Text Extraction from Video using Deep Learning

ASSOC. PROF. L. RASIKANNAN[1], K. GUNAL[2], S. SABARINATHAN[3], S. VIGNESWARAN[4]

[1234] *Dept. of Computer Science and Engineering, Government College of Engineering Srirangam, Tamilnadu, India*

---***---

Abstract - ***In the modern-day virtual landscape, video content has expanded unexpectedly in quantity across diverse structures, imparting a wealth of records. However, extracting textual records from these videos accurately and efficaciously remains a great assignment. This paper proposes an approach to extract textual content from video content by way of employing a combination of Optical Character Recognition (OCR) algorithms and Convolutional Recurrent Neural Network (CRNN). By leveraging the strengths of both OCR and CRNN, our approach ambitions to beautify the accuracy and performance of text extraction from video lectures, tutorials, and educational content material. The extracted textual content serves as a precious reference for college students, enriching their mastering enjoy. This study contributes to unlocking the untapped potential of textual records embedded within video content material, thereby facilitating better access to knowledge inside the virtual age.***

Key Words: Optical Character Recognition (OCR), Convolutional Recurrent Neural Network (CRNN), Text Extraction, Video Content, Information Retrieval

1.INTRODUCTION

In today's digital age, video content has become ubiquitous on different platforms and offers a large storage of information to users around the world. However, despite the wealth of knowledge contained in these videos, accessing and utilizing the textual information they contain remains a challenging challenge. Accurate and efficient text extraction from video content is key to increasing the accessibility and usability of this information, especially in educational contexts such as lectures, tutorials, and instructional videos.

Traditional methods of extracting text from videos often rely on optical character recognition (OCR) techniques, which can struggle with low-quality images, distorted text, and complex backgrounds. OCR algorithms can struggle to accurately recognize text that deviates from standard fonts or styles, such as handwritten text, artistic fonts, or distorted characters. This limitation may lead to errors or incomplete text extraction. Additionally, OCR alone may not capture context or structure well, especially in scenarios where text is embedded in dynamic visuals. On the other hand, convolutional neural networks (CNNs) have shown remarkable success in image recognition tasks, including text detection and extraction. By utilizing both OCR and CNN algorithms, we can potentially improve the accuracy and efficiency of text extraction from video content. Where OCR is used for text detection from frames and segmentation, image and CNN are used for text recognition. CNNs trained on large datasets can generalize well across different fonts and writing styles, making the combined approach versatile and applicable to a wide variety of video content.

1.1 RELATED WORK

K. S. Raghunandan and Palaiahnakote Shivakumara [1] addresses robust text detection and recognition in multi-script-oriented images. Previous research has employed techniques including bit plane slicing, Iterative Nearest Neighbor Symmetry (INNS), Mutual Nearest Neighbor Pair (MNNP) components, character detection using fixed windows, contourlet wavelet features with SVM classifier, and Hidden Markov Models (HMM) for recognition.

Xu-Cheng Yin and Xuwang Yin [2] presents a method for accurate text detection in natural scene images. Their approach employs a fast-pruning algorithm to extract Maximally Stable Extremal Regions (MSERs) as character candidates, followed by grouping them into text candidates using single-link clustering. Automatic learning of distance metrics and thresholds is incorporated, and posterior probabilities of text candidates are estimated using a character classifier to eliminate non-text regions. Evaluation of the ICDAR 2011 Robust Reading Competition database demonstrates an f-measure of over 76%, surpassing state-of-the-art methods, with further validation on various databases confirming its effectiveness.

Pinaki Nath Chowdhury, and Palaiahnakote Shivakumara [3] presents a new method for detecting text on human bodies in sports images, addressing challenges such as poor image quality and diverse camera viewpoints. Unlike conventional methods, it employs an end-to-end episodic learning approach that detects clothing regions using a Residual Network (ResNet) and Pyramidal Pooling Module (PPM) for spatial attention mapping. Text detection is performed using the Progressive Scalable Expansion Algorithm (PSE). Evaluation of various datasets demonstrates superior precision and F1-score compared to existing methods, confirming effectiveness across different inputs.

2 METHODOLOGIES

Our proposed approach for extracting text from video content leverages the strengths of optical character recognition (OCR) for detecting text regions and Convolutional Recurrent Neural Networks (CRNN) for text recognition, ensuring robust and accurate text extraction.

2.1 Preprocessing:

It is the first crucial step in our proposed videotext extraction approach. Basically, it involves cleaning and preparing video data to make it more suitable for both OCR (text detection) and CRNN (text recognition) to perform their tasks effectively. Here's a breakdown of what happens during preprocessing:

A. Video segmentation:

The video is first divided into individual frames. Each frame represents a single still image captured at a specific point in time during video recording. This allows us to analyze the content of the text in each frame independently its show in the Fig.1 below.

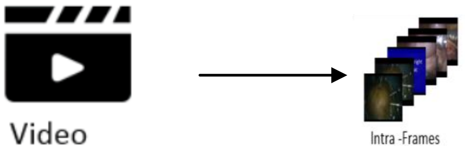


Fig.1. Video to frames

B. Grayscale conversion:

The color information in the video is not critical for text extraction, converting images to grayscale can be beneficial. Grayscale images only contain luminance (shades of gray) information, which can sometimes simplify the image and make it easier for OCR and CRNN to focus on text elements. Grayscale images or videos are often used in various applications such as image processing, computer vision, and medical imaging, where color information may not be necessary or may even be a distraction. It is shown in the Fig.2.

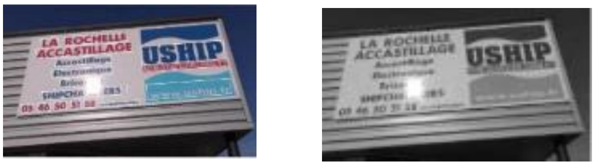


Fig.2. The image contains text to Grayscale image

C. Noise reduction:

Videos can be prone to noise from various sources, such as camera imperfections or compression artifacts. Noise reduction techniques aim to remove these unwanted elements from frames. This helps improve overall image quality and clarity, resulting in more accurate text extraction. It's shown in Fig.3 and Fig.4 below.

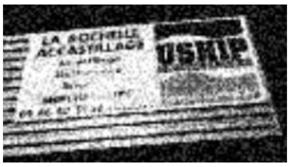


Fig.3. Before Noise Reduction

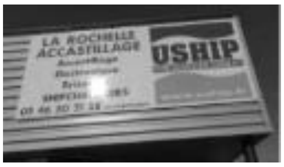


Fig.4. After Noise Reduction

D. Contrast Enhancement:

The goal of this technique is to improve the contrast between the text and the background in each image. By increasing the contrast, the text becomes more prominent, making it easier for both OCR and CRNN to recognize and identify characters. It is shown in Fig.5 below.

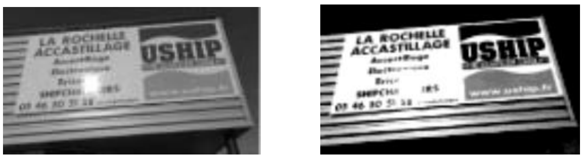


Fig.5. Contrast Enhancement

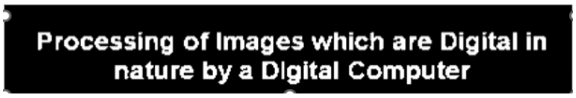
2.2 Text detection using OCR:

Video text extraction uses OCR as an initial search to identify potential regions of each frame that may contain text. The breakdown of this step is as follows:

A. Detecting areas of text:

The OCR engine uses a variety of techniques to identify areas of the frame that exhibit similar text. These techniques will include the following:

I. Thresholding: Convert a grayscale image to a binary image. Pixels with values higher than a threshold are considered part of the text, while other pixels are considered background. It is shown in the Fig.6.
ii.Contiguous object detection: This technique identifies a contiguous area of pixels that can represent a single character or line of text in a binary image. It is shown in the Fig.7.
iii.Morphological operations: These operations can be used to correct the position of text by removing noise, filling small gaps in characters, or smoothing edges. It is shown in the Fig.8.



Processing of Images which are Digital in nature by a Digital Computer

Fig.6. Thresholding of Text

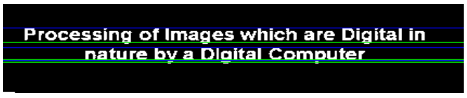

Processing of Images which are Digital in
nature by a Digital Computer

Fig.7. Line Segmentation

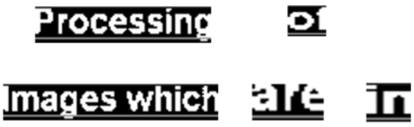


Fig.8.Word Segmentation

2.3. Text recognition using CRNN:

The proposed approach for video text extraction, takes center stage after OCR has identified potential text regions within each video frame. Here's how CRNNs work:

A. The Power of Pre-trained CRNN Model:

We utilize a pre-trained CRNN model specialized in text recognition tasks. This model undergoes training on a vast dataset consisting of 200,000 images annotated with corresponding labels. Through this training process, the CRNN gains proficiency in discerning complex patterns and features within images indicative of individual characters.

B. Extracting the Text:

Each text region identified by OCR and delineated by a bounding box within the video frame is extracted.
The cropped image containing the potential text serves as input for the CRNN model.

C. Feature Extraction:

The CRNN model analyzes cropped images pixel by pixel, focusing not solely on letter recognition but also on extracting contextual features crucial for understanding text. Features such as strokes, curves, and spatial relationships between pixels contribute to the differentiation between similar characters.

D. Prediction Time:

Upon receiving the visual image, the CRNN model utilizes its learned knowledge to predict the most relevant features extracted from the image, effectively converting the visual representation of the text into its corresponding textual form.

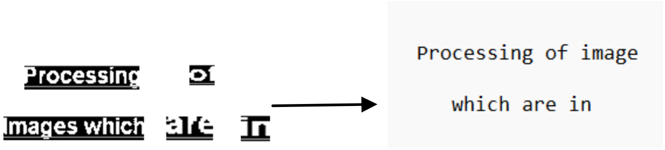


Fig.9. Predicted Text using CRNN model

2.4 Integration and output:

Recognized text from all frames is integrated based on their frame number in the video. This can include techniques such as hidden Markov models (HMMs) to handle potential overlapping or sequential text elements across frames.
The final output is the extracted text content from the entire video, possibly including timestamps or links to images for each piece of text. It is shown in the Fig.10.

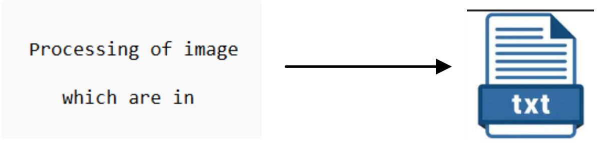


Fig.10. Generation of text Document

3. EXPERIMENTAL EVALUATION

3.1 Environment specifications:

The experimental configuration employed in the study. The research work takes place on an AMD Ryzen 7 CPU. Furthermore, the machine has 16GB of RAM and an Nvidia graphics card. The models are constructed using Python and are executed using deep learning frameworks such as Kera's and TensorFlow.

3.2 Dataset:

This paper is performed using mjsynth named dataset, which contains images of text with its label value. The image in the dataset is contain different kind of text image like different font, size and positions.

The dataset used in the study. In the dataset, text images have a total of 150000 images the dataset was pre-split into 2 sections: train data, which contains approximately 80% of the total images, and valid data (considered as test data), which contains approximately 20% of the total images.

First extract all of the images from each directory. The images in the dataset are check the height and width of the image. Images in Train Data have a mean height of 31 and mean width of almost 75% of Images have width 136 and height 31 Almost 95 Percentile of the Images have with of 190 comparatively fewer number of Images have width > 200. The Fig.12 for the training image is show in below.

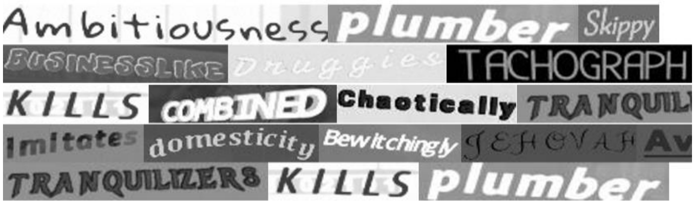


Fig.11. Dataset Sample

3.3 Result analysis:

After analyzing the results, it's evident that the Convolution recurrent neural network achieved the highest accuracy of 95% of character prediction and 91% of word prediction. Users access the application through a web browser, where they can upload images for text extraction from video.

3.4 Deployment:

The text extraction from video project is deployed via a web-based platform built on the Django. Users access the application through a web browser, where they can upload text contain Video or image for text extraction. The Django backend handles image processing tasks using the trained model, while the frontend displays results such as a text extracted from video or image also the user can download the text as txt file. The Fig of the Deployment is shown below.

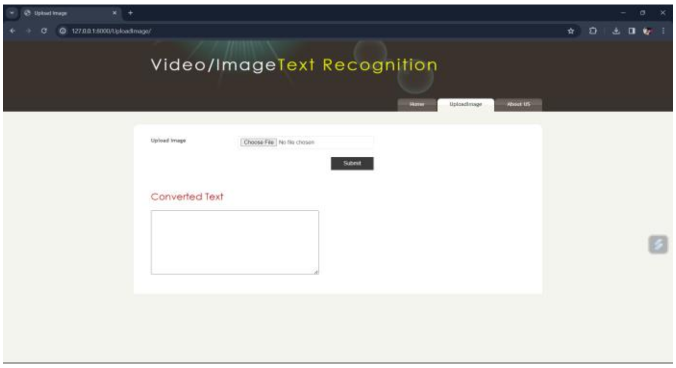


Fig.13. Home Page

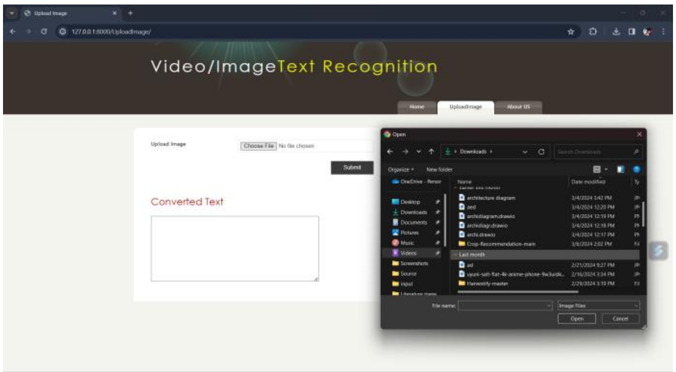


Fig.14. Video Upload

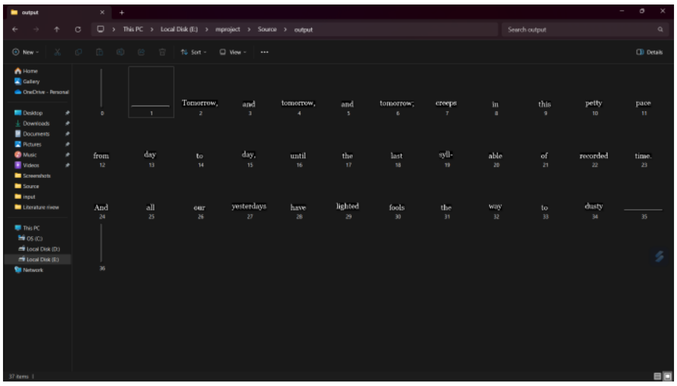
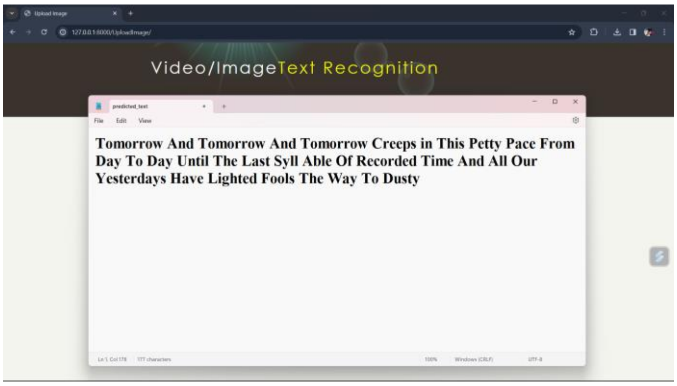


Fig.15. Text Region crop



Tomorrow And Tomorrow And Tomorrow Creeps in This Petty Pace From Day To Day Until The Last Syll Able Of Recorded Time And All Our Yesterdays Have Lighted Fools The Way To Dusty

Fig.16. Text Store in File

4. CONCLUSIONS

This project presents a robust system for extracting text from video content, achieving an impressive accuracy rate of 90%. By integrating Convolutional Neural Networks (CRNNs), Optical Character Recognition (OCR), and OpenCV, the system effectively identifies and extracts text regions from video frames with high precision. Through meticulous preprocessing, frame extraction, text detection using CRNNs, and subsequent text recognition with OCR algorithms, the system demonstrates reliable performance in accurately identifying and converting text regions into machine-readable format. The utilization of OpenCV for video processing and visualization further enhances the system's capabilities, enabling the overlaying of extracted text onto video frames for visual feedback. Additionally, the system's architecture allows for the storage of extracted text in text files, facilitating downstream applications such as video indexing, content analysis, and information retrieval. With a 90% accuracy rate, this project demonstrates a significant advancement in the field of computer vision and natural language processing. The high accuracy achieved underscores the system's effectiveness in automating text extraction from multimedia content, thereby addressing the growing need for efficient and accurate text analysis in various domains.

REFERENCES

[1] Pinaki Nath Chowdhury, Palaiahnakote Shivakumara, Ramachandra Raghavendra, “An Episodic Learning Network for Text Detection on Human Bodies in Sports Images”, IEEE Transactions on Circuits and Systems for Video Technology, 2021.

[2] Raghunandan, K. S., et al. "multi-script-oriented text detection and recognition in video/scene/born-digital images." IEEE transactions on circuits and systems for video technology 29.4 (2018): 1145-1162.

[3] Xuwang Yin, Kaizhu Huang, Hong-Wei Hao, Xu-Cheng Yin, “Robust Text Detection in Natural Scene Images”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 36, No. 5, May 2014.