Impact of Deep Learning on Localizing and Recognizing Handwritten Text in Lecture Videos

Lakshmi Haritha Medida[1]*
Research Scholar, CSE
JNTUA, Ananthapuramu
India

Kasarapu Ramani[2]
Soft Computing Research Centre, Department of IT
Sree Vidyanikethan Engg College (Autonomous)
Tirupati, India

***Abstract*—Now-a-days, the video recording technologies have turned out to be more and more forceful and easier to utilize. Therefore, numerous universities are recording and publishing their lectures online in order to make them reachable for learners or students. These lecture videos encapsulate the handwritten text written either on a paper or blackboard or on a tablet using a stylus. On the other hand, this mechanism of recording the lecture videos consumes huge quantity of multimedia data in a faster manner. Thus, handwritten text recognition on the lecture video portals has turned out to be an incredibly significant and demanding task. Thus, this paper intends to develop a novel handwritten text detection and recognition approach on the video lecture dataset by following four major phases, viz. (a) Text Localization, (b) Segmentation (c) Pre-processing and (d) Recognition. The text localization in the lecture video frames is the initial phase and here the arbitrarily oriented text on video frames is localized using the Modified Region Growing (MRG) algorithm. Then, the localized words are subjected to segmentation via the K-means clustering, in which the words from the detected text regions are segmented out. Subsequently, the segmented words are pre-processed to avoid the blurriness artifacts as well. Finally, the pre-processed words are recognized using the Deep Convolutional Neural Network (DCNN). The performance of the proposed model is analyzed in terms of the performance measures like accuracy, precision, sensitivity and specificity to exhibit the supremacy of the text detection and recognition in lecture video. Experimental results reveal that at Learning Percentage of 70, the presented work has the highest accuracy of 89.3% for 500 count of frames.**

Keywords*—Lecture video; text localization; segmentation; word recognition; Deep Convolutional Neural Network (DCNN)

I. Introduction

In the recent days, the professional lecture videos are abundant and the number is constantly growing in the web. These lecture videos are motivating the students towards tele-teaching and e-learning [1] [2] [3] [4]. It is more crucial for the students to quickly understand the subject by viewing the video rather than reading the text. The advanced analysis techniques help to automatically collect the relevant metadata from these videos and hence the video lectures are becoming an easiest technique of online course learning [5] [6] [7]. In the MOOCs, it is vital to understand the lecture videos for educational research as it has become synonymous with distance learning. The better understanding of the lecture video lies in the vital cues like the figures, images and text [8] [9] [10] [11]. Among these vital cues, the text is available in almost all lectures as it can be utilized for variety of tasks like the extracting class notes, generation of keywords, search enabling and video indexing.

In the lecture video, the handwritten text can be on a blackboard or a paper and this text can be written using a stylus on a tablet and displayed on a screen or font rendered text appearing in presentation slides (digital text). These lectures are usually documented with typically positioned cameras. In general, the identification of the text from the presentation slides is a bit easier while compared to the handwritten blackboard text, since they are more legible. On the other hand, the handwritten text recognition is a different beast considering the amount of variations and the character overlaps. The Characters can be small/large, stretched out, swooped, stylized, slanted, crunched, linked, etc. Digitizing handwritten text recognition is extremely exciting and is still far from solved - but deep learning is assisting us in improving the accuracy of the handwritten text recognition. The handwritten blackboard text recognition is additionally challenging and is not legible due to lower contrast, bad illumination, smaller size letter etc. Moreover, detection of the text on the blackboard or paper might be difficult and cluttered, if the lecture over-writes or writes over the figures and equations [12] [13] [14]. The Handwritten Text Recognition (HWR) focuses on the handwritten text in documents and it is practically inherent to complexity in case of different writing styles.

Over the decades, extensive research has been carried out in the field of text recognition on the lecture videos and a variety of methods and algorithms were developed. Word spotting is a key challenge and the majority of the up to date works uses DCNN for learning the features. DCNNs learn the features of the word from dissimilar attribute spaces and are invariant to diverse styles and degradations. Thus, with due interest to handwritten text identification on the lecture video with utmost accuracy, this work focuses on formulating a novel technique by specifically looking into the problems of existing works.

The major contribution of the current research work is highlighted below:

- A novel deep learning based handwritten text recognition approach for video lectures is developed.

*Corresponding Author

- Initially, the text in the collected video frames is localized with Modified Region Growing Approach and these texts are segmented with K-means clustering.
- The segmented words are pre-processed and recognized using DCNN.
- The performance of the proposed model will be analyzed in terms of certain performance measures like accuracy, precision, sensitivity and specificity.

The rest of the paper is organized as: Section II discusses about the literature works undergone under this subject. Section III portrays about the proposed handwritten textual recognition in lecture videos. The resultant acquired with the presented work is discussed in Section IV. Finally, a strong conclusion is given to the current research in Section V.

II. Literature Review

In 2014, Yang et al. [15] have developed a novel framework for video text detection and recognition. A Fast Localization-Verification Scheme (FLVS) with the Edge Based Multi-Scale Text Detection (EMS-TD) was constructed in the text detection stage. This algorithm consists of three main steps: text gradient direction analysis, seed pixel selection and seed-region growing. The novel video text binarization algorithm was employed for better text recognition. The potential text candidates were detected with high recall rate by the edge based multi-scale text detector. Then, the detected text lines of the candidate in the video were refined by using image entropy-based filter. Subsequently, the false alarms in the lecture video were discarded by the authors with the help of the Stroke Width Transform (SWT) and Support Vector Machine (SVM). In addition, a novel skeleton-based binarization method was constructed to disconnect text from compound backgrounds in the text recognition phase. The proposed text recognition model in lecture video was evaluated in terms of accuracy using the publicly available test data sets.

In 2018, Poornima & B. Saleena [16] has developed a new technique for successful repossession of the lecture videos from the database using the Correlated Naive Bayes (CNB) classifier. Here, the textual features as well as the image texture were extracted from the key frames with the help of the Tesseract Classifier (TC) and Gabor Ordinal Measure (GOM). The extracted feature dataset encloses three major types of features like the keywords, semantic words, and the image texture. On the basis of the similarity of the features, the authors grouped the video with K-means clustering. Finally, the texts were recognized from the lecture video on the basis of the correlation as well as the posterior probability. The proposed model was compared over the existing models in terms of precision and recall.

In 2018, Kota *et al*. [17] have constructed a Deep Learning based method for handwritten text, math expressions and sketches recognition in the online lecture videos. In the proposed model, the input from the whiteboard lecture video was recorded by the video processing pipeline using a still camera. Then, the summary of the handwritten elements on the whiteboard in the lecture was generated as keyframes over time. It suffers from the occluded content owing towards the motion of the lecturer. They implied the conflict minimization approach after spatio-temporal content associations with the aim of generating the summary of the key frames. In addition, the Coarse-Grained Temporal Refinement (CTR) was employed to the Content Bounding Boxes (CBB) to detect the variations in the detector output in terms of dissimilarity like the occlusions and illumination.

In 2015, Husain *et al*. [18] have projected a distributed system in which the lecture video frames were stored in the Hadoop's Distributed File System (HDFS) repository. Then, with the help of the HDFS, the processing operations and the highly concurrent images were processed. Further, the MapReduce framework was implied for reading text information as well as for counting the frequent appearance of the words. The proposed text recognition and word count algorithms were tested with the cluster size of 1 and 5 in the Hadoop framework. The resultant of the proposed model confirmed its application in the field of video processing and high-speed image processing.

In 2019, Dutta *et al*. [19] have investigated the efficiency of the traditional handwritten text recognition and word spotting methods on the lecture videos. The dataset was collected from LectureVideoDB having 24 different courses across science, management and engineering. Once the frames were stored, they were pre-trained using the TextSpotter. They localized the words in the video lecture using the deep Fully Convolutional Neural Network (FCNN) and to the output of FCNN; the Non-Maximal Suppression (NMS) was employed to detect the arbitrarily oriented text on the blackboards. Once, the location of the word is identified, the word was recognized using the Convolutional Recurrent Neural Network (CRNN) architecture and Convolutional Recurrent Neural Networks Based Spatial Transformer Network (CRNN-STN). Then, as a novelty they spotted the keywords in the video by extracting the features with two parallel streams of network and label information was concatenated using Pyramidal Histogram of Characters (PHOC) features.

In 2019, Miller [20] has designed a lecture summarization service model by leveraging Bidirectional Encoder Representations from Transformers (BERT) model. The initial contribution of this research work was based on the supervision of lecture transcript and summarizations, which help the users to edit, retrieve and delete the stored items. The second contribution of this research work was an inference from the BERT model with K-Means model in order to produce the embeddings for clustering. Further, on the basis of specified configuration, the summaries for users were generated by BERT model. Finally, the proposed BERT model was compared with the TextRank and the resultant exhibited no golden truth summaries, but there was improvement in the handling context words and was applicable to more lecture videos.

In 2015, Miller *et al*. [21] have constructed a new approach for Automated Video Content Retrieving (AVCR) within large lecture video archives. Initially, the audio lecture was separated from the video and the video was converted into image key-frame using Optical Character Recognition (OCR)

algorithm. Then, from the image, the keywords were extracted using the OCR algorithm. Subsequently, for the video content navigation, a visual guidance was provided by the key-frame detection as well as the automatic video segmentation model. Then, the video OCR was employed on the key-frames and Automatic Speech Recognition (ASR) in order to extract the textual metadata available in the lecture videos. Further, on the basis of the multimedia search diversification method, appropriate answers were collected on the basis of the words. The proposed model had provided more relevant information with more effectiveness to the users.

In 2019, Husain and Meena [22] have introduced a novel method for efficient Automatic Segmentation and Indexing of Lecture Videos (AS-ILV). The proposed model helps in faster reorganization of the specific and relevant content in the online lecture video. In the proposed model, the authors projected the automatic indexing of lecture videos from both slide text and audio transcripts with the extracted topic hierarchies in the video. On the basis of the slide text information, the authors have indexed the videos with higher character recognition rates in an accurate manner. As a novelty, the authors have overcome the problem of high Word Error Rate (WER) transcribed in the video due to the unconstrained audio recording with the semi-supervised Latent Dirichlet Allocation (LDA) algorithm. They have tested the proposed model with Coursera, NPTEL and KLETU classroom videos and the resultant of the evaluation exhibited average percentage improvement in F-Score, while compared to the existing one. Table I summarizes the above-mentioned works along with the methodology, features and challenges.

Regardless of massive amount of works on lecture videos and MOOCs, there are incredibly a small number of which distinctively come across this problem. Among them, the SVT and SVM approach in [1] has high robustness and High recall rate. Apart from these advantages, it requires improvement in the text recognition rate with the aid of the context- and dictionary-based post processing and the text detection result need to be improved with the help of the text tracking algorithms. Further, in CNB, tesseract classifier and GOM [2], the computational time is lower and the precision as well as recall are improved. This technique suffers from retrieval of text from large dataset and hence optimization technique needs to be implied. In [3], the conflict minimization approach gives higher text detection rate. This technique need to handle occlusions and temporal refinement for end-to-end detection of content in video frames. Then, HDFS and MapReduce in [4] is a fault tolerant distributed system and it is cost effective. The memory shortage problems are created by large datasets and hence memory optimization needs to be implied. Moreover, CRNN and CRNN-STN in [5] is Applicable for low resolution and complex images. But, this technique does not use Applicable for low resolution and complex images. A higher trade-off is achieved between the speed and inference Performance by BERT model. But here the automatic extractive summarization is not perfect. Further, OCR algorithm [7] is good in providing the relevant information in a better way and can collect the appropriate answer for the words. As a controversy to these advantages, it too suffers from low recognition rate and high cost. The F-Score is enhanced by LDA algorithm; however, the WER is not removed completely. The research works in this area should focus on one or more of these problems.

TABLE I. FEATURES AND CHELLENGES OF EXISTING LECTURE VIDEO RECOGNITION APPROACHES

Author [citation]	Methodology	Features	Challenges
Yang *et al*. [15]	SWT and SVM	✓ High recall rate. ✓ High robustness	× Requires improvement in text recognition rate × Requires improvement in text detection result
Poornima & B. Saleena [16]	CNB , tesseract classifier and GOM	✓ Better precision, recall ✓ Minimum computation ✓ Time	× Retrieval can be done with the optimization techniques × Cannot retrieve text from large databases
Kota *et al*. [17]	conflict minimization approach	✓ Better text detection ✓ Extract semantically meaningful information	× Need to handle occlusions × Need to investigate temporal
Husain *et al*. [18]	HDFS and MapReduce	✓ Great improvement in time of execution ✓ Cost effective	× Large datasets often creates memory shortage problems × Requires memory optimization
Dutta *et al*. [19]	CRNN and CRNN-STN	✓ Good contrast. ✓ Easier to recognize	× Need to use recognized text for larger video understanding problems.
Miller [20]	BERT model	✓ High tradeoff between speed and inference Performance ✓ Improves the quality of recognition	× Automatic extractive summarization is not perfect × Difficulty in handling context words
Miller *et al*. [21]	OCR algorithm	✓ Applicable for large lecture video archives ✓ Collects appropriate answer for the words	× High cost × Do not address the way of retrieving the appropriate information
Husain and Meena [22]	LDA algorithm	✓ Enhances the F-Score ✓ Provides indexing information	× Poor retrieval performance × WER is not removed completely

III. Handwritten Textual Information Recognition

The pictorial depiction of the adopted video handwritten text recognition approach is revealed in Fig. 1. In the presented scheme, a new handwritten textual image recognition approach is developed using an intellectual technique. The presented scheme comprises of four most important stages such as, Text Localization, Segmentation, Pre-processing and Recognition. At first, the video frames are acquired and the text within the video frames is localized using the Modified Region Growing Algorithm. Subsequently, the localized words are subjected to segmentation via the K-means clustering, in which the words from the detected text regions will be segmented out. Subsequently, the segmented words are pre-processed to avoid the blurriness artifacts as well. Finally, the pre-processed words are recognized using the DCNN. The resultant from DCNN exhibits the recognized textual information from the acquired lecture video.

A. Text Localization

The collected lecturer video frame encompasses the textual and the audio contents [23]. The textual contents in the images $I(i, j)$ are converted into binary image and the white regions are extracted from it. Further, these extracted white regions $I_w(i, j)$ are subjected to region growing approach for localizing the texts. The segmentation of $I_w(i, j)$ occurs via seed points that have to be regularized. A seed point is the commencement stage for region growing and its selection is significant for the segmentation solution. The stages of region growing approach are portrayed in the following steps.

Step 1: The input image $I_w(i, j)$ is split into a huge count of blocks P, in which all the blocks encompass one centre pixel and several vicinity pixels.

Step 2: Then, fix the Intensity threshold (R^i).

Step 3: For the entire block P, carry out the subsequent course of action in anticipation of the count of blocks that reaches the entire count of blocks for an image.

Step 3(a): Find out the histogram G of all pixels in P.

Step 3(b): The most recurring histogram of the P^{th} block, signified by U^h is fine-tuned.

Step 3(c): Select any pixel, as per U^h and distribute a pixel as seed point with intensity Int_u.

Step 3(d): The adjacent pixels are computed with respect to intensity Int_n.

Step 3(e): Find out the intensity variation of u and n (i.e.) $Dif_{Int} = \|Int_u - Int_n\|$.

Step 3(f): If $Dif_{Int} < R^i$ add the consistent pixel to the region, and hence the region would grow, or go to step 3(h).

Step 3(g): Authenticate if the whole pixels are added to the region. If yes, go to step 2 and then carry out step 3(h).

Step 3(h): Re-estimate the region and discover the new seed points and perform the procedure from step 3(a).

Step 4: Finish the whole process.

The textual information acquired from the region-based approach is $I_{text}(i, j)$, which is fed as input to k-means clustering for cropping the data from the texts.

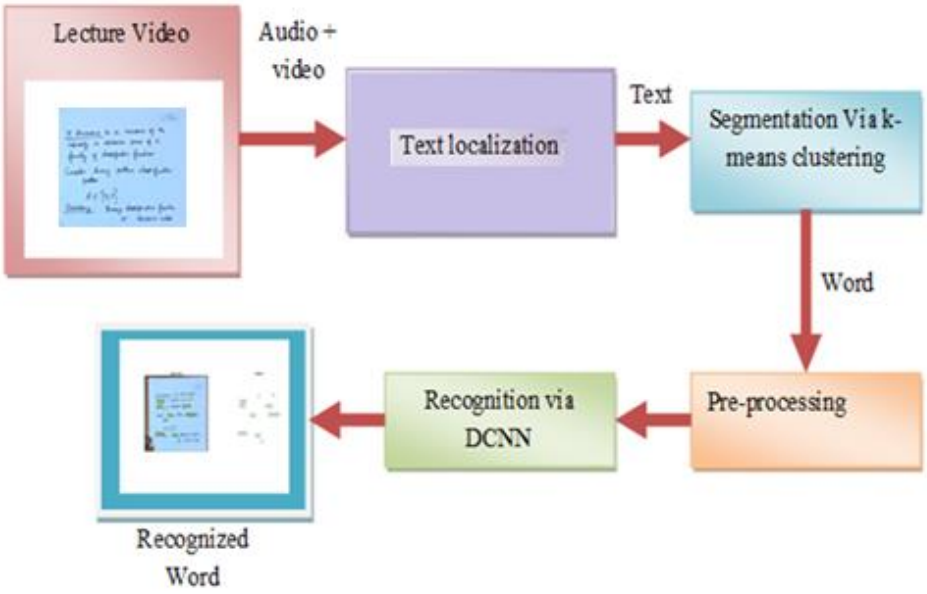

Fig. 1. Proposed Architecture for Handwritten Text Recognition In Lecture Video.

B. Segmentation

Clustering [24] is a scheme that split a group of data into a precise number of sets. A renowned technique among various clustering models is K-means clustering. The words in the texts $I_{text}(i,j)$ are segmented into k count of clusters.

Two separate stages are considered in this algorithm. In the initial stage k centers are selected randomly, in which the k value is previously fixed. Then, in the subsequent stage, every data object is moved to the closest center. Basically, the distance among every object of data and the cluster centers is determined using the Euclidean distance. This process is iterated until the termination criteria happen to a minimum. At the end of segmentation, the words $I_{word}(i,j)$ are extracted and they are subjected to pre-processing.

C. Pre-processing

The gathered words $I_{word}(i,j)$ are pre-processed for enhancing the accuracy of recognition. The steps involved in pre-processing are listed below:

Step 1: Initially, the collected words $I_{word}(i,j)$ are subjected to histogram equalization, which involves transforming the intensity values, i.e. stretching out the intensity range of the image.

Step 2: In order to convert the $I_{word}(i,j)$ image frames into binary image, black-and-white (B&W) images, the adaptive thresholding is used. The binarized image is denoted as $I_{binary}(i,j)$. Further, the weights of these binary images $I_{binary}(i,j)$ are subjected to recognition.

D. Recognition

The weights of the binarized images $I_{binary}(i,j)$ are subjected to recognition via DCNN [25]. Actually, DCNNs are CNNs which encompasses numerous layers and it follows a hierarchical principle. Usually, deep CNNs involve several wholly-connected layers, i.e., layers with dense weight matrix W. To do the recognition process, the outputs are exploited as inputs to a SVM or RF and the output phase can be a softmax function as specified in Eq. (1), in which 1 indicates a column vector of ones.

$$u = \sigma\left(I_{binary}(i,j)\right) = \frac{\exp\left(I_{binary}(i,j)\right)}{1^T \exp\left(I_{binary}(i,j)\right)} \tag{1}$$

Practically, the entire quantities are turned to be positive by exponential function and accordingly, the normalization assures that the entries of u adds up to 1. Generally, the softmax function is noticed as a multidimensional generalization of sigmoid function deployed in LR. This function termed as softmax is one of the $I_{binary}(i,j)_i$ entries, for instance, if $I_{binary}(i,j)_{b0}$ is superior over the others, then $I_{binary}(i,j)$ and therefore Eq. (2) is modeled. The above function acts as an indicator amongst the largest entry in x and thus, Eq. (3) is formulated.

$$u_{I_{binary}(i,j)_0} \approx 1 \text{ and } u_b \approx 0 \text{ for } b \neq b_0 \tag{2}$$

$$\lim_{\alpha\to\infty} I_{binary}(i,j)^T \sigma\left(\alpha . I_{binary}(i,j)\right) = \max\left(I_{binary}(i,j)\right) \tag{3}$$

In brief, DCNN performs the formulations as specified in Eq. (4)-Eq. (7), in which the output activation function f_x could be softmax, identity, or other function.

$$z^{(0)} = z \tag{4}$$

$$z^{(q)} = \pi\left(f\left(W^{(q)}\tilde{z}^{(q-1)}\right)\right) \quad for\ q = 1,....,Q_c \tag{5}$$

$$z^{(q)} = f\left(W^{(q)}\tilde{z}^{(q-1)}\right) \quad for\ q = Q_c + 1,....,Q \tag{6}$$

$$I_{binary}(i,j) = f_{I_{binary}(i,j)}\left(z^{(Q)}\right) \tag{7}$$

The matrix $W^{(q)}$ consists of $F^{(q-1)}+1$ columns and $F^{(q)}$ rows with $F^{(0)} = F$ and $F^{(q)}$ for $q > 0$ that is similar to the output count at q^{th} layer. The initial Q_c layers are convolution and the rest ones are wholly-connected. The diagrammatic representation of DCNN is exhibited in Fig. 2.

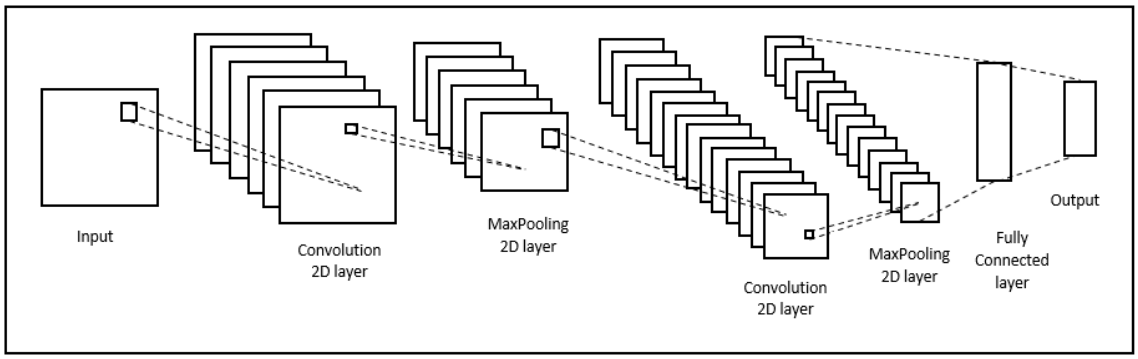

Fig. 2. The Proposed Architecture of DCNN for Handwritten Text Recognition.

IV. Results and Discussion

A. Simulation Procedure

The proposed video lecture recognition approach is implemented in PYTHON and the resultant acquired is noted. The dataset for the evaluation is downloaded from LectureVideoDB. In addition, two public datasets are utilized for pre-training the word recognition models.

- IAM Handwriting Database: It includes contributions from over 600 writers and comprises of 115,320 words in English.
- MJSynth: This is a synthetically generated dataset for scene text recognition. It contains 8 million training images and their corresponding ground truth words. The sample image collected and its segmented images are depicted in Fig. 3.

This evaluation is accomplished by varying the learning percentage (LP=60, 70) in terms of positive measures. The accuracy, sensitivity, specificity and precision come under the positive measures.

Accuracy(Acc): The accuracy indicates the accurate detection process. The mathematical formula for accuracy is expressed in Eq. (8).

$$Acc = \frac{TrP + TrN}{TrP + TrN + FrP + FrN} \quad (8)$$

PPV or precision: It represents the proportion of positive samples that were correctly classified to the total number of positive predicted samples. It is mathematically shown in Eq. (9).

$$PPV = \frac{TrP}{TrP + FrP} \quad (9)$$

Sensitivity: It is "positive correctly classified samples to the total number of positive samples". Mathematically, it is expressed in Eq. (10).

$$Sensitivity = \frac{TrP}{TrP + FrN} \quad (10)$$

Specificity: It is the "ratio of the correctly classified negative samples to the total number of negative samples". This can be mathematically defined in Eq. (11).

$$Specificity = \frac{TrN}{FrP + TrN} \quad (11)$$

where,

TrP - true positive.

TrN - true negative.

FrP - false positive and.

FrN - false negative.

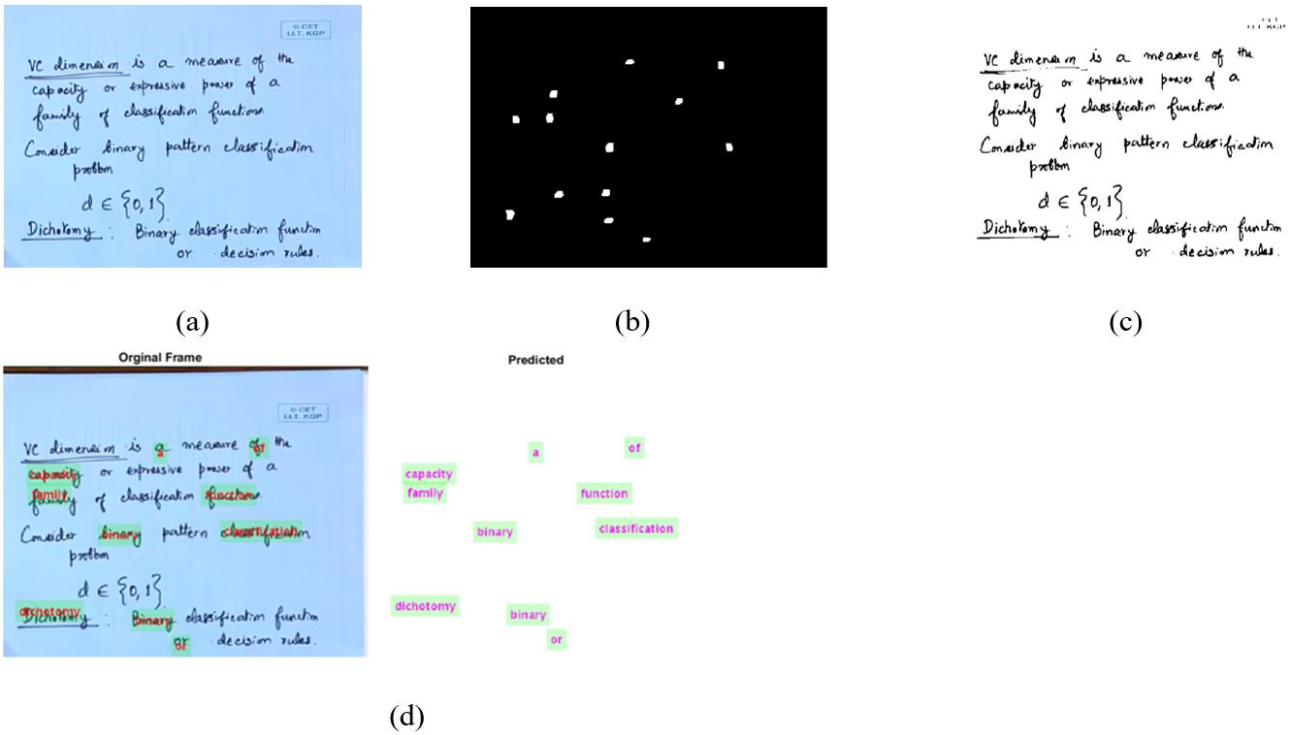

Fig. 3. Sample Images Showing (a) Input Frame (b) Black and White Image Frame (c) Text Localized Frame (d) Recognized Image Frame.

B. Evaluation

The evaluation is done by varying the training percentage (TP). The resultant acquired in terms of positive measures for diverse count frames is shown graphically. Table II and Fig. 4 shows the accuracy of the presented work. It is observed that the presented work has the highest accuracy as 89.3 for 500 count of frames corresponding to both LP =60 and 70. The resultant values of precision acquired are tabulated in Table III and is exhibited graphically in Fig. 5. The highest precision of 95 is obtained for LP=70 for 500 count of frames. The sensitivity of the presented work is highest for both LP=60 and LP=70 at 500 count of frames and the resultants acquired represented in Table IV and Fig. 6. The highest value of sensitivity obtained for the presented work at 500 count of frames is 91.2. The specificity of the presented work for LP=60 and LP=70 is exhibited graphically in Table V and Fig. 7. The specificity of the presented work at LP=70 has the highest value of 91.2 for 500 count of frames and it is higher for LP=70 for every variation in count of frames. The experimental results show that the modern DCNN model shows promising recognition accuracy. On a whole, it is observed the detection rate is higher for LP=70.

TABLE II. EVALUATION ON ACCURACY OF PRESENTED WORK FOR LP=60 AND LP=70

No. of frames	Accuracy for LP=60	Accuracy for LP=70
100	84.7	85.6
200	85.6	86.4
300	86.1	87.3
400	87.3	88.1
500	89.3	89.3

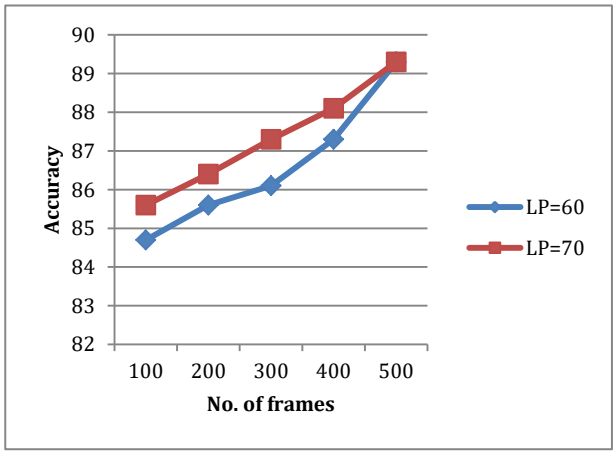

Fig. 4. Evaluation on Accuracy of Presented Work for LP=60 and LP=70.

TABLE III. EVALUATION ON PRECISION OF PRESENTED WORK FOR LP=60 AND LP=70

No. of frames	Precision for LP=60	Precision for LP=70
100	86.5	85
200	87	90
300	93.1	92.5
400	93.9	93
500	94.9	95

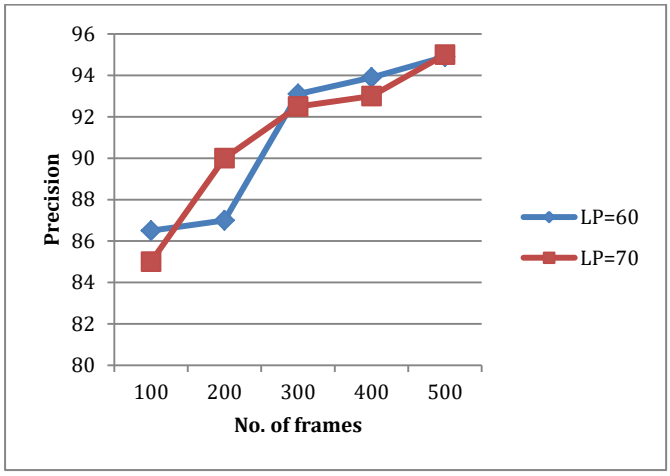

Fig. 5. Evaluation on Precision of Presented Work for LP=60 and LP=70.

TABLE IV. EVALUATION ON SENSITIVITY OF PRESENTED WORK FOR LP=60 AND LP=70

No. of frames	Sensitivity for LP=60	Sensitivity for LP=70
100	81.5	86.5
200	87.2	87
300	87.9	88
400	88.5	89
500	91.2	91.2

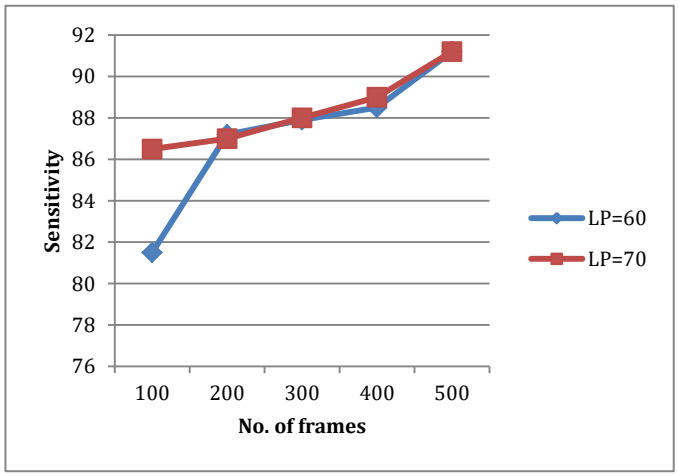

Fig. 6. Evaluation on Sensitivity of Presented Work for LP=60 and LP=70.

TABLE V. EVALUATION ON SPECIFICITY OF PRESENTED WORK FOR LP=60 AND LP=70

No. of frames	Specificity for LP=60	Specificity for LP=70
100	87.1	85.6
200	87.9	88.4
300	88.3	89.1
400	89.9	90.7
500	90.8	91.2

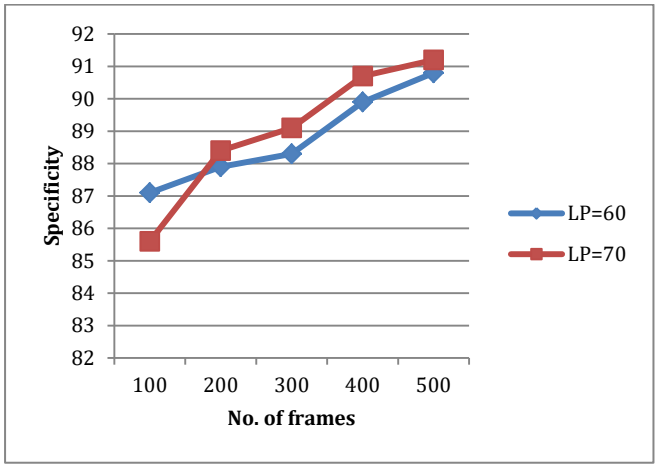

Fig. 7. Evaluation on Specificity of Presented Work for LP=60 and LP=70.

V. Conclusion

Although OCR has been considered as a solved problem, Handwritten Text Recognition a crucial component of OCR is still a challenging problem statement. The huge discrepancy in handwriting styles across different people and the poor quality of the handwritten text as compared to the typed or printed text pose substantial hurdles in converting the handwritten text into machine readable text. However, working on this crucial problem is important due to its pertinence in multiple industries such as healthcare, insurance and banking. This paper presented a novel text detection and recognition approach on the video lecture dataset by following four major phases, viz. (a) text localization, (b) segmentation and (c) pre-processing and (d) recognition. In the initial phase, the text localization in the lecture video frames were accomplished using the MRG algorithm. Then, the localized words were subjected to segmentation via the K-means clustering, in which the words from the detected text regions were segmented out. Subsequently, the segmented words will be pre-processed to avoid the blurriness artifacts as well. Finally, the pre-processed words are recognized using the DCNN. The performance of the proposed model is analysed in terms of certain performance measures like accuracy, precision, sensitivity and specificity to exhibit the supremacy of the proposed text detection and recognition in lecture video. Experimental results reveal that at LP=70, the presented work has the highest accuracy as 89.3 for 500 count of frames. In future, some fusion-based DCNN models will be explored for further achieving more accurate detection of handwritten text recognition. Also, a more convincing and robust training could be applied with added preprocessing techniques. We would focus on developing a more comprehensive model with a reduced amount of training time.

Acknowledgment

This work is supported by University Grants Commission (UGC) under Minor Research Project titled "Fast Content Based Search, Navigation and Retrieval system for E-Learning". Project Id: F.No:4-4/2015(MRP/UGC-SERO).

References

[1] Siti N. H. Hadie, Anna A. Simok, Shamsi A. Shamsuddin, Jamilah A. Mohammad, "Determining the impact of pre-lecture educational video on comprehension of a difficult gross anatomy lecture", Journal of Taibah University Medical Sciences, vol.14, no.4, pp.395-401, August 2019. https://doi.org/10.1016/j.jtumed.2019.06.008.

[2] Zhongling Pi, Yi Zhang, Fangfang Zhu, Ke Xu, Weiping Hu, "Instructors' pointing gestures improve learning regardless of their use of directed gaze in video lectures", Computers & Education, vol.128, pp.345-352, January 2019. https://doi.org/10.1016/j.compedu.2018.10.006.

[3] Hamidreza Aghababaeian, Ladan Araghi Ahvazi, Ahmad Moosavi, Sadegh Ahmadi Mazhin, Leila Kalani, "Triage live lecture versus triage video podcast in pre-hospital students' education", African Journal of Emergency Medicine, vol. 9, no. 2, pp. 81-86, June 2019. https://doi.org/10.1016/j.afjem.2018.12.001.

[4] Alexander R. Toftness, Shana K. Carpenter, Sierra Lauber, Laura Mickes, "The Limited Effects of Prequestions on Learning from Authentic Lecture Videos", Journal of Applied Research in Memory and Cognition, vol. 7, no. 3, pp.370-378, September 2018. https://doi.org/10.1016/j.jarmac.2018.06.003.

[5] Aydın Sarıhan, Neşe Colak Oray, Birdal Güllüpınar, Sedat Yanturali, Berna Musal, "The comparison of the efficiency of traditional lectures to video-supported lectures within the training of the Emergency Medicine residents", Turkish Journal of Emergency Medicine, vol.16, no.3, pp.107-111, September 2016. 10.1016/j.tjem.2016.07.002.

[6] Marco Furini, "On gamifying the transcription of digital video lectures", Entertainment Computing, vol.14, pp. 23-31, May 2016. https://doi.org/10.1016/j.entcom.2015.08.002.

[7] I-Chun Hung, Kinshuk, Nian-Shing Chen, "Embodied interactive video lectures for improving learning comprehension and retention", Computers & Education, vol.117, pp. 116-131, February 2018. https://doi.org/10.1016/j.compedu.2017.10.005.

[8] Kep Kee Loh, Benjamin Zhi Hui Tan, Stephen Wee Hun Lim, "Media multitasking predicts video-recorded lecture learning performance through mind wandering tendencies", Computers in Human Behavior, vol. 63, pp. 943-947, October 2016. https://doi.org/10.1016/j.chb.2016.06.030.

[9] Andrew T. Stull, Logan Fiorella, Richard E. Mayer, "An eye-tracking analysis of instructor presence in video lectures", Computers in Human Behavior, vol.88, pp. 263-272, November 2018. https://doi.org/10.1016/j.chb.2018.07.019.

[10] Juan Daniel Valor Miró, Joan Albert Silvestre-Cerdà, Jorge Civera, Carlos Turró, Alfons Juan, "Efficiency and usability study of innovative computer-aided transcription strategies for video lecture repositories", Speech Communication, vol. 74, pp. 65-75, November 2015. https://doi.org/10.1016/j.specom.2015.09.006.

[11] Rabab El-Sayed Hassan El-Sayed, Samar El-Hoseiny Abd El-Raouf El-Sayed, "Video-based lectures: An emerging paradigm for teaching human anatomy and physiology to student nurses", Alexandria Journal of Medicine, vol. 49, no. 3, pp. 215-222, September 2013. https://doi.org/10.1016/j.ajme.2012.11.002.

[12] Feng Wang, Chong-Wah Ngo, Ting-Chuen Pong, "Structuring low-quality videotaped lectures for cross-reference browsing by video text analysis", Pattern Recognition, vol. 41, no. 10, pp. 3257-3269, October 2008. https://doi.org/10.1016/j.patcog.2008.03.024.

[13] Karl K. Szpunar, Helen G. Jing, Daniel L. Schacter, "Overcoming overconfidence in learning from video-recorded lectures: Implications of interpolated testing for online education",Journal of Applied Research in Memory and Cognition, vol. 3, no. 3, pp.161-164, September 2014. https://doi.org/10.1016/j.jarmac.2014.02.001.

[14] Alendra Lyons, Stephen Reysen, Lindsey Pierce, "Video lecture format, student technological efficacy, and social presence in online courses" Computers in Human Behavior, vol.28, no.1, pp.181-186, January 2012. https://doi.org/10.1016/j.chb.2011.08.025.

[15] Haojin Yang,Bernhard Quehl,Harald Sack,"A framework for improved video text detection and recognition",Multimedia Tools and Applications, vol.69,no.1,pp 217–245,March 2014. 10.1007/s11042-012-1250-6.

[16] N. Poornima & B. Saleena,"Multi-modal features and correlation incorporated Naive Bayes classifier for a semantic-enriched lecture

video retrieval system",The Imaging Science Journal, Jan 2018. 10.1080/13682199.2017.1419549.

[17] B. Urala Kota, K. Davila, A. Stone, S. Setlur and V. Govindaraju, "Automated Detection of Handwritten Whiteboard Content in Lecture Videos for Summarization," 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), Niagara Falls, NY, pp. 19-24, 2018. 10.1109/ICFHR-2018.2018.00013.

[18] M. Husain, Meena S M, A. K. Sabarad, H. Hebballi, S. M. Nagaralli and S. Shetty, "Counting occurrences of textual words in lecture video frames using Apache Hadoop Framework," 2015 IEEE International Advance Computing Conference (IACC), Banglore, pp. 1144-1147, 2015. 10.1109/IADCC.2015.7154882.

[19] Kartik Dutta, Minesh Mathew, Praveen Krishnan and C.V. Jawahar,"Localizing and Recognizing Text in Lecture Videos",sep 2019. 10.1109/ICFHR-2018.2018.00049.

[20] Derek Miller,"Leveraging BERT for Extractive Text Summarization on Lectures",june 2019. arXiv:1906.04165**.**

[21] Surabhi Pagar, Gorakshanath Gagare,"Multimedia based information retrieval approach for lecture video indexing based on video segmentation and Optical Character Recognition", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4,no. 12, December 2015. 10.17148/IJARCCE.2015.41273.

[22] M. Husain and S. M. Meena, "Multimodal Fusion of Speech and Text using Semi-supervised LDA for Indexing Lecture Videos," 2019 National Conference on Communications (NCC), Bangalore, India , pp. 1-6, 2019. 10.1109/NCC.2019.8732253.

[23] Junhua Chen, Miao Tian, Xingming Qi, Wenxing Wang, Youjun Liu,"A solution to reconstruct cross-cut shredded text documents based on constrained seed K-means algorithm and ant colony algorithm",Expert Systems with Applications, vol.127,pp.35-46,2019. https://doi.org/10.1016/j.eswa.2019.02.039.

[24] Feng Cheng, Shi-Lin Wang, Xi-Zi Wang, Alan Wee-Chung Liew, Gong-Shen Liu,"A global and local context integration DCNN for adult image classification",Pattern Recognition, vol.96,December 2019. https://doi.org/10.1016/j.patcog.2019.106983.

[25] U.-V. Marti and H. Bunke, “The IAM-database: an english sentence database for offline handwriting recognition,” IJDAR, 2002.10.1007/s100320200071.