DOI: 10.15849/IJASCA.240730.08

Int. J. Advance Soft Compu. Appl, Vol. 16, No. 2, July 2024
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Deep Learning Based Text Detection Model for Lecture Video Analysis: Impact of Covid 19 in Education Sector

Geetabai S Hukkeri[1], R H Goudar[2], Gururaj H L[3], Shilpa Ankalaki[4*],

[1]Department of Computer Science and Engineering, Manipal Institute of Technology Bengaluru, Manipal, India.
e-mail: geetabai.hukkeri@manipal.edu

[2]Department of Computer Science and Engineering, Visvesvaraya Technological University, Belagavi, India.
e-mail: rhgoudar.vtu@gmail.com

[3]Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India, Department of Information Technology, gururaj.hl@manipal.edu

[4]Department of Computer Science and Engineering, Manipal Institute of Technology Bengaluru, Manipal, India.
e-mail:shilpa.ankalaki@manipal.edu

*Corresponding author- Shilpa Ankalaki, e-mail:shilpa.ankalaki@manipal.edu

Abstract

The global education system is impacted by the corona virus. Therefore, online learning is a way to keep the educational system going. Learners who are suffering from health issues cannot sit for long time to watch lecture videos. It is not possible to navigate directly to the precise needed point of topic in the video when students are just interested in seeing the desired chunk of topics from the lengthy lecture video. Therefore, we are presenting a study on lecture video indexing to quickly and non-linearly access the topic of interest in a long lecture video, which can help learners who cannot go to the offline classes and who are suffering from health issues. Key frame representation is an effective method for getting accurate video content. Thus, a faster R-CNN-based improved algorithm is proposed to detect text from the key-frames for index points generation. The experimental findings demonstrate that the accuracy of keyframe extraction of proposed method is 90%, and the optimized Faster R-CNN algorithm paradigm significantly increases the detection accuracy to 93.4%, which is the best compared to other algorithms and minimizes the skipped detection performance.

Keywords: *Online learning, Video Lectures, Keyframes, Video Content analysis, Faster R-CNN.*

Received 13 February 2024; Accepted 21 June 2024

1. Introduction

The global education system is impacted by the corona virus. Higher education institutions, universities, and schools are closed to prevent the corona virus from spreading. Parents, instructors, and kids all experience challenges when schools are closed. Therefore, online learning is a way to keep the educational system going [17]. As a result, lecturers in remote learning would record their lectures and send them to the students. It is not possible to navigate directly to the precise needed point of topic in the video when students are just interested in seeing the desired chunk of topics from the lengthy lecture video. This forces the viewers to watch the full long video in order to get to the short topic. Therefore, we are presenting a study on lecture video indexing to quickly and non-linearly access the topic of interest in a long lecture video, which can help learners who cannot go to the offline classes and who are suffering from health issues. The educational system will incorporate video technologies such as video lectures, e-classrooms, virtual classrooms, etc. in this multi-media era. We need a lot of storage space and extra access time to manage the material for the video lectures. Such information might not be retrieved instantly. The direction of improvement for the future educational system is an intelligent education system within a big data environment. Intelligent learning includes intelligent video analysis as a crucial component. Universities and colleges have benefited from the digitization of lecture materials by recording them as videos in order to enhance their teaching abilities. Due to their accessibility on online platforms, learners like learning materials in video format. The strength of the learning resources has thereby increased as a result of the lecture videos. Video combines text, image, and sound. Therefore, using lecture videos as study material gives students the opportunity to engage in live learning. Universities can also publish study materials on their portals so that students can easily access them. In recent years, most colleges have adopted the practise of using lecture recordings as study material. The number of video files has greatly expanded since the advent of contemporary technology. Video retrieval from the vast video database has become a difficult issue as a result of the massive storage requirements and high intricacy of these video files. The method of locating the necessary video footage from a sizable database is known as video retrieval. The length and unstructured nature of the university video recordings from around the world is evident. The browsing, access, and location of online video material have become more challenging as a result.

The current method focuses on keyframe extraction from videos in a huge collection to make indexing, retrieving, and browsing simple. In order to give a practical way to rapidly and thoroughly comprehend relevant data, key-frame extraction is a crucial component of the static video summarising methodology. It characterises the main video content using indicative frames extracted from the video. The general belief is that the objective is to effectively and efficiently extract video summaries. Three types of key frame extraction techniques can be distinguished: content-based, cluster-based, and shot-based [18]. Keyframe collection has been a regularly employed method for video indexing for years. A video's contents can be presented to the user in a simplified manner with the help of effective keyframe selections. When there is a limited amount of bandwidth available and the user wants to get pertinent films as quickly as possible, this type of video summary is quite helpful.

Numerous applications, including mobile computer vision systems and transformation, can make use of text in images. Text detection and recognition research has been carried out for years in order to take advantage of this text data. Convolutional neural network (CNN) deployment has significantly increased the accuracy of both word and specific

character recognition[21]. It is challenging to distinguish between different types of text in photos due to the different colours, fonts, and sizes of the text.

OCR is working well with some applications, but text extraction (detection and recognition) still has a lot of obstacles to overcome, like the pre requisites in various settings. The detection of text from key-frames of the lecture video is the main focus of this work. The fact that an image typically has a high resolution and several textual items but just one textual entity comprises a very small space presents the biggest obstacle to applying a text prediction model to an image. When interpreting such a large image, it takes longer to train and train the hypotheses and more memory to keep the model's variables. Resizing the huge image to a tiny scale is a frequent [11] [22] solution to this issue.

This study examines how to choose keyframes from videos for quick indexing. The detection of text from images is discussed using a deep learning method. This method divides the text recognition process into two phases: training and detection. The following is a summary of the contributions made by this work:

- To identify the key frames of the lecture video with the least amount of storage and access time possible, a deep learning method for text detection from key-frames is provided.
- The key-frame extraction method is highlighted in this work to summarize and analyze the video lectures to help online learners. A deep learning-based text detection model that integrates Faster R-CNN is used to increase the efficiency and accuracy of the video content extraction system. In adaptive lecture video analysis, a faster R-CNN text detection method is applied.
- Prior to changing the existing NMS (Non maximum Suppression) process with Soft-NMS to cope with the coincidence of text, we first obtained the key-frames to input into the proposed model. After that, we introduced a hole convolution to the system to remove unwanted features in the high-resolution video frame, making it perfect for key-frame text detection.
- We compared the results of key-frame extraction to the existing approach.

The remainder of the text is structured as follows. A discussion of the literature review is provided in Section 2. It covers the study area, a description of the methodology and the explanation for their choice. Section 3 is of two parts, one is key-frame extraction and other is deep-learning based text detection. Section 4 provides a thorough analysis of the findings, their interpretation, and how the model's performance compared to earlier studies. Conclusion is in Section 5.

2. Materials and Methods

In general, lecture videos contain a variety of information, such as talking heads, slides, writing on whiteboards, illustrations, etc. There are numerous transitions among audiovisual views during the creation of a lecture video, such as going from a talking head perspective to a slide or from a slide to an illustrated example. A machine learning technique for slide identification and a slide transition prediction algorithm are presented in [1]. The colour difference histogram is the primary feature, and the support vector machine (SVM) is the classifier in the slide identification system.

2.1 Literature Review

Various works relating to the identification of key frames in video streams are presented. Users will be able to browse and choose multimedia content of their choosing thanks to automatic summary and indexing procedures. To cluster smaller dimensional data and find the cluster's centre, density maximum clustering [23] was introduced. Integrating several types of video footage enables key frame retrieval for diverse types of video. Typically, selected keyframes must also illustrate the video content and have the least amount of extraneous material. On the basis of keypoint coordination, standards of addition and excess are used in the evaluation procedure. Slide Based Lecture Videos (SBLV) [24], which make up a sizable component of online educational videos, were given special consideration when designing an easily verified and navigable system for video content. As per visual components, text, and formulas included in learning material, lectures recorded, as well as mouse and pointer pointing actions recorded during a lecture, the functionality comprehensively originates versatile conceptual hints for video content tagging and visual aid creation. A framework for an efficient video summarization was suggested [4], much like a video movement description. Capsule Net (Capsule Networks) is initially set up as an extractor of data sets, and an intraclass movement bend is created based on the data sets' features. As a result, it is suggested to use a progress impact detection technique to automatically split the video content into frames. Finally, a conscious model is familiar with some key-frame groupings within the photos; as a result, key static images are selected as the video's rundown of information, and optical flows can be identified as the synopsis of movement. It was advised to use a two-phase User-Directed Video Segmentation framework [13] that includes measurement reduction and transient clustering.

Some research attempts to automatically extract video features from CNN. Using a substantial 3DCNN network [2], With this method, CNN's dimension is increased at various levels, and the frame serves as its input. The 3DCNN is fundamentally unable to prevent video segment recognition since it is constrained by the length of the neural network and cannot deal with videos of varying lengths. Authors in [6] developed a more efficient RNN structure known as LSTM to combine the features derived from each frame and introduced the LRCN framework, depending on the use of CNN to obtain the best result from separate video frames. LSTM and 3DCNN require a significant amount of training time and memory size. Authors in [9] presented a dual-stream approach to derive timestamps. In order to obtain the features of a particular image from the input images and numerous frames of the optical flow image and blend them at the final play level, this process uses sparse optical flow as an understudy and two separate convolutional neural networks (CNN). Visual object detection and feature extraction are the main sources of inspiration for recent text detection research [7][19]. These techniques can be divided into three groups: segmentation-based techniques, bounding box estimation techniques, and mixed techniques. Each textual region is treated as a distinct type of object by bounding box correlation coefficients algorithms [12], which predict the bounding box and categorization of each textual region. Text sections are attempted to be separated from backgrounds using segmentation-based techniques [15], which then produce the final bounding boxes in accordance with the segmented results. Similar to Mask R-CNN [7], combined approaches [14] employ bounding box modeling in addition to segmentation for improved performance. Nevertheless, because there are more steps in combined methods, they take longer. The most widely used of the three types of approaches for text detection is the bounding box linear interpolation methodology, which we also use.

There are two types of bounding box parameter estimation methods. The output of first-phase approaches [12] at various grids corresponds to the precise places on feature maps. These techniques frequently move more quickly but with less precision. The CNN (Convolutional Neural Network) is used in the first stage of second-phase approaches [22] to get features and create a collection of candidate suggestions that are intended to include all texts and filter away the majority of unfavourable choices. In the second phase, each potential proposal is categorised into a single class in the second phase, and learning bounding box analysis is used to conduct a more precise localization. In order to recognise sequential characteristics in text, the ConnectionistText Proposal Network (CTPN) integrates CNN with RNN (Recurrent Neural Network). A further two-stage analyzer, An Efficient and Accurate Scene Text Detector (EAST)[20], uses an FCN-based (Fully Convolutional Networks) workflow to combine the characteristics from each convolution operation. EAST additionally produces text angular position and quadrilateral coordinates in addition to the given class and coordinate axes parameters. Our text detector has a two-stage approach as well. But unlike these techniques, which strive to capture natural scene photographs.

2.2 Keyframe Extraction

The most important aspect of each shot in the video is represented by the keyframe. As a result, precisely extracting the key frame from each photo can significantly speed up retrieval operations while also increasing retrieval accuracy. There are two steps in this plan. The first step is to create a flexible key frame extraction and key-frame text detection system based on convolutional neural networks (CNN).

The video streams' frame contents contain noise and distortion. Preprocessing must be done on the video streams to get rid of these noise components. The key-frames can be extracted from the video recordings' frames using the preprocessing technique. The key-frames contain the majority of the video's information, which lowers the computational burden and boosts the effectiveness of the procedure for retrieving video content. Each sequence in the video may be broken down into a number of views, and each view has a number of key frames.

Information from a video includes the plot, features, temporal domain, and spatial domain. Manually extracting and indexing video features requires a great deal of work, memory, and processing power. The scene analysis, video identification, and information clip retrieval will all suffer if the obtained key frames are not accurate. As was previously indicated, numerous scholars have put forth a variety of key frame extraction strategies. However, these techniques still have drawbacks. For instance, many computational solutions use the first frame of every shot as the key-frame. However, this solution is simple to lose a significant set of visual information from the prism, the unpredictability is high, and the scale is difficult to understand. Alternatively, some methods choose key frames by listing and making comparisons between each frame of the prism. There is also a lot of strain on storage space and computer power. This study suggests a novel algorithm, as shown in figure 1, in light of the key frame extraction technology, which demands the qualities of high accuracy and quick calculation speed [5].

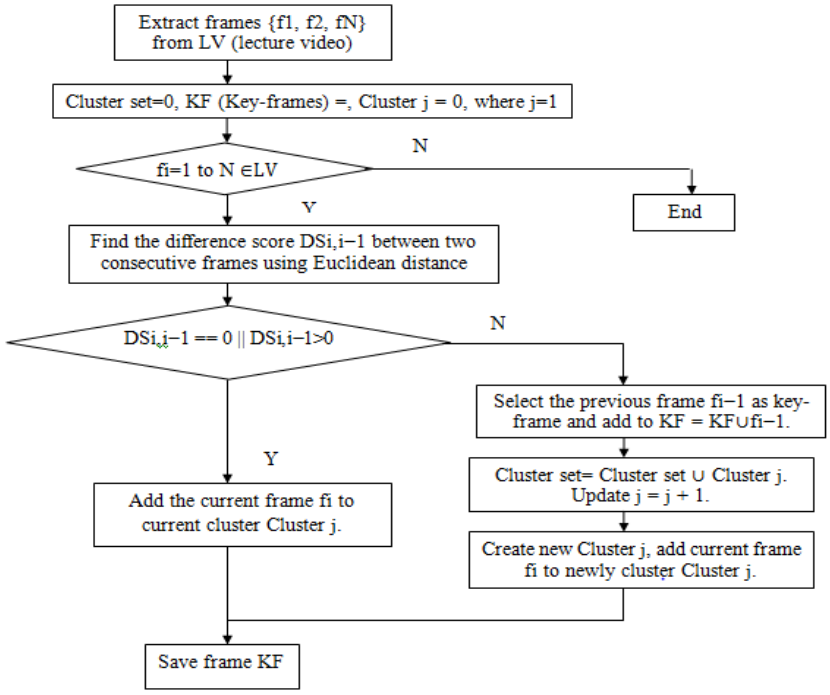


Fig 1. Flowchart of Key-frame extraction

Although the visual content of the subsequent frames in video lectures may vary significantly, analysing these frames requires processing in all 3 dimensions. All the coloured frames in a video are converted into grayscale frames to lower the processing cost. Then, a 1-dimensional vector is scaled into each frame.

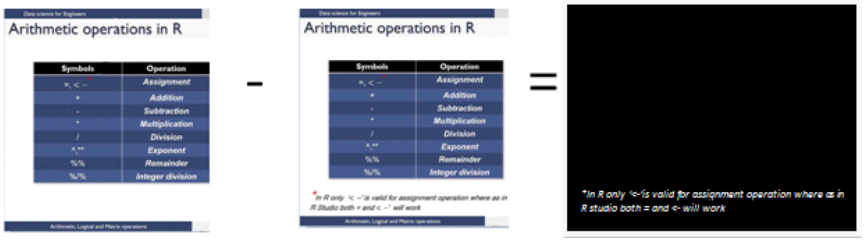


Fig 2. Inter-frame difference example

To retrieve image features from a lecture, we used the inter-frame difference method. Estimating the difference between two succeeding frames is what the inter-frame difference technique does (as shown in Fig. 2). We have used binarized subtracting (as shown in eq.1) to estimate the difference score Di,i−1 using the Euclidean Distance between the current frame Fi and the preceding frame Fi−1:

$$D_{i,i-1} = ||F_i - F_{i-1}||_2 = \left(\sum_{j=1}^{n} |e_{j,i} - e_{j,i-1}|\right)^{1/2} \quad (1)$$

where ej,i is jth constituent of Fi. When a lecturer is writing on a board, we've seen that the content is added frame by frame, and then when the board is completely full, the content should always be removed. To retain the entire lecture's visual material, the proper information frame should indeed be saved first before any content is deleted. These frames are retrieved and stored as the key-frames of the video lectures. In the next section, we will detect text regions from the obtained key-frames using a deep learning approach.

2.3 Deep Learning Based Text Detection

Text plays a vital role in many situations and serves as one of the primary means by which people communicate with one another. Recently, computer vision and record analysis have been active research areas focused on the detection and recognition of text from images. Text localization is considered to be a more unpredictable problem than classification, which can also recognise a text but cannot tell you with certainty where it is located in the image and is ineffective for images with numerous objects. The task of text recognition is to make visible all relevant text in the image and determine its precise classification and location. One of the main issues in the field of computer vision is this. Target placement has continually been the most intriguing problem in the field of computer vision, despite the fact that many items have distinct looks, postures, and forms along with the resistance of obstruction, illumination, and various components throughout imaging. With the increasing use of deep learning-based processes, text localization efforts have shown promising results. In order to address the problem of text localization, current processes consider content to be the exact substance and look into general text localization systems. The below figure 3 depicts the overall structure of the Faster-RCNN network. Each layer of this network has key functionalities as discussed below.

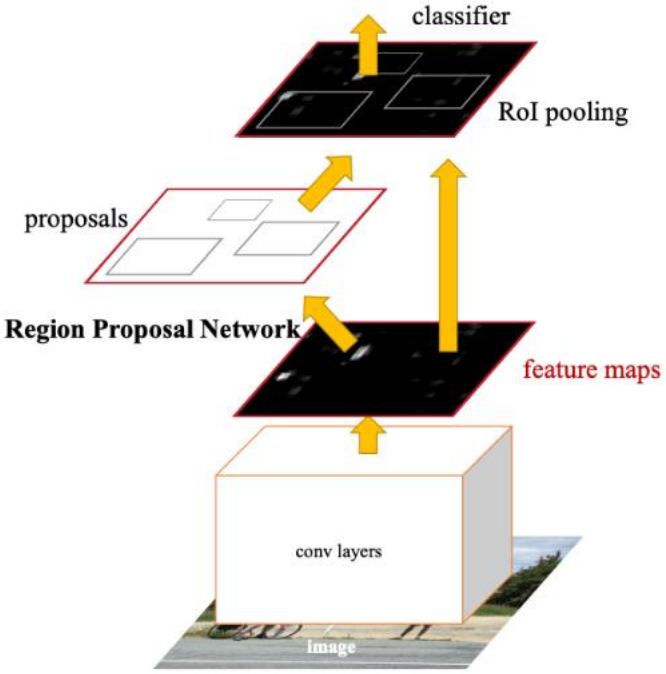


Fig 3. Faster RCNN network [25]

Convolutional (Conv) layers: In order to extract the feature maps from the input image for use in the following Region Proposal Networks (RPN) layer and fully connected layer, Faster R-CNN first employs a set of fundamental convolutional layers.

RPN : The primary function of the RPN network is to produce regional proposals. A large number of anchor boxes are first created. After trimming and sifting the anchors, Softmax is used to decide whether they belong in the background or the foreground, or in

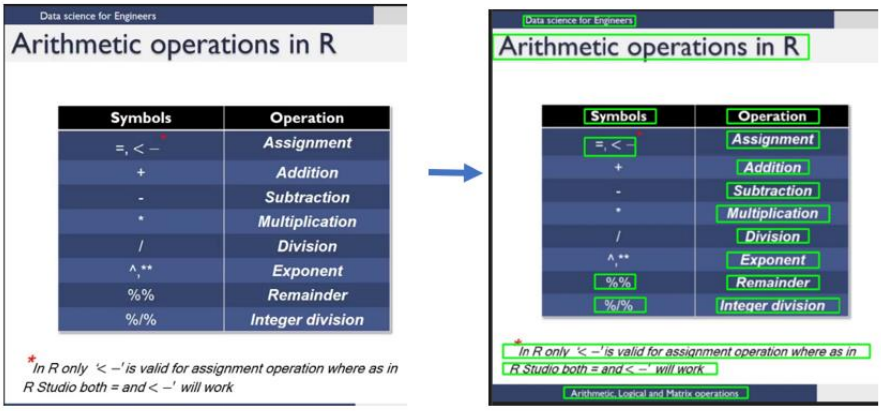
other words, whether they are objects or not. Since this is a two-category problem, another division bounding box predictor modifies the anchor box to create a more precise proposal.

Region of Interest (ROI) pooling: This layer creates a rectified proposal feature map using the RPN proposals and the softmax layer from the previous layer of VGG16. It can execute target location and recognition once entering it by using the entire connection operation.

Classifier: In order to perform a full connection process, the ROI pooling layer will be converted into a rectified feature map. Softmax will be used to categorise particular categories, and L1 Loss will be used to complete the bounding box prediction process to determine the exact location of the text. High-resolution images may result in duplicated feature information when Faster R-CNN is performed on a lecture video. Table 1 summarises several of the text detection algorithms. The first three methods belong to the region of the proposal, and the next three are semantic segmentation based methods.

Table 1:Text detection algorithms**.**

Methods	Description
CTPN [22]	It employs VGG16 for extracting features and uses Faster-anchor RCNN's regression technique, enabling the RPN to identify objects of various sizes using a solitary sliding window as well as to experiment with parameters. Yet, CTPN has some drawbacks, such as its poor ability to identify non-horizontal text.
SegLink [16]	SegLink's central concept is to first identify individual text line segments, subsequently, when all segments are identified, combine the information of every image feature, and end up joining the individual segments to make a comprehensive text line
ABCNet [10]	Bezier curves can be utilised to detect curved text, and BezierAlign is employed for extracting features and text region correction. This makes it possible to find text lines with strange shapes in scenes from nature.
EAST [20]	Eliminates the laborious process of candidate consolidation and lexical analysis among them by using an unified neural network(NN) to immediately detect text in an image.
PixelLink [3]	The model is mostly built on CNN models that have been adjusted to assess whether each pixel point contains text (or not) and to forecast whether it is connected to the eight directions that are immediately surrounding that pixel.



(a) Input (b) Output

Fig 4. Text detection Model

The process of the proposed text detection (as shown in figure 4) model contains two stages, namely, training and detection (as shown in figure 5).

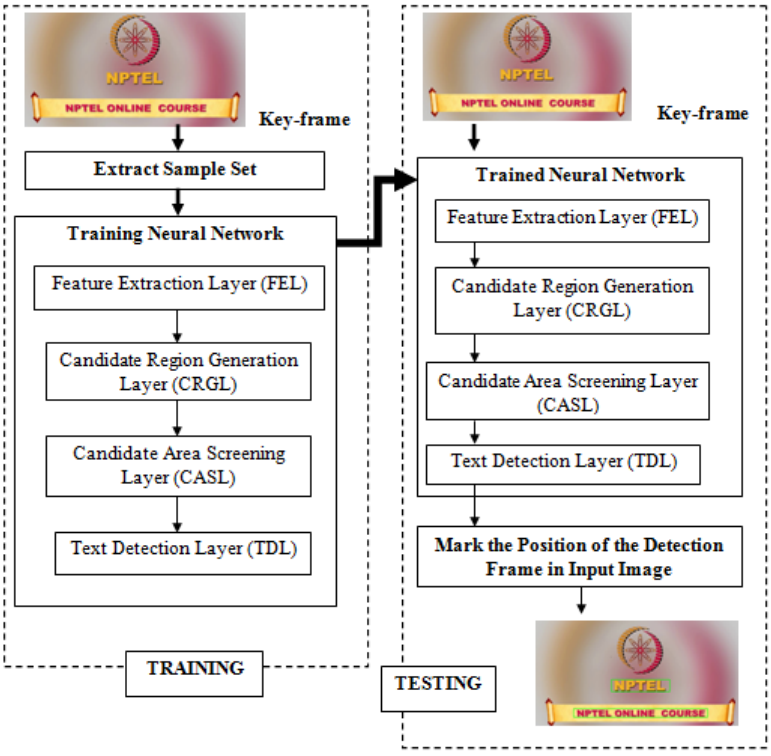


Fig 5. Training and Detection phases of Text Detection Model

FEL: To create a feature map, feature extraction is performed on the supplied image;

CRGL: Utilise anchor process to produce preliminary areas and 3×3 hole convolutional having expansion factor r=2 to remove redundant features;

CASL: Soft-Nonmaximum Suppression (NMS) is a milder screening strategy for target frames when compared to NMS. The preliminary areas are therefore coarsely screened using Soft-NMS;

TDL: ROI pooling should be done on the target region, a fixed length should be realised and supplied to the fully connected layer (FCL), and then Soft NMS should be linked once more to display the text detection screen and produce the text detection output.

The training phase does btaining the training sample set and creating the trained model. After numerous rounds, the trained text detection layer is produced by feeding the sample trained model further into the NN for training. The Execution steps are:

1. To create a frame set R, one must manually filter crucial frames, extract frames from lecture videos, and get rid of similar frames.
2. To create the VOC (Visual Object Classes) style specimen set RL, manually mark the text's outer margin on the frame set R, create mark data L, then apply VOC style to R and L.
3. Create the feature extraction model from scratch and feed it the RL test set. Subsequently, a single image Ri makes use of the region formation network's Anchor process to establish numerous first text regions A.
4. We obtain a number of training specimen sets S = {+, -}, in which + is a positive specimen and - is a negative specimen. If the IoU between both the first recommended region Ai and the real boundary of the labelled text is B, then the first recommended region is a positive specimen when B= Bmax or B>= 0.6, and a negative specimen when B< 0.3.
5. Learn to change the system parameters by using test frame set S loss function till the training is finished and indeed the system parameter M is acquired.

The detection phase does directly enter the input key-frame into the trained NN and get the precise location of the text's area in the frame and Get the result of text detection, as shown in Figure 5. Execution steps are:

1. Input the key-frame K to the trained network of text detection.
2. The K in the conv (convolutional) layer is subjected to feature extraction to produce a convoluted feature map, which will then be fed to the candidate area producing network to produce a candidate area set Rs.
3. After screening the candidate region set Rs, certain patterns that go beyond the frame boundary are suggested, as well as the candidate region B′ is discovered.
4. After screening, the candidate region Rs′ is sent into the detection phase, fix the text's bounding location, and get the actual text location.
5. Obtain text outcomes.

Loss Function: To maintain the correctness of data processing, the actual model must be adjusted in accordance with various environments. By defining a multi-task loss function, such as the following, the NN is trained:

$$L\left(\{PR_i\}\{l_i\} = \frac{1}{M_c}\sum_i L_c\,(PR_i PR_i^*) + \beta\frac{1}{M_r}\sum_i PR_i^*\,L_r(l_i PRl_i^*)\right) \quad (2)$$

where I is the address of the ith candidate frame in batch production processing. Mc, Mr, and balance the standardized rates of classification loss (equation 3) and prediction loss (equation 4), respectively. It is likely that the ith candidate box will contain the target.

PR*i = 1 if the ith candidate box has a candidate target; else, PR*i = 0.

$$L_c(PR_i PR_i^*) = -\log[PR_i PR_i^* + (1 - PR_i^*)(1 - PR_i)] \quad (3)$$

$$L_r\,(T_i T_i^*) = R(T_i - T_i^*) \quad (4)$$

where the smoothL,1 function R is used. The vector-prediction weighted potential frame coordinates are Ti = Tx,Ty,Tw, and Th while the coordinate matrix of actual limits is T*i = Tx*, Ty*, Tw*, and Th*. We define Ti and T*i as follows:

$$Tx = (x-xj)/wj \quad , \quad Ty = (y-xyj)/hj \quad , Tw=\log(w/wj), Th=\log(h/hj),$$
$$Tx^*= (x^*-xj)/wj \ , Ty^*=(y^*-yj^*)/hj \ , Tw^*=\log(w^*/wj), Th^*=\log(h^*/hj) \qquad (5)$$

where (w,h), (wj,hj), and ($w*$,$h*$) are the width and height of anticipated areas, candidate areas, and formal areas, respectively, and (x,y), (xj,yj), and ($x*$,$y*$) are the predicting region, candidate region, and formal state capital coordinates, respectively.

3. Results and Discussions

The data is an image of text with simple or complex patterns or backgrounds in a natural landscape or document. We can obtain a visual representation of the text using a digital camera and a handheld scanner. There are various sorts of text image repositories that can be utilised to conduct research.

We will go over the quantitative and qualitative analysis of the proposed system. From the NTPEL Project, we chose a total of 50 videos from five distinct courses: 10 videos on machine learning (ML), 10 videos on cloud computing (CC), 10 videos on computer networks (CN), 10 videos on cryptography (Crypt), and 10 videos on database management systems (DBMS). We provide ground truth reports so that our model can be evaluated. All of the videos are utilised in various experiments and evaluations that are carried out on a desktop computer with an i8 processor, 8 GB of RAM, 512 GB of storage, and an HDD (Hard Disk Drive). We prepare ground truth results to compare the proposed model to the existing model. Figure 6 shows that the proposed key-frame extraction method provides more key-frames than existing models. Key-frame extraction has been extensively used with the goal of producing the finest video summary. However, there is no ideal method for assessing their performances yet. The key-frames of video produced by various algorithms are compared to key-frames chosen by humans based on the three assessment parameters: Precision, Recall, and F-measure in order to determine the quality of each model.

$$\text{Precision (Pr)} = \frac{\text{TruePositive(TP)}}{\text{TP+FalsePositive(FP)}} \qquad (6)$$

$$\text{Recall(Re)} = \frac{\text{TP}}{\text{TP+TrueNegetive(TN)}} \qquad (7)$$

$$\text{F1} - \text{Score} = \frac{2*\text{Pr}*\text{Re}}{\text{Pr+Re}} \qquad (8)$$

Fig 6. Key-frames of ML data set.

When Recall and Precision are equally weighted, F1 becomes the harmonic mean between Re and Pr, often known as a symmetrical F-score. As a result, the model with the highest F1 score is the most accurate. We used the threshold value $\delta = (H \times W)/10$, where H and W are the height and width of a frame, respectively. The below table 2 shows the total number of key-frames obtained.

Table 2: Total key-frames obtained.

Lecture videos	Keyframes			
	EVS	PM	Manual	Total slides
ML	35	38	41	42
CC	72	75	76	77
CN	68	72	75	75
Crypt	47	51	56	57
DBMS	119	125	146	147

Table 3 shows the quantitative evaluation of our method using state-of-the-art datasets, with the best evaluation findings shown in bold. Here, we found that our model outperformed existing model in terms of recall across all five datasets. Some strategies work better in the setting of Precision, but others could work better in the context of Recall. Our model's F-measure on five data sets is likewise the highest compared to leading models, indicating that the proposed keyframe extraction model performs better.

Table 3: A comparison of proposed method with existing model

Dataset (LV)	Method	Pr (%)	Re (%)	F1-Score (%)
ML	EVS	95.7	76.7	85.1
	PM	89.3	86.8	88.0
CC	EVS	96.4	78.4	86.5
	PM	88.7	90.0	89.3

CN	EVS	97.9	77.4	86.3
	PM	92.1	87.2	89.6
Crypt	EVS	98.4	77.9	87.0
	PM	92.8	88.0	90.3
DBMS	EVS	96.9	75.8	85.0
	PM	94.9	91.7	93.0

Below figure 7 shows performance results of the Elastic Video Summarization Algorithm (EVS of 85.9 %) technique and the proposed method (PM of 90.0%) [8].

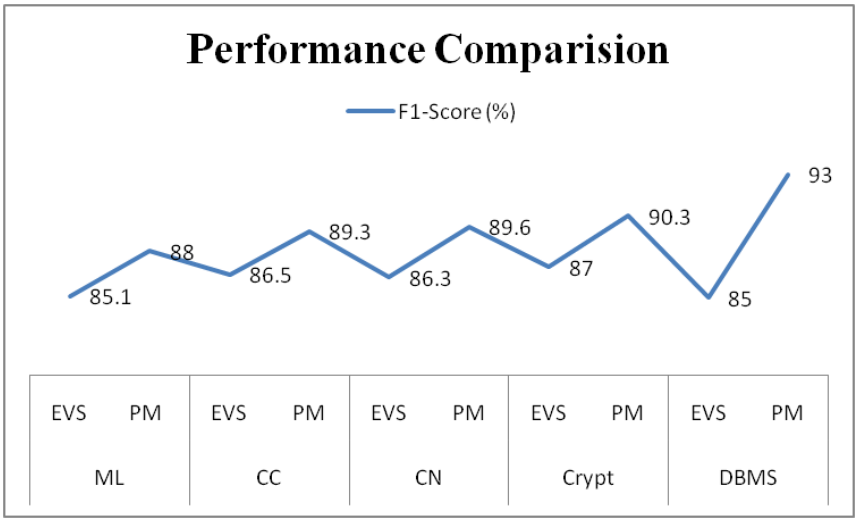


Fig 7. Performance comparision of key-frame extraction methods

Comparison of text detection accuracy between proposed model and other different methods have been shown in the below figure 8.

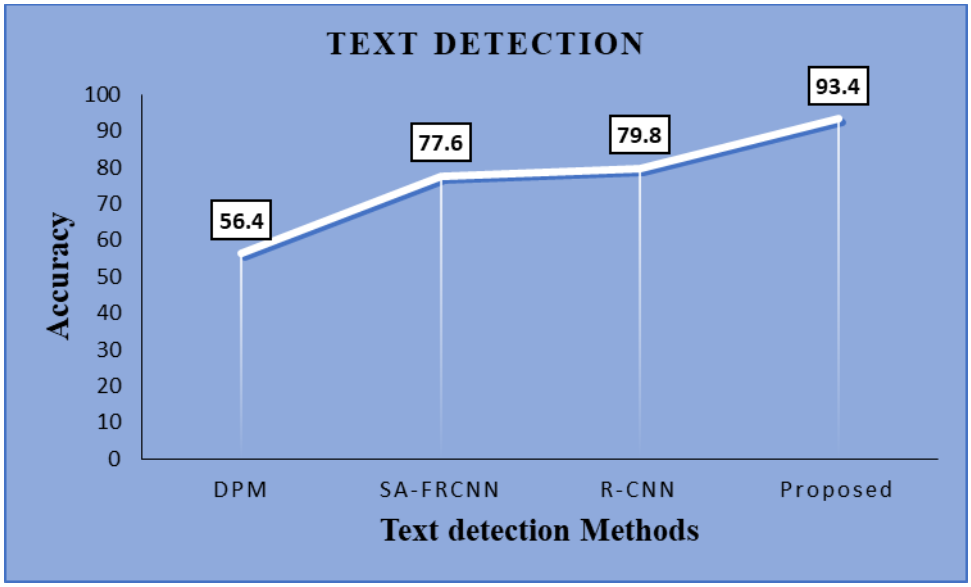


Fig 8. Text Detection accuracy between different methods

3.1 Computational Complexity

To reduce the cost of computing, we use the online clustering method. Our method requires an average of 4668 seconds to process 50 videos, or 93.36 seconds per video. As a result, the presented method is almost 50% quicker than current models [1][23]. Thus, given a video lecture lasting roughly 30 minutes, the proposed method can extract the key-frames in an average of 1.5 minutes. As a result, it may satisfy the needs of real-time practical applications such as video lectures, slide summaries, sports highlights, video surveillance, etc. Table 4 compares the proposed key-frame extraction model with the most recent approaches to determine the best performing one.

Table 4: Computational cost comparison.

Methods	Sampling Rate (frames per sec)	Average Time per videos (sec)
EVS	30	180.00
PM	25	93.36

4. Conclusion

The key-frame extraction method is highlighted in this work as a way to summarise and analyse the video lectures in oreder to help online learners. A deep learning-based text detection model that integrates Faster R-CNN is suggested in order to increase the efficiency and accuracy of the video content extraction system. In adaptive lecture video analysis, a faster R-CNN text detection method is applied. Prior to changing the existing NMS process with Soft-NMS to cope with the coincident of text, we first obtained the key-frames to input into the proposed model. After that, we introduced a hole convolution to the system to remove unwanted features in the high resolution video frame, making it perfect for key-frame text detection. The key-frame extraction results are best compared to EVS, and the text detection findings demonstrate that the modified model enhances the detection accuracy of 93.4%, which is the best compared to other algorithms and also minimizes the skipped detection performance.

As a future work we are working on other powerful machine learning and deep learning methods to improve the OCR.

References

[1] Adcock, J., Cooper, M., Denoue, L., Pirsiavash, H., & Rowe, L. A. (2010, October). Talkminer: a lecture webcast search engine. In *Proceedings of the 18th ACM international conference on Multimedia* (pp. 241-250).

[2] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 4489-4497).

[3] Deng, D., Liu, H., Li, X., & Cai, D. (2018, April). Pixellink: Detecting scene text via instance segmentation. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

[4] Huang, C., & Wang, H. (2019). A novel key-frames selection framework for comprehensive video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(2), 577-589.

[5] Liang, J. S., & Wen, H. P. (2019). Key frame abstraction and retrieval of videos based on deep learning. *Control Eng. China*, *26*, 965-970.

[6] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2625-2634).

[7] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961-2969).

[8] Kumar, K., Shrimankar, D. D., & Singh, N. (2016, November). Equal partition-based clustering approach for event summarization in videos. In *2016 12th International conference on signal-image technology & internet-based systems (SITIS)* (pp. 119-126). IEEE.

[9] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, *27*.

[10] Liu, Y., Chen, H., Shen, C., He, T., Jin, L., & Wang, L. (2020). Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9809-9818).

[11] Liao, M., Shi, B., Bai, X., Wang, X., & Liu, W. (2017, February). Textboxes: A fast text detector with a single deep neural network. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 31, No. 1).

[12] Liao, M., Shi, B., & Bai, X. (2018). Textboxes++: A single-shot oriented scene text detector. *IEEE transactions on image processing*, *27*(8), 3676-3690.

[13] Peng, X., Li, R., Wang, J., & Shang, H. (2019). User-guided clustering for video segmentation on coarse-grained feature extraction. *IEEE Access*, *7*, 149820-149832.

[14] Lyu, P., Liao, M., Yao, C., Wu, W., & Bai, X. (2018). Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 67-83).

[15] Long, S., Ruan, J., Zhang, W., He, X., Wu, W., & Yao, C. (2018). Textsnake: A flexible representation for detecting text of arbitrary shapes. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 20-36).

[16] Shi, B., Bai, X., & Belongie, S. (2017). Detecting oriented text in natural images by linking segments. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2550-2558).

[17] Tadesse, S., & Muluye, W. (2020). The impact of COVID-19 pandemic on education system in developing countries: a review. *Open Journal of Social Sciences*, *8*(10), 159-170.

[18] Li, W., Qi, D., Zhang, C., Guo, J., & Yao, J. (2020). Video summarization based on mutual information and entropy sliding window method. *Entropy*, *22*(11), 1285.

[19] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14* (pp. 21-37). Springer International Publishing.

[20] Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., & Liang, J. (2017). East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 5551-5560).

[21] Liu, X., Kawanishi, T., Wu, X., & Kashino, K. (2016, March). Scene text recognition with high performance CNN classifier and efficient word inference. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1322-1326). IEEE.

[22] Tian, Z., Huang, W., He, T., He, P., & Qiao, Y. (2016). Detecting text in natural image with connectionist text proposal network. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14* (pp. 56-72). Springer International Publishing.

[23] Zhao, H., Wang, T., & Zeng, X. (2018). A clustering algorithm for key frame extraction based on density peak. *Journal of Computer and Communications*, *6*(12), 118-128.

[24] Zhao, B., Xu, S., Lin, S., Wang, R., & Luo, X. (2019, July). A new visual interface for searching and navigating slide-based lecture videos. In *2019 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 928-933). IEEE.

[25] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28.

Notes on contributors

Dr. Geetabai S. Hukkeri is Assistant Professor in the Department of Computer Science and Engineering at the Manipal Institute of Technology, Bengaluru, India. Her research interests include Artificial Intelligence, Deep Learning, Machine Learning, Big Data, Computer Vision, Computer Networks and Multimedia Information Retrieval. She has published several papers in International Journals, International Conferences, and International Book Chapters. She had published a book on "Understanding Big Data Technologies-A simple Approach". She is a member of ACM. She holds a Ph.D. Degree in Computer Science and Engineering from Visweswaraya Technological University, Belagavi, India.

Dr. R H Goudar is currently working as an Associate Professor, Department of Computer Science and Engineering, Visvesvaraya Technological University, Belagavi, Karnataka. He has published over 150 papers in International Journals, International Conferences, and International Book Chapters. He is a Reviewer of IEEE Transactions on Knowledge and Data Engineering and Editor/ Editorial Member of various Computer Science Journals. He has received various awards, such as the Outstanding Faculty Award, the Research Performance Award, the Young Research Scientist Award from VGST Karnataka, and the Eminent Engineer Award from the Honorable Chief Minister of Karnataka.

Dr. Gururaj H L is currently working as Associate Professor in the Department of Information Technology at Manipal Institute of Technology Bengaluru, India. He has received Young Scientist International Trevel Grant from ITS-SERB, DST, Government of India. He is Faculty Sponsor of MITB ACM Student Chapter. His research interests include QoS aware network congestion control, Cyber Security, Blockchain Technology, network security, Cloud computing and Machine Learning.

Dr. Shilpa Ankalaki is currently working as Assistant Professor, Department of Computer Science and Engineering, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal. Her research interests include Machine Learning, Deep Learning, Data Mining and Artificial Intelligence applications. She holds a Ph.D. Degree in Computer Science and Engineering from the Visweswaraya Technological University, Belagavi, India.

**Corresponding author: Shilpa Ankalaki (shilpa.ankalaki@manipal.edu)*