International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 07 | July - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

VIDEO TEXT DETECTION AND CONVERSION TO AUDIO FILE

SUDEEP GOWDA S R 1, VIDYA S 2

1 Student, Dept. of MCA, Bangalore Institute of Technology, Karnataka, India

2 Assistant Professor, Dept. of MCA, Bangalore Institute of Technology, Karnataka, India

---

Abstract -Video text detection and recognition play a crucial role in extracting meaningful information from multimedia content for various applications such as video indexing, content retrieval, and automatic captioning. This paper provides a comprehensive study on the application of a R-CNN and CRNN algorithm for video text detection and recognition. In an era dominated by video content, the fusion of Optical Character Recognition (OCR), faster R-CNN for Text Detection and Convolutional Recurrent Neural Network (CRNN) for text recognition holds the promise of transforming the way we interact with and extract insights from visual media. This project aims to provide a comprehensive system for video text detection, OCR implementation, R-CNN and CNN meaningful text output generation. The core objective is to bridge the gap of visual information and textual insights, enabling automated extraction, analysis, and classification of text within video frames. By seamlessly integrating advanced OCR techniques, the system extracts textual content from video frames, overcoming challenges posed by diverse fonts, sizes, and backgrounds. The extracted text is then subjected to pre-processing to ensure accuracy and relevance.

***Key Words*:** Binirizationan, frame-extraction, Grey scale conversion, Noise reduction Dilation, Convolutional Recurrent Neural Network (CRNN), Faster R-CNN

1.INTRODUCTION

In today's rapidly evolving digital landscape, the extraction and interpretation of textual information from videos have emerged as a fundamental aspect of various applications, ranging from video content analysis to information retrieval and beyond. This intricate process, which involves the synergistic utilization of Optical Character Recognition (OCR), Faster R-CNN for text detection and Convolutional Recurrent Neural Network for text recognition techniques, holds the key to unlocking the latent potential of visual data present within videos. Optical Character Recognition, or OCR, stands as the cornerstone of this endeavor, enabling the transformation of visual content into machine-readable text. By meticulously analyzing individual frames extracted from videos, OCR algorithms decipher text embedded within images, creating a bridge between the visual and textual realms. This transformative step does not only facilitate the extraction of textual data but also sets the stage for subsequent analysis and classification.

2. LITERATURE SURVEY

Optical character recognition is the process of converting printed or handwritten text into a machine-readable format so that it can be edited, searched, and indexed. With the help of illustrative examples, the performance of the current OCR explains and displays the actual faults and image defects in recognition. This paper aims to create an application interface for OCR using an artificial neural network as a back end to achieve a highly accurate rate of recognition [1].

Inputting data into a computer is most commonly done by keyboarding, which is also the most labor and time-intensive procedure. It is a method of digitalizing printed texts for use in machine processes and electronic searching. It translates the photos into text that is machine encoded and useful for text mining, text-to-speech, and machine translation. This paper describes a straightforward, cost-effective, and efficient method for building OCR for reading any document with a handwritten or set font style and size [2].

Identifying characters from scanned photos is an extremely difficult operation. To carry out several manipulation procedures for record-keeping, we need all of the data in digital format. The primary challenge in character identification is the variety of writing styles and typefaces used in the text. We put forth a prominent method for character recognition from scanned photos by combining the ideas of artificial neural networks and nearest neighbor techniques [3].

One of the most active and difficult study topics in the fields of image processing and pattern recognition has been handwriting recognition. It has several applications, which include a reading assistance for the blind, bank cheques, and the conversion of any handwritten document into structural text form. In this paper, a multilayer feed-forward neural network is used to try handwritten character recognition for English alphabets without feature [4].

3. EXISTING SYSTEM

Every year, several approaches for scene text recognition (STR) in the present system are put out, greatly improving the field's performance. These techniques, however, have not kept up with more significant developments in text analysis, speech recognition, image identification, and detection. The effectiveness of several deep learning approaches for the encoder part of the Transformer in STR is assessed in this research.

Initially, the encoder's baseline feed-forward network (FFN) module is altered to include cross-stage partial (CSP)-FFN or squeeze-and-excitation (SE)-FFN architectures. Second, other encoder architectures are investigated, including Conformer structures and local dense synthesizer attention (LDSA). Of them, the Conformer encoder performs better in terms of test accuracy in a variety of experimental configurations, while SE or CSP-FFN modules compete competitively in terms of parameter efficiency. Qualitative insights into the performance of various encoder combinations can be obtained through the visualization of attention maps produced by them.

4. PROPOSED SYSTEM

The developed project will contribute fast and accurate solution for real-time video text recognition, powered by the latest state-of-the-art computer vision and OCR technologies. It uses multi-step processing, initiated with video frame extraction using Open CV, followed by end-to-end image pre-processing to enhance the visibility of texts and reduce noise. It detects text using a Faster R-CNN to provide robustly identify intrinsic regions of the text. From another end, a CRNN is used to text recognition that makes an accurate character-level analysis. The speech synthesis for the recognized text makes use of Google Text-to-Speech in accessibility to obtain an audio output. The whole workflow fits into a very user-friendly Flask web interface: subsequent video uploads, its real-time processing with textual and auditory feedback. This system delivers a system that corrects the many issues involved in text orientation, complex backgrounds, and motion blur but extends to broad uses in accessibility, video indexing, and real-time video subtitling.

5.FLOW DIAGRAM

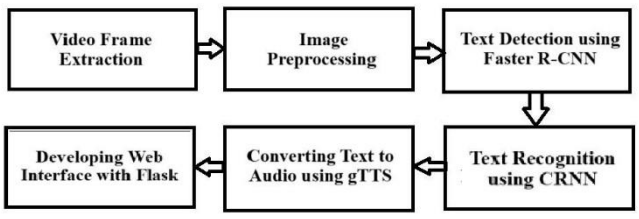


Fig 1. flow diagram

6.METHODOLOGY

1. **Vedio input**: Using a web interface, provide a way for users to upload video
2. **Frame extraction**:To extract frames from the video at predetermined intervals for text detection and identification use Open CV.
3. **video pre-processing**: It includes grayscale conversion, binarization techniques such as Otsu's thresh holding, simple thresh holding, and adaptive thresh holding, techniques for reducing noise and text's structure is improved by morphological processes including dilatation.
4. **Text detection with faster R-CNN**: To recognize text areas in pictures, use Faster R-CNN it finds regions of interest (ROIs) that have text in them. It includes,
   - **Region-Proposal-Network (RPN):** The RPN creates region proposals, or possible bounding regions where text might be found, by scanning input image *i*. At every point on the feature map, the RPN creates *k* anchor boxes.

$$\{(x_i, y_i, w_i, h_i)\}_{i=1}^{k}$$

   - **Bounding Box Regression**: In this stage, the bounding box are improved so that the text portions are surrounded more precisely. For every suggestion, the

$$\hat{x} = x + \Delta x, \quad \hat{y} = y + \Delta y, \quad \hat{w} = w \cdot e^{\Delta w}, \quad \hat{h} = h \cdot e^{\Delta h}$$

regression model produces adjustments (Δx, Δy, Δw, Δh)

5. **Text Recognition with CRNN**: Apply a Convolutional Recurrent Neural Network to obtain an architecture that can recognize the text within the ROIs.It includes
   - **Convolutional Layers:** These layers encode spatial data into feature maps *F* by extracting features from the text regions that Faster R-CNN has identified.F=Conv(I)

     F=Conv(I)

     where *F* represents the feature maps.

   - **Recurrent Layers (RNN):** In particular, sequential dependencies the text are captured by Long-Short-Term Memory (LSTM) units, which help the network comprehend character context and order. The concealed state at time t can be represented by the symbol h t.

$$\mathbf{h}_t = \mathrm{LSTM}(\mathbf{F}_t, \mathbf{h}_{t-1})$$

6. **Text-to-Speech Conversion using gTTS API**: Sends the identified text that have been recognized to the gTTS API to convert them into audio.
   i. **Text Input:** The gTTS API receives the identified text *T*.

   Audio=gTTS(T)……(5)

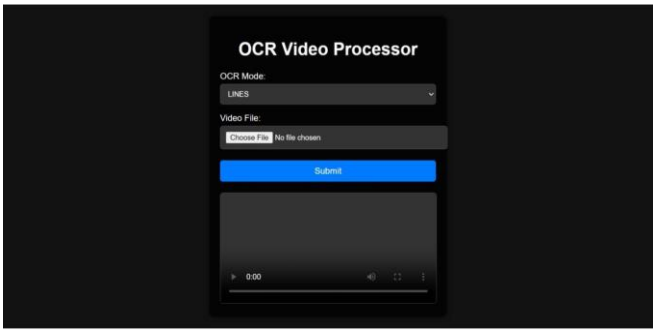
   ii. **Audio Output:** The user can play back an audio file created by the API

7. **Backend Integration Using Flask:** Setting up Flask Application, create endpoint.

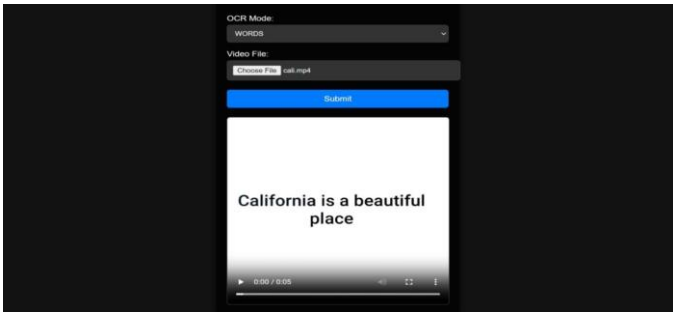
7.RESULT

With the application of OCR techniques, we were able to extract frames from videos, process the input, and create a real-time text recognition system, that could detect and recognize text. The system integrates a Flask-based user interface for user interaction, using Open CV for picture pre-processing and Pytesseract for text recognition

8.SCREEN SHOT



Screen shot 1. Home page



Screen shot 2. Video uploading page

Screen shot 3. output page

CONCLUSION

The main conclusions, contributions, and future directions of our work on computer vision and optical character recognition (OCR) based real-time text recognition in videos are outlined in this chapter. The developed video text detection and recognition system has shown improved results in accurately detecting and recognizing text content within video sequences. By employing a combination of pre-processing techniques, including frame extraction, edge detection, and region-of-interest extraction, the system effectively identifies text regions of interest within the video frames. Utilizing Optical Character Recognition (OCR) technology further enhances the system's capability to accurately recognize and extract text content from these regions. We investigated several approaches during this study, put in place a working system, and assessed its effectiveness.

REFERENCES

[1] B. Vanni, M. Shyni, and R. Deepalakshmi, “High accuracy optical character recognition algorithms using learning array of ANN” in Proc. 2014 IEEE International Conference on Circuit,pp (1474-1479),2014.

[2] Anirudh Sanjay Patil[1], Jashneet Singh Mong “Optical_Character_Recognition_using Artificial Neural Networks” in AAAI,pp.12216-12224,2020.

[3] Honey_Mehta, Sanjay_Singla, Aarti_Mahajan” Optical-Character-Recognition-(OCR) System for-Roman Script & English Language-using-Artificial Neural Network (ANN) Classifier””, International Conference on Document Analysis and Recognition (ICDAR). IEEE, pp. 781–786,2019.

[4] J.Pradeep,E.Srinivasan, and S.Himavathi, “Diagonal Based Feature Extraction For Handwritten Alphabets Recognition System Using Neural Networks”, International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, pp. 364-368 No 1, Feb 2011.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, JakobUszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and IlliaPolosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, pp. 5998–6008, 2017.

[6] Xiaoxue Chen, LianwenJin, Yuanzhi Zhu, CanjieLuo, and Tianwei Wang, “Text recognition in the wild: A survey,” ACM Computing Surveys (CSUR), vol. 54, no. 2, pp. 1 35, 2021.

[7] Tianwei Wang, Yuanzhi Zhu, LianwenJin, CanjieLuo, Xiaoxue Chen, Yaqiang Wu, Qianying Wang, and MingxiangCai, “Decoupled attention network for text recognition.,” in AAAI, pp. 12216–12224,2020.

[8] JeonghunBaek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, SeongJoon Oh, and Hwalsuk Lee, “What is wrong with scene text recognition model comparisons? dataset and model analysis,” in Proceedings of the IEEE International Conference on Computer Vision, pp. 4715–4723,2019.

[9] Fenfen Sheng, Zhineng Chen, and Bo Xu, “Nrtr: A no recurrence sequence-to-sequence model for scene text recognition,” in 2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, pp. 781–786,2019.

[10] Sk Asif Akram, MousumiSaha and Tamasree Biswas, "Evaluation of descriptive answer sheet using machine learning", in International Journal of Engineering Sciences & Research Technology (IJESRT), pp. 184-186, April 2019.