DOI: https://doi.org/10.24113/ijoscience.v5i7.218

CNN Filter Based Text Region Segmentation from Lecture Video and Extraction using Neuro OCR

Ashima Godha
M.Tech (CTA)
RKDF School of Engineering
Indore, Madhya Pradesh, India
ashimagodha@maill.com

Puja Trivedi
Assistant Professor (CSE)
RKDF School of Engineering
Indore, Madhya Pradesh, India
pujatrivedi13@yahoo.com

Abstract

Lecture videos are rich with textual information and to be able to understand the text is quite useful for larger video understanding/analysis applications. Though text recognition from images have been an active research area in computer vision, text in lecture videos has mostly been overlooked. In this paper, text extraction from lecture videos are focused. For text extraction from different types of lecture videos such as slides, whiteboard lecture videos, paper lecture videos, etc. The text extraction, the text regions are segmented in video frames and extracted using recurrent neural network based OCR. And finally, the extracted text is converted into audio for ease of convenience. The designed algorithm is tested on different videos from different lectures. The experimental results show that the proposed methodology is quite efficient over existing work.

***Keywords*—** Text Detection, Contrast Enhancement, CNN Filter, Bounding Box, RNN, Optical Character Recognition

I. INTRODUCTION

Visual text is one in all the foremost necessary strategies of communication utilized by human beings and is wide utilized in our everyday life. Hence, interpreting this textual data is of great significance. Human beings inherently get the ability to find and acknowledge the textual content in their surroundings whereas it's a challenging problem for computer systems. The field of text detection focuses on detecting text embedded in images and videos with the help of computer systems[1].

Researchers have made significant progress in detecting the text from images of machine printed documents; on the other hand detecting the text in natural scenes is still a new topic for research.

Text detection and recognition in unconstrained environments is a challenging computer vision problem. Such functionality can play valuable role in numerous real-world applications, ranging from video indexing, assistive technology for the visually impaired, automatic localization for businesses, and robotic navigation. In recent years, the problem of scene text detection and recognition in natural images has received increasing attentions from the computer vision community. As a result, the domain has enjoyed significant advances on an increasing number of datasets of public scene text benchmark.

In the last decade tele-teaching and lecture video portals have become more and more popular. The amount of lecture video data available on the World Wide Web (WWW) is constantly growing. Thus, the challenge of finding lecture video data on the WWW or within lecture video libraries has become a very important and challenging task. We focus our research on such lecture recordings having been produced by state-of-the-art lecture recording systems. With this kind of a system, we are able to record the lecture video such that we combine two video streams: the main scene of lecturers which is recorded by using a video camera, and the second which captures the images projected onto the screen during the lecture through a frame grabber. We synchronize the frame grabber output with the video camera so that each slide can be linked to a video recording segment. In this way, indexing two-part lecture videos can be performed by indexing the slide videos only.

Content-based retrieval within video data can be performed by automated extraction of textual metadata. Techniques from OCR, which focus on high resolution scans of printed (text) documents, have to be improved and adapted to be also applicable for video OCR. In video OCR, the text within video frame has to be automated localized and separated from its background and the image quality enhancement have to be applied before deepOCR procedures can process the text successfully.

II. Related Work

Yin et al. [2] proposed a Maximally Stable Extremal Regions (MSERs) based method to extract text form images. Extremal region can be viewed as a connected component in an image whose pixels can have either lower or higher intensity than its outer boundary pixels. In this method, first character candidates, i.e., pixels containing text are extracted based on their difference in variance. Extremal regions are extracted in the form of a rooted tree for the whole image. Second, character candidates are merged into text candidates by the single-link clustering algorithm. Distance metric learning algorithm can automatically learn about distance weights, i.e., distance between two text pixels and clustering threshold and these are used in clustering character candidates. Third, non-text candidates are eliminated by using character classifier. The width, height, smoothness, and aspect ratio of Text region are used by character classifiers.

Le et al. [3], present a learning-based approach which involves three steps. First, in preprocessing the image is binarized using Otsu binarization and the connected component are extracted based on various features like elongation, solidity, height, and width of the connected component, Hue moment which describes the shape of the connected component, stroke width of connected components for discriminating text from non-text. These features provide shape and location information of connected components. Then Adaboosting Decision Trees is used to label these connected components into non-text or text component. Then post processing to correct some connected components which are labeled incorrectly by the classifier.

Vidyarthi et al. [4] purposed a method in which first colored image is converted into gray-level image and histogram is drawn. Otsu method uses this histogram to find global threshold value. This value is further used to find out connected components. The components which are close to each other are merged to form one component. The height histogram is constructed by using the heights of the bounding boxes. Variance is selected for verification of the text and non-text. Finally, filtered components are verified a text on text.

Zhong et al. [10] purposed system to localize text in the color images, using DCT. Image patches of high horizontal spatial intensity variation are detected as text components, morphological operations are applied to those patches and thresholding spectrum energy is used to verify the regions.

Khodadadi et al. [17] introduced a method for extraction of text from images based on stroke filters and color histogram. First, stroke filter is applied to the input image and then local and global threshold values are calculated on stroke filter output. The image is then divided into text blocks by making use of horizontal and vertical projections of binary pixels. The text blocks with overlap or close blocks are merged to form a new text block. It is assumed that the text in every block has the same color, so histogram of color channels are used for extraction of the text characters in the candidate blocks for the text and background areas.

Kumar et al. [18] suggested an algorithm that consists of three stages: first, images are converted into an edge map by using line edge detector which uses vertical and horizontal masks. Second, for identification of the text regions in images, the vertical and the horizontal projection profiles are analyzed and the edged image is then converted into the localized image. After localization, in order to obtain text area segmentation process is done on the localized image by using median filters.

III. Proposed Methodology

A novel Connecting Character based text recognition and extraction algorithm is designed which uses convolution neural network based for test candidate segmentation and its recognition. To allow for detecting small letters in video lectures of limited resolution or blurred Image, the complimentary properties of Lucy-Richardson Algorithm and canny edge Algorithm is used. Further CNN filter is used and applied to efficiently obtain more reliable results. Finally, texts are clustered into word and neural network trained OCR is used to extract text and further convert into audio. The proposed algorithm, illustrated in Figure 1, is divided into basic steps i.e. text area detection and text recognition.

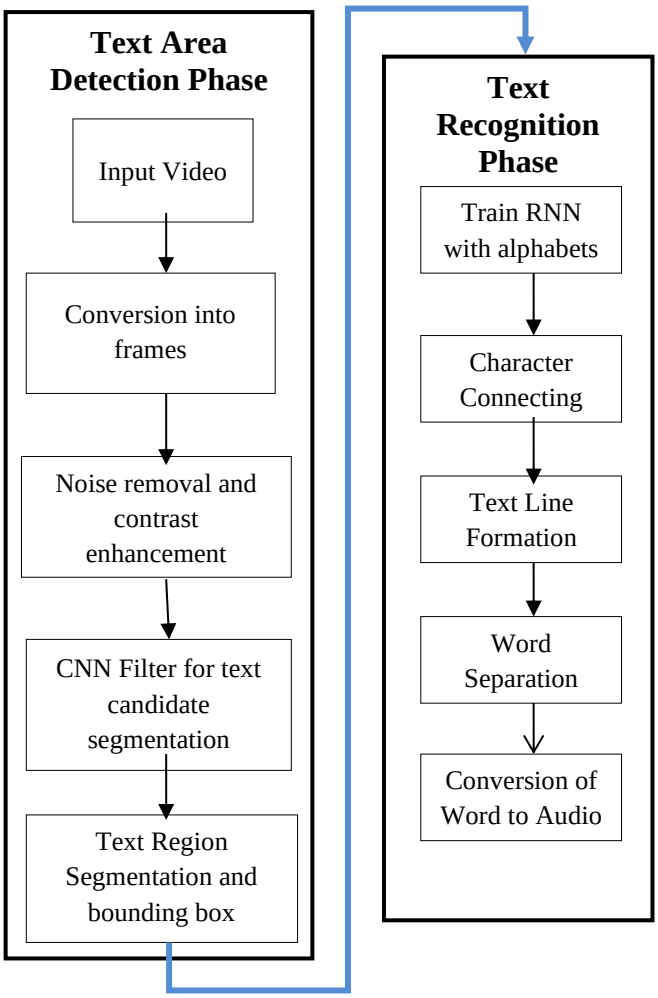


Figure 1: Flow chart of proposed Algorithm

The overall flow of the proposed algorithm is divided into following stages:

- *Conversion of Video into Frames:* Inn this stage the input video are converted into frames. Different types of videos are taken such as presentation slide videos, white board videos, white paper videos, etc.
- *Segmentation and Detection of Text Candidates:*

  Step 1: Noise Removal and De-blurring Image

  Due to imperfections in the image capturing procedure, on the other hand, the recorded image invariably represents a corrupted version of the original image. The degradation results in blurring of image, which affects identification and retrieval of the essential information in the image frames. It can be the result of relative motion between the camera and the original image frame, by an out of focus of optical system, atmospheric disturbances and deviation in the optical system. Noise introduced by the medium through which the image is created can also cause degradation. The degradation phenomenon of the acquired images results severe cost-effective loss. Consequently, restoring the corrupted images is an urgent task in order to expand uses of the images. In this step the proposed algorithm uses Lucy-Richardson Algorithm is used for noise removal and de-blurring the blurred image.

  Step 2: Contrast Adjustment and Conversion RGB image to Grayscale Image

  Image enhancement techniques are used to improve an image. Intensity adjustment is an image enhancement technique that maps an image's intensity values to a new range. In this step, contrast or brightness level of the input image is enhanced. Further in this step RGB Image is converted into gray scale Image. The rgb2gray method function transforms RGB images to grayscale by removing the information of hue and saturation at the same time as retaining the luminance.

  Step 3: CNN Filter based Text Candidates Recognition

  In this step, CNN filter is used to segment text region. Word region is spotted to localize specific words that are given in a lexicon. End-to-end recognition concerns both detection and recognition. o not match any of the given words.

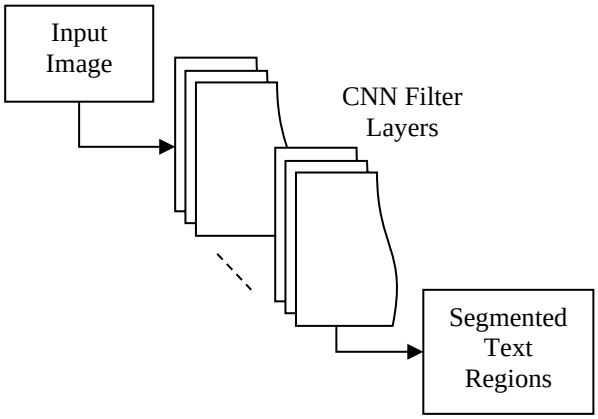


Figure 2: Text Region Segmentation using CNN Filter

This paper had adopted the CNN model as our text candidate recognizer. CRNN uses softmax as its output layer, which estimates sequence probability conditioned on input image, i.e. p(t|I), where I is an input image and t represents a text region.

Softmax layer generates the probability as a matching score, which measures the compatibility

of an image to a particular word. The detection score is then the maximum score among all.

And thus, filtered candidates are further connected using bounding box technique. The layers are connected and converted into convolutional layers by parameters and pooled together and output layer is connected to softmax layer.

- *Extraction of Text and Conversion to Audio*

  Step 1: Text Extraction and recognition

  Text Extraction and recognition by trained neural network OCR. The recurrent neural network is used to design the trained OCR. RNN is trained with different style alphabets and numbers. This trained network is used to read the characters in the image and convert into text.

  Step 2: Text line formation and Word separation

  Subsequent in this stage of the algorithm locates lines of text among the detected candidate regions. This allows the total number of CCs to be reduced, removing non-character CCs and therefore raising the probabilities for higher accuracy. As a final step, text lines are split into individual words by classifying, by Neural network OCR, the inter letter distances into two classes: the character spacing and the word spacing.

  Step 3: Conversion to Text to Audio

  Further the word separated by above steps are further converted into audio for easy understanding.

IV. RESULT ANALYSIS

Different styles/modalities of text present in the dataset are the following:

Slides: This set includes the frames where a presentation slide is shown. The text in this case is mostly born digital text, which is relatively more legible and free from distortions.

Whiteboard: This category generally encompasses frames where either the instructor is using a physical white board along with markers to explain a concept or is using a digital pad to write on a personal computer.

These videos are created from youtube [20], nptel [21] and khan academy [22]. Some of the video frames are shown in figure 2.

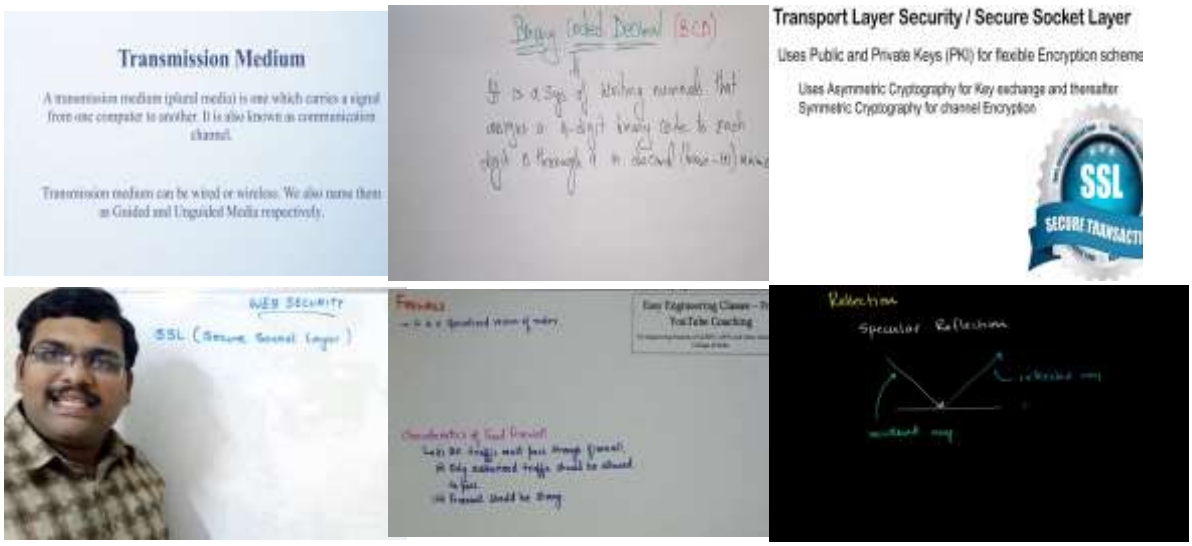
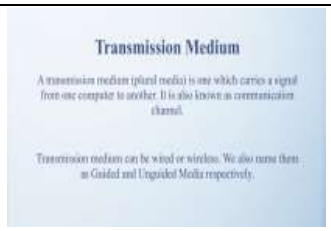
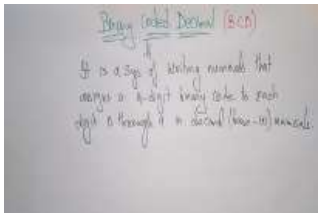
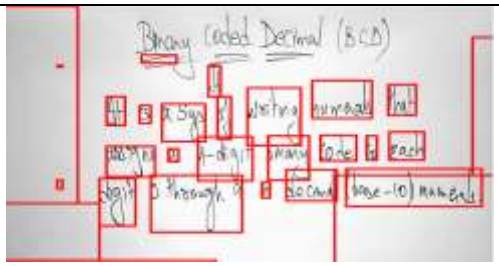
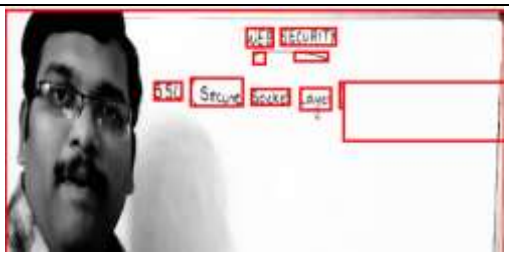
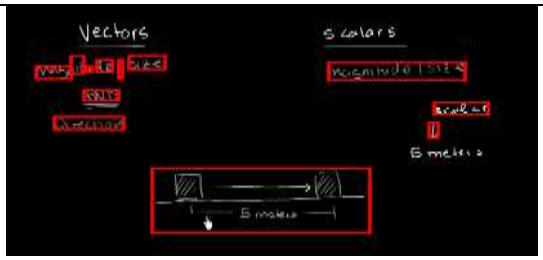


Figure 3: Sample frames from the Lecture video dataset

Table I: Sample Text Segmented Cases

Input Video Frames	Segmented and Bounded Text Candidates
	
	
	
	

After saving the frames, CNN filter model is used to detect the text regions and create the word bounding boxes for all the saved frames. Figure 3 shows a few cropped word images that are part of this research work. As one can notice, the word images posses different styles and also contains blurriness artifacts.

Table I shows the performance of the proposed methodology on different modalities of the Lecture Videos. In most of the cases, the detector split a single word into multiple bounding boxes and hence increasing the number of false positives. This occurs more in case of handwritten text where breaks in the cursive writing are confused with spaces between words. Table II-IV Shows the performance analysis of CNN segmented images in terms of recall, precision and f_measure with existing work.

Table II: Word Segmentation Recall Performance of Various Video Samples

Data Types	Dutta [19]	Proposed
Slides	69	99.98
Whiteboard	37	66.2
Paper	51	66.58
Blackboard	62	86.61

Table III: Word Segmentation Precision Performance of Various Video Samples

Data Types	Dutta [19]	Proposed
Slides	96	100
Whiteboard	47	95.48
Paper	67	94.78
Blackboard	69	57.79

Table IV: Word Segmentation F_Score Performance of Various Video Samples

Data Types	Dutta [19]	Proposed
Slides	80	99.99
Whiteboard	41	78.19
Paper	58	78.21
Blackboard	66	69.32

V. CONCLUSION

In this paper, the work is focused on text detection and recognition on the new lecture video dataset, by using training set of scene text and handwritten text. The dataset is prepared from different resources and CNN filter based text candidate regions are segmented and extracted using recurrent neural network. In future, we plan to work towards developing methods which can work well on settings where text of multiple modalities appear together in complex and low resolution images.

REFERENCES

[1] Keechul Jung, Kwang In Kim, Anil K. Jain "Text information extraction in images and video: a survey", Elsevier, Pattern Recognition 37 (2004).

[2] Xu-Cheng Yin, Member, IEEE, Xuwang Yin, Kaizhu Huang, and Hong-Wei Hao "Robust Text Detection in Natural Scene Images" IEEE transactions on pattern analysis and machine intelligence, Vol. 36, No. 5, (2014).

[3] Viet Phuong Le, Nibal Nayef, Muriel Visani, Jean-Marc Ogier and Cao De Trant "Text and Non-text Segmentation based on Connected Component Features" IEEE, 13th International Conference on Document Analysis and Recognition (ICDAR), (2015).

[4] Ankit Vidyarthi, Namita Mittal, Ankita Kansal, "Text and Non-Text Region Identification Using Texture and Connected Components", International Conference on Signal Propagation and Computer Technology (ICSPCT), IEEE (2014).

[5] Yingying Zhu, Cong Yao, Xiang Bai "Scene text detection and recognition: recent advances and future trends" Front. Comput. Sci., (2016).

[6] Qixiang Ye, and David Doermann, "Text Detection and Recognition in Imagery: A Survey" IEEE transactions on pattern analysis and machine intelligence, Vol. 37, No. 7, (2015).

[7] N. Senthilkumaran and R. Rajesh, "Edge Detection Techniques for Image Segmentation – A Survey of Soft Computing Approaches", International Journal of Recent Trends in Engineering, Vol. 1, No. 2, (2009).

[8] Zhong Y, Karu K, Jain A K. "Locating text in complex color images." in Proceedings of the 3rd IEEE Conference on Document Analysis and Recognition, pp-146–149, IEEE (1995).

[9] Kim K I, Jung K, Kim J H. "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm." IEEE Transactions on Pattern Analysis and Machine Intelligence, pp-1631–1639, IEEE(2003).

[10] Y. Zhong, H. J. Zhang, and A. K. Jain, "Automatic caption localization in compressed video," IEEE Trans. Pattern Anal. Mach. Intell., Vol. 22, No. 4, pp. 385–392, IEEE (2000).

[11] Li H, Doermann D, Kia O. "Automatic text detection and tracking in digital video.", 9(1): 147–156, IEEE Transactions on Image Processing, (2000).

[12] K. I. Kim, K. Jung, and H. Kim, "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm," IEEE Trans. Pattern Anal. Mach. Intell., Vol. 25, No. 12, pp. 1631–1639, (2003).

[13] Chong Yu, Yonghong Song, Quan Meng, Yuanlin Zhang, Yang Liu, "Text detection and recognition in natural scene with edge analysis", IET Comput. Vis., Vol. 9, Issue. 4, pp. 603– 613, (2015).

[14] Chucai Yi, Ying Li Tian, "Text String Detection From Natural Scenes by Structure-Based Partition and Grouping", Vol. 20, No. 9, IEEE Transactions on Image Processing (2011).

[15] Shijian Lu, Tao Chen, Shangxuan Tian, Joo-Hwee Lim, Chew-Lim Tan, "Scene text extraction based on edges and support vector regression", 18:125–135, IJDAR (2015).

[16] K. C. Kim, H. R. Byun, Y. J. Song, Y. W. Choi, S. Y. Chi, K. K. Kim, Y. K. Chung, "Scene Text Extraction in Natural Scene Images using Hierarchical Feature Combining and Verification", 17th International Conference on Pattern Recognition (ICPR'04), IEEE (2004).

[17] Mohammad Khodadadi, and Alireza Behrad, "Text Localization, Extraction and Inpainting in Color Images", IEEE, 20th Iranian Conference on Electrical Engineering, (ICEE2012), (2012).

[18] Anubhav Kumar "An Efficient Text Extraction Algorithm in Complex Images", IEEE, (2013).

[19] Kartik Dutta, Minesh Mathew, Praveen Krishnan and C.V. Jawahar, "Localizing and Recognizing Text in Lecture Videos", International Conference on Frontiers in Handwriting Recognition, 2018.

[20] https://www.youtube.com/

[21] https://nptel.ac.in/

[22] https://www.khanacademy.org/