IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 4, August 2026, pp. 3537~3545
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i4.pp3537-3545

E-TEXTLOC: efficient text localization and extraction in real-world video scenes

Dayananda Kodala Jayaram[1,2], Puttegowda Devegowda[1]
[1]Department of Computer Science and Engineering, ATME College of Engineering affiliated to Visvesvaraya Technological University, Belagavi, India
[2]Department of Computer Science and Engineering, Vidyavardhaka College of Engineering (Autonomous Institute) affiliated to Visvesvaraya Technological University, Belagavi, India

Article Info

Article history:

Received Sep 6, 2025
Revised May 22, 2026
Accepted Jul 9, 2026

Keywords:

Accuracy
Machine learning
Text detection
Text localization
Text recognition

ABSTRACT

Localization of an accurate text is quite a complicated issue, especially when attempting to extract from a complex video. It is mainly due to the limitation of resources, dynamic background, and text variability. There is various artificial intelligence based methods adopting machine learning for addressing such issues encounter lower positional accuracy and incur maximized computational cost. Therefore, the proposed system introduces efficient text localization and extraction for comprehensive positional accuracy in complex videos (E-TEXTLOC). Different from conventional approaches, E-TEXTLOC facilitates sampling of video frames while MobileNetV2 is deployed towards faster localization of text. The outcome of recognized text is further refined by a verifier module, which provides a self-supervised response. Assessed on the YouTube video dataset, the proposed model accomplishes 98.2% accuracy with 29.6 ms towards generating analytical outcomes. It means the proposed model accomplishes 6-10% accuracy enhancement with a 40-50% reduction of speed in contrast to the existing system. The implications of the proposed study can be stated towards surveillance system, autonomous vehicles, and assistive devices that works in real-time.

Corresponding Author:

Dayananda Kodala Jayaram
Department of Computer Science and Engineering, ATME College of Engineering
affiliated to Visvesvaraya Technological University
Bannur Rd, Mysuru, Karnataka – 570028, India
Email: dayananda.kem@gmail.com

1. INTRODUCTION

The term text localization refers to the mechanism of determining the position of textual content present within any form of multimedia object. There are two other equivalent terms known as text detection and text recognition, which are closely connected to text localization, where the former one relates to finding the text element within an image or video frame, while the latter relates to the method of decoding the context of actual text from a localized region [1]. The significance of text localization is that it acts as a foundation for text recognition, while they are quite essential in various real-world applications, viz., assistive technology, augmented reality, document digitization, and autonomous vehicles [2]–[4]. Text localization also assists in efficient processing as the system doesn't require executing optical character recognition (OCR) on the complete image, as localization regions are only subjected to further processing. There are various research-based approaches towards text localization, viz., texture-based, semantic segmentation, and region proposal method [5]–[7]. Irrespective of various research-based ongoing solutions,

there are various ongoing research challenges, viz., text variability, visual variation, generalization, resource limitation, and layout complexity [8].

Machine learning-based methods have evolved as one of the promising solution towards text localization [9]. It is deployed for automated feature learning where hierarchical features are learned autonomously from data. Various machine learning models are used specifically towards localization and recognition, viz., character region awareness for text detection (CRAFT), efficient and accurate scene text detector (EAST), and connectionist text proposal network (CTPN). These models are quite capable of identifying both non-standard and non-linear layouts of text regions. Various lightweight models, like variants of MobileNet, facilitate real-time deployment on resource-constrained devices and thereby optimize latency. Adoption of machine learning also provides the system to undergo learning operations with invariant representations, which can generalize quite well on diverse conditions. Various transformer-based models (e.g., vision transformers) provide modelling of global context and offer explicit attentions. This approach potentially assists in emphasizing the relevant region of text for any given condition of scenes. Although machine learning and deep learning offer significant contributions towards solving text localization problems, there are various potential setbacks too that are consistently being a part of ongoing investigation. The primary bigger challenge in the deployment of machine learning approaches is real-world variability, which is characterized by a dynamic environment with unseen layouts, scripts, and fonts. Usually, models are trained on a cleaner version of the dataset, which fails when subjected to different unforeseen environments. Various dynamic conditions like background noise, motion blur, occlusion, and lighting can easily degrade the performance of the model. Various high-performance model of machine learning (e.g., transformer) demands support of graphical or tensor processing unit. This not only consumes extensive resources but also degrades the balance among accuracy, speed, and power demands.

Various related works have been studied to understand the adoption of machine learning towards text localization. Existing systems have been dominantly witnessed to adopt convolution neural network (CNN) towards localization of text contents by integrating CNN with various other deep learning models, e.g., transformer and recurrent neural network [10]–[14]. Although these CNN-based approaches are known to address issues pertaining to the integration of text into the detection of objects, shapes of irregular text, and recognition of bilingual text, they demand massive annotated data and are highly challenging to handle occluded instances of text. There are various research models designed using histogram of oriented gradients (HOG) and support vector machine (SVM) towards scene text detection [15]–[19]. These approaches are known to accomplish a decent accuracy of recognition that is found suitable for non-English textual contents with simplified feature extraction. However, HOG schemes with SVM are also witnessed with shortcomings that are reported to limit the performance in contrast to modern times of deep learning approaches. HOG has also been integrated with another efficient machine learning algorithm of random forest towards accomplishing better performance of text localization [20]–[26]. These studies perform the extraction of unique features where HOG is utilized for extracting gradient-oriented features, while they are also integrated with various types of moments. However, such approaches are not found to withstand complex variation in scene text.

The research problems identified are i) majority of the existing system is found to working fine with their dataset with static video scene but not much evidence is put forwarded to address inaccurate localization of text for complex scenes of video, ii) almost all the approaches gives more importance to solve detection-based problem but not much emphasis is towards investigating location-based information viz. bounding coordinates, size, and orientation, iii) irrespective of exhibits of some higher accuracy scores, majority of existing models adopts machine learning approaches that will definitely result in higher computational cost and increased inference time, and iv) conventional framework that are trained on particular dataset doesn't generalize optimally to multimedia dataset with an inclusion of different language, motion, and fonts. The problem of accurate localization of text is addressed in the proposed system by presenting a unique feature extraction scheme that adaptively performs text localization capable of fine-tuning itself to different types of background of video. The issue pertaining to positional information is addressed in the proposed system by considering a bounding box formulated dynamically, facilitating the model with enriched spatial context apart from performing detection and recognition. The problem of extensive computational cost and increased inference time is controlled by the proposed system by adopting an optimized architecture of a neural network, making it suitable for real-time applications.

The research aim of the proposed study model is towards introducing an efficient text localization framework guided by an intelligent verifier for complex forms of video content. The proposed model is called as E-TEXTLOC, which will mean an efficient text localization and extraction for comprehensive positional accuracy in complex videos. The value-added contribution of E-TEXTLOC are as follows: i) E-TEXTLOC presents a framework for accurate localization of text by adopting contour-based region proposal technique for highlighting areas enriched with text with increased spatial saliency, ii) the model

implements deep convolution neural network (DCNN) for verification of candidate region rapidly optimizing the accuracy performance with minimal overhead, iii) the proposed framework targets to reduce the inference time for each video frame adopting optimized feature selection for facilitating E-TEXTLOC in real-time environment, and iv) the model implements a novel and yet simplified feedback mechanism using verifier for assessing outcome of OCR followed by incorporating model learning for optimized accuracy performance.

The proposed model is different from conventional localization methods such as EAST and CRAFT, which essentially depend upon dense pixel-level forecasting and demand post-processing that is actually computationally intensive. On the contrary, a lightweight contour-based region proposal method has been presented by E-TEXTLOC for minimizing the search space before validating it by deep learning. Similarly, transformer-based methods are good in modelling contextual dependencies and yet they need an extensively large labelled dataset and need computational resources. On the contrary, E-TEXTLOC focuses on computational efficiency using verifier modelling of CNN and feedback methods using OCR. In short, a self-supervised verified modelling has been presented by E-TEXTLOC, which refines the localization result using the confidence score of an OCR. The next section presents a discussion of the adopted methodology to accomplish the study objectives.

2. METHOD

The proposed system of E-TEXTLOC adopts a simplified mathematical modelling targeting text localization, which further consists of both detection as well as recognition of text residing within the video frames. Figure 1 highlights the adopted study architecture consisting of various operation modules to accomplish the study goal. The prime module involved in architecture is responsible for the generation of region as part of accomplishing the goal of localization while the DCNN module and the OCR module carry out the task of detection and recognition. The optimal structure and integration of each module are mathematically formulated.

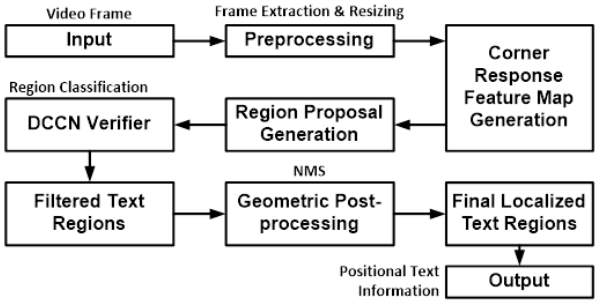

Figure 1. Architecture of E-TEXTLOC

2.1. Corner response feature map region proposal

This module is responsible for generating the candidate region with a high possibility of the presence of textual content in the video frames. The notion of corner response feature map (CRFM) is based on the computation of local variation associated with intensity, followed by the detection of structures with contours. Consider adopting a gradient edge detector as (I_x, I_y)=(∂I/∂x, ∂I/∂y) where $I(x, y)$ represents grayscale intensity at location (x, y) in the video frame. The model defines curvature-based interest point detector contour response $R(x, y)$ as (1).

$$R(x,y) = \det\left(\begin{bmatrix} I_x^2 & I_xI_y \\ I_xI_y & I_y^2 \end{bmatrix} * G_\sigma\right) - k.(trace\begin{bmatrix} I_x^2 & I_xI_y \\ I_xI_y & I_y^2 \end{bmatrix})^2 \quad (1)$$

In the (1), the structure inside the determinant *det* function is a structure tensor associated with each pixel, while G_σ represents a Gaussian kernel that is used for smoothening (along with standard deviation), and the variable k is an empirical constant with value [0.04, 0.06]. A closer look into the first component of (1) will be represented in the form of $\{I_x^2I_y^2\text{-}(I_xI_y)^2\}$ while the second component of (1) will represent $(I_x^2+I_y^2)$. Further, *trace*() represents the summation of all diagonal elements present within the structure tensor. CRFM also defines $C(x, y)$ as a binary contour map that is represented as (2).

$$C(x,y) = \begin{cases} 1 & R(x,y) > T_c \\ 0 & otherwise \end{cases} \quad (2)$$

According to the (2), the valuation of the binary contour map completely depends upon contour response $R(x, y)$, where T_c represents its cut-off value for determining the potential contour responses. The model then applies morphological operations subjected to a binary contour map $C(x, y)$ that can now be represented as $C'(x,y) = dilate(close(C(x,y)))$. The system then uses labelling of all connected components to acquire the contours. $(R_i)_{i-1}^{N}$ from connected regions of $C'(x, y)$. Finally, a specific region with location information with a valid aspect ratio is retained by the filtering step as (3).

$$B_i \in S = D_1 \in [\alpha_1, \alpha_2], A_i \in [A_{min}, A_{max}] \quad (3)$$

In the (3), the variable B_i represents a bounding box specific to contour region R_i, while S represents a set of all candidate regions of text i.e., $S = \{Bi\}$. The variable D_i represents (w_i/h_i), where w_i and h_i represent the width and height of the bounding box and area A_i is represented as (w_i. h_i). The valid range of area is further designated as (A_{min}, A_{max}). The novelty of this part of the implementation is associated with the introduction of low-level feature extraction along with the mechanism of region proposal. This method not only contributes towards highly granular and lightweight text localization but also towards faster identification of potential candidate regions.

2.2. Deep convolution neural network verifier

The key purpose of this module is towards enriching the outcome of the prior CRFM module by removing all the false positives where regions may be detected with enriched contours, but they lack the possession of any textual content. As the CRFM module cannot afford to miss any text, hence DCCN verifier ensures that the text region with a higher confidence rate should be passed on for its consecutive stages of recognition. E-TEXTLOC uses the DCCN verifier in order to minimize the possibilities of outliers from CRFM, which is empirically represented as (4).

$$f_\theta = R^{H \times W \times 3} [0, 1] \quad (4)$$

In the (4), f_θ represents a binary classifier that takes the value of either 1 or 0 based on whether the region x is found to have textual contents or otherwise, respectively, while the suffix θ is a network parameter. The model considers a lightweight model that consists of convolution layers followed by batch normalization, activation of rectified linear unit (ReLU), and a layer with a fully connected network. A dataset $db = \{(x_i, y_i)\}$ is used to train the network, where x_i and y_i represent a cropped patch of an image and label (0, 1), respectively. The system computes binary cross-entropy loss as (5).

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}[B_1 + B_2] \quad (5)$$

In the (5), it can be noted that computation of loss function $L(\theta)$ depends upon two entities, i.e., B_1 and B_2, which represent $y_i . log\ f_\theta(x_i)$ and $(1 - y_i) . log(1 - f_\theta(x_i))$ respectively. Finally, the inference is carried out with $\hat{y}_i$ that undertakes the value of either 1 or 0 for $f_\theta(x_i) > T_p$ or otherwise, respectively; where the attribute T_p represents a binary cut-off of classification confidence. It should be noted that the only region with a detected bounding box as a part of S has with value of $\hat{y}_i$ as unity as forwarded to the next stage of operation.

2.3. Recognition module with optical character recognition

The purpose of this module is to transform the confirmed region of text into a specific form of text that is machine-readable. Resizing is carried out for the region passed via the DCNN filter and then subjected to a pretrained OCR processing unit. In order to carry out transcription of textual contents, the regions of retained image are forwarded to a processing unit which executes recognition operation as $\Phi R^{H \times W \times 3} \rightarrow \Sigma^*$, where the notation Σ* represents a string associated with an alphabet Σ. Considering $x \in S$ be an input region, the recognised text t is represented as (6).

$$t = \Phi(\mathrm{x}) \in \Sigma^* \quad (6)$$

In (6) infers that when t≠ ∅ than the identified region is confirmed to possess textual content. The proposed system uses Tesseract OCT for implementing Φ. E-TEXTLOC also implements non-maximum suppression (NMS). It is done to control the possible occurrences of redundant bounding boxes. The computer vision metric (CVM) used for this purpose is as in (7).

$$CVM(B_i, B_j) = \frac{|B_i \cap B_j|}{|B_i \cup B_j|} \quad (7)$$

In the (7), the computation of CVM is performed by implementing intersection over union, which computes the degree of overlapping of predicted and ground truth scores. The bounding box with CVM more than cut-off τ is subjected to suppression, while the bounding box with maximum area or confidence of OCR is retained. The value of cut-off τ can be retained as 0.2-0.6.

3. RESULT AND DISCUSSION

The assessment of the E-TEXTLOC is performed on a highly controlled simulation environment where the scripting is carried out in Python. The model is assessed with a dataset constructed using YouTube videos. The dataset consists of an .mp4 video file with 25 frames per second with 354 kbps as data rate. The video processing, extraction of features, and formation of a bounding box, is carried out using OpenCV, while deep learning modules are implemented using PyTorch, where MobileNetV2 is used for feature extraction and the DCCN classifier is used for training. OCR is implemented using Pytesseract for acquiring readable text, while mathematical operation is performed by NumPy. The implementation of the baseline machine learning model is performed using scikit learning. The hardware configurations are as follows: the system is implemented on a normal Intel i5 processor with 32 GB RAM and NVIDIA GEFORCE GTX installed on a Windows 11 machine. The proposed study is compared with existing approaches, as there are various reports stating the usage of machine learning and OCR-based methodologies, which are nearly equivalent to E-TEXTLOC [27], [28].

3.1. Accomplished result

Table 1 highlights the accomplished study outcomes using various libraries and software tools. The proposed system is compared with three baseline models of TinyCNN, HOG integrated with SVM, and HOG integrated with random forest. It is to be noted that TinyCNN is a slightly lower version of the adopted deep learning model, while the extraction of shape-oriented features is carried out by HOG integrated SVM model that can differentiate both textual and non-textual regions. Further, a similar feature is also used by HOG when integrated with random forest; however, the regions are classified via a consensus method participated in by multiple decision trees. The accomplished outcome is analyzed using accuracy, precision, recall, F1-score, and average inference time per frame.

Table 1. Accomplished benchmarked outcome

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Avg. inference time per frame (ms)
E-TEXTLOC	98.2	95.4	94.1	94.7	29.6
TinyCNN	92.3	87.6	84.5	86.0	52.3
HOG + SVM	89.7	81.4	77.8	79.5	38.5
HOG + random forest	88.1	82.1	73.2	77.4	40.1

The study outcome shows that E-TEXTLOC accomplishes 98.2% accuracy, which infers its appropriateness with respect to all decisions towards classification on all video frames. The precision and recall rate of the proposed model are found to be 95.4% and 94.1% respectively, exhibiting a significant reduction of false positives as well as false negatives. Further, the proposed system is noted to accomplish 94.7% of F1-score under different conditions of video for localizing the position of text. Finally, the inference time for E-TEXTLOC is noted to be just 29.6 ms, proving it as the fastest operation suitable for real-time applications.

3.2. Discussion

Based on the accomplished numerical outcome, the analysis has been extensively carried out to find that the range of the accuracy metric resides between 97-99% while the enhancement in the average rate of precision and recall is between 15 and 12% respectively. Similarly, the F1-score is improved to 13-18% in contrast to baseline models, while potential improvement in average inference time is noted to be 40-50%. There are various rationales to be attributed to this accomplished outcome (Figure 2). E-TEXTLOC uses pseudo-labelling as well as intelligent sampling, which ensures the selection of only potential frames either during training or during inference generation. These operations lead to a significant reduction in computational overhead as well as noise. E-TEXTLOC also introduces an efficient technique of region proposal, along with the implementation of feature extraction carried out by MobileNetV2, without sacrificing any richness in features. Apart from this, it is also noted that conventional models usually treat the

problems of localization in a separate way, in contrast to problems related to detection and recognition. E-TEXTLOC integrates a smart and intelligent verifier for assessing the outcome of OCR, followed by feeding it back to training. This operation results in consistent improvement along with superior robustness of the model. The deployment of the confidence scores of OCR either for accepting or rejecting the prediction potentially assists in addressing the dissemination of errors while inference is in the process of generation, as shown in Figure 2(a), leading to higher accuracy. The conventional mechanism adopted in either the adopted baseline model or any others are also proven not to possess inclusion of contextual learning as well and lacks any feedback process. The baseline models are not witness capable towards filtering out outliers effectively, in contrast to E-TEXTLOC, as they often consider each region or frame uniquely, resulting in an increased rate of errors as well as sub-optimal generalization performance when subjected to complex or noisy scenes of video.

Better precision and recall performance (Figure 2(b)) and their balanced operation noted in Figure 2(c) showcase that training carried out by the verifier facilitates E-TEXTLOC for controlling various types of challenging conditions in real-time, viz., motion blur, occlusion, and illumination. Based on the outcome accomplished in the study, it can be inferred that the modular design of E-TEXTLOC facilitates it to be deployed over any resource-constrained ecosystem exhibiting increasing scalability (Figure 2(d)). Due to minimal inference time, E-TEXTLOC ensures high suitability towards assistive reading devices, autonomous vehicles, and surveillance systems. There are various publicly available benchmark datasets, e.g., multi-lingual scene text (MLT), common context-text (COCO-Text), and International Conference on Document Analysis and Recognition (ICDAR), that are frequently adopted towards assessing text localization from scenes (static images). Such datasets are witnessed not to sufficiently extract the real-time constraints, frame-level redundancies, and temporal dynamics connected with the localization of video-based text. The E-TEXTLOC model is essentially crafted for continuous streams of video influenced by multiple attributes, e.g., latency, occlusion, motion blur, and frame sampling. The study uses a curated YouTube video dataset for carrying out the evaluation that represents a deployment event in the real world. Hence, assessment of E-TEXTLOC adopting different public benchmarks could be carried out in future lines of work for evaluating the generalization across different datasets.

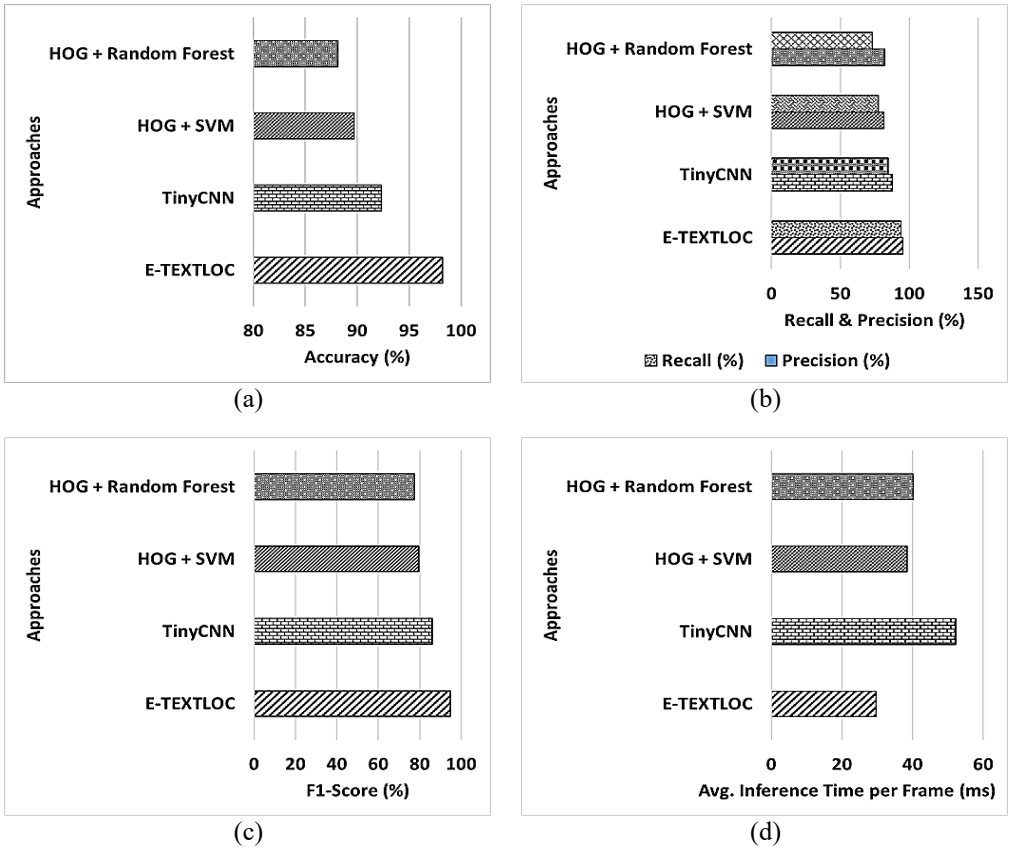

Figure 2. Accomplished study outcomes (a) accuracy, (b) precision and recall, (c) F1-score, and (d) inference time

3.3. Ablation study

The effectiveness of the proposed model is further assessed by adopting an ablation study, where the key motive is to study the contribution of individual components towards accomplishing the stated performance in the prior section. Table 2 highlights the numerical outcome of the ablation study that is obtained by selectively permitting and disabling particular components during the complete operation. In the first configuration of an evaluation, a study is carried out by eliminating the CRFM module, while the CNN verifier can be directly applied to a complete stream of video. The outcome of this configuration is noted with maximized event of outliers and a sudden increase in inference duration. This is mainly due to the voluminous search space representing the significance of CRFM towards the generation of candidate regions efficiently. In the second configuration of an evaluation, the study involves removing the DCCN verifier while the system directly forwards the CRFM-generated region to the OCR component. This configuration results in a notable degradation in precision as the model detects all non-text regions with contour as text regions incorrectly. Hence, this displays the importance of the verifier model towards controlling outliers. In the third configuration of an evaluation, the study model disables the feedback mechanism controlled by OCR while prior modules of DCNN and CRFM are retained. The outcome shows the comparable recall score, while the precision is gradually found to be declining in the presence of a complex scene of video scene. The analysis highlights the importance of refinement using feedback in retaining better consistency.

Table 2. Numerical outcomes of ablation study

Configuration	Inference time (ms)	Precision (%)	Accuracy (%)	OCR feedback	DCNN verifier	CRFM
Full E-TEXTLOC	29.6	95.4	98.2	✔	✔	✔
Without OCR feedback	30.4	88.6	95.1	✘	✔	✔
Without DCNN verifier	33.9	82.4	92.8	✔	✘	✔
Without CRFM	61.8	86.1	91.5	✔	✔	✘

4. CONCLUSION

The proposed study of text localization is proven to successfully address the limitations associated with baseline models by collaborative usage of confidence-based filtering, feedback using a verifier, and pseudo-labelling. The contribution towards advancement in E-TEXTLOC can be attributed to an effective sampling of video frames, minimizing the duration of inference for each frame, maximizing the richness of feature extraction and improving the robustness of text localization from complex video. The study outcome is in support of advocating E-TEXTLOC deployment in real-time usage as well as in an environment with reduced resource availability. Cumulatively, E-TEXTLOC contributes to a practical and innovative approach towards localization of text from video content. However, there are certain limitations associated too. A potential performance is exhibited by E-TEXTLOC on English video text, while the study finds that the generalization ability towards multilingual scripts could demand potential fine-tuning and extra training. It is also expected that the adoption of diversified scenarios with faster scene changes or extremely cluttered backgrounds could eventually challenge the localization accuracy.

FUNDING INFORMATION

The authors state no funding is involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Dayananda Kodala Jayaram	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Puttegowda Devegowda		✓				✓		✓	✓	✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The author declares that there are no known conflicts of interest associated with this publication. There are no financial or personal relationships that could inappropriately influence or bias the content of this work.

INFORMED CONSENT

Not applicable. This study did not involve human participants, human data, or any personally identifiable information. All data used were either publicly available, fully anonymized, or derived from non-human sources, and therefore no informed consent was required from individuals.

ETHICAL APPROVAL

Not applicable. This research did not involve human subjects, human biological materials, or experimental procedures on animals. The work was conducted solely on computational models, publicly available datasets, or non-sensitive data that did not require intervention with living organisms. Therefore, ethical approval from an institutional review board or animal ethics committee was not necessary for this study.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [DKJ], upon reasonable request.

REFERENCES

[1] S. Golla, B. Sujatha, and L. Sumalatha, "TIE- text information extraction from natural scene images using SVM," *Measurement: Sensors*, vol. 33, Jun. 2024, doi: 10.1016/j.measen.2023.101018.

[2] L. Kettle and Y.-C. Lee, "Augmented reality for vehicle-driver communication: a systematic review," *Safety*, vol. 8, no. 4, Dec. 2022, doi: 10.3390/safety8040084.

[3] M. Bonanno *et al.*, "Assistive technologies for individuals with a disability from a neurological condition: a narrative review on the multimodal integration," *Healthcare*, vol. 13, no. 13, Jul. 2025, doi: 10.3390/healthcare13131580.

[4] M. P. Kantipudi, S. Kumar, and A. K. Jha, "Scene text recognition based on bidirectional LSTM and deep neural network," *Computational Intelligence and Neuroscience*, vol. 2021, no. 1, Jan. 2021, doi: 10.1155/2021/2676780.

[5] J. Tang, L. Wang, J. Huang, A. Shi, and L. Xu, "Image semantic recognition and segmentation algorithm of colorimetric sensor array based on deep convolutional neural network," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–16, Sep. 2022, doi: 10.1155/2022/2439371.

[6] D.-L. Li, S.-K. Lee, and Y.-T. Liu, "Printed document layout analysis and optical character recognition system based on deep learning," *Scientific Reports*, vol. 15, no. 1, Jul. 2025, doi: 10.1038/s41598-025-07439-y.

[7] K. Mohsenzadegan, V. Tavakkoli, and K. Kyamakya, "A smart visual sensing concept involving deep learning for a robust optical character recognition under hard real-world conditions," *Sensors*, vol. 22, no. 16, Aug. 2022, doi: 10.3390/s22166025.

[8] V. Kadha, B. B. Duddeti, K. Srinadh, S. K. Buddepu, L. Janjanam, and K. Medhi, "From pixels to text: a deep learning survey of scene text detection and recognition," *Computers and Electrical Engineering*, vol. 135, Jul. 2026, doi: 10.1016/j.compeleceng.2026.111139.

[9] Z. Liu, R. Song, K. Li, and Y. Li, "From detection to understanding: a systematic survey of deep learning for scene text processing," *Applied Sciences*, vol. 15, no. 17, Aug. 2025, doi: 10.3390/app15179247.

[10] B. M. Albalawi, A. T. Jamal, L. A. A. Khuzayem, and O. A. Alsaedi, "An end-to-end scene text recognition for bilingual text," *Big Data and Cognitive Computing*, vol. 8, no. 9, Sep. 2024, doi: 10.3390/bdcc8090117.

[11] Q. Ding, E. Zhang, Z. Liu, X. Yao, and G. Pan, "Text-guided object detection accuracy enhancement method based on improved YOLO-world," *Electronics*, vol. 14, no. 1, Dec. 2024, doi: 10.3390/electronics14010133.

[12] M. Sinthuja, C. G. Padubidri, G. S. Jayachandra, M. C. Teja, and G. S. P. Kumar, "Extraction of text from images using deep learning," *Procedia Computer Science*, vol. 235, pp. 789–798, 2024, doi: 10.1016/j.procs.2024.04.075.

[13] Z. Raisi and J. Zelek, "Visual place recognition from end-to-end semantic scene text features," *Frontiers in Robotics and AI*, vol. 11, Sep. 2024, doi: 10.3389/frobt.2024.1424883.

[14] A. F. Rizky, N. Yudistira, and E. Santoso, "Text recognition on images using pre-trained CNN," Feb. 2023, *arXiv: 2302.05105.*

[15] R. Noviyanti and E. Sinduningrum, "Braille character recognition with histogram of oriented gradients (HOG) and SVM-based image processing," *Syntax Literate: Jurnal Ilmiah Indonesia*, vol. 9, no. 8, pp. 4411–4419, Aug. 2024, doi: 10.36418/syntax-literate.v9i8.16951.

[16] M. M. D. Savio, T. Deepa, A. Bonasu, and T. S. Anurag, "Image processing for face recognition using HAAR, HOG, and SVM algorithms," *Journal of Physics: Conference Series*, vol. 1964, no. 6, Jul. 2021, doi: 10.1088/1742-6596/1964/6/062023.

[17] S. Sharma, L. Raja, V. Bhatnagar, D. Sharma, S. N. Bhagirath, and R. C. Poonia, "Hybrid HOG-SVM encrypted face detection and recognition model," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 25, no. 1, pp. 205–218, Jan. 2022, doi: 10.1080/09720529.2021.2014141.

[18] B. Liu, F. Gao, P. Zhao, Y. Li, and W. Li, "Marker identification of disc workpiece based on HOG-SVM classifier," *Mobile Information Systems*, vol. 2022, pp. 1–9, Jul. 2022, doi: 10.1155/2022/6058602.

[19] L. M. Francis and N. Sreenath, "TEDLESS – text detection using least-square SVM from natural scene," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 3, pp. 287–299, Mar. 2020, doi: 10.1016/j.jksuci.2017.09.001.

[20] A. T. Maung, S. Salekin, and M. A. Haque, "A hybrid approach to Bangla handwritten OCR: combining YOLO and an advanced CNN," *Discover Artificial Intelligence*, vol. 5, no. 1, Jun. 2025, doi: 10.1007/s44163-025-00251-7.
[21] W. Azizah and S. Agustin, "Feature extraction using histogram of oriented gradients and moments with random forest classification for Batik pattern detection," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 14, no. 1, pp. 8–14, Jan. 2025, doi: 10.32736/sisfokom.v14i1.2225.
[22] C. Mukku and M. Santhosh, "Tri-stage offline Telugu character recognition system based on fusion of HOG and ULBP," *Measurement: Sensors*, vol. 32, Apr. 2024, doi: 10.1016/j.measen.2024.101059.
[23] J. Liu, Z. Liu, Q. Li, W. Kong, and X. Li, "Multi-domain controversial text detection based on a machine learning and deep learning stacked ensemble," *Mathematics*, vol. 13, no. 9, May 2025, doi: 10.3390/math13091529.
[24] E. Pintelas, A. Koursaris, I. E. Livieris, and V. Tampakas, "TextNeX: text network of experts for robust text classification—case study on machine-generated-text detection," *Mathematics*, vol. 13, no. 10, May 2025, doi: 10.3390/math13101555.
[25] H. Kohli, J. Agarwal, and M. Kumar, "An improved method for text detection using Adam optimization algorithm," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 230–234, Jun. 2022, doi: 10.1016/j.gltp.2022.03.028.
[26] N. Jalal, A. Mehmood, G. S. Choi, and I. Ashraf, "A novel improved random forest for text classification using feature ranking and optimal number of trees," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 2733–2742, Jun. 2022, doi: 10.1016/j.jksuci.2022.03.012.
[27] B. Avinash, I. A. Shaik, and R. Syed, "Real-time text detection and recognition based on optical character recognition (OCR)," *Journal of Science & Technology*, vol. 9, no. 4, pp. 1–10, 2024, doi: 10.46243/jst.2024.v9.i4.pp1-10.
[28] S. M. Patil, V. S. Malemath, S. Muddapur, and P. M. Dhulavvagol, "Enhanced text detection in natural scenes using advanced machine learning techniques," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 22114–22118, Apr. 2025, doi: 10.48084/etasr.10029.

BIOGRAPHIES OF AUTHORS

Dayananda Kodala Jayaram is working as assistant professor in the Department of Computer Science and Engineering, Vidyavardhaka College of Engineering (Autonomous Institute) affiliated to Visvesvaraya Technological University India. He has received his M.Tech. and B.E. from Visvesvaraya Technological University, Belagavi, Karnataka, India. He is pursuing his Ph.D. from Visvesvaraya Technological University, Belagavi, Karnataka, India. His teaching and research interests are in the field of data mining, machine learning, and image processing. He has around total teaching experience of 11 years. He can be contacted at email: dayananda.kem@gmail.com.

Dr. Puttegowda Devegowda is working as professor and head in the Department of Computer Science and Engineering, ATME College of Engineering affiliated to Visvesvaraya Technological University, India. He has received his Ph.D. from Mysuru University, Mysuru, Karnataka, India. His teaching and research interests are in the field of data mining, video processing, machine learning, and image processing. He has around total teaching experience of 20 years. He can be contacted at email: puttegowda.77@gmail.com.