Augmented Human Research (2021) 6:14
https://doi.org/10.1007/s41133-021-00051-5

ORIGINAL PAPER

Deep Learning and Particle Swarm Optimisation-Based Techniques for Visually Impaired Humans' Text Recognition and Identification

Binay Kumar Pandey[1,4] · Digvijay Pandey[2] · Subodh Wariya[3] · Gaurav Aggarwal[4] · Rahul Rastogi[4]

Received: 14 May 2021 / Revised: 6 August 2021 / Accepted: 12 October 2021 / Published online: 29 October 2021

Abstract

Blind people can benefit greatly from a system capable of localising and reading comprehension text embedded in natural scenes and providing useful information that boosts their self-esteem and autonomy in everyday situations. Regardless of the fact that existing optical character recognition programmes seem to be quick and effective, the majority of them are not able to correctly recognise text embedded in usual panorama images. The methodology described in this paper is to localise textual image regions and pre-process them using the naïve Bayesian algorithm. A weighted reading technique is used to generate the correct text data from the complicated image regions. Usually, images hold some disturbance as a result of the fact that filtration is proposed during the early pre-processing step. To restore the image's quality, the input image is processed employing gradient and contrast image methods. Following that, the contrast of the source images would be enhanced using an adaptive image map. The stroke width transform, Gabor's transform, and weighted naïve Bayesian classifier methodologies have been used in complicated degraded images to segment, feature extraction, and detect textual and non-textual elements. Finally, to identify categorised textual data, the confluence of deep neural networks and particle swarm optimisation is being used. The text in the image is transformed into an acoustic output after identification. The dataset IIIT5K is used for the development portion, and the performance of the suggested come up is evaluated using parameters such as accuracy, recall, precision, and F1-score.

Keywords Text extraction · Deep Neural Network · Complex video · Image · Humans

Introduction

An intricate deteriorated image document is a widely accessible and useful medium that contains crucial, helpful data. These data are composed of pixels, from which important information is extracted from complex images according to the requirements of computer vision [1]. Complex videos and complex image text messages produce quality, just like text messages in images are used in many complex image learning [2–4] and indulging implementation processes, including language translators, book digitisation, video or image recovery [2]. Considerable devotion has been paid recently to the use of pre-programmed text detection and text recognition [5]. In history, an assortment of explorations were carried out to remove text from complex scenes, and this process of text acquisition is one of the most important parts of optical character recognition [6]. After text detection and further binarisation, OCR [7] is used to recognise text from images [8].

The visual images provide accurate and appropriate details for oblivious direction-finding, image perceptive, and retrieval methods, respectively [9]. This complex image frequently incorporates a variety of fonts and other properties [10]. The primary focus of the complicated corrupted images could include distinct-designed characters, information exposed in digital signposts, and information displayed

✉ Binay Kumar Pandey
binaydece@gmail.com

1 Department of Information Technology, College of Technology, Govind Ballabh Pant University of Agriculture and Technology, Pantnagar, Uttarakhand, India

2 Department of Technical Education, IET, Dr. A.P.J. Abdul Kalam Technical University, Lucknow, Uttar Pradesh, India

3 Department of Electronics Engineering, Institute of Engineering & Technology, Lucknow, India

4 Department of Computer Science and Engineering, Invertis University, Bareilly, India

on a monitor. This is a very common task for traditional OCR: identifying textual information with a different appearance. The texts are sprinkled throughout this complex, degraded image, and the preceding information about their position is not presented. The input documents from the camera recognise the line spacing, number of characters, but the complex image text does not contain any formatting rules, so it is not possible to directly apply the segmentation approach to complex images [11]. The process of detecting text in complex degraded images consists of two major steps: classification of textual images and non-textual images, and character recognition [12].

A primary goal of textual information extraction [13] from images would be generally used for directing visitors, locating vehicle locations, visually impaired people, and so on. Throughout this approach, an effective DNN is incorporated with the purpose of fetching the text from tarnished images [14]. In advance of that, the distorted frame of the image must be evaluated to determine whether or not the usable image contains any relevant information. To achieve this, a weighted [15] naïve-based classifier-based method is used before using deep learning-based [16] recognition of characters.

The Weighted naïve-based classifier distinguishes between textual and non-textual images [17]. The errors caused by reassessment mostly during the text classification process can be decreased by further using an optimisation algorithm. Finally, an image containing the text is fed into the deep neural network (DNN) algorithm for text extraction. Furthermore, the load parameter available in DNN [18] during the text extraction method reduces the likelihood of reliability in DNN. To avoid such a situation, a hybrid approach combining particle swarm optimisation and DNN has been used. That further results in an optimal weight parameter for DNN. Thereafter, the obtained text data are given as input to the text-to-speech synthesizer (TTS) that scans and reads each character and number of textual data and changes it into voices. This kind of system helps visually impaired people to interact effectively through vocal interface.

The following is how this paper was prepared. The review of literature is described in "Literature Review" Section. "Problem Statement and Motivation" Section is concerned with the problem statement and motivation. "Proposed Methodology" Section discusses the suggested methodology. Finally, "Results and Discussion" Section summarises the suggested methodology's results.

Literature Review

In this section, numerous approaches to textual data classification and character identification have been explored, and the effectiveness of all these approaches has also been evaluated.

In [19], the detection of bowed and multifaceted textual data in a unified type from complex degraded images frames via evolving a new-mask region-based convolutional neural network-based text detection method [4]. However, according to [20], a new method for data recognition and detection has been a connected component-focused method that makes use of maximally stable extreme regions. The multiple blurs produced by motion and defocus make the text detection process a challenging one.

In [21], a method for text identification in a distorted or non-distorted image is discussed. The contrast variants experienced in nearby pixels were identified in this method to evaluate the blur degree. Furthermore, the low-pass filter was used for de-blurring. Mostly, this approach gives pixels under consideration for de-blurring images. The process of detecting the scene text from videos attains high value in several information removal hinged audiovisual applications, like video-frame recovery and investigation. [4] gives a text tracking and recognition approach for frames of videos [22]. The public scene text video was included in this method, which outperforms the other existing methods.

According to [23], texture data were retrieved and analysed from complex degraded images using an improved algorithm. To begin, the Discrete Wavelet Transform (DWT) was used to detect edges in images. Following that, the textual regions were located using the Ada Boost classification model and connected component clustering. [24] introduced morphological reconstruction, which is based on a technique known as the geodesic transform, which emphasizes artefacts in the image's centre while erasing light and dark constructions that are problematic near the image's boundary. This binarisation technique has been found to be far superior to other text binarisation techniques.

The novel methodology of determining automobile trajectory would use an optimum boundary for both the vehicles presented in [25]. Footage taken from a mounted camera at such a junction is being used to recover automobile trajectories, which would be built on a convolutional neural network (CNN). Firstly, the YOLOv2 model has been used for actual vehicular object recognition. That represents among the most accurate object identification methods based on CNN. To compensate for the imperfection of a YOLOv2 vehicle position, its trajectories were validated with a vehicle tracking method including a Kalman filter and an intersection-over-union (IOU) tracker. An effort is being made, specifically, to rectify vehicle trajectories by identifying a centre point based on the geometrical properties of the travelling vehicle pertaining to the basic bounding box. All quantitative and qualitative assessments show the designed system detects mobile

vehicular paths better than the conventional approach. Regardless of the fact that perhaps the centre points of the bounding boxes acquired using the previous traditional method seem to be frequently beyond an automobile owing to spatial deformation of the camera, a suggested methodology could even minimise spatial inaccuracies as well as retrieve the ideal bounding box to evaluate vehicular coordinates.

As stated in [26], an examination of the present-state deep learning methods from both methodologies like Fast R-CNN, Faster R-CNN, RetinaNet, and YOLOv3 was conducted as well as a thorough analysis of the benefits but mostly limits of the methods. In particular, it tested modelling of various bases using various data containing multi-scale entities to determine which kinds of entities, as additions to bases, were appropriate for every system. Comprehensive actual testing was carried out on two image datasets, specifically the small object data and the filter data, using PASCAL VOC 2007. Furthermore, comparison findings but mostly analyses were provided.

According to [27], the aim of the studies would have been to carry out an analysis of the state-of-the-art in relation to the efficiency of pre-trained modelling techniques for object recognition in order to compare such methodologies in terms of effectiveness, accurateness, time processed, and problems identified. The model referenced is typically built using the Python programming language, as well as frameworks dependent on Tensor-Flow, OpenCV, and open image datasets. These methods are not just concerned with identification and detection of elements in images, but primarily with their position inside them. Creating a bounding box behind them is the appropriate way. For all of this study, multiple pre-trained systems of object identification, such as R-CNN, R-FCN, SSD (single-shot multibox), and YOLO (You Only Look Once), were compared using various extractors for properties, including VGG16, ResNet, Inception, and MobileNet. As a consequence, it is not advisable to conduct immediate and analogous analyses of the various architectural designs or even modelling techniques, since every other issue seems to have a unique solution to each challenge. The objectives of this article would be to create an estimated concept of an experiment which has been conducted and also to envision a reference point in its use that is intended to be provided.

Fruit identification is described as a critical component of a robotic cultivation device throughout [28]. After all, irregular environmental circumstances, including branch as well as leaf and stem deformation, lighting variability, tomato groupings, shadows, and so forth, have also made fruit identification extremely difficult. To tackle such issues, the modified YOLOv3 model, known as the YOLO-Tomato framework, has been used to identify tomatoes throughout complicated climate change factors. The densely architectural design virtues, hierarchical feature consolidation, as well as Mish function stimulation were applied to the modified YOLOv3 framework, or even the YOLO-Tomato designs: YOLO-Tomato-A at an average precision of 98.3% with a classification accuracy of 48 ms, YOLO-Tomato-B at an average precision of 99.3% with an identification time of 44 ms, and YOLO-Tomato-C outperformed other cutting-edge technologies with an AP of 99.5% with an average precision of 52 ms.

The research [29] examines three important image processing techniques: single-shot detection (SSD), faster region-based convolutional neural networks (Faster R-CNN), and You Only Look Once (YOLO) to determine which would be the quickest and perhaps most efficient. Using Microsoft COCO (Common Object in Context) datasets, the comparison assessment measures the efficiency of these three methods and then analyses their strengths and limits depending upon metrics like correctness, precision, and F1-score. As per the findings of the investigation, the usefulness of all the techniques to the other two has been governed in large part by the use scenarios for which they were implemented. YOLO-v3 surpasses SSD and Faster R-CNN in a similar evaluation setting, giving it the strongest of the three technologies.

The identification of tomato abnormalities in the complex natural ecosystem constitutes a significant scientific topic throughout the development of plant science, according to [30]. With its tiny size and complex backdrop, automated detection of tomato abnormalities remains a difficult challenge. To address the challenge of tomato outlier recognition in such a complicated natural environment, a new YOLO-Dense approach predicated on a one-stage deep detection YOLO architecture is presented. The suggested model's network inference time could be significantly enhanced by simply incorporating the densely connected modules into the network infrastructure. Using the K-means method for clustering the anchoring boxes, nine alternative dimensions of anchor boxes containing probable items to be recognised were generated. A multi-scale training strategy has been used to increase its identification accuracy for objects at multiple scales. The experiment findings showed that perhaps the average precision and recognition duration of a simple picture of a YOLO-Dense system are 96.41% and 20.28 ms, respectfully. When contrasted to SSD, Faster R-CNN, and the classic YOLOv3 networks, the YOLO model performed best in tomato outlier detection in complicated natural surroundings.

The purpose of this [31] study would be to describe and localise healthcare facial mask artefacts in actual photographs. Using a healthcare facial mask in public places provides protection from the spread of COVID-19. The developed framework is comprised of two parts. The first

element is intended for extracting features using the ResNet-50 deep transfer learning system. The second element, depending on YOLO v2, is intended for identification of healthcare face masks. Two datasets of medical face masks have been integrated into a single dataset for this study. As per the findings, an Adam optimizer earned a high average correctness percentage of 81% as just a detection. Furthermore, at the conclusion of the research, a comparison outcome using based tasks was provided. The suggested detectors outperformed existing relevant studies in terms of accuracy and precision.

Yuan [32] discussed the utilisation of information edge for identifying textual blocks from greyscale images. Its primary goal is to detect text in noisy images and distinguish it from graphical images. In this case, an algorithm was created to extract features from various objects and then classify those feature points to identify textual regions. Directionally placed text blocks can be easily obtained by using methods such as line of approximation and layout categorisation. In the final step, a feature-based connected component is merged with similar textual areas that exist inside the bounding rectangles. The methods anticipated here yield promising results, demonstrating the method's effectiveness.

In [33], a binarisation technique for colour images is discussed, and it is discovered that the traditional method, which is based on thresholding, does not produce better results for images that contain both foreground and background colours. To begin, features of the image under consideration are obtained based on luminance distribution. Binarisation [34] was then performed using a decision-tree-based method that chose different features of colour images to binarize images. If it is discovered that the colours in a colour image are intense within a defined colour range during the feature extraction process [35], an effective saturation is put into the image. In addition, if the image colours in the foreground are more dominant, luminance is one of the most important parameters to consider. Finally, luminance was supposed to apply when the colouring of the image's background appears to be strenuous within a specified boundary, and saturation was supposed to apply when the number of pixels to limited luminosity would be less than 60. However, both luminance and saturation are enforced. The analysis reported in this article includes 519 colour images in total. The majority of those are itemised receipts as well as name-card images. The suggested binarisation method outperforms others in terms of shape as well as connected component in this study.

As said by [36], detecting text in images, frames, or videos is thought to be an important step in retrieving any multimedia information. The author proposed an improved algorithm for detecting, localising, and retrieving side to side oriented textual data in image frames with degraded backgrounds in this paper. The proposed method is based on colour reduction techniques, an edge detection-based method, and text region localisation [37] using a projection profile, which evaluates geometric attributes of colour images. This same algorithm generates a series of text-boxes with a very simple background that are prepared to be fed into an OCR engine for consequent character recognition. Promising research findings for a set of image frames and videos illustrate the approach's effectiveness.

As stated in [38], a texture-based method was used to detect textual data in the image frame. Support vector machines are utilised to evaluate the textual characteristics of texts. Rather than using a different method to retrieve textual data features, the SVM [39]-based classifier is given the intensities of the raw pixels that comprise the textual pattern. Another method involves using a continuously adaptive mean shift algorithm to analyse textual data and identify textual areas. CAMSHIFT, a combination of SVMs, improved text detection results.

According to [40], it provides information about textual information available in images and video frames for annotation, indexing, and image construction. Identification, localisation, tracking, fetching, enhancement, and recognition of text data from an image are all processes involved in retrieving such information from an image. Variations in textual data due to various parameters, on the other hand, may cause problems in automatic text extraction. Based on [41], a two-phase noise removing scheme based on a two-phase noise removing technique from images such as salt and pepper is presented. In the first phase, an adaptive median filter is used to identify picture elements that are most probably influenced by noise. In the second phase, the image is revamped again using a regularisation function that is applied to the selected noisy images. Edge perpetuation and noise reduction, as well as their regained image frames, provide a considerable improvement over the nonlinear filter.

Peng-Lang Shui [42] proposed that using local Wiener filtering in conjunction with the wavelet domain is an effective image noise removal method for non-severely degraded images. In this paper, the author proposes a doubly local wiener filtering algorithm-based method, which uses an elliptic directional window for various sub-bands to perform the calculation on the variances of signal for noisy wavelet coefficients, and the two other procedures of local wiener filtering are accomplished on the noisy containing images. The outcomes obtained after the experiment showed that the algorithm proposed in this work may improve the denoising performance.

As per [43], a novel adaptive method for binarisation and improvement of tarnished images is presented. The approach described does not show any user-changeable requirements parameters and can easily handle image

distortion, which is commonly caused by shadows, uneven illumination, low contrast, extremely signal-dependent noise, smearing, and strain.

According to [44], a morphological component analysis procedure is based on sparse expression of signals. The morphological component analysis is built on the hypothesis that each signal's elemental behaviour must be segregated and that there is a dictionary for doing so using sparse representation. Following that, the pursuit algorithm for sparse representation can be used to obtain the desired separation. The paper also includes several image content application results as well as some theoretical results that explain the separation process.

Starck et al. [45] presented a novel method that is focused on the addition of the basis pursuit denoising (BPDN) algorithm along with total-variation regularisation methodology used for the separation of texture features and cartoon parts from the image. The author suggests using two dictionaries for the representation of textures and natural scene parts, respectively. Both dictionaries prompt sparse representations of images over single images. The main use of the basis pursuit denoising gives a method for desired separation as well as noise removal. The separation process is directed using a TV regularisation scheme. It is also used for removing ringing artefacts. A highly improved numerical scheme describes a method for providing the solution to a combined optimisation problem and several investigational outcomes that validated the fulfilment of the proposed algorithm.

Vese et al. [46] proposed the modelling of textured images using function and partial differential equation minimisation. In this work, the image is decomposed into a summation of two procedures represented by u + v, where u represents a bounded variation procedure, while v is a procedure denoting the texture or noise. The algorithm proposed uses a differential equation and is simple to solve. It also explained that the method can be used for texture discrimination of textures and texture segmentation.

Guo et al. [47] gives a theory of Marr's primal sketch that integrates three components: a texture model, a generative model along with image primitives, and a Gestalt field. It also describes the meaning of "stretchability," which helps to divide images into texture and geometry, after studying and examining two different types of models, i.e., the detailed Markov random field model and the generative wavelet/sparse coding model.

Problem Statement and Motivation

As an active field of research, a large scientific community has concentrated on pattern recognition and computer vision, detection of textual data, and recognition of textual data. Text detection and recognition may be considered an active research area due to the recent development of various portable devices and smartphone-based applications. Nowadays, text detection is a difficult and complicated task since there is a clear distinction between segregating textual and non-textual regions, as well as segregating every character technique [48] from the frame of reference. This makes the textual data fetching process much more difficult to streamline. Furthermore, the intensity of illumination is the most important part of what makes textual data detection and recognition in natural scene images complicated.

The brightness of photographs is also affected by available darkness and different lighting in the environment, but the intricate backgrounds of photographs are typically obtained from outdoor images, making the text extraction process more difficult to automate. As a result, proper text detection filtering techniques are required [19]. They usually extract sub-images from the primary image and categorise them as textual or non-textual. They do the same thing with sliding-windows of various sizes. An efficient classifier that identifies the textual and non-textual parts of natural scene images [49] with less error classification is required to eliminate this repetitive process. The textual part that has been classified is then given to a deep neural network for character recognition. A deep neural network requires optimal parameter (weight) selection, which is accomplished using an optimisation algorithm.

Proposed Methodology

The intended approach has been described in this part. The intended methodology is split among three key elements, particularly outlined in Fig. 1. A first section separates distortions mostly from the input image by inserting an edge-aware weighted filtering mechanism into the guided image filter (GIF) utilised for smoothness as well as mist elimination, increasing the overall clarity of the source image. The marker-based watershed technique is being used to preserve visual variability, minimise distortion, and improve segmentation of inside images. The Marker watershed segmentation technique initially used bilateral filtration for image analysis, which would be better for reducing the slight effect of distortion mostly in post-processing, but instead uses distance transform as well as shape rebuilding techniques for image recognition, resulting in the upcoming segmentation outcomes with extra high-accuracy capacity. The watershed transformation relies on markers to efficiently retrieve a much more precise shape of an area within a textual image. Following that, textual feature vectors are retrieved directly from text images using Gabor filtering that has been developed,

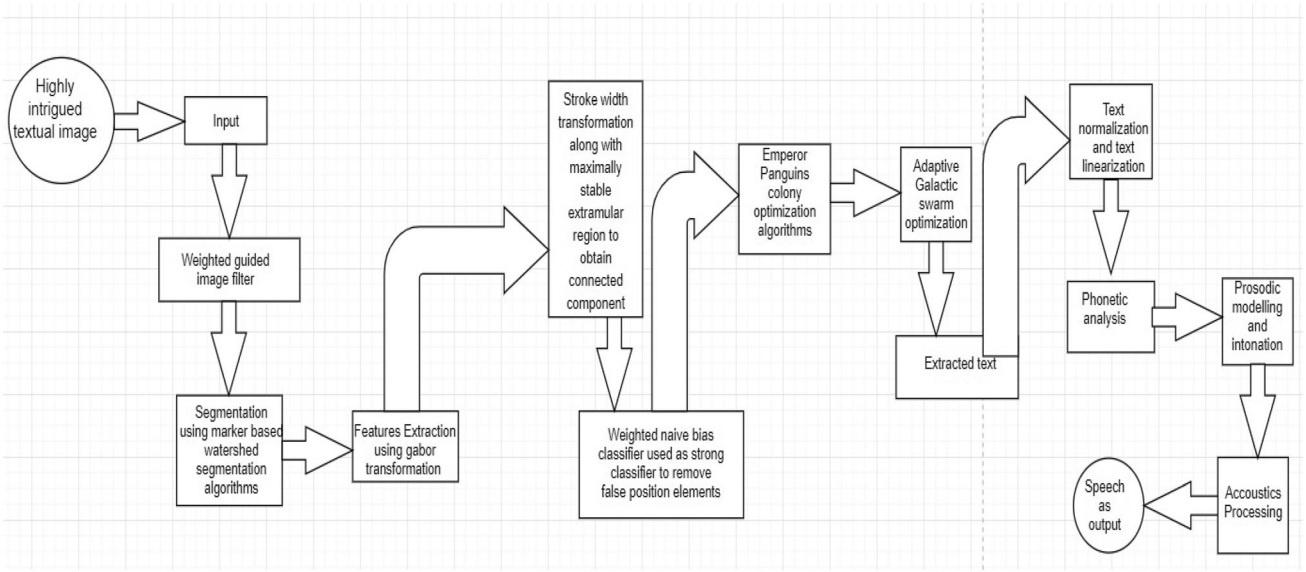

Fig. 1 Workflow of suggested text extraction process

particularly using statistical data about character structure. To improve efficiency on reduced images, all outcomes from Gabor filtering are subjected to an adaptable sigmoid function. The stroke width transformation is applied for segmentation of picture characteristics. Whenever an edge pixel becomes located in this stroke width transform, a perfect line is searched for in the direction of the edge gradient orientation of that pixel that locates another edge picture element with the reverse edge gradient orientation. If this pixel were discovered, every pixel here between two would be given a stroke width equal to the Euclidean distance between any two pixels. This procedure is typically performed on all pixels in order to obtain a mapping of the stroke widths among all picture elements. Text characters get the greatest constant size in the picture characteristics following the stroke width transform.

The collected characteristics were then input into a weighted naïve Bayes classifier to differentiate between textual and non-textual images. To further reduce errors, an emperor penguin optimum methodology is adopted in conjunction with such a weighted naïve classification algorithm. The characters within the words have been separated and then fed into a deep neural network with a particle swarm optimisation technique to identify letters in textual images. When a character is identified, a text-to-speech transformation mechanism is triggered, and audio can be acquired as an outcome by combining audio messages that are kept in a repository to turn a certain arbitrary or selected voice into text.

The term "speech synthesis" means the process of converting text to speech. Every letter is assigned a phonetic transcript throughout this procedure. Aggregation and linearisation of texts is a normalisation method for converting textual letters into phonetic forms. The text normalisation procedure converts upper-case and lower-case letters as well as eliminates punctuation. This is most effective when comparing letters that convey the very same concept. Standardisation includes the processes of acronym translation, textual data segmentation, numerical conversion, and abbreviation transformation, after which the content is linearised.

Afterwards, prosodic modelling and intonation prosody were carried out. Timing of voice, tone and tempo of voice, and pauses within words are all examples of prosodic modelling and assessment. Intonation of voice refers to the variety in tone that occurs whenever a word or sentence is spoken. Phonetic analysis p hone seems to be the tiniest component of sound and a grouping of numerous phones into phonemes. Lastly, acoustical analysis is carried out, wherein phonemes and prosody are both employed to generate a speech waveform for every phrase and word. Table 1 shows the methodology that is used to recover textual data from textual images and afterwards transform the extracted textual data into corresponding speech.

Results and Discussion

The presented work has been designed and tested throughout the personal computer environment on a processor with an Intel (R) Core (TM) i7-8565U CPU @ 1.80 GHz, RAM of 8.00 GB (7.88 GB usable), 64-bit operating system, × 64-based processor, and an operating system with Windows 10 Pro 64-bit with dimension = 15,

Table 1 Algorithm of proposed methodology for entire text extraction process

Input: Natural scene image with text Output: Extracted text
Step 1: Incorporate filtration technologies to enhance the smoothness, erase noise, and re-establish the image data
Step 2: If the input image is a colour image, it'll be transformed to a greyscale
Step 3: The input image is processed utilising gradient image and contrast image techniques
Step 4: An Adaptive contrast map will come into action to improve the contrast of the input image
Step 5: Segment the contrast enhanced image using a marker-based watershed segmentation algorithm
Step 6: Features present in the segmented image are retrieved by using the Gabor's transform and stroke width transform
Step 7: Based on this extracted feature, the weighted naïve Bayes classifier identified the textual and non-textual parts
Step 8: The error that occurs during the process of classification is minimised by using the Emperor Penguin optimisation algorithm by providing an optimal solution, and it also prevents the solution from falling into the local optimum
Step 9: The classified textual part is then given to a deep neural network for character recognition [59]
Step 10: Optimal parameter (weight) selection is necessary for deep neural networks, which is achieved by particle swarm optimisation
Step 11: The classification errors that occur during text extraction are minimised by determining the Manhattan distance between the strings
Step 12: Accomplish a Lexicon search; if the Manhattan proximity is zero, the text or string is the same; otherwise, if the proximity is one, the optimised word is gained

population-size = 3, iteration 1 = 100, iteration 2 = 100, and epoch number = 5. A 300-image set of data has been used to evaluate the suggested technique. A collection of distinct letters in appropriate font sizes has been used to train a deep neural network. The IIIT5K sample dataset was chosen as one of the standard datasets for implementing the suggested methodology. This has been the most difficult and important data, which has been further utilised by a wide range of studies. These photographs throughout this collection vary considerably in terms of distortions, the presence as well as design of blur, colour, noise, typography, and luminance. The dataset includes 6000 text images, including complex scenes with born-digital images with textual images. 1,000 of these 6000 images were chosen during learning, while 2000 images were chosen during testing. The suggested approach has proven to be far more efficient than other conventional techniques. Figure 2 depicts a textual image that is used in this suggested work. Figures 3, 4, and 5 show the images that have been acquired following the pre-processing of textual images, filtering of images, and segmentation of textual images. Figure 6 depicts characters that are

Fig. 2 Input textual image

Fig. 3 Pre-process image

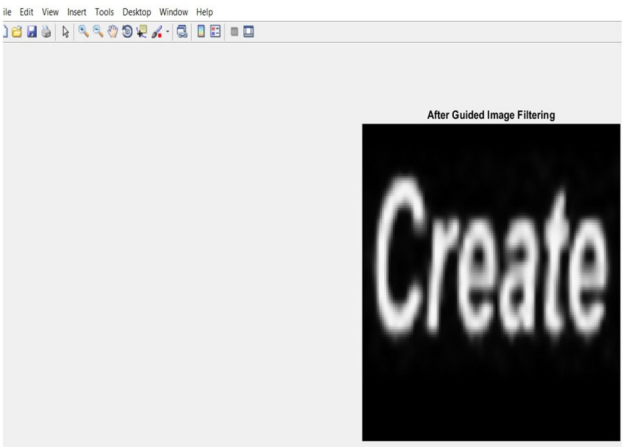

Fig. 4 Image after applying guided image filter

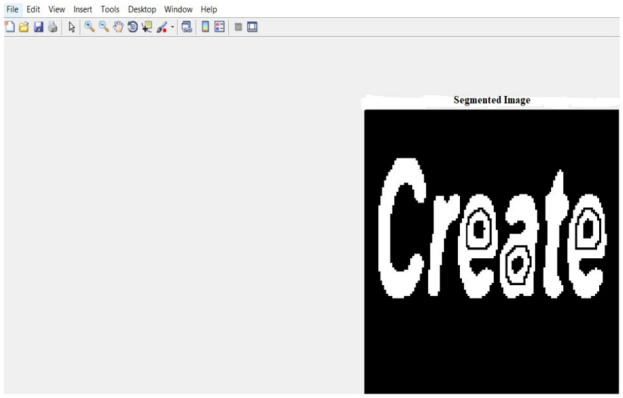

Fig. 5 Image after segmentation

retrieved from such relevant input images, and Fig. 7 shows the corresponding sound wave of input textual data.

Some of the images that are taken for the implementation process and their respective output from each stage are represented in Fig. 8.

The suggested fusion of textual extraction [50] and text-to-speech conversion technique's efficacy is measured by evaluation criteria such as accuracy, recall, precision, and F1-score. Ansari et al. [51], He et al. [52], Almaz'an [53], Khlif [11], Zhu, Zhang, R-FCN, FasterR-CNN (Ren et al. [52]). The outcomes are measured using four parameters: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

- True positives indicate those textual parts that have been appropriately recognised as textual.
- A false negative determines a textual component which is wrongly classified as a non-textual part.
- . True negative determines the non-textual part which is appropriately classified as non-text.
- False positives indicate the non-textual portion of data which are erroneously detected as textual data.

Fig. 7 Sound waves withdrawn from this available input image

Accuracy

The most basic and important metric used in evaluating the effectiveness of classification and recognition is accuracy. The formula for determining the precision is given in Eq. (1).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

Accuracy is perhaps the most fundamental and crucial metric used to assess the success of categorisation and identification. Equation (1) gives a formula for determining accuracy. As shown in Table 2, the proposed strategy outperforms the existing text retrieval methodology in terms of accuracy. Ansari [51], He [52], Almaz'an [53], SSD, and YOLOV3 are examples of well-known techniques. Figure 8 depicts the calculated graphs relying upon the aforementioned values. The value of accuracy is found to be significantly better than that of the other existing techniques.

Figure 7 depicts the results of comparing the accuracy of this deep neural network-based text retrieval procedure against existing approaches. The character recovered from this input image is accurately extracted by a deep neural network-based classifier. Whenever a misclassification problem occurs, the Manhattan distance must be included to correct this classification error. YOLO and SSD, which also classify artefact identification as just a minor problem with regression whilst having to take images and learn likelihoods of a class but also bounding box coordinates, achieve reduced costs of accuracy and therefore are

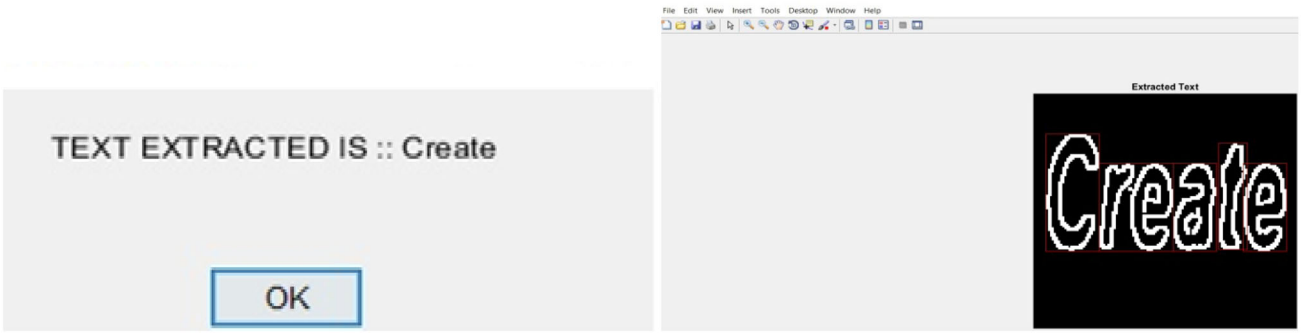

Fig. 6 Text withdrawn from this available input image

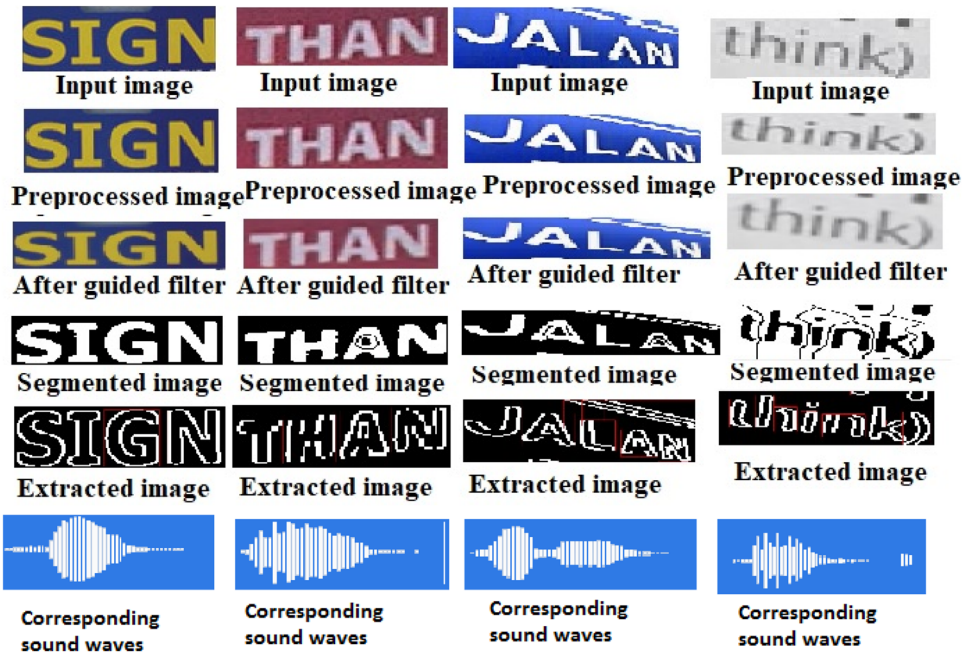

Fig. 8 Dataset taken for implementation process and its respective output from each stage

Table 2 Computation of accuracy for proposed and prevailing methods

Methods	Accuracy (%)
Proposed approach	97.88
Ansari et al. [51]	95.29
He et al. [52]	93.25
Almaz'an et al.[53]	90.63
SSD [60]	88.87
YOLOV3 [60]	83.46

significantly faster than Ansari et al., He et al., and Almaz'an et al. (Fig. 9).

Precision

Precision is calculated as the proportion of true positives to complete detection. Precision can be expressed mathematically (2). As shown in Table 3, the precision effectiveness of the proposed approach is considerably higher than that of the existing method of text extraction for small-sized textual images. A few of the known approaches

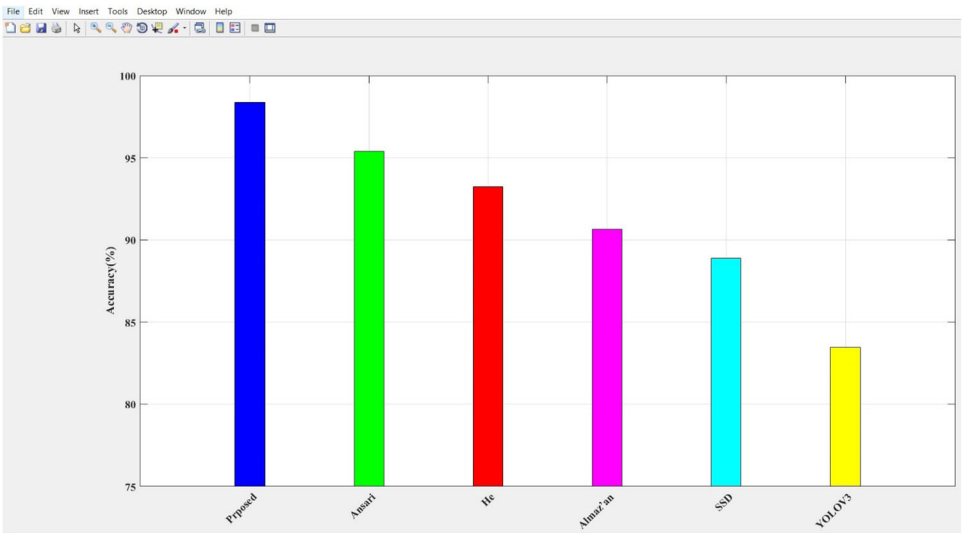

Fig. 9 Accuracy (%) of proposed and existing methods

Table 3 Precision of proposed and prevailing methods

Methods	Precision (%)
Proposed approach	93.79
Khlif [54]	91.84
R-FCN [57]	78.85
FasterR-CNN [58]	89.75
Zhu [55]	89.95
Zhang [56]	83
YOLOV3 [60]	82
SSD [60]	78

Table 4 Recall of proposed and prevailing methods

Methods	Recall (%)
Proposed approach	95.94
Khlif [54]	90.82
Zhu [55]	83.28
Zhang [56]	84
R-FCN [57]	88
FasterR-CNN [58]	76
YOLOV3 [60]	74
SSD [60]	75

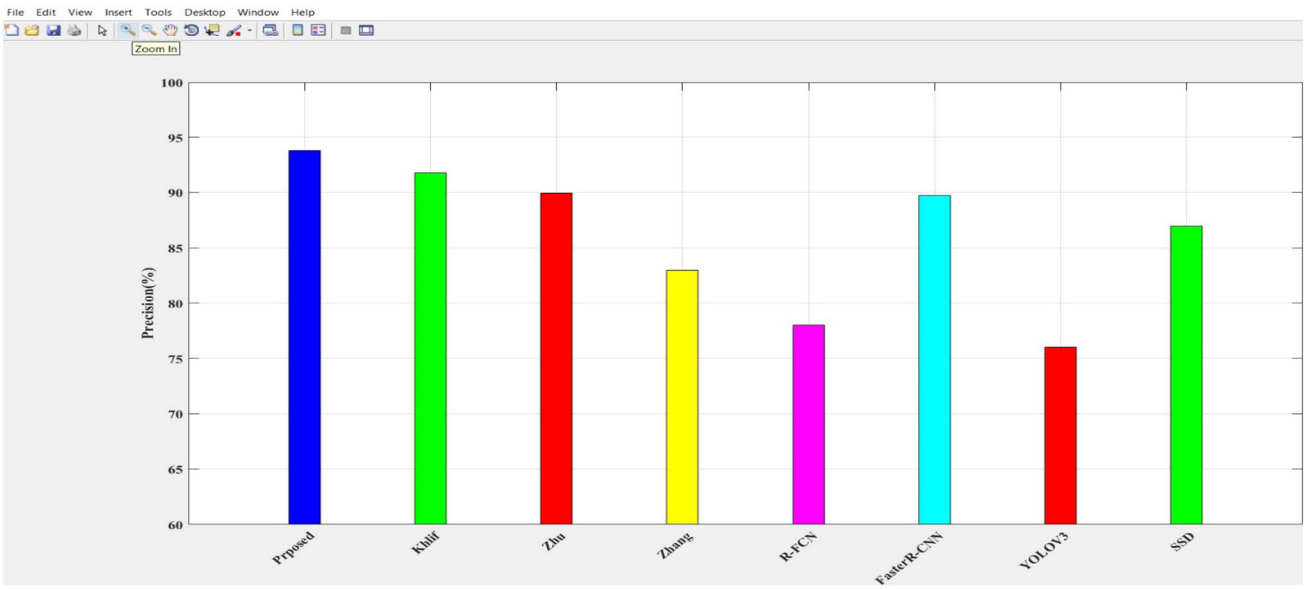

Fig. 10 Precision (%) of proposed and existing methods

such as Khlif [54], Zhu [55], Zhang [56], R-FCN [57], FasterR-CNN [58], YOLOV3, and SSD are the existing methods. Based on these values, the below graph is plotted as shown in Fig. 10.

$$p = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2}$$

Recall

The proportion of the reported true positive textual data to the entire identified true positive text as well as false negative textual data is determined by recall. The recall metric-score [45] can be mathematically expressed by Eq. (3). As shown in Table 4, the proposed method has an elevated recall value than that of the existing method of textual data fetching.

Few of the known approaches like Khlif [54], Zhu [55], Zhang [56], R-FCN [57], FasterR-CNN FasterR-CNN [58], YOLOV3, and SSD are taken as the existing method. Based on these values, the below graph is plotted shown in Fig. 3 [59].

$$r = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

F1-Score

Among the various performance parameters, the F1-score is considered an essential one. It functions as a metric for the proposed method. Precision and recall are both utilised in the F1-score estimation. The F1-score [46] can be expressed mathematically by Eq. (4). The F1-score value of the proposed method is better than the existing method of text extraction as shown in Table 5.

A few of the known approaches are Khlif [54], Zhu [55], Zhang [56], R-FCN [57], Liu et al. (2015), [58] and

Table 5 F1-score of proposed and prevailing methods

Methods	F1-score (%)
Proposed approach	95.271
Khlif [54]	91.340
Zhu [55]	86.481
Zhang [56]	83.497
R-FCN [57]	82.203
FasterR-CNN [58]	82.114
SSD [60]	76.470
YOLOV3 [60]	77.794

FasterR-CNN (Ren et al.). Based on these values, the below graph is plotted as shown in Figs. 3, 11, and 12.

$$F_1 = 2 \times \frac{p \times r}{p + r} \tag{4}$$

Comparative Analysis

Figure 13 depicts overall precision, recall, and F1-score efficacy estimates of existent and suggested textual data retrieval algorithms. The outcomes of the outcome measures were reported as just a proportion. According to the existing results, these suggested methodologies surpass previous known strategies in recovering text information from arbitrarily complicated distorted images. There seem to be eight types of textual extraction, comprising suggested methods such as khilf, RFCN, Faster R-CNN, Zhu, Zhang, YOLOV3, and SSD. Models such as khilf RFCN, Faster R-CNN, Zhu, Zhang, and others have the best average accuracy and are comparatively the slowest. The issue with R-CNN is that it takes a long time to train the classifier since it needs to properly categorize 2000

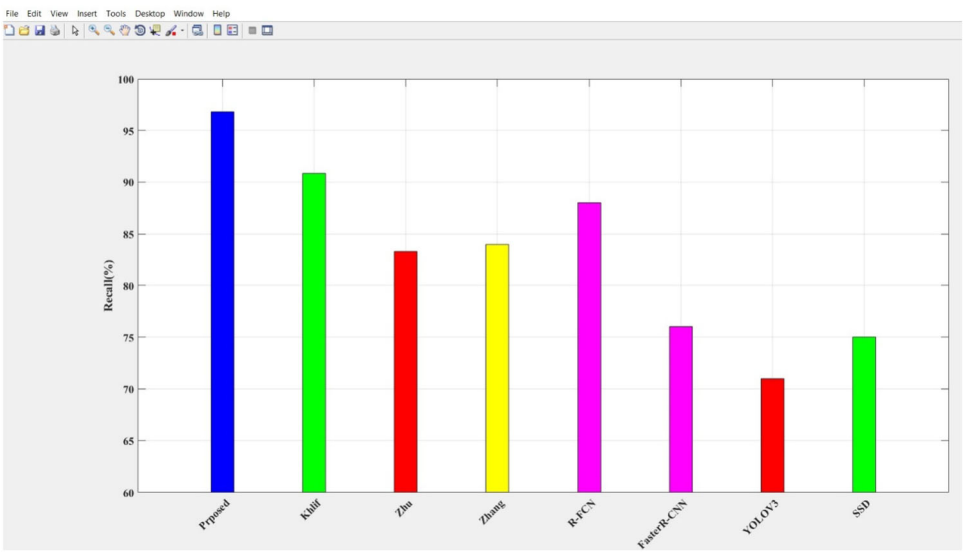

Fig. 11 Recall (%) of proposed and existing methods

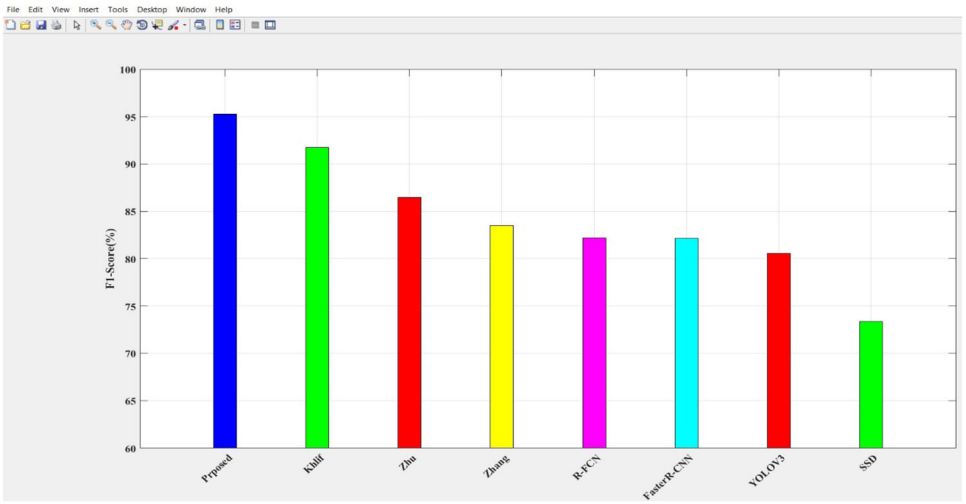

Fig. 12 F1-score (%) of proposed and existing methods

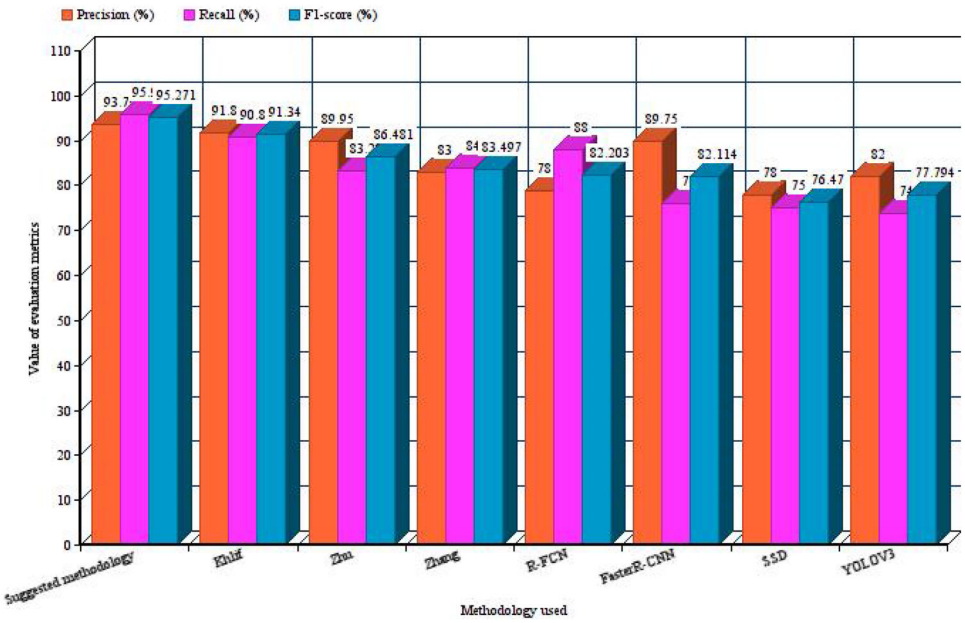

Fig. 13 The comparative analysis precision, recall and F1-score for various methodologies

conceptual frameworks with each image, preventing real-world application. As a result, no learning is required.

This could lead to the formation of inadequate regional initiatives. However, there seem to be single-stage suggestions, like YOLO and SSD, that classify artefact identification as just a simple linear regression problem by having taken images and learning the probability distribution of a class as well as bounding box coordinates. These designs accomplish reduced cost of precision, although they are significantly faster than two-stage object identification.

Conclusion

This suggested methodology can recognize text on any document, identify any notice board, identify medication, locate device robbery, identify past and present individuals, as well as interpret menu options to support visually impaired persons. Such individuals could indeed fix several of their everyday struggles with the assistance of a third party and integrating different strategies for text detection and extraction outcomes in such a method that functions faster and more efficiently using a specific strategy for the entire system. To retrieve text from exquisitely degraded images, text identification accompanied by recognition using efficacious deep learning has been used.

To begin, smoothing, noise removal, and restoration techniques are applied to the input images, and if the input image is coloured, it is converted to a greyscale. The input image is then processed using gradient and contrast image techniques to regain the image's quality. Following that, the input image's contrast will be improved using an adaptive contrast map. Stroke width transform, Gabor's transform, and weighted naïve Bayes classifier techniques are also used to segment, extract features, and detect textual and non-textual components in complex degraded images.

Finally, a combination of deep neural network and particle swarm optimisation is being used to recognise classified text. The dataset IIIT5K is used for the development portion, and while high performance is achieved with parameters such as accuracy, recall, precision, and F1-score, characters may occasionally deviate. Alternatively, the same character is frequently extracted multiple times, which may result in incorrect textual data being extracted from natural images. A minor modification in the architecture of used deep neural networks can improve the detection rate for the case of very small textual image. As a result, an efficient technique for avoiding such flaws in the text retrieval process must be implemented in the future.

Author Contributions All authors approve the final manuscript.

Funding The author(s) received no financial support for the research, authorship, and/or publication of this article.

Availability of data and material The data samples have been taken using MATLAB software.

Declarations

Conflict of interest The author declare that they have no conflict of interest.

Ethical approval Not Applicable (as the results of studies does not involve any human or animal).

Consent to participate Not Applicable (as the results of studies does not involve any human or animal).

Consent for Publication Not Applicable (as the results of studies does not involve any human or animal).

References

1. Yin XC, Yin X, Huang K, Hao HW (2014) Robust text detection in natural scene images. IEEE Trans Pattern Anal Mach Intell 36(5):970–983. https://doi.org/10.1109/TPAMI.2013.182
2. Wang L, Uchida S, Zhu A, Sun J (2018) Human reading knowledge inspired text line extraction. Cogn Comput 10(1):84–93. https://doi.org/10.1007/s12559-017-9490-4
3. Wang Y, Shi C, Xiao B, Wang C, Qi C (2018) CRF based text detection for natural scene images using convolutional neural network and context information. Neurocomputing 295:46–58. https://doi.org/10.1016/j.neucom.2017.12.058
4. Wang Y, Wang L, Su F (2018) A robust approach for scene text detection and tracking in video. In: lecture notes in computer science Pacific Rim conference on multimedia. Cham, Germany: Springer, 303–314. https://doi.org/10.1007/978-3-030-00764-5_28
5. Sain A, Bhunia AK, Roy PP, Pal U (2018) Multi-oriented text detection and verification in video frames and scene images. Neurocomputing 275:1531–1549. https://doi.org/10.1016/j.neucom.2017.09.089
6. Paul S, Saha S, Basu S, Saha PK, Nasipuri M (2019) Text localization in camera captured images using fuzzy distance transform based adaptive stroke filter. Multimed Tools Appl 78(13):18017–18036. https://doi.org/10.1007/s11042-019-7178-3
7. Ma J, Shao W, Ye H, Wang L, Wang H, Zheng Y, Xue X (2018) Arbitrary-oriented scene text detection via rotation proposals. IEEE Trans Multimed 20(11):3111–3122. https://doi.org/10.1109/TMM.2018.2818020
8. Ghai D, Jain N (2019) Comparative analysis of multi-scale wavelet decomposition and k-means clustering based text extraction. Wireless Pers Commun 109(1):455–490. https://doi.org/10.1007/s11277-019-06574-w
9. Dutta IN, Chakraborty N, Mollah AF, Basu S, Sarkar R (2019) Multi-lingual text localization from camera captured images based on foreground homogenity analysis. Adv Intell Syst Comput. https://doi.org/10.1007/978-981-13-1280-9_15
10. Ahmed SB, Naz S, Razzak MI, Yusof RB (2019) A novel dataset for English-Arabic scene text recognition (EASTR)-42K and its evaluation using invariant feature extraction on detected extremal regions. IEEE Access 7:19801–19820. https://doi.org/10.1109/ACCESS.2019.2895876
11. Khare V, Shivakumara P, Raveendran P, Blumenstein M (2016) A blind deconvolution model for scene text detection and recognition in video. Pattern Recogn 54:128–148. https://doi.org/10.1016/j.patcog.2016.01.008
12. Mehmood Z, Mahmood T, Javid MA (2018) Content-based image retrieval and semantic automatic image annotation based on the weighted average of triangular histograms using support vector machine. Appl Intell 48(1):166–181. https://doi.org/10.1007/s10489-017-0957-5
13. Tian C, Xia Y, Zhang X, Gao X (2017) Natural scene text detection with MC–MR candidate extraction and coarse-to-fine filtering. Neurocomputing 260:112–122. https://doi.org/10.1016/j.neucom.2017.03.078
14. Pandey D, Pandey BK, Wairya S (2021) Hybrid deep neural network with adaptive galactic swarm optimization for text extraction from scene images. Soft Comput 25(2):1563–1580. https://doi.org/10.1007/s00500-020-05245-4
15. Chen CT, Chen LG (1996). A self-adjusting weighted median filter for removing impulse noise in images. In Image processing. In: proceedings, international conference on (Vol 1, pp 419–422). IEEE Publications
16. Manne R, Kantheti SC (2021) Application of artificial intelligence in healthcare: chances and challenges. Curr J Appl Sci Technol 40(6):78–89. https://doi.org/10.9734/cjast/2021/v40i631320
17. Antonini M, Barlaud M, Mathieu P, Daubechies I (1992) Image coding using wavelet transform. IEEE Trans Image Process 1(2):205–220. https://doi.org/10.1109/83.136597
18. Ahmed SB, Naz S, Razzak MI, Rashid SF, Afzal MZ, Breuel TM (2016) Evaluation of cursive and non-cursive scripts using recurrent neural networks. Neural Comput Appl 27(3):603–613. https://doi.org/10.1007/s00521-015-1881-4
19. Huang Z, Zhong Z, Sun L, Huo Q (2019) Mask R-CNN with pyramid attention network for scene text detection. In: IEEE winter conference on applications of computer vision (WACV) (pp 764–772). IEEE Publications
20. 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: advances in intelligent systems and computing international conference on intelligent human systems integration. Cham, Germany: Springer, 42–48. https://doi.org/10.1007/978-3-319-73888-8_8
21. Xue M, Shivakumara P, Zhang C, Lu T, Pal U (2019) Curved text detection in blurred/non-blurred video/scene images. Multimed Tools Appl 78(18):25629–25653. https://doi.org/10.1007/s11042-019-7721-2
22. 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. https://doi.org/10.1016/j.imavis.2005.01.004
23. Kumuda T, Basavaraj L (2017) Edge based segmentation approach to extract text from scene images. In: 7th international advance computing conference (IACC). IEEE Publications, Institute of Electrical and Electronics Engineers. pp 706–710
24. Trémeau A, Fernando B, Karaoglu S, Muselet D (2011) Detecting text in natural scenes based on a reduction of photometric effects: problem of text detection. In: lecture notes in computer science international workshop on computational color imaging. Berlin, Heidelberg: Springer, 230–244. https://doi.org/10.1007/978-3-642-20404-3_18.
25. Seong S, Song J, Yoon D, Kim K, Choi J (2019) Determination of vehicle trajectory through optimization of vehicle bounding boxes using a convolutional neural network. Sensors 19:42–63. https://doi.org/10.3390/s19194263
26. Nguyen ND, Do T, Ngo TD, Le DD (2020) An evaluation of deep learning methods for small object detection. J Elect Comput Eng. https://doi.org/10.1155/2020/3189691
27. Sanchez SA, Romero HJ, Morales AD (2020) A review: comparison of performance metrics of pretrained models for object detection using the TensorFlow framework. In: InIOP conference series: materials science and engineering, 844, 012024, https://doi.org/10.1088/1757-899x/844/1/012024
28. Lawal O (2021) Tomato detection based on modified YOLOv3 framework. In Nature Research Scientific Reports. 11. https://doi.org/10.1038/s41598-021-81216-5
29. Srivastava S, Divekar AV, Anilkumar C et al (2021) Comparative analysis of deep learning image detection algorithms. J Big Data. https://doi.org/10.1186/s40537-021-00434-w

30. Wang X, Liu J (2021) Tomato anomalies detection in greenhouse scenarios based on YOLO-dense. InFrontiers Plant Sci. https://doi.org/10.3389/fpls.2021.634103
31. Loey M, Manogaran G, Taha MHN, Khalifa NEM (2021) Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection. Sustain Cities Soc 65:102600. https://doi.org/10.1016/j.scs.2020.102600
32. Yuan Q, Tan CL (2001) Text extraction from gray scale document images using edge information. In: document analysis and recognition. Sixth international conference on, 2001. Proceedings (pp. 302–306). IEEE Publications
33. Tsai CM, Lee HJ (2002) Binarization of color document images via luminance and saturation colorfeatures. IEEE Trans Image Process 11(4):434–451. https://doi.org/10.1109/TIP.2002.999677
34. Sauvola J, Pietikäinen M (2000) Adaptive document image binarization. Pattern Recogn 33(2):225–236. https://doi.org/10.1016/S0031-3203(99)00055-2
35. Sobottka K, Kronenberg H, Perroud T, Bunke H (2000) Text extraction from colored book and journal covers. Int J Doc Anal Recogn 2(4):163–176
36. Gllavata J, Ewerth R, Freisleben B (2003) A robust algorithm for text detection in images. In: image and signal processing and analysis, 2003. Proceedings of the 3rd international symposium on, 2. IEEE p 2003. Illinois School Psychologists Association
37. Andrew TD (1998) Representing multiple region of interest with wavelets. In: proceedings of the SPIE, 3309, visual communications and image processing '98, 975
38. Kim KI, Jung K, Kim JH (2003) Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm. IEEE Trans Pattern Anal Mach Intell 25(12):1631–1639. https://doi.org/10.1109/TPAMI.2003.1251157
39. Francis LM, Sreenath N (2019) Robust scene text recognition: using manifold regularized twin-support vector machine. J King Saud Univ Comput Inf Sci. https://doi.org/10.1016/j.jksuci.2019.01.013
40. Jung K, Kim KI, Jain AK (2004) Text information extraction in images and video: a survey. Pattern Recognit 37(5):977–997. https://doi.org/10.1016/j.patcog.2003.10.012
41. Chan RH, Ho CW, Nikolova M (2005) Salt-and-pepper noise removal by median-type noise detectors and detail-preserving regularization. IEEE Trans Image Process 14(10):1479–1485. https://doi.org/10.1109/tip.2005.852196
42. Peng-Lang Shui PL (2005) Image denoising algorithm via doubly local Wiener filtering with directional windows in wavelet domain. IEEE Signal Process Lett 12(10):681–684. https://doi.org/10.1109/LSP.2005.855555
43. Gatos B, Pratikakis I, Perantonis SJ (2006) Adaptive degraded document image binarization. Pattern Recogn 39(3):317–327. https://doi.org/10.1016/j.patcog.2005.09.010
44. Starck JL, Elad M, Donoho D (2004) Redundant multiscale transforms and their application for morphological component separation. Adv Imaging Electron Phys 132:287–348. https://doi.org/10.1016/S1076-5670(04)32006-9
45. Starck JL, Elad M, Donoho DL (2005) Image decomposition via the combination of sparse representations and a variational approach. IEEE Trans Image Process 14(10):1570–1582. https://doi.org/10.1109/tip.2005.852206
46. Vese LA, Osher SJ (2003) Modeling textures with total variation minimization and oscillating pattern in image processing. J Sci Comput 19(1/3):553–572. https://doi.org/10.1023/A:1025384832106
47. Guo C, Zhu S, Wu Y (2003) Towards a mathematical theory of primal sketch and Sketchability. In: proceedings of the ninth IEEE international conference on computer vision (ICCV), (Nice, France)
48. Tang Y, Wu X (2018) Scene text detection using super pixel-based stroke feature transform and deep learning based region classification. IEEE Trans Multimed 20(9):2276–2288. https://doi.org/10.1109/TMM.2018.2802644
49. Yin XC, Pei WY, Zhang J, Hao HW (2015) Multi-orientation scene text detection with adaptive clustering. IEEE Trans Pattern Anal Mach Intell 37(9):1930–1937. https://doi.org/10.1109/TPAMI.2014.2388210
50. Ali A, Pickering M, Shafi K (2018) Urdu natural scene character recognition using convolutional neural networks. In: 2nd international workshop on arabic and derived script analysis and recognition (ASAR), IEEE, 2018, (pp 29–34). IEEE publications
51. Ansari GJ, Shah JH, Yasmin M, Sharif M, Fernandes SL (2018) A novel machine learning approach for scene text extraction. Futur Gener Comput Syst 87:328–340. https://doi.org/10.1016/j.future.2018.04.074
52. He P, Huang W, Qiao Y, Loy CC, Tang X (2016) Reading scene text in deep convolutional sequences. In: thirtieth AAAI conference on artificial intelligence
53. Almazán J, Gordo A, Fornés A, Valveny E (2014) Word spotting and recognition with embedded attributes. IEEE Trans Pattern Anal Mach Intell 36(12):2552–2566. https://doi.org/10.1109/TPAMI.2014.2339814
54. Khlif W, Nayef N, Burie JC, Ogier JM, Alimi A (2018) Learning text component features via convolutional neural networks for scene text detection. In: 13th IAPR international workshop on document analysis systems (DAS) (pp. 79–84). IEEE Publications
55. Zhu A, Uchida S (2017) Scene text relocation with guidance. In: 14th IAPR international conference on document analysis and recognition (ICDAR), 1 (pp. 1289–1294). IEEE Publications
56. Zhang Z, Zhang C, Shen W, Yao C, Liu W, Bai X (2016) Multi-oriented text detection with fully convolutional networks. In: proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4159–4167)
57. Dai J, Li Y, He K, Sun J (2016) R-fcn: object detection via region-based fully convolutional networks. In: advances in neural information processing systems, 379–387
58. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. In: advances in neural information processing systems, 91–99
59. Bhunia AK, Kumar G, Roy PP, Balasubramanian R, Pal U (2018) Text recognition in scene image and video frame using Color Channel selection. Multimed Tools Appl 77(7):8551–8578. https://doi.org/10.1007/s11042-017-4750-6
60. Morera Á, Sánchez Á, Moreno AB, Sappa ÁD, Vélez JF (2020) SSD vs. YOLO for detection of outdoor urban advertising panels under multiple variabilities. Sensors 20(16):4587. https://doi.org/10.3390/s20164587