Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2022.018030
Article

Tech Science Press

ResNet CNN with LSTM Based Tamil Text Detection from Video Frames

I. Muthumani[1,*], N. Malmurugan[2] and L. Ganesan[3]

[1]Department of Electronics & Communication Engineering, Government College of Engineering, Thanjavur, 613402, India
[2]Department of Electronics & Communication Engineering, Mahendra College of Engineering, Salem, 636106, India
[3]Department of Computer Science & Engineering, AC Govt. College of Engineering & Technology, Karaikudi, 630004, India
*Corresponding Author: I. Muthumani. Email: muthugce14@gmail.com
Received: 22 February 2021; Accepted: 03 June 2021

Abstract: Text content in videos includes applications such as library video retrievals, live-streaming advertisements, opinion mining, and video synthesis. The key components of such systems include video text detection and acknowledgments. This paper provides a framework to detect and accept text video frames, aiming specifically at the cursive script of Tamil text. The model consists of a text detector, script identifier, and text recognizer. The identification in video frames of textual regions is performed using deep neural networks as object detectors. Textual script content is associated with convolutional neural networks (CNNs) and recognized by combining ResNet CNNs with long short-term memory (LSTM) networks. A residual mapping underlies the network. In ResNet, skipping condenses the network into few layers to help it learn more quickly. Bidirectional LSTM enables networks to always have information about the sequence, backward as well as forward. For fast pattern learning processes, this combination uses the residual learning process. In comparison to the existing approaches, the proposed approach's results show improved accuracy, precision and F-measures. The combination of ResNet CNNs and bidirectional LSTMs has high recognition rates for detecting video texts in Tamil cursive script.

Keywords: Cursive text detection; deep neural networks (DNNs); ResNet convolutional neural networks (ResNet CNNs); long short-term memory (LSTM) networks

1 Introduction

The amount of content from video archives and live streams has significantly increased in recent years. According to the most recent statistics [1], 300 hours of video are posted every minute on YouTube. This huge rate is due to the availability of cost-effective smartphones that come with cameras. Efficient and effective recovery methods are needed to allow users to obtain their desired content. Typically, videos with identification tags are stored, as they are easier to recognize later. Tags are merely annotations or keywords assigned by users. A keyword is matched with these tags in the query phase to find the corresponding information. Tags normally cannot contain rich video content, which results in limited restoration capability. Therefore, to search inside the content, i.e., video recovery from content (CBVR),

is easier and more accurate than matching the tags. CBVR systems have long been studied and developed, and they can find the desired material more intelligently. The contents may be visual (video or people), audio (spoken words), or textual (news lines, VJ names, scorecards). This paper focuses on text, using an intelligent recovery system. Video text can be classified as text or as a description of a scene (closed captioning). During video recording, scene text such as in Fig. 1 is recorded by the camera and is typically used for applications like visually impaired assistance systems and robot navigation, while the artificial texts in Fig. 2 are superimposed or collated as a second layer on the video.

Figure 1: Examples of scene text

Figure 2: Example of caption text

The caption text is usually connected to the video content, and hence it is typically used for semantic retrieval of videos. This study is focused on the detection and recognition of the caption text. Fig. 2 demonstrates this clearly. The caption texts are within the red circle, and the scene texts are not encircled; thus, the caption texts are distinguished from the scene text appearing inside the video.

The main blocks for the index and retrieval of text content systems include text area recognition, context removal of text, script detection, and text recognition (e.g., optical character recognition, OCR). Text recognition can be achieved by unattended, monitored, or hybrid methods. The unmonitored approach utilizes image processing to distinguish text from non-text areas. In a hybrid approach, unattended methods authenticate the text regions recognized by a monitored method. If text is detected, then a recognition system can be provided. If the videos have text in multiple scripts, another module is required to recognize different scripts through OCR. Although ABBYY Fine Reader and Tesseract are reconnaissance systems proven to identify Romanic scripts, cursive texts are difficult to recognize [2].

Furthermore, while documents are usually scanned in high-resolution, video text resolution is generally low (see Fig. 3). In addition, text in video is presented against complex backgrounds, which challenges recognition.

This paper explores a broad framework in multi-script history for the detection and recognition of text in video. The main elements of the work are as follows:

- Text script identification in video frames using a convolutional neural network (CNN)
- Video Tamil text recognition by integrating ResNet CNN and LSTM
- Suggested validation by broad data collection
- Validation

Figure 3: Examples of low-resolution text in video frames

The remainder of the paper is structured as follows. Section 2 provides background on text identification methods, identity, and script recognition. Section 3 presents different network architectures. Section 4 explains the proposed mechanism, and Section 5 discusses our findings. Section 6 relates our conclusions.

2 Background

The detection and recognition of texts in videos, images, documents, and natural scenes have been researched for more than four decades. We summarize the important contributions in their detection in images and video frames.

2.1 Text Detection

Text content translation is referred to as image text detection. Its techniques are usually divided into unattended and monitored approaches in the literature. Unattended techniques differentiate text from background and use learning algorithms to distinguish text and non-text regions following training. Monitored methods use algorithms trained in text and non-text portions after pixel values and related features have been extracted. Over the years, classification systems such as naïve Bayes [3], Support Vector Machine (SVM) [4], Artificial Neural Network (ANN) [5], and deep learning networks (DLNs) [6] have been applied. DLNs have become known for their outstanding success relative to conventional methods in many classification problems. The contribution of Krizhevsky et al. [7] to the ImageNet Large Scale Visual Recognition Competition (ILSVRC) [8] was to build and apply CNNs with decreasing error rates. CNNs are considered one of the most advanced methods for the extraction and classification of features [9,10], and are used in many recognition applications. Region-based convolutional networks [11], such as Fast R-CNN [12] and Faster R-CNN [13], are used as object detectors, and traditional CNNs are

typically used for object classification. New architectures such as You Only Look Once (YOLO) [14] and the single shot detector (SSD) [15] have also been suggested for real-time object detection. In the development of this application, the availability of a benchmark database in English has provided momentum. However, the lack of datasets in Tamil remains a challenge of generic text detection.

2.2 Script Recognition

Jieru et al. [16] took a CNN and recurrent neural network (RNN) as a single end-to-end network for learning and acknowledgment in script detection, with high recognition rates on many datasets of different scripts. The identification of scripts with much less labeled data was examined for a collection of mid-level characteristics, where the identification of CVSI datasets exceeding 96% was recorded [17]. Gomez et al. [18] used anaïve Bayes classifier with convolution features in unconstrained video scene text to identify scripts. Patch classification using CNNs was extended [19]. AlexNet and VGG-16 have been considered for script recognition in several publications, along with transference learning and refining [20]. A bag of visual words model was explored [21] by learning convolutional characteristics from image patches in triplet shape. A CNN-LSTM model was investigated by Bhunia et al. [22] to derive local and global characteristics from four public datasets. While good detection technology has been introduced, the low resolution in video pictures and the detection of more similar scripts make this work difficult.

2.3 Text Recognition

Text recognition is a standard technique in pattern recognition, and it has been studied for decades. Recognition systems for printed and scanned handwritten documents, picture text, and subtitle text inserted in images has been explored. ABBYY Fine Reader records nearly 100% recognition rates on texts for various scripts in modern recognition systems, such as Google's Tesseract [23,24]. Recognition of cursive scripts is still difficult, especially in text videos with little resolution. In video scenes, various variables, as opposed to document images, play a major role in the identification of text. During video capture, there are various camera locations, uniform lights, and more complex and dynamic backgrounds. Few common text descriptors have been studied for the detection of texts in natural scenes in invariant feature transform (SIFT) and oriented gradient histogram [25–27]. Recent advances in text recognition include the combination of CNNs and RNNs, aiming to extract efficient features from text images, where the RNN provides transcription in the feature sequences. In this relation, the CNN component extracts effective features from the text image. Apart from the traditional C-RNN combination, improvements in scene text recognition have been suggested in the architecture [28]. From the perspective of acknowledging text, lowtext resolution is the key problem. Many papers address this as a preprocessing stage. A high-resolution image can be generated and supplied to a detector by integrating the information from multiple frames [29]. A few authors have tried to detect Tamil text and extract videos *via* ANN [30], convolutional sequences [31], and angular gradient [32].

3 Network Architectures

3.1 ResNet CNN

The ImageNet challenge sees increasing numbers of layers in DNNs in order to reduce the error rate. These introduce the issue of gradient disappearance in deep learning, and the rates of training and test errors increase. A new architecture using skip connections was developed to circumvent the issue of the vanishing gradient. The newly added block is called a skip residual block. Training is skipped from a few layers and directly linked to the output. A residual mapping is the underlying mapping of the network. In ResNet, skipping condenses the network to a few layers, which can enable quick learning. All layers are

extended during training, and the remaining parts of the network explore more and more of the feature space of the source picture. ResNet CNN is shown in Fig. 4 as a compressed diagram.

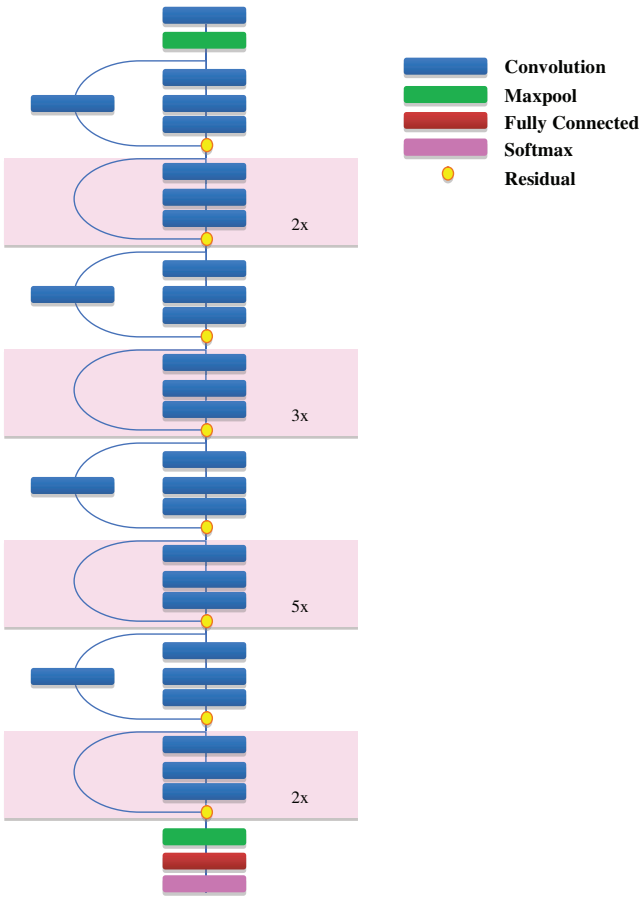

Figure 4: Architecture of standard ResNet-CNN

Huang et al. [33] suggested an approach of randomly falling layers and testing of all layers within the network. The remaining block is regarded as the network building block. Therefore, the residual block entry flows from the identity shortcut to the weight layers when the residual block is enabled during workout or if the input only flows through the identity shortcut. Each layer has a chance of survival, and one is randomly dropped. All blocks are enabled and adjusted according to their probabilities of survival during the test period in training.

The output of the n^{th} residual block during training is given by

$$H_n = \mathrm{ReLU}(b_n * w_n(H_{n-1}) + \mathrm{id}(H_{n-1})), \tag{1}$$

where w_n is the weighted mapping of the n^{th} block, and b_n is a Bernoulli random variable that takes a value of 1 or 0 to indicate the active state of the block. If $b_n = 1$, the block turns into a normal residual block; if $b_n = 0$, Eq. (1) becomes

$$H_n = \mathrm{ReLU}(\mathrm{id}(H_{n-1})). \tag{2}$$

The above equation is the output of a ReLU identity layer. Since H_{n-1} is nonnegative, the above equation becomes an identity layer which passes only the input through to the next layer,

$$H_n = id(H_{n-1}). \tag{3}$$

Let p_n be the survival probability of layer n. Then Eq. (1) becomes

$$H_n = ReLU(p_n * w_n(H_{n-1}) + id(H_{n-1})). \tag{4}$$

A linear decay rule is applied to the survival probability of each layer. Since low-level features extracted by former layers will be used by later layers, the former layers should not be dropped. The rule used thus becomes

$$p_n = 1 - \frac{n}{N}(1 - p_N), \tag{5}$$

where N is the total number of blocks and p_N is the survival probability of the last residual block, which is presumed as 0.5 for better performance. Veit et al. [34] suggested dropping some layers of a trained ResNet, and performance was comparable. This makes the ResNet architecture even more attractive. ResNet blocks are two layers deep in small networks and three layers deep in large networks. The block shown in Fig. 4 is three layers deep.

3.2 Bidirectional Long Short-term Memory (Bi-LSTM)

Bidirectional RNNs combine two RNNs, enabling networks to always have information about the sequence, backward as well as forward. Inputs are moved in two directions, from past to future and future to past.

Compared to unidirectional LSTM, in Bi-LSTM future knowledge is preserved and information from the past and future is maintained at any step, using the two hidden states combined. Thus Bi-LSTMs are effective in better recognizing the meaning of the text within the scene than unidirectional LSTMs. Fig. 5 shows the Bi-LSTM internal block diagram. Both activations (forward, backward) are taken to calculate the output ŷ of Bi-LSTMs at a given time t:

$$\hat{y}(t) = g\left(W_y[\overleftarrow{a}(t), \overrightarrow{a}(t)] + b_y\right).$$

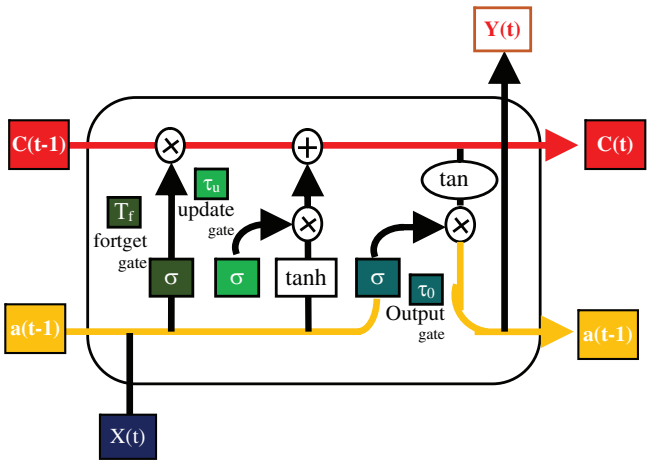

Figure 5: Architecture of Bi-LSTM

4 Proposed Method

We present the details of the proposed framework, as summarized in Fig. 6. The proposed model consists of a text detector, script identifier, and text recognizer.

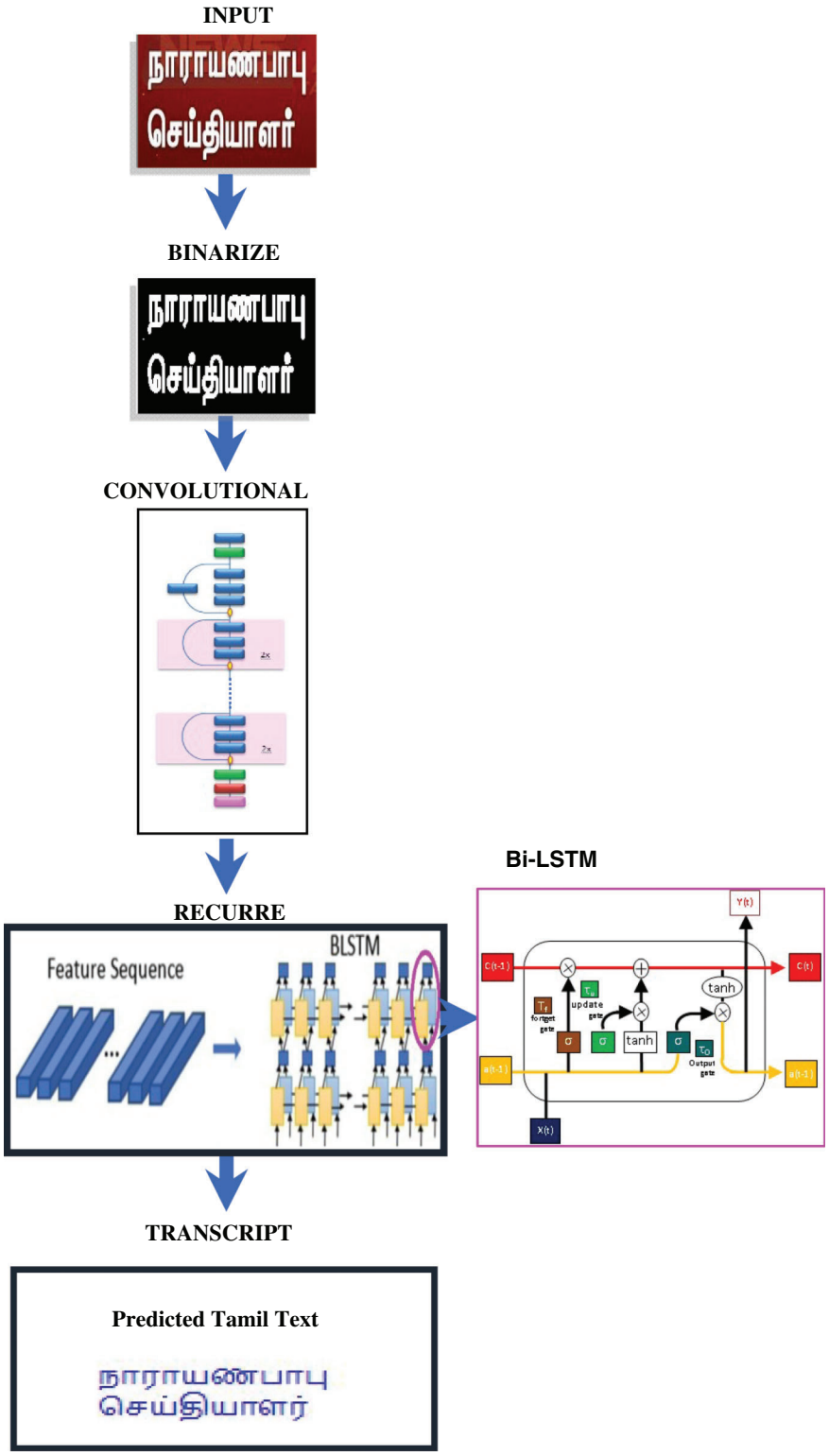

Figure 6: Proposed method of Tamil text recognition of video scenes

The text detector defines and confines the text content in a video frame. The script recognition module must be filled in with the identified texted regions, because the text can be entered in more than one script within the same frame from the previous input stage. The text lines are divided into various scripts. The work includes English and Tamil text. Each script in each text is identifiable in its respective recognition module. The proposed model has an entry size of 30 to 300 years and is based on a model ResNet-50 that discharges entirely linked layers. This model receives an input video and locates the text areas. Once the text region is located, the script identification module is applied for script recognition.

The model divides the text lines into various scripts, either Tamil or English, after practicing. A detected text line is transmitted for final recognition to the appropriate recognition module. Google's off-the-shelf OCR engine is used for English text recognition. However, this engine does not perform well for cursive text in Tamil, but these text lines are recognized by the LSTM RNN-based recognition engine. Preprocessing consists of height standardization and image binarization before feature extraction. A seven-layer CNN is supplied with the preprocessed image. For text line image input, the CNN assigns a function map. A sequence of the recurrent layers of the device is provided with the function map. Two layers of two-way LSTMs are used to create the RNN. After completing the pre-frame projections, the outputs of LSTM are transferred to the search table. The transcription of the output is provided by the search table, as shown in Fig. 6.

5 Results and Discussion

The first step is to collect videos and 60 videos from five news channels by capturing live streams. Both videos are captured at 900 to 600 resolution and 25 fps. Although the videos contain mainly Tamil text content, a few are in Tamil as well as English.

A frame per two seconds is extracted for labelling in order to prevent redundant details in successive frames of the film. In particular, different textual contents in collected frames are guaranteed. In our proposed work, we use 550 frames with over 10,000 text lines. Efficiency with precision, precise, recall and F-measure is calculated quantitatively. Since Tamil text detection with deep learning networks does not contain any work in the Tamil textual literature, CNN and RCNN networks with or without LSTM are built for comparison with the proposed ResNet CNN with the LSTM network model. Tabs. 1 and 2 summarises the results of both of these methods for identification and acknowledgement of the Tamil wording. We must note that we test various methods on our own data set because no standardised data set for Tamil video scenes is available. Fig. 7 displays Tamil text recognition from various video algorithms in the same video scenes, LSTM LSN CNN, LSTM R-CNN and LSTM ResNet proposed CNN. The text "நிவர்புயல் எச்சரிக்கை காரணமாக 3 துறைமுகங்களில் மூன்றாம்எண் புயல் எச்சரிக்கை கூண்டு ஏற்றம்" is correctly recognised by the proposed algorithm, although certain characters are missing or misrecognized by two other algorithms. Fig. 8 shows the Accuracy and Precision and Fig. 9 shows the Recall and F-score of the of the proposed system, all the measures were with existing algorithms.

Table 1: Comparison of accuracy and precision from different algorithms

Accuracy							Precision					
No. of samples	CNN	R-CNN	ResNet CNN	CNN + LSTM	R-CNN + LSTM	ResNet CNN+ LSTM	CNN	R-CNN	ResNet CNN	CNN + LSTM	R-CNN + LSTM	ResNet CNN + LSTM
20	50.3	52.4	55.6	58.1	61.1	64.1	49	52	56	58	62	65
40	55.7	58.3	62.3	63.8	66.6	70.1	53	57	61	62	67	71
60	61.6	64.1	68.7	69.3	73.3	77.6	58	63	67	67	72	76
80	66.2	69.6	73.1	74.4	77.7	81.9	63	68	72	72	76	81
100	71.1	74.2	78.4	81.1	84.1	88.2	68	72	77	79	84	88
120	75.3	79.0	83.3	85.2	89.2	93.2	73	78	83	83	88	93
140	79.6	82.4	86.0	89.2	92.5	95.8	78	83	87	87	91	95

Table 2: Comparison of recall and F score from different algorithms

Recall							F Score					
No. of samples	CNN	R-CNN	ResNet CNN	CNN + LSTM	R-CNN + LSTM	ResNet CNN+ LSTM	CNN	R-CNN	ResNet CNN	CNN + LSTM	R-CNN + LSTM	ResNet CNN + LSTM
20	53	58	62	65	70	74	50	54	59	61	65	69
40	58	63	67	71	76	80	54	59	64	66	71	76
60	63	69	73	75	81	84	60	64	69	73	78	84
80	68	73	77	79	84	88	65	70	75	76	80	85
100	73	79	84	84	89	93	71	76	81	82	87	92
120	79	85	89	89	92	94	76	81	86	87	92	96
140	85	90	93	95	97	99	81	86	90	92	95	98

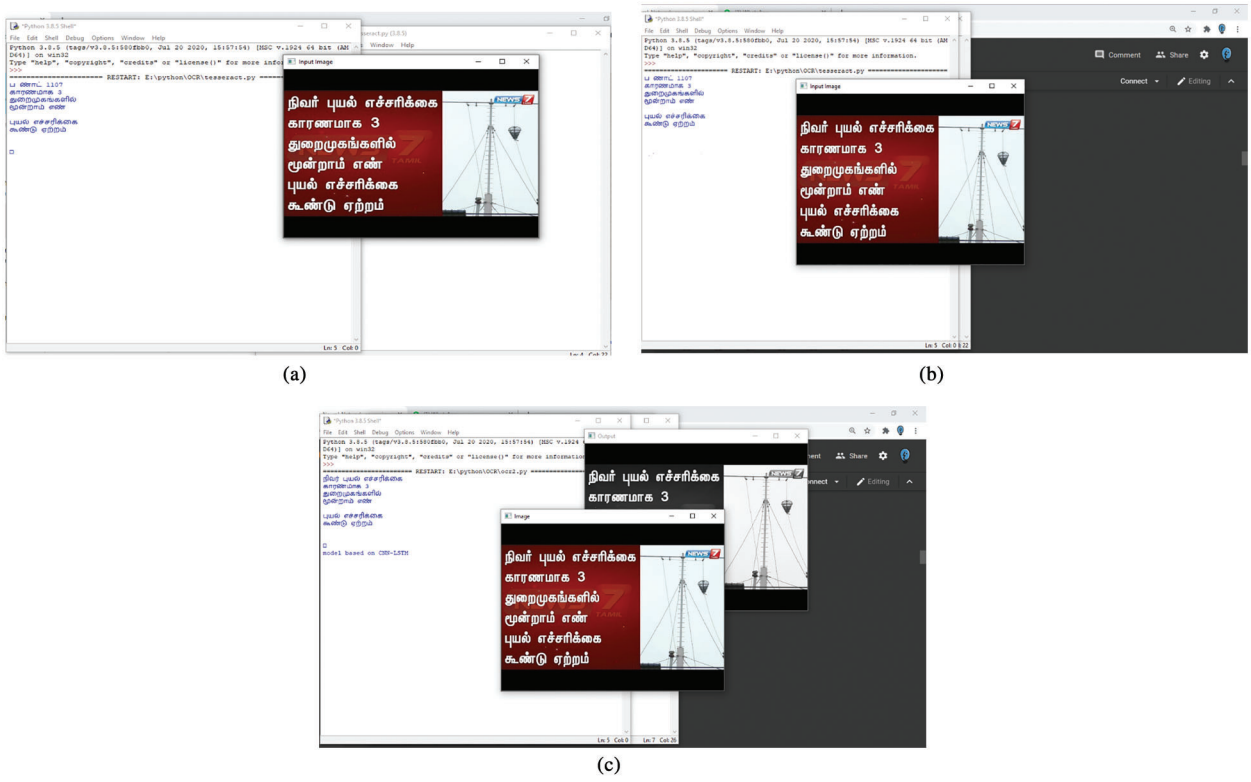

Figure 7: Results of recognition of Tamil text from video scenes from (a) CNN with LSTM (b) R-CNN with LSTM and (c) ResNet CNN with LSTM

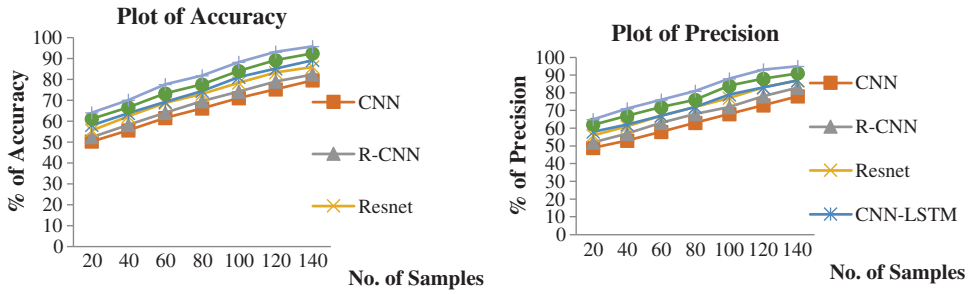

Figure 8: Plots of % of accuracy and precision *vs*. number of samples from different algorithms

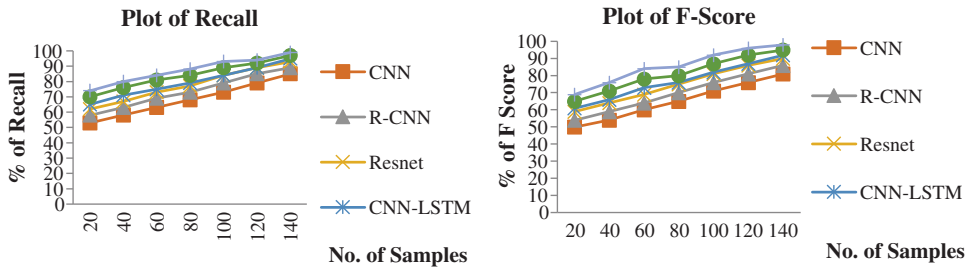

Figure 9: Plots of % of recall and F score *vs*. number of samples from different algorithms

6 Conclusion

A detailed structure was provided to identify and recognize text in video frames in English and Tamil. State-of-the-art object detectors were built and implemented. A ResNet CNN model was built with LSTM to recognize Tamil script in detected textual regions, and a ResNet CNN was implemented to distinguish areas of the text. The combination of ResNet CNNs and bidirectional LSTMs, with high recognition rates for demanding video texts in Tamil cursive script, is a major contribution of this work. The precision, recall, accuracy, and F-score of the proposed algorithm are improved from 5% to 7%, compared to the existing approaches. However, the changes are 10%–12% and 8%–10% higher relative to LSTM's CNN and LSTM's RCNN.

Acknowledgement: We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. Burgess and J. Green, *YouTube: Online Video and Participatory Culture*, 2nd ed., Cambridge: Wiley, pp. 13–14, 2018.

[2] R. Smith, "An overview of the tesseractocr engine," in *Proc. Ninth IEEE Int. Conf. on Document Analysis and Recognition (ICDAR 2007)*, Parana, Brazil, pp. 629–633, 2007.

[3] 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.

[4] W. Zhen and W. Zagiqiang, "A comparative study of feature selection for SVM in video text detection," in *Proc. 2nd IEEE Int. Sym. on Computational Intelligence and Design*, Changsha, China, pp. 552–556, 2009.

[5] X. C. Yin, X. Yin and K. Huang, "Robust text detection in natural scene images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 970–983, 2013.

[6] Y. L. Cun, Y. Bengio and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[7] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. on Neural Information Processing Systems - Volume 1, NIPS'12*, Nevada, USA, pp. 1097–1105, 2012.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh *et al.,* "ImageNetlarge scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. on Learning Representations, ICLR 2015*, San Diego, USA, pp. 40–53, 2015.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed *et al.,* "Goingdeeper with convolutions," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, CVPR 2015*, Boston, MA, USA, pp. 1–9, 2015.

[11] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Columbus, USA, pp. 580–587, 2014.

[12] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. on Computer Vision*, Las Condes, Chile, pp. 1440–1448, 2015.

[13] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Advances in Neural Information Processing Systems, Neural Information Processing Systems Foundation, Inc. (NIPS2015)*. Quebec, Canada, pp. 91–99, 2015.

[14] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, Clark, pp. 779–788, 2016.

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed *et al.,* "SSD: Single shot multibox detector," in *Proc. European Conf. on Computer Vision*, Amsterdam, Netherlands, Springer, pp. 21–37, 2016.

[16] J. Mei, L. Dai, B. Shi and X. Bai, "Scene text script identification with convolutional recurrent neural networks," in *Proc. 23rd IEEE Int. Conf. on Pattern Recognition (ICPR 2016)*, Cancun, Mexico, pp. 4053–4058, 2016.

[17] A. K. Singh, A. Mishra, P. Dabral and C. Jawahar, "A simple and effective solution for script identification in the wild," in *Proc. 12th IAPR Workshop on Document Analysis Systems, (DAS 2016)*, Santorini, Greece, pp. 428–433, 2016.

[18] L. Gomez and D. Karatzas, "A fine-grained approach to scene text script identification," in *Proc. 12th IAPR Workshop on Document Analysis Systems, (DAS 2016)*, Santorini, Greece, pp. 192–197, 2016.

[19] L. Gomez, A. Nicolaou and D. Karatzas, "Improving patch-based scene text script identification with ensembles of conjoined networks," *Pattern Recognition*, vol. 67, no. 12, pp. 85–96, 2017.

[20] M. Tounsi, I. Moalla, F. Lebourgeois and A. M. Alimi, "CNN based transfer learning for scene script identification," in *Proc. Int. Conf. on Neural Information Processing*, California, USA, Springer, pp. 702–711, 2017.

[21] J. Zdenek and H. Nakayama, "Bag of local convolutional triplets for script identification in scene text," in *Proc. 14th IAPR Int. Conf. on Document Analysis and Recognition (ICDAR)*, vol. 1, IEEE, Kyoto, Japan, pp. 369–375, 2017.

[22] K. Bhunia, A. Konwer, A. K. Bhunia, A. Bhowmick, P. P. Roy *et al.,* "Script identification in natural scene image and video frames using an attention based convolutional-LSTM network," *PatternRecognition*, vol. 85, pp. 172–184, 2019.

[23] C. Yi and Y. Tian, "Scene text recognition in mobile applications by character descriptor and structure configuration," *IEEE Transactions on Image Processing*, vol. 23, no. 7, pp. 2972–2982, 2014.

[24] T. Jayasankar, J. Arputha Vijayaselvi, A. Balaji Ganesh and D. Kumar, "Word and syllable based concatenative model of text to speech synthesis of Tamil language," *International Journal of Applied Engineering Research*, vol. 9, no. 24, pp. 23955–23966, 2014.

[25] S. Tian, U. Bhattacharya, S. Lu, B. Su, Q. Wang *et al.,* "Multilingual scene character recognition with co-occurrence of histogram of oriented gradients," *Pattern Recognition*, vol. 51, pp. 125–134, 2016.

[26] 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.

[27] Z. Lei, S. Zhao, H. Song and J. Shen, "Scene text recognition using residual convolutional recurrent neural network," *Machine Vision and Applications*, vol. 29, no. 5, pp. 1–11, 2018.

[28] Y. Gao, Z. Huang and Y. Dai, "Double supervised network with attention mechanism for scene text recognition," in *Proc. IEEE Visual Communications and Image Processing (VCIP)*. Sydney, Australia, pp. 1–4, 2019.

[29] T. Q. Phan, P. Shivakumara, T. Lu and C. L. Tan, "Recognition of video text through temporal integration," in *Proc. 12th Int. Conf. on Document Analysis and Recognition (ICDAR)*, Washington, USA, IEEE, pp. 589–593, 2013.

[30] A. Thilagavathy, K. Aarthi and A. Chilambuchelvan, "Text detection and extraction from videos using ANN based network," *Proc. Int. Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI)*, vol. 1, no. 1, pp. 19–28, 2012.

[31] P. Shivakumara, N. Sharma, U. Pal, M. Blumenstein and C. L. Tan, "Gradient-angular-features for word-wise video script identification," in *Proc. 22nd Int. Conf. on Pattern Recognition (ICPR '14)*, Stockholm, Sweden, pp. 3098–3103, 2014.

[32] S. Nag, P. K. Ganguly, S. Roy, S. Jha, K. Bose *et al.,* "Offline extraction of Indic regional language from natural scene image using text segmentation and deep convolutional sequence," in *Methodologies and Application Issues of Contemporary Computing Framework*, 1st ed., Springer, pp. 49–68, 2018.

[33] G. Huang, Y. Sun, Z. Liu, D. Sedra and K. Q. Weinberger, "Deep networks with stochastic depth," in *Proc. European Conf. on Computer Vision (ECCV 2016)*, Amsterdam, The Netherlands, pp. 646–661, 2016.

[34] A. Veit, M. Wilber and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," in *Proc. 13th Annual Conf. on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain, pp. 550–558, 2016.