

Detection and identification of un-uniformed shape text from blurred video frames

Ravikumar Hodikehosahally Channegowda¹, Raghavendra Srinivasaiah², Santosh Kumar Jankatti³, Meenakshi⁴, Niranjana Shravanabelagola Jinachandra⁵, Raveendra Kumar Tavarekere Hombegowda⁶

¹Department of Electronics and Communication Engineering, Dayananda Sagar Academy of Technology and Management, Bengaluru, India

²Department of Computer Science and Engineering, CHRIST Deemed to be University, Bengaluru, India

³Department of Computer Science and Technology, Dayananda Sagar University, Bengaluru, India

⁴Department of Computer Science and Engineering, RNS Institute of Technology, Bengaluru, India

⁵Department of Mechanical Engineering, CHRIST Deemed to be University, Bengaluru, India

⁶Department of Electronics and Communication Engineering, Ghousia College of Engineering, Ramanagara, India

Article Info

Article history:

Received Dec 26, 2023

Revised May 16, 2024

Accepted Jun 1, 2024

Keywords:

Deep neural networks
Optical character recognition
Region convolution neural networks
Region of interest
Text detection
Text recognition

ABSTRACT

The identification and recognition of text from video frames have received a lot of attention recently, that makes many computer vision-based applications conceivable. In this study, we modify the picture mask and the original identification of the mask region convolution neural network and permit detection in three levels, including holistic, sequence, and at the level of pixels. To identify the texts and determine the text forms, semantics at the pixel and holistic levels can be used. With masking and detection, existences of the character and the word are separated and recognised. In addition, text detection using the results of 2-D feature space instance segmentation is done. Moreover, we explore text recognition using an attention-based optical character recognition (OCR) method with mask region convolution neural networks (R-CNN) to address and detect the problem of smaller and blurrier texts at the sequential level. Using attribute maps of the word occurrences in sequence to seq, the OCR method calculates the character sequence. At last, a fine-grained learning strategy is proposed to constructs models at word level using the annotated datasets, resulting in the training of a more precise and reliable model. The well-known benchmark datasets ICDAR 2013 and ICDAR 2015 are used to test our suggested methodology.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Ravikumar Hodikehosahally Channegowda

Department of Electronics and Communication, Dayananda Sagar Academy of Technology and Management
Bengaluru, India

Email: raviec40@gmail.com

1. INTRODUCTION

The huge and expanding pools of photographs and videos on the social nets and internet have seen a rapid increase in size as a result of the explosive expansion of smartphones and online social media. For efficient content-based indexing and retrieval, understanding such databases is even more important. The manual indexing-based video search tool takes up too much time and money. Being the main sources of sensory information in our daily lives, imaging and video analysis is now in high demand. Text has gotten more attention since it is a visible and open source of info in the film. A caption or other piece of text in a film gives crucial information about the scenes' content and description [1], [2]. The language in the video's captions often includes information on the location and timing of the events as well as any participants [3]. An index of the multimedia database may be created using such text. Scene locations, speaker bios, programme names, sports results, dates, and times are all included in the indexing data.

Text may be used to convey information and is one of the most expressive forms of communication. This is carried out in a readable manner. Text is essential to our daily life. Scene text and graphics text have been assigned to the text. Graphic text is the machine-printed text that has been visually superimposed and is used for video and born-digital picture captions, subtitles, and commentaries. Graphics text is divided into two types: embedded text that is embedded and superimposed on the frame and layered text that is always printed on a backdrop layer that has been created especially for the purpose. Simultaneously, scene text is organically incorporated into objects in scenes and is a component of the camera pictures [4]–[6].

It is now possible for picture capture and processing to detect text under different situations due to the increasing accessibility of high-end mobile phones with imaging and computing capabilities [7], [8]. Images and videos of any natural scene can have text recognised using image processing algorithms. At present, text recognition and identification in photos and videos has drawn more attention [9], [10]. It is now more conceivable to tackle complex issues because to advancements in pattern recognition and computer vision. Text discovery and identification have also been utilised for real-time investigation, including aiding the blind to move around freely on highways, guiding visitors to reach the endpoints, improving driving skills, routing cars depending on licence plate recognition and datamining, data collection from the sports tape, and recognising participants in marathon trials [11].

Many applications arise from text recognition, but their core objectives are to find the presence of text in the image and, if it is, to locate, perceive, and recognise it. The steps of these vital tasks are denoted to by various names in the literature, such as text localization, that intends to identify the image locations of candidate text, and text finding, that uses localization and verification procedures to determine the presence of text, and text data abstraction [12], [13], that concentrates on both binarization and localization. The majority of the present algorithms [14], [15] also emphasis on caption text in scene text and video in natural photos but do not emphasis and not perform well with natural image and video. As a result of this acceptable accurateness both the natural image and videos is an interesting topic in the area of pattern recognition and image processing.

Publicly available optical character recognitions (OCRs) for the problem of text detection, and recognition perform nicely for smooth background and excessive contrast pictures, but they fall short for natural images and video and that have low resolution, blurred, uneven lighting, low contrast, perspective distortion, random orientations, variable font categories and font sizes, and numerous colours [13], [16]–[18]. Among all other abnormalities, text recognition and identification distortion caused by motion blur is a serious problem since blur interacts with the assembly of the modules, changing the shape of the features and producing unsatisfactory results [19], [20]. Images get blurry due to poor illumination or movement between the camera and the scene. In real-time the working of text recognition and identification systems must be enhanced by developing a method for deblurring obscured vision.

Video frames often have a lesser quality than documents and scene pictures, and scene text is always blurry in videos [21]. When video frames are blurry, it is exciting to perceive and recognize text from them without deblurring. To solve this problem, we are using our deblurring approach [22] to deblur the blurry video frames. Deep learning (DL) and instance/semantic segmentation as well as object recognition have advanced. The identification and recognition of scene text has drawn more attention lately. Their method is inspired by the procedures that are used. Their strategy is specifically influenced by the mask region convolution neural networks (R-CNN) instance segmentation technique [23]. The mask branch strategies presented in masks built on R-CNN have important distinctions, however. Their method may be used to perceive an instance series inside character maps instead of analysing the mask of the item, since the mask division do not section the areas of text or approximates the possibility of character maps and text series. By taking into account three semantics levels, this work proposes unified strategies to detect and recognise irregular forms. Text detection may be considered a sample job for segmentation using the image mask text spotter and the mask R-CNN [24].

2. LITERATURE SURVEY

Text identification, text localization, and text the extraction process, and OCR are the four stages that make up text extraction in images and video frames [5], [25]. To determine if a frame of an image or video includes text information, text detection is utilized. In order to establish a border zone and locate the exact placement of the text within the picture or frame, text localization is used. To gather and categorize the text for OCR is the goal of text extraction. In this section, we look at the research on deblurring and super-resolution techniques as well as text identification and extraction from complex images and videos. Texture-based, linked component-based, gradient- and edge-based approaches may all be found in the research on text identification in pictures and video [4], [14], [26]. These approaches do not perform effectively for high-contrast writing, such as captions in videos. The scene text varies as well. A technique for recognition of words that utilizes video corners was proposed by Zhao *et al.* [27]. This approach suggests using dense corners to spot potential text possibilities. The technique creates text sections using morphological procedures starting from the corners.

The approach may not provide the desired dense corners when there is blur in the photos or videos because there is a loss of character component forms as a result of blur artefacts.

The use of texture-based approaches is advised in order to overcome the disadvantage of connected component-based solutions. The methods are subject to font distortion and modification. Liu *et al.* [28] proposed a method for word detection in video using a set of texture information and k-means clustering. Shivakumara *et al.* [29] developed a technique that uses colour spaces and the Fourier transform to recognize text in films. These techniques call for additional computations and are unable to handle blurred pictures because blur distorts the texture property that is intended to be used for text detection. To lighten the load on the computer, texture-based approaches and methods that combine edge and gradient features are suggested. Epshtein *et al.* [30] demonstrate how to use their proposed image operator by calculating the value of the width of the stroke for every single pixel to detect text in real-world pictures. As the approach assumes that the stroke width will remain constant for each edge component, blur in the video or images may prevent the requirement from being satisfied. A Bayesian classification and boundary growth were suggested by Shivakumara *et al.* [31] to distinguish multi-oriented scenario text in video. This method combines Sobel and Laplacian techniques to improve the textual information. This method is prone to blur [31] because it relies on Sobel and Laplacian techniques, which can fail in offering crisp edges for fuzzy regions.

Several techniques improve low-contrast text pixels by using temporal information for text identification in photos or videos. Based on the video's temporal redundancy, Huang suggested a technique for finding text in motion pictures. A single frame of video is used to implement text detection in order to collect potential text areas. The created motion picture ultimately only keeps the candidate text parts that experience motion as the final scene text [32]. Congjie *et al.* [33] suggested a method for extracting text that relies on several frames. Edge properties are examined with similarity metrics to identify text candidates. Because retrieved features are susceptible to blur, the technique performs poorly when the blur region appears in the video. With the use of neural networks and color-enhanced contrasting extremal areas, Sun *et al.* [34] suggested a reliable approach for word identification in photos of natural scenes. With this technique, the traits of text components that were retrieved to distinguish between real and fake text candidates are eliminated. Low-contrast and non-horizontal text graphics do poorly with this strategy. In order to identify text in natural scene pictures with confidence, Yin *et al.* [35] recommended employing maximally stable extremal regions (MSER). The approach chooses suitable MSERs-detecting character candidates using a pruning algorithm. This approach could be more effective for text pictures with several orientations and blurry text. An adaptive clustering strategy for multi-oriented scene text identification was put out by [36]. They used a single distance metric learning paradigm for adaptive clustering based on hierarchy to overcome the concerns highlighted above. This method performs admirably in images with outstanding contrast and no blur impacts, but it suffers with pictures with weak contrast and complex backgrounds.

Liang *et al.* [37] first proposed the innovative idea of Laplacian with wavelet sub-bands convolving at different levels in the frequency domains for enhancing low-resolution text pixels and finding prospective text regions. However, more accuracy than that of analysing documents is needed. By using a technique based on the evaluation of sharp gradients profiles, Favorskaya and Buryachenko [38] enhanced the identification of damaged text fragments. This method includes automatic text detection in entirely or partially blurred frames of an unpredictable video sequence. A convolution neural networks model and text line entropy are combined in a cascaded manner. The performance of text identification is greatly improved as a result of the usage of it to validate text candidates, however certain multi-orientation text lines are not successfully identified [39]. Inverse rendering-based text segmentation is a brand-new technique that Zhou *et al.* [40] explain. The method employs iterative optimization to address rendering factors such as light source, material characteristics and blur kernel size. These techniques call for uniform, high-contrast backgrounds, which might not be appropriate for photographs of videos or real-world scenes.

Shi *et al.* [41] suggested employing sliding window classification and deformable part-based models to locate and identify characters in scene photos. The direct part markings (DPMs) can reliably identify characters that have been distorted and that use a range of typefaces. Strokelets is a learned representation that Yao *et al.* [42] proposed for character recognition. Strokelets, which can be as little as a bar, arc, or corner to as large as a whole character, indicate the structural characteristics of characters at different sizes. The bag-of-strokelets histogram feature is produced by binding the Strokelets, and the resulting random forest is used for training it to recognize patterns. This method is comprehensive enough to cover several languages and strong against distortion. Phan *et al.* [43] described a semiautomatic method for producing the reality for video text identification and recognition. The recommended method's efficiency is lower since most nonhorizontal messages are scene texts, that are more challenging to distinguish from horizontal texts and might produce false positives. Roy *et al.* [44] proposed a novel method for binarizing text in the video while preserving the limitations of conventional binarization approaches. When aberrations such disconnections and information loss caused by blur and lighting effects are present in the pictures, it performs worse. To solve the issue, Tian *et al.* [45] suggested restoring the character's form using the video's medial axis points. The procedure

finds the medial axis utilising a ring radius transform idea, and then it applies a fresh restoration technique based on medial axis points. When there is blur, the outlines may no longer be geometrically coherent [45].

Pan *et al.* [46] recommended an effective L0-regularized prior determined by intensity and gradient for text image deblurring. L0-norms are np-hard problems, hence using them in applications involving video is difficult and costly in terms of temporal complexity. Multi-scale dictionaries and a flexible non-uniform deblurring technique are two methods Cao *et al.* [47] recommended. The amount of data of the dictionary and kernel estimation in different contexts affect the technique's efficacy. A new approach built around Gaussian Weighted L1 with different reduction was proposed by Khare *et al.* [48] to enhance edge strength and reduce blur in the video or image. The proposed deblur model improves the performance of text identification and detection systems in uniform blur regions, but it still needs to be more effective in non-uniform blur areas. In order to address text recognition issues in blurred photos or video frames, the majority of models suggested for generic deblurring of images and various deblurring approaches are addressed in this study. Few models have been created for identifying and extracting text from blurred picture and video frames, according to a survey of the literature.

3. PROPOSED METHODOLOGY

Detecting and identifying text in video frames using the Mask R-CNN architecture involves several steps, from training a model to handling frame extraction and recognition. Mask R-CNN is an extension of the Faster R-CNN model, designed for instance segmentation. It performs the following tasks: Object detection, identifies objects within an image and outputs bounding boxes and class labels. Instance segmentation, generates masks for each detected object, indicating the exact pixels that belong to the object. Bounding boxes, represent regions containing text and masks, identifies the exact pixels where text is present within the bounding boxes. We then identify a precise learning approach for mastering our expertise.

3.1. The architecture of the proposed method

A network of links for extracting image characteristics, recognizing text in an image to identify character and word instances, and recognizing text in an image to segment the results of text identification are the four primary parts outlined in this section as shown in Figure 1, which also demonstrates the methodology introduced. The next step is to choose a specific learning strategy for mastering our skill. In order to classify, acknowledge and identify character occurrences, text identification employs a network of context refining to discover the meaning of the holistic level. As the image text mask evaluates pixel-level semantics for recognised occurrences, segmentation is performed. The image text form is then employed for initial recognition, and the character mask is generated using an image sample mask. To grasp the semantics of seq-level for exact identification as hazy and shorter characters are projected to suffer through failures, we encompassed 2-D feature maps of word usage into text detection. According on the editing distance between the findings and the given lexicon, a major recognition outcome and recognition output are then linked.

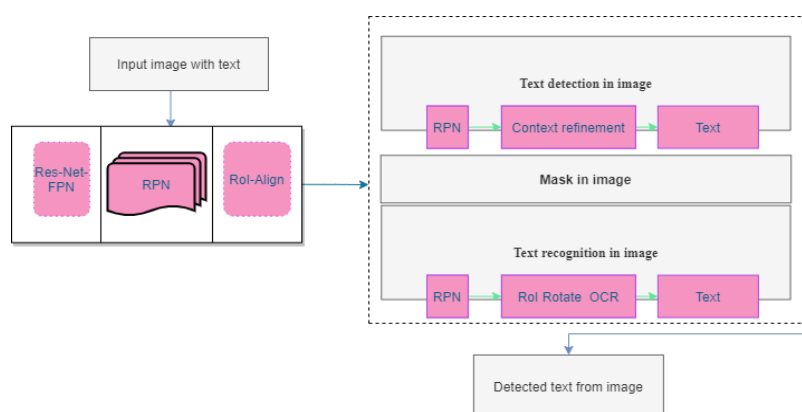


Figure 1. An architecture of proposed work

3.2. Connected network

This network includes the backbone network as well as region proposal network (RPN). The image feature is removed from a shared basis with a future network. The feature-pyramid networks employ the top-down architecture to construct the feature map of the next level at all scales with the help of each input picture and the marginal cost. As a result, we use the feature pyramid network (FPN) and Res-Net as the

network's backbone. The various area sizes and ratios of area for the image mask, R-CNN, and Mask-R-CNN are generated using the RPN. We use the RPN to create the area for the three subsequent operations in our technique. Based on their sizes, we distribute different stages. Every single stage uses a different aspect ratio. region of interest (RoI) pooling causes the misalignment problem by allocating a floating integer to various feature maps. The RoI-align procedure is then used because it generates image-based features on floating-type coordinates using a bilinear interpolation technique, resulting in exact area alignments and valuable features for further operations.

3.3. Text detection in image

By predicting their classes and anticipating that the bounding box's regression would change their coordinates, the recognition uses region techniques built with the aid of RPN to generate regions. This approach is stated in terms of determining the presence of holistic-level semantic for regional approaches. This usually requires estimating two classes: non-text and text. Text recognition aims to identify character occurrences and words in earlier work [49]. Additionally, we use character data to train identification techniques that help develop discriminative representation and enhance identification performance in both non-text and text. Based on the spatial relationship among character and word occurrences, a character identification result may be utilized for text identification. This is equivalent to carrying out text identification while using the character-based text identification approach.

However, if approaches overlap with an actual term, refining tactics that already exist may not work. Since techniques do not provide enough information to allow for an assessment of a whole object, the OCR module is dependent on the text localization outcomes. As a result, we suggest a context-refinement package to enable the in-text identification that has previously been refined. As a result of context enhancement, context data gathered from the surrounding locations is added to the depiction of a uniform context to enhance identification performance. The definition of the text loss functions identification is given in (1).

$$Loss_d = Loss_c(p, q) + \lambda[u \geq 1]Loss_l(t^u, v) \quad (1)$$

The confidence score for the provided class is roughly equal to the log loss for the true class where the boundary of the real regression targets is expressed as a tuple and defined as an estimated tuple by masking the text, one may study the pixel-level semantics of individual characters and words as well as perform semantic segmentation. Predicted word masks are utilised to offer accurate word placements in place of the conventional method. Character masks are also employed to distinguish persons in 2-D. We construct the FCN, which makes use of an image mask similar to the R-CNN mask, as illustrated in Figure 2. The provided RoI is created by the discovered method, and the appropriate RoI features are removed from fixed sizes and feature maps using the RoI-align method.

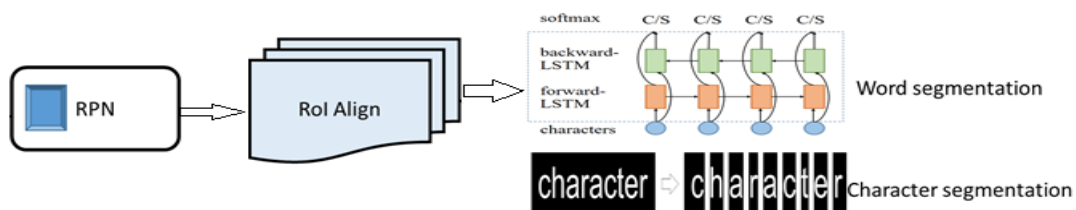


Figure 2. Image mask text

The loss function of the image text mask may be calculated as an average loss of binary cross-entropy using the mask R-CNN, as shown in (2). Where the N is the number of pixels, b_n is label of the pixel ($b_n \in (0,1)$), a_n is termed as predictable result, and the sigmoid function $S(a_n)$.

$$Loss_m = -\frac{1}{N} \sum_{n=1}^N [1 - b_n \times \log(S(a_n)) + (1 - b_n) \times \log(1 - S(a_n))] \quad (2)$$

3.4. Text recognition in image

In order to acquire sequential semantics and anticipate letter sequence using word samples from feature maps, the text recognition software intends to develop an OCR module. It is assumed that the task of Seq2Seq is text recognition. Furthermore, the attention-based technique has shown the necessary conditions for successful modelling in the absence of distances in an input sequence. As a consequence, we developed a text recognition

OCR module. A framework for text recognition includes OCR and FCN. We use FCN, which comprises of the deconvolution layer and four layers of convolution, to enlarge the maps of feature size.

We must convert an input from a 2-D feature map into a feature sequence before calculating character sequence. Before supplying the data to the OCR module, we must next apply the RoI-rotate. We can construct a rectangle for RoI with the rotation angle by figuring out the least area of the rectangle mask, which is made possible by an image mask. We may alter the RoI of the feature map horizontally using the RoI-rotate according to the rotation angle. The aspect ratio is unaffected by padding since the feature maps have a set height. Given that the same approach is used for testing as well as training, we evaluate the little impact on performance, given the fact that it inevitably deceives the picture attributes. We set up an experiment using a constant ratio aspect, but results weren't very good. As a result of the feature map's concision, which lessens network oscillation throughout the training stage and speeds up network coverage, this was done. We create an OCR module with a decoder and encoder, then a traditional Seq2Seq model. The encoder converts the feature-map into feature sequence (FS), and the decoder then advises utilizing FS to assess the character sequence. The encoder transforms a feature map to the FS using seven layers of convolution and up to four pooling layers. The multilayered bi-directional long short-term memory (LSTM) is then fed FS to find high-range FS dependencies. Last but not least, an encoder provides context via output FS. In addition to introducing the LSTM, we use recurrent neural network (RNN) to construct a decoder that seeks to predict the outcome of a character series using the seq2seq model [50]. Let ground truth for w be $A^w = \{a_0^w, a_1^w, a_2^w, \dots, a_{W+1}^w\}$ and the decoder be $B^w = \{b_0^w, b_1^w, b_2^w, \dots, b_{W+1}^w\}$ for the output sequence. In (3) calculates the loss of recognition.

$$Loss_r = -\frac{1}{N} \sum_{n=1}^N \sum_{a=1}^{W+1} \log b_x(a_x) \quad (3)$$

Here N text number to be trained and $b_x(a_x)$ describes assessed turnout probability. As shown in (4) give the multi-task loss function.

$$Loss = Loss_{rpn} + Loss_d + Loss_m + Loss_r \quad (4)$$

Where RPN loss is $Loss_{rpn}$. Three separate semantic stages, including a sequence, holistic, and pixel-level semantics, are taken into consideration by the unified framework in order to recognise and detect texts. Additionally, it offers two recognition outcomes. One is the outcome of 1-D text recognition, while the other is the result of 2-D holistic text spotting employing image masking and text identification. The outcome can withstand inaccurate localisation while text is being detected. As a result, we choose a term that is closer to the lexicon as the final recognition outcome in an effort to improve recognition.

3.5. Fine-grained learning method

Weak learning to recognise the texts allows us to get a promising result using the suggested strategy. We provide a fine-grained learning strategy that attempts to develop an accurate and dependable text spotting technique by gaining fine-grained learning with the help of trained experts. It is initially necessary to train the model and weak learning using just word annotations. This model is trained with the proposed methodology and is fully grained including character and word annotations. In the W dataset, each picture has a short footnote in the word phase and is characterized by a collection of polygons. Figure 3 shows how the suggested fine-grained learning is illustrated; then, use the learned model on the word annotations dataset W . In (5) gives the candidate set sample R .

$$\mathbb{R} = \{(p_0, c_0, \ell_0, m_0, r_0), (p_1, c_1, \ell_1, m_1, r_1), \dots, (p_x, c_x, \ell_x, m_x, r_x), \dots\} \quad (5)$$

where p_x, ℓ_x, r_x, c_x and m_x represents the anticipated category, bounding box, recognition outcome, bounding box, and image mask outcome of x^{th} sample of a potential candidate c_x . Using (6), we are able to obtain positive character samples with word level and confidence threshold annotations.

$$\mathbb{P} = \{(p_x, c_x) | p_x \in C \text{ and } c_x > S \frac{m_x \cap g_x}{m_x} > W\} \quad (6)$$

The letter C here stands for any detachable character, S describes the confidence scores threshold that is applied to perceive samples of constructive character, $m_x \cap g_y$ denotes an intersection overlap-of the candidate character r_x with the level of word ground truth g_y and the threshold is W to choose the samples of positive character. The word-level annotations' limitations, which are crucial for retaining the variety of the samples, allow for a lower confidence score threshold S . In order to create a precise and trustworthy model of the text spot, discovered samples of positive characters W can be used as character annotations and paired with word-level annotations G .

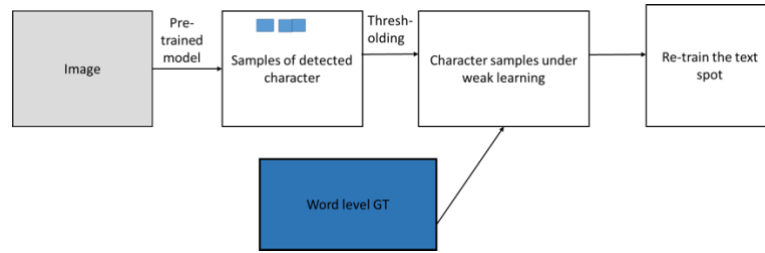


Figure 3. Framework of fine-grained learning method

4. RESULTS AND ANALYSIS

The findings and analysis of our suggested technique are presented in this part. Blurred video frames were taken into consideration. Our deblurring approach has successfully deblurred the blurry video frames [22]. The ICDAR-15 video dataset's frames have been taken into consideration. 24 films, each lasting between 12 to 15 seconds, totaling 10,800 frames, make up the ICDAR-15 video dataset [48], [51]. The test configuration consists of a Pytorch platform, an NVIDIA GTX 2080 Ti GPU, an Intel i5-7th generation CPU, 8 GB of memory, and a 1 TB SSD. With a starting learning rate of 10-3 and an ending learning rate of 10-4 after 100 epochs, the network is tweaked using stochastic gradient descent. The work done on the ICDAR-15 video dataset for hazy deblurring, text identification, and text recognition is shown in Figures 4 to 6.



Figure 4. Applied deblurring, text detection and recognition approach at blurred frames



Figure 5. Applied deblurring, text detection and recognition approach at blurred frames

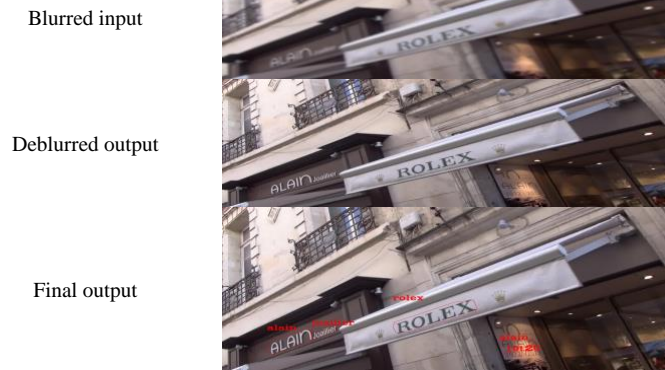


Figure 6. Applied deblurring, text detection and recognition approach at blurred frames

4.1. Comparison of the proposed model with the state of art techniques

Table 1 compares the recall, accuracy, and F-score using cutting-edge methods on the ICDAR-2015 video dataset. Table 2 compares recognition performance with cutting-edge techniques. Our approach works well compared to previous approaches.

Table 1. Text detection performance for ICDAR-15 video frames data set

Methods	Recall	Precision	F-score
Zhao <i>et al.</i> [27]	76.61	77.93	77.26
Liu <i>et al.</i> [28]	50.48	58.95	54.68
Shivakumara <i>et al.</i> [29]	81.22	81.73	81.47
Ephstein <i>et al.</i> [30]	51.4	59.75	55.03
Huang [32]	52.3	54.62	53.43
Congjie <i>et al.</i> [33]	61.86	42.48	50.37
Yin <i>et al.</i> [35]	55.7	66.3	60.5
Shivakumara <i>et al.</i> [52]	55.7	59.8	57.6
Liao <i>et al.</i> [53]	68.8	62.3	65.3
Dey <i>et al.</i> [54]	52.5	42.6	47.0
Raghunandan <i>et al.</i> [55]	67.6	62.8	66.1
Proposed method	85.6	88.45	87.01

Table 2. Text Recognition performance for ICDAR-15 video frames data set

Methods	ICDAR-13
Ostu's [56]	63.7
Bernsen [57]	59.89
Sauvola and Pietikäinen [58]	43.6
Wolf <i>et al.</i> [59]	56.82
Bhunia <i>et al.</i> [60]	75.41
Proposed method	79.82

5. CONCLUSION

Due to its potential applications in several fields, scene recognition and text detection has taken on increasing importance. We suggested an improved detection and identification strategy in this research. To find character and word occurrences in video frames, text detection is used. Text recognition in video frames for segmenting text identification results. Additionally, we pinpoint a fine-grained learning strategy for achieving the greatest results. The most common benchmark dataset, the ICDAR 2015 video dataset, is used to evaluate the proposed model and to compare it against state-of-the-art methods. Recall, accuracy, F-score, and performance with recognition are used to evaluate candidates. The recommended model has performed better when compared to cutting-edge techniques. We will incorporate a text tracking approach into our system for recognition and detection in further work in order to successfully track text in videos.

REFERENCES




- [1] P. Shivakumara, S. Bhowmick, B. Su, C. L. Tan, and U. Pal, "A new gradient based character segmentation method for video text recognition," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2011, pp. 126–130, doi: 10.1109/ICDAR.2011.34.

- [2] X. C. Yin, Z. Y. Zuo, S. Tian, and C. L. Liu, "Text detection, tracking, and recognition in video: a comprehensive survey," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2752–2773, 2016, doi: 10.1109/TIP.2016.2554321.
- [3] Y. Zhong, H. Zhang, and A. K. Jain, "Automatic caption localization in compressed video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 4, pp. 385–392, 2000, doi: 10.1109/34.845381.
- [4] C. H. Papadimitriou and J. D. Ullman, "Text detection and recognition in imagery: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 4, pp. 639–646, 2015.
- [5] K. Jung, K. I. Kim, and A. K. Jain, "Text information extraction in images and video: a survey," *Pattern Recognition*, vol. 37, no. 5, pp. 977–997, 2004, doi: 10.1016/j.patcog.2003.10.012.
- [6] P. Shivakumara, A. Dutta, U. Pal, and C. L. Tan, "A new method for handwritten scene text detection in video," in *Proceedings - 12th International Conference on Frontiers in Handwriting Recognition, ICFHR 2010*, 2010, pp. 387–392, doi: 10.1109/ICFHR.2010.67.
- [7] X. Chen, J. Yang, J. Zhang, and A. Waibel, "Automatic detection and recognition of signs from natural scenes," *IEEE Transactions on Image Processing*, vol. 13, no. 1, pp. 87–99, 2004, doi: 10.1109/TIP.2003.819223.
- [8] X. Liu and D. Doermann, "A camera phone-based currency reader for the visually impaired," in *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility*, 2008, pp. 305–306.
- [9] S. M. Lucas *et al.*, "ICDAR 2003 robust reading competitions: entries, results, and future directions," *International Journal on Document Analysis and Recognition*, vol. 7, no. 2–3, pp. 105–122, 2005, doi: 10.1007/s10032-004-0134-3.
- [10] A. Shahab, F. Shafait, and A. Dengel, "ICDAR 2011 robust reading competition challenge 2: reading text in scene images," in *The International Conference on Document Analysis and Recognition, ICDAR*, 2011, pp. 1491–1496, doi: 10.1109/ICDAR.2011.296.
- [11] C. Yao, X. Bai, and W. Liu, "A unified framework for multioriented text detection and recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4737–4749, 2014, doi: 10.1109/TIP.2014.2353813.
- [12] R. Lienhart and A. Wernicke, "Localizing and segmenting text in images and videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 4, pp. 256–268, 2002, doi: 10.1109/76.999203.
- [13] J. Zhang and R. Kasturi, "Extraction of text objects in video documents: recent progress," *2008 The Eighth IAPR International Workshop on Document Analysis Systems*, Nara, Japan, 2008, pp. 5–17, doi: 10.1109/DAS.2008.49.
- [14] J. Zhang and R. Kasturi, "A novel text detection system based on character and link energies," *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 4187–4198, 2014, doi: 10.1109/TIP.2014.2341935.
- [15] C. Yi and Y. Tian, "Localizing text in scene images by boundary clustering, stroke segmentation, and string fragment classification," *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 4256–4268, 2012, doi: 10.1109/TIP.2012.2199327.
- [16] D. Crandall, S. Antani, and R. Kasturi, "Extraction of special effects caption text events from digital video," *International Journal on Document Analysis and Recognition*, vol. 5, no. 2–3, pp. 138–157, 2003, doi: 10.1007/s10032-002-0091-7.
- [17] D. Chen and J. M. Odobez, "Video text recognition using sequential Monte Carlo and error voting methods," *Pattern Recognition Letters*, vol. 26, no. 9, pp. 1386–1403, 2005, doi: 10.1016/j.patrec.2004.11.019.
- [18] L. Wu, P. Shivakumara, T. Lu, and C. L. Tan, "Text detection using delaunay triangulation in video sequence," in *Proceedings - 11th IAPR International Workshop on Document Analysis Systems, DAS 2014*, 2014, pp. 41–45, doi: 10.1109/DAS.2014.28.
- [19] P. Shivakumara, Z. Yuan, D. Zhao, T. Lu, and C. L. Tan, "New gradient-spatial- structural features for video script identification," *Computer Vision and Image Understanding*, vol. 130, pp. 35–53, 2015.
- [20] X. Cao, W. Ren, W. Zuo, X. Guo, and H. Foroosh, "Scene text deblurring using text- specific multi scale dictionaries," *IEEE Transactions on Image Processing*, vol. 24, no. 4, pp. 1302–1314, 2015.
- [21] C. Yang *et al.*, "Tracking based multi-orientation scene text detection: a unified framework with dynamic programming," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3235–3248, 2017.
- [22] H. C. Ravikumar and P. Karthik, "Optimized and efficient deblurring through constraint conditional modelling," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 3, pp. 1503–1512, 2021, doi: 10.11591/ijeecs.v21.i3.pp1503-1512.
- [23] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298–2304, 2017, doi: 10.1109/TPAMI.2016.2646371.
- [24] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
- [25] S. Antani, D. Crandall, and R. Kasturi, "Robust extraction of text in video," in *Proceedings-International Conference on Pattern Recognition*, 2000, vol. 15, no. 1, pp. 831–834, doi: 10.1109/ICPR.2000.905537.
- [26] D. Chen, J. M. Odobez, and H. Bourlard, "Text detection and recognition in images and video frames," *Pattern Recognition*, vol. 37, no. 3, pp. 595–608, 2004, doi: 10.1016/j.patcog.2003.06.001.
- [27] X. Zhao, K. H. Lin, Y. Fu, Y. Hu, Y. Liu, and T. S. Huang, "Text from corners: a novel approach to detect text and caption in videos," *IEEE Transactions on Image Processing*, vol. 20, no. 3, pp. 790–799, 2011, doi: 10.1109/TIP.2010.2068553.
- [28] C. Liu, C. Wang, and R. Dai, "Text detection in images based on unsupervised classification of edge-based features," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2005, pp. 610–614, doi: 10.1109/ICDAR.2005.228.
- [29] P. Shivakumara, T. Q. Phan, and C. L. Tan, "New fourier-statistical features in RGB space for video text detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1520–1532, 2010, doi: 10.1109/TCSVT.2010.2077772.
- [30] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2963–2970, doi: 10.1109/CVPR.2010.5540041.
- [31] P. Shivakumara, R. P. Sreedhar, T. Q. Phan, S. Lu, and C. L. Tan, "Multioriented video scene text detection through bayesian classification and boundary growing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 8, pp. 1227–1235, 2012, doi: 10.1109/TCSVT.2012.2198129.
- [32] X. Huang, "A novel approach to detecting scene text in video," in *Proceedings - 4th International Congress on Image and Signal Processing, CISP 2011*, vol. 1, pp. 469–473, doi: 10.1109/CISP.2011.6099945.
- [33] M. Congjie, X. Yuan, L. Hong, and X. Xiangyang, "A novel video text extraction approach based on multiple frames," in *2005 Fifth International Conference on Information, Communications and Signal Processing*, 2005, vol. 2005, pp. 678–682, doi: 10.1109/icip.2005.1689133.
- [34] L. Sun, Q. Huo, W. Jia, and K. Chen, "A robust approach for text detection from natural scene images," *Pattern Recognition*, vol. 48, no. 9, pp. 2906–2920, 2015, doi: 10.1016/j.patcog.2015.04.002.
- [35] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, "Robust text detection in natural scene images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 5, pp. 970–983, 2013.
- [36] X. C. Yin, W. Y. Pei, J. Zhang, and H. W. Hao, "Multiorientation scene text detection with adaptive clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1930–1937, 2015, doi: 10.1109/TPAMI.2014.2388210.




- [37] G. Liang, P. Shivakumara, T. Lu, and C. L. Tan, "Multi-spectral fusion based approach for arbitrarily oriented scene text detection in video images," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4488–4501, 2015, doi: 10.1109/TIP.2015.2465169.
- [38] M. Favorskaya and V. Buryachenko, "Scene text deblurring in non-stationary video sequences," *Procedia Computer Science*, vol. 96, pp. 744–753, 2016, doi: 10.1016/j.procs.2016.08.259.
- [39] Y. Zheng, Q. Li, J. Liu, H. Liu, G. Li, and S. Zhang, "A cascaded method for text detection in natural scene images," *Neurocomputing*, vol. 238, pp. 307–315, 2017, doi: 10.1016/j.neucom.2017.01.066.
- [40] Y. Zhou, J. Feild, E. Learned-Miller, and R. Wang, "Scene text segmentation via inverse rendering," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2013, pp. 457–461, doi: 10.1109/ICDAR.2013.98.
- [41] C. Shi, C. Wang, B. Xiao, Y. Zhang, S. Gao, and Z. Zhang, "Scene text recognition using part-based tree-structured character detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2961–2968, doi: 10.1109/CVPR.2013.381.
- [42] C. Yao, X. Bai, B. Shi, and W. Liu, "Strokelets: a learned multi-scale representation for scene text recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 4042–4049, doi: 10.1109/CVPR.2014.515.
- [43] T. Q. Phan, P. Shivakumara, S. Bhowmick, S. Li, C. L. Tan, and U. Pal, "Semiautomatic ground truth generation for text detection and recognition in video images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 8, pp. 1277–1287, 2014, doi: 10.1109/TCSVT.2014.2305515.
- [44] S. Roy, P. Shivakumara, P. P. Roy, U. Pal, C. L. Tan, and T. Lu, "Bayesian classifier for multi-oriented video text recognition system," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5554–5566, 2015, doi: 10.1016/j.eswa.2015.02.030.
- [45] S. Tian, P. Shivakumara, T. Q. Phan, T. Lu, and C. L. Tan, "Character shape restoration system through medial axis points in video," *Neurocomputing*, vol. 161, pp. 183–198, 2015, doi: 10.1016/j.neucom.2015.02.044.
- [46] J. Pan, Z. Hu, Z. Su, and M. H. Yang, "Deblurring text images via L0-regularized intensity and gradient prior," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2901–2908, doi: 10.1109/CVPR.2014.371.
- [47] X. Cao, W. Ren, W. Zuo, X. Guo, and H. Foroosh, "Scene text deblurring using text-specific multi scale dictionaries," *IEEE Transactions on Image Processing*, vol. 24, no. 4, pp. 1302–1314, 2015.
- [48] V. Khare, P. Shivakumara, P. Raveendran, and M. Blumenstein, "A blind deconvolution model for scene text detection and recognition in video," *Pattern Recognition*, vol. 54, pp. 128–148, 2016, doi: 10.1016/j.patcog.2016.01.008.
- [49] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, "FOTS: fast oriented text spotting with a unified network," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5676–5685, doi: 10.1109/CVPR.2018.00595.
- [50] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *Advances in Neural Information Processing Systems*, pp. 577–585, 2015.
- [51] D. Karatzas et al., "ICDAR 2015 competition on robust reading," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2015, vol. 2015-Novem, pp. 1156–1160, doi: 10.1109/ICDAR.2015.7333942.
- [52] P. Shivakumara, L. Wu, T. Lu, C. L. Tan, M. Blumenstein, and B. S. Anami, "Fractals based multi-oriented text detection system for recognition in mobile video images," *Pattern Recognition*, vol. 68, pp. 158–174, 2017, doi: 10.1016/j.patcog.2017.03.018.
- [53] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "TextBoxes: a fast text detector with a single deep neural network," in *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, 2017, pp. 4161–4167, doi: 10.1609/aaai.v31i1.11196.
- [54] S. Dey et al., "Script independent approach for multi-oriented text detection in scene image," *Neurocomputing*, vol. 242, pp. 96–112, 2017, doi: 10.1016/j.neucom.2017.02.061.
- [55] K. Raghunandan, P. Shivakumara, S. Roy, G. H. Kumar, U. Pal, and T. Lu, "Mutiscript-oriented text detection and recognition in video/scene/born digital images," *IEEE transactions on circuits and systems for video technology*, vol. 29, no. 4, pp. 1145–1162, 2019, doi: 10.1109/TCSVT.2018.2817642.
- [56] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979, doi: 10.1109/TSMC.1979.4310076.
- [57] J. Bernsen, "Dynamic thresholding of gray-level images," in *Proceedings - International Conference on Pattern Recognition*, 1986, pp. 1251–1255.
- [58] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, no. 2, pp. 225–236, 2000, doi: 10.1016/S0031-3203(99)00055-2.
- [59] C. Wolf, J. M. Jolion, and F. Chassaing, "Text localization, enhancement, and binarization in multimedia documents," in *Proceedings - International Conference on Pattern Recognition*, 2002, vol. 16, no. 2, pp. 1037–1040, doi: 10.1109/icpr.2002.1048482.
- [60] A. K. Bhunia, G. Kumar, P. P. Roy, R. Balasubramanian, and U. Pal, "Text recognition in scene image and video frame using color channel selection," *Multimedia Tools and Applications*, vol. 77, no. 7, pp. 8551–8578, 2018, doi: 10.1007/s11042-017-4750-6.

BIOGRAPHIES OF AUTHORS






Dr. Ravikumar Hodikeyhosahally Channegowda    completed his Ph.D. from VTU, Belagavi in 2021. He has done his masters in VLSI design and embedded systems from VTU Extension Centre, PESCE, Mandya. His areas of interest are image processing, machine learning, pattern recognition, and multimedia concepts. He is currently working as Assistant Professor at Dayananda Sagar Academy of Technology and Management, Bengaluru. He can be contacted at email: raviec40@gmail.com.






Dr. Raghavendra Srinivasaiah    is currently working as Associate Professor in the Department of Computer Science and Engineering at CHRIST Deemed to be University, Bangalore. He completed his Ph.D. degree in Computer Science and Engineering from VTU, Belgaum, India in 2017 and has more than 18+ years of teaching experience. His interests include data mining, artificial intelligence, and big data. He can be contacted at email: raghav.trg@gmail.com.






Dr. Santosh Kumar Jankatti    is currently working as Associate Professor in the Department of Computer Science and Technology at Dayananda Sagar University, Bangalore. He completed his Ph.D. degree in Computer Science and Engineering from VTU, Belgaum, India in 2022 and has more than 12 years of teaching experience and 3 years of IT Industry experience. His interests include data mining, artificial intelligence, and big data. He can be contacted at email: sjankatti@gmail.com.






Mrs. Meenakshi    is currently working in RNSIT Bengaluru which is affiliated to VTU Belagavi in the Department of Computer Science and Engineering, she secured VTU 8th Rank, she completed masters from VTU, she has 4 years of teaching experience and her area of interest are big data analytics, data mining. She can be contacted at email: meenakshib437@gmail.com.



Dr. Niranjana Shravanabelagola Jinachandra    completed his Ph.D. from VTU, Belagavi in 2022. He has done his masters in Machine design from VTU, Belagavi. His areas of interest are image processing, machine learning, and fluid dynamics. He is currently working as Associate Professor in the Department of Mechanical Engineering at CHRIST Deemed to be University. He can be contacted at email: sjniranjan86@gmail.com



Raveendra Kumar Tavarekere Hombegowda    currently pursuing Ph.D. in VTU, Belagavi. He has done his masters in VLSI Design and Embedded System from SJCE, Mysore, VTU. His areas of interest are signal and image processing, machine learning, and multimedia concepts. He is currently working as an assistant professor at Ghousia College of Engineering, Ramanagra. He can be contacted at email: thraveendra@gmail.com.