MSER for degraded social media images

FIGURE 3 | MSER for component detection for degraded social media images.

natural selection and reproduction, genetic algorithms can produce excellent solutions for a range of problems, including search, optimization, and learning. Due to their similarity to natural evolution, genetic algorithms can solve some problems that traditional search and optimization algorithms find difficult, especially when dealing with problems that have a large number of parameters and intricate mathematical representations. Genetic algorithms maintain a population of potential solutions to a particular problem, or individuals, whereas Darwinian evolution maintains a population of individual specimens. A new generation of solutions is created by repeatedly reviewing these prospective solutions. Candidates who are more adept at handling this challenge stand a better possibility of being selected and imparting their knowledge to the following batch of applicants. As generations go by, potential remedies become more effective in resolving the current issue. The goal of genetic algorithms is to identify the best solution to a problem. To implement genetic algorithms in programming we define operators namely population, fitness function, selection, crossover, and mutation. For our method, we trained a genetic programming model that returns a GP tree comprising of convolution operations, that was then applied to the target image component to highlight the text features and suppress the non-text features.

Population: Genetic algorithm maintains a population of individuals—a collection of candidate solutions to the problem at hand–at any given time. In our method, the population consisted of randomly generated GP trees consisting of convolution operations as their nodes. In our experiment, a typical node is a convolution operation having a kernel with its size ranging from 1×1 to 5×5 and values in the range $[-3, 3]$ followed by the ReLU activation function. This step outputs a feature map as its output. Figure 4 shows an example of a GP tree from the initial population.

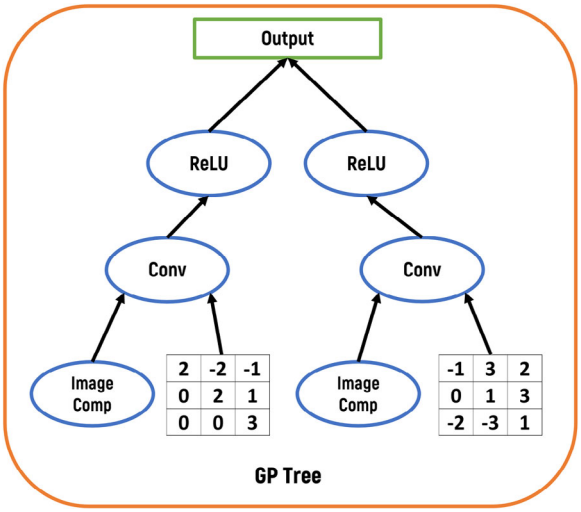


FIGURE 4 | GP Tree formed using convolution operations with initial random filters. This GP tree would function as an individual of the population in the Genetic Algorithm implementation.

Fitness Function: Individuals are evaluated using a fitness function at each iteration of the process (also called the target function). This is the function to solve the problem or improve the results. Individuals with higher fitness scores are more likely to be picked to reproduce and represent the next generation. The quality of the solution increases with time and the fitness value rises, the process can be stopped once a suitable fitness value is found.

For the calculation of the fitness value, we applied the convolution operation obtained from the population of the GP tree on a set of MSER components obtained from few random images from different datasets. Since the output of the GP tree is a feature vector of 10,000 dimensions, we applied Principal Component Analysis (PCA) to reduce its dimensions. This is further given to the SVM classifier for its training. Once done, the testing dataset is used to calculate the recall for the trained model which is eventually used as the fitness value for the respective individual. This means that the GP tree that results in a higher recall value than others is a fitter individual than the others. Figure 5 shows all the steps performed to calculate the fitness value for each individual.

Selection: Following the calculation of each individual's fitness score, a selection procedure is employed to determine which individual in the population should be selected for genetic operators and would constitute the next generation. Individual fitness scores are used in this selection procedure. Those with better scores are more likely to be chosen for the successive generations. Individuals with low fitness values can still be chosen, but their chances are slimmer. Their genetic material is not fully excluded in this way.

For our method, we used the process of tournament selection to select the best individuals. In this we choose 'k' individuals and hold a tournament among them. The fittest candidate among the chosen (based on their fitness values) is picked and transferred to the next generation. Similarly, several such tournaments are held, and we eventually come up with a final list of candidates for the subsequent generations.

Crossover: The crossover operation generates two new trees (offspring) from two selected trees (parents). The operation begins by selecting two nodes at random from the parent trees and then swapping the subtrees to create a new tree based on the selected nodes. Through various experimentation and trials, we applied a crossover of 50% to our population. Figure 6a below shows how the crossover operation takes place between two GP trees by swapping their subtrees from their nodes.

Mutation: Based on a single chosen tree, the mutation operation creates a new tree. The goal of mutation is to refresh the population at random intervals and introduce random changes. For our model, we introduced a mutation of 35% in the population.

The operation of the final evolved GP tree obtained from the above steps on text and non-text components of both original and degraded images is depicted in Figure 6b.

After the application of the final GP tree to the MSER components, they were then classified as text or non-text using EfficientNetB0. The schematic architecture of the classifier is shown in Figure 7 which consists of two dense layers with 1280 and 512 units with batch normalization and ReLU as activation functions along with a dropout of 0.4. Finally, the confidence score is obtained using a sigmoid function. The confidence score obtained for text components is higher as compared to the confidence score for non-text components. This is nothing but text component detection.

3.3 | Text Localization

Mathematical morphology is used to determine the shape of elements in an image. Dilation, erosion, opening, and closing are the four morphological operations available. Dilation and erosion were originally defined just for sets, but they have since been extended to functions, and in our text detection approach, we used erosion and dilation as two of these four operations.

Dilation is a technique that expands or thickens elements in a binary image. Two pieces of data are fed into the dilation operator. First is the image that is to be dilated and the second is a structuring element, which is a (typically small) group of coordinate points. This is also known as a kernel. The structuring element specifies the precise effect of dilation on the input image, that is, it controls the manner and extent of thickening or expansion of image elements. These are represented computationally by a matrix of 1's and 0's. Mathematically, it is defined in terms of set operations. $A \oplus C$ denotes the dilation of 'A' by 'C', which is defined as in Equation (2).

$$A\,C = \{z|(\hat{C})z \cap A \neq \phi\} \tag{2}$$

This equation is based on getting the reflection of 'C' denoted as '$\hat{C}$' about its origin and shifting it by 'z' indicated as $(\hat{C})z$ where set C is generally referred to as the structuring element in dilation.

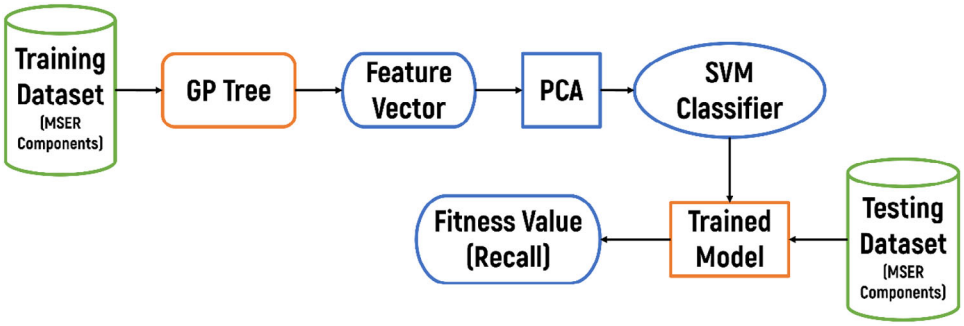


FIGURE 5 | Block diagram for fitness function. Each GP Tree is tested on the testing dataset of MSER component images, and the recall obtained for the model is considered as the fitness value for that individual.

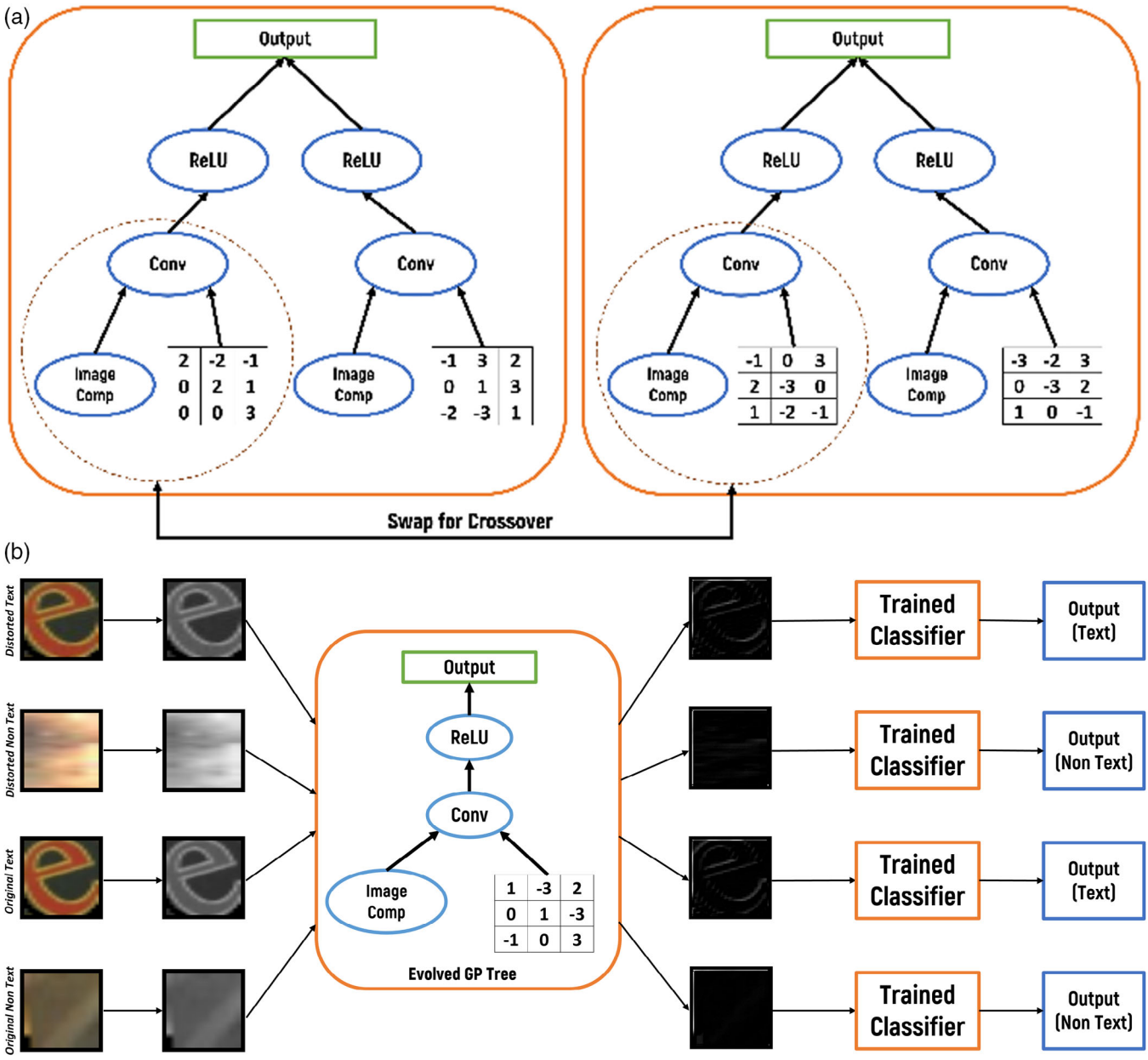


FIGURE 6 | (a) Depiction of crossover operation on two GP trees. (b) Operation of evolved GP tree on the text and non-text components of degraded and original image components.

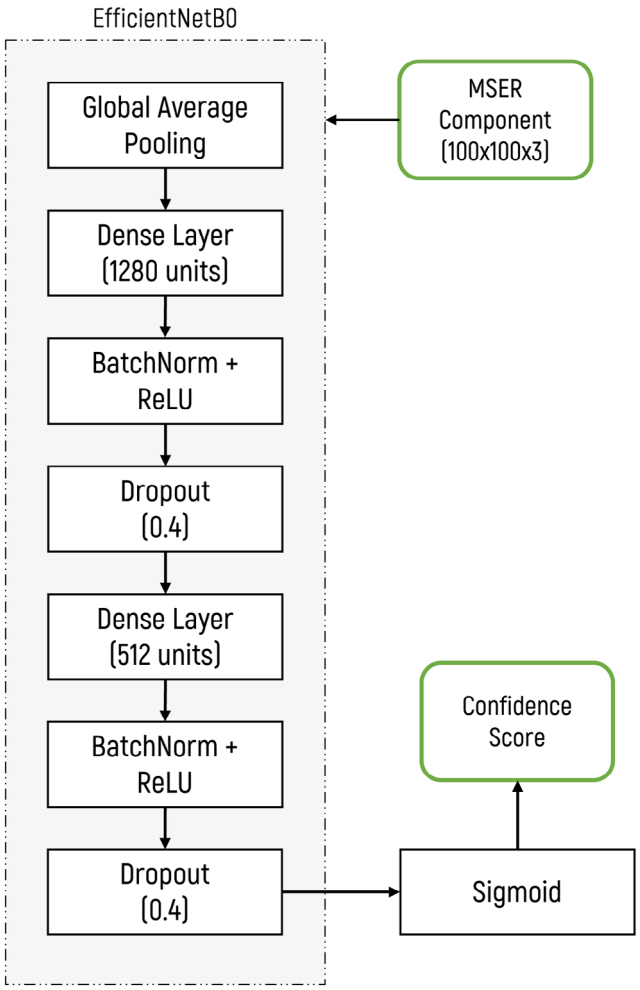


FIGURE 7 | Architecture of EfficientNet for classification of text components.

FIGURE 8 | The process of merging text component into words.

For fixing the bounding box for the words of any orientation, we perform a merging operation based on the nearest neighbor criterion, which considers the spacing between characters. Based on the spacing, the structuring element for dilation has been determined. To obtain such boxes, we use the technique of iterative dilation as shown in Figure 8, where the classified text components can be seen on a character level and the detection result after identification using dilation.

The effect of text localization for normal scenes and degraded social media images can be seen in Figure 9, where it can be confirmed that the proposed model fixes proper bounding boxes irrespective of orientation and type of input images. The suggested model is therefore universal and resistant to degradations brought on by uploading and downloading on various social media platforms.

4 | Experimental Results

Since it is the first work for text localization in degraded social media images, the dataset for experimentation is not available in the literature. Therefore, we construct our dataset by uploading and downloading the images of standard scene text detection datasets, namely, ICDAR 2013, MSRATD-500, CUTE, CTW1500, and Total-Text on different social media platforms, namely, WhatsApp, Telegram, and Instagram. We evaluate the proposed model on both normal and degraded social media image datasets. The dataset and code are available at the link provided in the footnote[1].

4.1 | Dataset and Evaluation

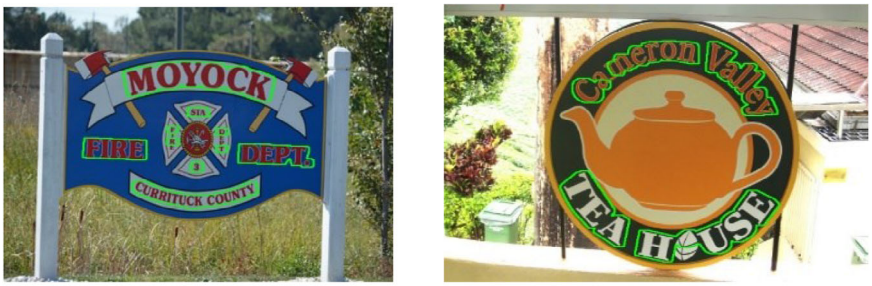
We used established datasets containing text images captured in real-world settings, specifically targeting natural images that depict text in various contexts (e.g., street signs, advertisements, shopping boards, etc.). However, to better simulate the conditions found on social media, we introduced distortions that reflect common image degradation encountered on these platforms.

To repurpose these commonly available datasets and introduce distortions in them, we developed a pipeline that involved sequentially sending the original images through multiple social media platforms. We specifically used WhatsApp, followed by Telegram, and finally Instagram in our workflow. This multi-step process allowed us to capture the cumulative effects of compression, resizing, and other quality reductions typically applied by these platforms.

This approach ensures that the degraded datasets closely resemble the variety of (text) images one might encounter on social media. Given the open nature of these platforms, users often upload images with varying levels of quality, which reflects a wide spectrum of distortions. By employing this methodology, we create a dataset that is not only diverse but also representative of actual social media image conditions.

The accuracy of a model is usually quantified as the ratio of correctly classified samples to the total number of samples, although there are other options. The confusion matrix provides the foundation for several widely used accuracy metrics, including precision, support, confidence, recall, sensitivity, specificity, and others. True positives (TP) are positive instances that have been classified as positive, false negatives (FN) are positive instances that have been classified as negative, false positives (FP) are negative instances that have been classified as positive, and true negatives (TN) are negative instances that have been classified as negative.

The classification results were compared and tested against Precision, Recall, and F1-Score as the performance metrics while recall was given a high priority for text detection.



(a) Text localization of the Proposed model for normal images

(b) Text localization of the Proposed model for the degraded social media images

FIGURE 9 | Text localization of the proposed model for both normal and degraded social media images.

Here,

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{F1} - \text{Score} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{6}$$

We compare the suggested model to the most recent models [14, 18–21] to demonstrate its efficacy. Because the models are resistant to degradations and take into account the difficulties of arbitrarily shaped text detection as the proposed method, the aforementioned methods should be taken into consideration for comparison analysis. The models used deep learning in several ways to provide the best scene text detection outcomes.

Implementation Details: It is noted that the combination of Genetic Algorithms and EfficientNet might introduce overfitting. However, to overcome this problem, we use techniques such as training data augmentation, cross-validation, and regularization. Based on our experiments, the following values are determined for hyperparameters to achieve the best results. Delta = 0.4 (least count of varying threshold intensities from 0 to 255), max_variation = 0.1 (threshold that controls the stability of the detected regions based on their variation in intensity), min_area = 60 (prunes the area that is smaller than this), and min_diversity = 0.2 (it controls how diverse the detected regions can be based on their stability score. It helps in suppressing the regions that are too similar to each other, ensuring only distinct regions are returned).

4.2 | Ablation Study

To address the challenges of text localization in degraded social media images, the proposed method involves the key steps of MSER for component detection, GA for choosing the best components, and EfficientNet for the classification of text components. Previous studies used MSER for text detection but using it in degraded images poses a challenge because unwanted extra components are obtained in degraded images due to compression and loss of information. This is showcased quantitatively in Table 1. Therefore, MSER alone may not work well for degraded social media images. Thus, this work adapts MSER to improve text localization performance for both original and degraded scene text images with the help of GA and EfficientNet.

TABLE 1 | Number of MSER components before and after introducing distortion.

Dataset	MSER components on	
	Original images	Degraded images
ICDAR 2013	390	414
MSRA	650	671
CUTE	697	788
CTW-1500	739	844
Total-Text	842	870

To assess the effectiveness of each key step, the following experiments are conducted on the Total-Text dataset. The experiments include both normal and degraded social media images of the Total-Text dataset and the results are reported in Table 2. (i) In this experiment, MSER components are fed to EfficientNet directly without GA for text localization. This is to show the effectiveness of GA in tackling the challenges of text localization. (ii) The EfficientNet is replaced by RestNet50 for text localization. This is to show that EfficientNet is better than ResNet50 for text localization in normal and degraded social media images. (iii) The proposed model includes MSER + GA and EfficientNet for text localization in both original as well as degraded social media

TABLE 2 | Assessing the contribution of the key steps of the proposed model using the Total-Text dataset.

	Original images			Degraded social media images		
Exp	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
(i)	0.56	0.38	0.41	0.50	0.37	0.38
(ii)	0.63	0.52	0.54	0.64	0.52	0.55
(iii)	0.83	0.60	0.69	0.82	0.62	0.70

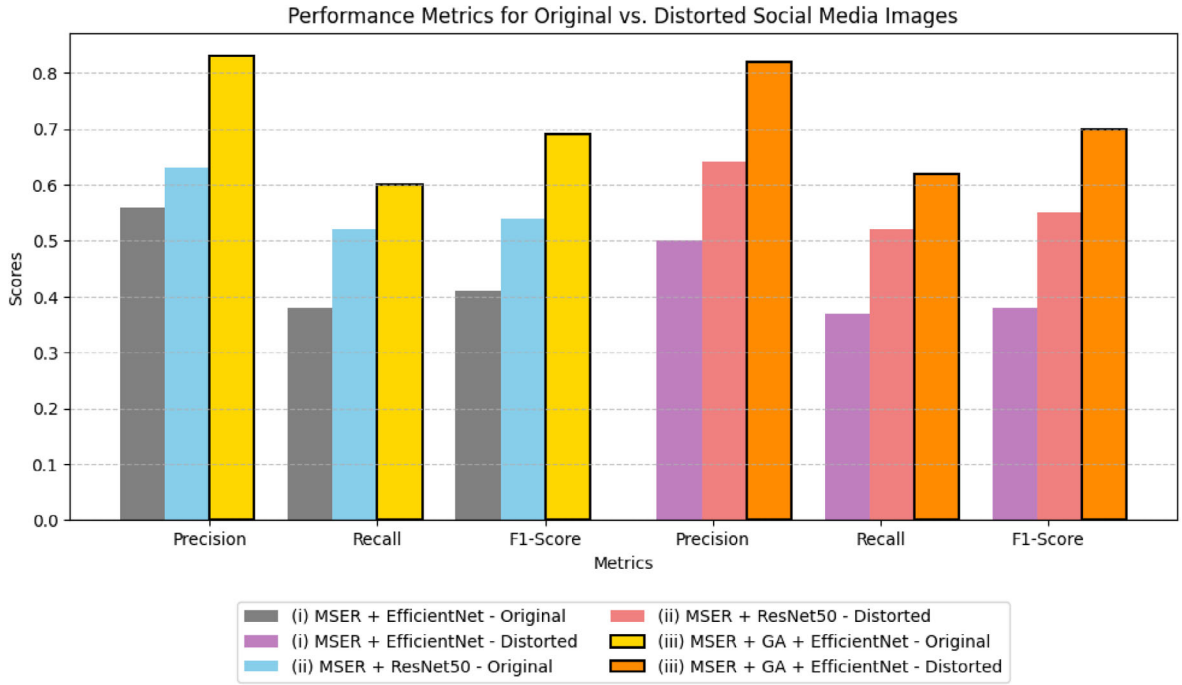


FIGURE 10 | Comparison of performance with different combinations of methods.

images. It is observed from Table 2 that each key step is effective and contributes equally to achieving the best results for both normal and degraded social media images. It is also noted from Table 2 that the results of each key step are not higher than the proposed model (experiment (iii)). Therefore, one can infer that all the key steps are essential for achieving the best results. The same inferences can be observed in Figure 10, where we can visualize the performance of the different combinations of key steps.

In summary, we can conclude that although MSER is fairly good at identifying potential text regions, the presence of distortions often introduces much noise, that is, non-text components, that can lead to false positives. To overcome this, our implementation of GA refines the component selection process, ensuring that only the most relevant features are considered for classification. This iterative optimization enhances the model's resilience to the noise inherent in degraded images.

4.3 | Experiments on Text and Non-Text Component Detection

Qualitative results of text and non-text components classification for original and degraded social media images can be seen in Figure 11a,b, respectively. It is noted from Figure 11a,b that the classification step does not miss text components. However, it can be noted that some of the non-text components are misclassified as text components. This is due to the complex background and cluster of pixels that share the same properties.

Quantitative results of the classification of text components are reported in Table 3 on original and degraded social media images of different benchmark datasets. For all the datasets, the proposed text component classification steps achieve almost the same results for both original and degraded social media images. This demonstrates the consistency, stability, and independence of the proposed model from both high and low-quality photos. Please note that we manually counted in order to calculate measurements because there is no ground truth for text components.

4.4 | Experiments on Text Localization

Similarly, qualitative results of the proposed model for text localization on different benchmark datasets are shown in Figure 12, where it is noted that the proposed model fixes proper bounding boxes for any oriented text irrespective of original and degraded social media images. Therefore, one can conclude that the

(a) Text component detection for the original images

(b) Text component detection for degraded social media images

FIGURE 11 | Sample text component detection with the proposed model.

TABLE 3 | Performance of the proposed text component detection on different datasets.

	Original images			Degraded social media images		
Datasets	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
ICDAR2013	0.704	0.795	0.742	0.703	0.791	0.742
MSRA	0.718	0.771	0.719	0.781	0.815	0.781
CUTE	0.573	0.679	0.621	0.645	0.691	0.667
CTW1500	0.771	0.686	0.726	0.769	0.724	0.745
Total-Text	0.721	0.717	0.718	0.735	0.726	0.730

proposed model is capable of handling the challenges of images affected by distortions caused by different social media platforms. Figure 13 shows qualitative results of the state-of-the-art models [14, 18–21], where it is observed that the models do not fix proper and tight bounding boxes. In addition, some of the models miss text [18]. Although state-of-the-art models were developed to address the challenges of arbitrarily shaped text localization, the models are not effective for fixing closed bounding boxes, especially for irregularly shaped text with complex backgrounds. This is due to conflict between the features extracted from background objects and text. On the other hand, the proposed models can cope with the challenges of irregularly shaped texts as well as degraded text as shown in Figure 12. This is because the proposed combination of MSER, GA, and EfficientNet can differentiate text and non-text components accurately.

Quantitative results of the proposed and existing models on original and degraded social media images of different standard datasets, namely, CUTE, ICDAR32013, Total-Text, CTW1500, and MSRA-TD500 datasets are reported in Tables 4–8, respectively. It is observed from Tables 4–8 that the proposed model is the best at recall for the degraded images compared to existing models. In the same way, the performance of existing models degrades for degraded social media images compared to original images of all the above-mentioned datasets. On the other hand, the proposed model obtains an almost consistent F-score for both original and degraded social media images of all the datasets. This shows that the proposed model is stable and reliable compared to existing models.

Furthermore, for the original images of CUTE and ICDAR 2013 datasets, the proposed model achieves the best precision compared to existing models. However, for the original images of Total-Text, CTW1500, and MSRA-TD500 datasets, the proposed model does not achieve the best results compared to existing models. The existing models were developed to address the

(a) Text localization for the original images

(b) Text localization for degraded social media images

FIGURE 12 | Sample text localization of the proposed model.

TABLE 4 | Experiments on CUTE dataset.

	Original images			Degraded images		
Method	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
PaddleOCR [18]	0.58	0.55	0.56	0.61	0.56	0.58
TextSnake [19]	0.48	**0.89**	0.62	0.51	0.80	0.62
DRRG [20]	0.66	0.86	0.74	0.69	0.82	0.74
DBNet++ [14]	0.35	0.62	0.44	0.39	0.59	0.46
FCENet [21]	0.70	0.94	**0.80**	0.69	0.81	0.74
Proposed method	**0.73**	0.85	0.78	**0.70**	**0.84**	**0.76**

TABLE 5 | Experiments on ICDAR 2013 dataset.

	Original images			Degraded images		
Method	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
PaddleOCR [18]	0.81	0.65	0.72	0.81	0.64	0.71
TextSnake [19]	0.63	0.65	0.63	0.65	0.65	0.65
DRRG [20]	0.74	0.63	0.68	0.75	0.64	0.69
DBNet++ [14]	0.84	**0.86**	**0.84**	0.72	0.69	0.70
FCENet [21]	0.78	0.74	0.75	0.73	0.70	0.71
Proposed method	**0.90**	0.79	**0.84**	**0.83**	**0.72**	**0.77**

FIGURE 13 | Example of text localization of existing methods.

TABLE 6 | Experiments on Total-Text dataset.

	Original images			Degraded images		
Methods	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
PaddleOCR [18]	0.74	0.39	0.51	**0.88**	0.45	0.59
TextSnake [19]	0.74	0.82	0.77	0.71	0.58	0.63
DRRG [20]	0.80	0.85	0.82	0.84	0.54	0.65
DBNet++ [14]	**0.88**	0.83	**0.85**	0.67	0.61	0.63
FCENet [21]	0.82	**0.89**	**0.85**	0.65	0.49	0.55
Proposed method	0.83	0.60	0.69	0.82	**0.62**	**0.70**

TABLE 7 | Experiments on CTW1500 dataset.

Method	Original images			Degraded images		
	Precision	Recall	F1-score	Precision	Recall	F1-score
PaddleOCR [18]	0.86	0.68	0.75	**0.88**	0.67	0.76
TextSnake [19]	0.85	0.67	0.74	0.76	0.60	0.67
DRRG [20]	0.80	0.83	0.81	0.87	0.68	0.76
DBNet++ [14]	**0.88**	0.82	**0.84**	0.39	0.56	0.45
FCENet [21]	0.83	**0.87**	**0.84**	0.80	0.74	0.76
Proposed method	0.80	0.71	0.75	0.83	**0.75**	**0.78**

TABLE 8 | Experiments on MSRATD-500 dataset.

Method	Original images			Degraded images		
	Precision	Recall	F1-score	Precision	Recall	F1-score
PaddleOCR [18]	**0.95**	0.73	0.82	0.80	0.65	0.71
TextSnake [19]	0.73	0.83	0.77	0.79	0.70	0.74
DRRG [20]	0.82	**0.88**	0.84	0.84	0.72	0.77
DBNet++ [14]	0.91	0.83	0.86	0.49	0.60	0.53
FCENet [21]	0.92	0.85	**0.88**	**0.89**	0.75	**0.81**
Proposed method	0.76	0.76	0.76	0.70	**0.77**	0.73

challenges of scene text localization; it is obvious that the existing models are better than the proposed model, especially for original images. However, the same models report poor performance for the degraded social media images of all the above datasets in terms of recall compared to the proposed model. Since the main objective of the proposed work is to achieve stable and reliable results for both original and degraded social media images of several benchmark datasets, the poor results on a few datasets may not be regarded as a major weakness of the proposed work. The poor results of the existing models reported for degraded social media images show that the models are not effective for the degraded images compared to the original images. However, the proposed model is capable of handling both degraded and original images. This is because the combination of MSER, GA and EfficientNet is robust enough to differentiate text and non-text components.

Since text localization is a pre-processing step for recognition, the proposed method is capable of handling the images of different scripts. The steps of component detection, and feature extraction using the Genetic algorithm and the EfficientNet for classification work well irrespective of scripts. This is evident from the results on the MSRA-TD-500 dataset because this dataset includes English and Chinese text images. However, the shape and style of the character change from one script to another, which may affect feature extraction for separating text and non-text components. This leads to poor performance. Therefore, the results of the proposed method are poor for MSRA-TD-500 datasets compared to the existing methods. This shows that the work requires further investigation to improve the results for multiscript scene images.

4.5 | Limitation of the Proposed Model

The key reasons for not reporting high results for both original and degraded social media images of all the datasets are as follows. The proposed MSER is slightly sensitive to contrast, and degradations as shown in Figure 14a, where the MSER misses text components. Therefore, the recall is low. Sometimes, for complex backgrounds, the MSER detects non-text as text components, and hence precision is low. In addition, when the spacing between characters is arbitrary as shown in Figure 14b, the steps of fixing bounding boxes do not work. This is also another reason for obtaining low results with the proposed model. With this discussion, one can understand that there is a scope for improvement. Therefore, our next aim is to develop a transformer to address the challenges of text localization in both original and degraded social media images. The reason to explore the transformer is that transformers use self-attention to capture pixel-level distortion, this could mitigate the problems faced with our proposed method to overcome the above challenge. Therefore, our idea is to use a pre-trained transformer model and then fine-tune it on the distorted image dataset to achieve the best results for the complex background and degraded social media images. Furthermore, if the text components are not classified with MSER in the first place, the text localization step does not work. This leads to poor performance of the proposed method. Thus, we plan to explore the combination of language model and transformer to predict the missing text in the images.

One more limitation is that proposing a generalized approach irrespective of the number of social media is a big question mark. However, we believe that the distortions or degradations caused

(a) MSER step fails to output text components

(b) Non-uniform spacing between characters causes problems.

FIGURE 14 | Example of unsuccessful text localization of the proposed model.

by social media considered in this work include all the possible distortions. In other words, even if we add images from another or new social media, the effect of such distortions is minimal. However, this fact should be investigated further, and it is beyond the scope of the proposed work.

5 | Conclusion and Future Work

This research proposes a unique model for text localization in degraded social media images that is based on a genetic algorithm (GAN-TLDSMI). The suggested approach uses MSER to identify text components in the input image. The detected components are supplied to the Genetic Algorithm (GA) to choose the candidate text component detection. Text component classification and text localization have both been investigated using the EfficientNet. First of its kind, text localization in Degraded social media images using MSER, GA, and EfficientNet. Experimental results on original and degraded social media images of different benchmark datasets show that the proposed model is stable and consistent compared to the existing models. However, there are some limitations, especially the use of MSER for text component detection in multi-script images. This is due to the variations in character shapes and styles of different scripts. Therefore, the proposed model does not report very high results for all the datasets. To overcome this challenge, we explore a combination of language models and transformers for text spotting in degraded social media images.

Author Contributions

Shivakumara Palaiahnakote: supervision, validation, writing – original draft. **Chandrahas Pavan Kumar:** formal analysis, methodology. **Pranjal Aggarwal:** conceptualization, data curation, formal analysis, methodology. **Shubham Sharma:** software, validation, visualization. **Pasupuleti Chandana:** formal analysis, software, validation. **Mahadveppa Basavanna:** visualization. **Umapada Pal:** writing – review and editing.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

CUTE: The data supporting this study's findings are openly available at: https://drive.google.com/drive/folders/1K5HeCcGyyE-t0lBVwZQWiawHn7KdblSQ?usp=sharing. ICDAR-2013: The data supporting this study's findings are openly available at: https://drive.google.com/drive/folders/1K5HeCcGyyE-t0lBVwZQWiawHn7KdblSQ?usp=sharing. CTW-1500: The data supporting this study's findings are openly available at: https://drive.google.com/drive/folders/1K5HeCcGyyE-t0lBVwZQWiawHn7KdblSQ?usp=sharing. Total-text: The data supporting this study's findings are openly available at: https://drive.google.com/drive/folders/1K5HeCcGyyE-t0lBVwZQWiawHn7KdblSQ?usp=sharing. MSRA-TD500: The data that support the findings of this study are openly available in MSRA Text Detection 500 Database (MSRA-TD500) at: https://drive.google.com/drive/folders/1K5HeCcGyyE-t0lBVwZQWiawHn7KdblSQ?usp=sharing.

Endnotes

[1] Dataset Link: https://drive.google.com/drive/folders/1K5HeCcGyyE-t0lBVwZQWiawHn7KdblSQ?usp=sharing Code Link: https://github.com/Shubhamm097/TLDSMI.

References

1. J. Zhou, B. Huang, W. Fan, Z. Cheng, Z. Zhao, and W. Zhang, "Text-Based Person Search via Local-Relational-Global Fine Grained Alignment," *Knowledge-Based Systems* 262 (2023): 110253.

2. Z. Wang, A. Zhu, J. Xue, et al., "SUM: Serialized Updating and Matching for Text-Based Person Retrieval," *Knowledge-Based Systems* 248 (2022): 108891.

3. K. Biswas, P. Shivakumara, U. Pal, T. Chakraborti, T. Lu, and M. N. B. Ayub, "Fuzzy and Genetic Algorithm Based Approach for Classification of Personality Traits Oriented Social media Images," *Knowledge-Based Systems* 241 (2022): 108024.

4. F. Gao, Y. Ge, S. Lu, and L. Weng, "Vehicle Tire Reader: Text Spotting and Rectifying for Small, Curved and Rotated Characters," *IEEE Transactions on Instrumentation and Measurement* 70 (2021): 1–12.

5. F. Gao, S. Li, H. You, S. Lu, and G. Xiao, "Text Spotting for Curved Metal Surface: Clustering, Fitting and Rectifying," *IEEE Transactions on Instrumentation and Measurement* 70 (2021): 5000212.

6. R. Zhang, Z. Bahrami, T. Wang, and Z. Liu, "An Adaptive Deep Learning Framework for Shipping Container Code Localization and Recognition," *IEEE Transactions on Instrumentation and Measurement* 70 (2021): 2501013.

7. R. Zhang, Z. Bahrami, and Z. Liu, "A Vertical Text Spotting Model for Trailer and Container Codes," *IEEE Transactions on Instrumentation and Measurement* 70 (2021): 1–13.

8. H. Li, X. Hu, and H. Lu, "T-Skeleton: Accurate Scene Text Detection via Instance-aware Skeleton Embedding," *IET-Image Processing* 18 (2024): 1491–1503.

9. H. Chen, P. Chen, Y. Qiu, and N. Ling, "FARNet: Fragmented Affinity Reasoning of Text Instances for Arbitrary Shape Text Detection," *IET-Image Processing* 17 (2023): 1959–1977.

10. M. Moradi and S. Mozaffari, "Hybrid Approach for Farsi/Arabic Text Detection and Localization in Video Frames," *IET-Image Processing* 7 (2013): 154–164.

11. M. Huang, Y. Liu, Z. Peng, et al, "SwinTextSpotter: Scene Text Spotting via Better Synergy Between Text Detection and Text Recognition," in Proc. CVPR (IEEE, 2022), 4593–4603.

12. X. Zhang, Y. Su, S. Tripathi, and Z. Tu, "Text Spotting Transformers," in Proc. CVPR (IEEE, 2022), 9519–9528.

13. Y. Kittenplon, I. Lavi, S. Fogel, Y. Bar, R. Manmatha, and P. Perona, "Towards Weakly-Supervised Text Spotting Using a Multi-Task Transformer," in Proc. CVPR (IEEE, 2022), 4604–4613.

14. M. Liao, Z. Zou, Z. Wan, C. Yao, and X. Bai, "Real-Time Scene Text Detection With Differentiable Binarization and Adaptive Scale Fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, no. 1 (2022): 919–931.

15. J. Matas, O. Chum, and T. Pajdla, "Robust Wide-Baseline Stereo From Maximally Stable Extremal Regions," *Image and Vision Computing* 22 (2004): 761–767.

16. P. G. Espejo, S. Ventura, F. Herrera, "A Survey on the Application of Genetic Programming to Classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 40, no. 2 (2010): 121–144.

17. M. Tan and Q. Le, "Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks", in Proc. ICML (International Machine Learning Society, 2019), 6105–6114.

18. Y. Du, C. Li, R. Guo, et al, "PP-OCR: A Practical Ultra-Lightweight OCR System," https://github.com/PaddlePaddle/PaddleOCR.

19. S. Long, J. Ruan, W. Zhang, X. He, W. Wu, and C. Yao, "TextSnake: A Flexible Representation for Detecting Text of Arbitrary Shapes," in Proc. ECCV (Springer, 2018), 19–35.

20. S. X. Zhang, X. Zhu, J. B. Hou, et al, "Deep Relational Reasoning Graph Network for Arbitrary Shape Text Detection," in Proc. CVPR (IEEE, 2020), 9699–9708.

21. Y. Zhu, J. Chen, L. Liang, Z. Kuang, L. Jin, and W. Zhang, "Fourier Contour Embedding for Arbitrary-shaped Text Detection," in Proc CVPR (IEEE, 2021), 3123–3131.

22. M. Xing, H. Xie, Q. Tan, et al., "Boundary-Aware Arbitrary-Shaped Scene Text Detection With Learnable Embedding Network," *IEEE Transaction on Multimedia* 24 (2022): 3129–3143.

23. X. Rong, C. Yi, and Y. Tian, "Unamigous Text Localization, Retrieval, and Recognition for Cluttered Scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (2022): 1638–1652.

24. P. Dai, Y. Li, H. Zhang, J. Li, and X. Cao, "Accurate Text Detection via Scale-Aware Data Augmentation and Shape Similarity Constraint," *IEEE Transactions on Multimedia* 24 (2022): 1883–1895.

25. Y. Zhong, X. Cheng, T. Chen, J. Zhang, Z. Zhou, and G. Huang, "PRPN: Progressive Region Prediction Network for Natural Scene Text Detection," *Knowledge-Based Systems* 236 (2022): 107767.

26. P. Keserwani and P. P. Roy, "Text Region Conditional Generative Adversarial Network for Text Concealment in the Wild," *IEEE Transaction on Circuits and Systems for Video Technology* 32 (2022): 31522–33163.

27. P. Dai, S. Yao, Z. Li, Z. Zhang, and X. Cao, "ACE: Achor-Free Corner Evolution for Real-Time Arbitrary-Oriented Object Detection," *IEEE Transactions on Image Processing* 31 (2022): 4076–4089.

28. M. Akallouch, K. S. Boujemaa, A. Bouhoute, K. Fardousse, and I. Berrada, "ASAYAR: A Dataset for Arabic-Latin Scene Text Localization in Highway Traffic Panels," *IEEE Transactions on Intelligent Transportation Systems* 23 (2022): 3026–3036.

29. L. Nandanwar, P. Shivakumara, R. Ramachandra, et al., "A New Deep Wavefront-Based Model for Text Localization in 3D Video," *IEEE Transactions on Circuits and Systems for Video Technology* 37 (2022): 33375–33388.

30. P. N. Chaowdhury, P. Shivakumara, R. Raghavendra, et al., "An Episodic Learning Network for Text Detection on Human Bodies in Sports Images," *IEEE Transactions on Circuits and Systems for Video Technology* 32 (2022): 2279–2289.