IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

TEXT EXTRACTION USING MACHINE LEARNING

Authors

Mr.C. Murale, P. Guna , R. Keerthi , T S. Keerthivela , V. Vikram

Coimbatore Institute of Technology

Abstract Our project is specifically tailored to provide efficient support for students by saving valuable time. The primary objective is to empower users through flexible text extraction capabilities, particularly in recognizing handwritten documents. This initiative caters to students seeking a more accessible method for extracting text and notes from tutorial videos in the mp4 format and text format. Additionally, the project aims to convert scanned handwritten documents into searchable formats, offering a valuable tool for individuals with visual impairments. A pivotal aspect of the project is its commitment to data accuracy, ensuring that the extracted information aligns precisely with the source material. In essence, our project revolves around the central objective of optimizing text recognition to create an inclusive and time-efficient experience for the student community. Moreover, our project also aims to convert extracted text into mp3 format, providing further accessibility options for students.

Keywords – Video Processing, Text extraction ,Audio file, visually impaired, handwritten documents.

I. INTRODUCTION

In today's world, our project helps bridge accessibility gaps in education by revolutionizing text recognition technology. Focusing on handwritten documents and tutorial videos in mp4 and text formats, we aim to provide students with a more accessible method for extracting crucial information. By converting scanned handwritten documents into searchable formats, we empower individuals with visual impairments and enhance overall inclusivity. Our commitment to data accuracy ensures that the extracted information precisely mirrors the source material, guaranteeing reliability and trustworthiness. With a central objective of optimizing text recognition, our initiative prioritizes creating an inclusive and time-efficient experience for the student community. Through innovative technology, we strive to make learning more accessible, efficient, and empowering for all.

Text extraction from video processing involves utilizing computer vision techniques to identify and extract textual information embedded within video content. This advanced technology employs optical character recognition algorithms to recognize and convert text present in the video frames into machine-readable data. Video text extraction finds applications in various fields such as video analytics, content indexing, and accessibility services. It enables the automatic extraction of valuable information from video sources, aiding in tasks like subtitling, keyword tagging, and content summarization.

Audio conversion from video processing involves the extraction or conversion of audio content from video files. This process utilizes various audio processing techniques to separate the audio track from the video, enabling users to obtain standalone audio files. Commonly used methods include demuxing, where the audio and video streams are separated, and transcoding, which involves converting the audio format to a desired output.

Optical Character Recognition, is a technology that converts different types of documents, such as scanned paper documents, PDFs, or images captured by a digital camera, into editable and searchable data. It uses algorithms to recognize text characters within these images and translates them into machine-readable text.

The Scale-Invariant Feature Transform (SIFT) is a key computer vision algorithm designed for detecting and describing distinctive features within images, regardless of their scale, orientation, or illumination variations. SIFT has become a fundamental tool for tasks like image recognition, object detection, and matching. SIFT identifies keypoints, which are unique and stable points in an image, using a process that is invariant to transformations like scaling and rotation.

FLANN (Fast Library for Approximate Nearest Neighbors) is a powerful and efficient library designed for fast and approximate nearest neighbor search. FLANN provides a collection of algorithms and data structures to accelerate the process of finding approximate nearest neighbors in high-dimensional spaces. It is particularly valuable in applications such as computer vision, image matching, and machine learning, where efficiently identifying similar data points is crucial.

II. PROPOSED METHODOLOGY

The proposed system for our project encompasses a comprehensive suite of features designed to enhance support for students by streamlining text extraction processes. At its core, the system will leverage advanced optical character recognition (OCR) technology to accurately recognize and extract text from various sources, including handwritten documents and tutorial videos in MP4 format.

Through the integration of flexible text extraction capabilities, users

will benefit from a time-saving solution that facilitates efficient access to essential information. Moreover, the system will prioritize data accuracy, employing robust algorithms to ensure precise alignment between the extracted text and its source material. By emphasizing accuracy, the system aims to provide students with reliable and trustworthy extracted content, enhancing the overall effectiveness of their academic endeavors.

Furthermore, the proposed system will address the diverse needs of students by offering accessibility options such as converting extracted text into MP3 format. This feature will enable individuals with visual impairments to access and engage with the extracted content more effectively, thereby fostering inclusivity within the student community. Additionally, the system will support the conversion of scanned handwritten documents into searchable formats, further enhancing accessibility for students who rely on handwritten notes.

By catering to a range of user preferences and accessibility requirements, the system will empower students to navigate and utilize educational materials more efficiently, ultimately contributing to a more equitable learning environment.

III ARCHITECTURE OF THE PROPOSED SYSTEM :

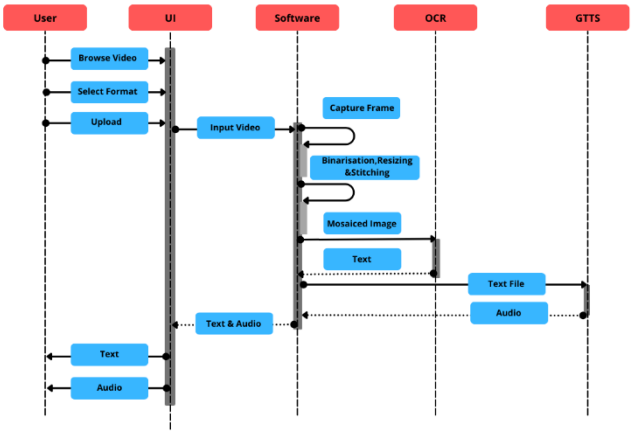

Figure 1 Architecture of the Text Extraction

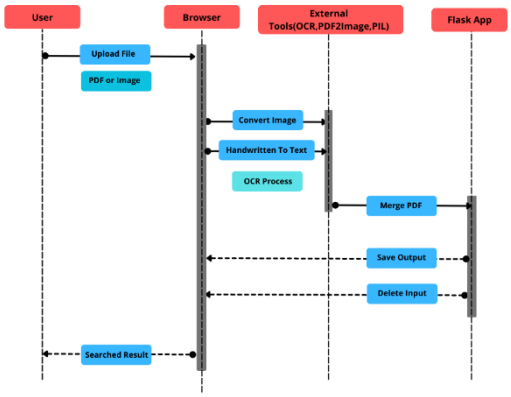

Figure 2 Architecture of the Image Genreration

METHODOLOGY

For implementing the stealth steganography we have done three module like text extraction and text generation.

Text Extraction:

The proposed system offers a comprehensive solution for extracting textual information from videos containing book or article content. It employs video processing techniques to extract frames, subsequently utilizing Optical Character Recognition (OCR) algorithms for accurate text extraction. Each frame serves as a snapshot, enabling tasks such as object detection and scene recognition, alongside OCR. The system employs the Scale-Invariant Feature Transform (SIFT) algorithm for robust feature extraction, facilitating content matching between frames and generating panoramas for improved readability.

Following frame extraction, the system applies further processing including binarization, resizing, and stitching to isolate and extract text lines, producing a mosaic image. This image undergoes text line extraction to generate a final output with organized textual content. The system ensures clear and readable text output, leveraging Text-to-Speech (TTS) technology to generate audio from the processed text, enhancing accessibility.

The system offers a sophisticated pipeline for converting visual information into textual data. It prioritizes accuracy through OCR integration and SIFT-based feature extraction, ensuring coherent alignment and readability of extracted content. By enabling text extraction and audio generation, the system caters to diverse user preferences and accessibility needs, ultimately empowering students with efficient access to educational materials.

Text Generation:

The system facilitates the extraction and processing of textual content from PDF files or handwritten images. Users initiate the process by uploading their chosen file format, which undergoes PDF-to-image conversion, saving each page as an image file like JPEG or PNG. Optical Character Recognition (OCR) technology is then employed to recognize individual words within the input text, ensuring accuracy in word recognition. To maintain confidentiality or security, a method is implemented to conceal recognized words in a hidden form.

Subsequently, the system merges the processed images into a cohesive PDF document while preserving the original layout and formatting. Special attention is given to accurately retaining the extracted handwritten text within the PDF. The resulting PDF is then generated to support text search functionality, enhancing usability by enabling users to efficiently locate specific words or phrases within the document. Moreover, the system directs users to the specified hidden form during searches, facilitating easy retrieval of concealed information.

The system offers a streamlined workflow for converting PDFs and handwritten images into searchable PDF documents. Through PDF generation and OCR capabilities, it ensures accurate extraction and retention of textual content while maintaining confidentiality where necessary. By enabling text search functionality and facilitating access to concealed information, the system enhances usability and efficiency for users interacting with digitized documents.

RESULT AND DISCUSSION :

This culmination of our project is a multifaceted solution tailored to revolutionize student support by optimizing text extraction processes and enhancing accessibility. Through advanced optical character recognition (OCR) technology, our system adeptly recognizes and extracts text from various sources, including handwritten documents and tutorial videos in both MP4 and text formats. This capability enables students to efficiently extract essential information and notes, saving valuable time and streamlining their learning experience.

Moreover, our project extends its reach by converting scanned handwritten documents into searchable formats, providing a crucial tool for individuals with visual impairments to access educational materials with ease. Central to our endeavor is a steadfast

commitment to data accuracy, ensuring that the extracted information aligns precisely with the source material, thereby fostering trust and reliability in the system.

By prioritizing accuracy, flexibility, and inclusivity, our project endeavors to create an inclusive and time-efficient experience for the student community, empowering them to navigate and utilize educational resources effectively. Additionally, our project aims to further enhance accessibility by converting extracted text into MP3 format, catering to diverse learning preferences and ensuring that educational content is accessible to all students, regardless of their individual needs or abilities.In essence, the culmination of our project represents a significant leap forward in leveraging technology to optimize text recognition and create a more inclusive and accessible learning environment for students worldwide.

Fig 3 Home Page

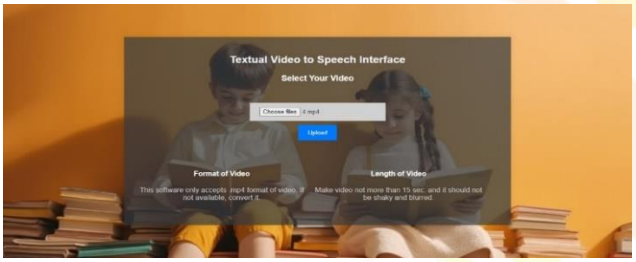

Fig.4 Textual Video to Speech

Fig.5 Text Video to Speech Interface

Fig.6 Search Generation

CONCLUSION :

In conclusion, our project represents a holistic endeavor to revolutionize text extraction processes, particularly catering to the needs of students while emphasizing inclusivity and efficiency. By leveraging advanced technologies such as Optical Character Recognition (OCR) and video processing techniques, we aim to empower users with flexible text extraction capabilities, allowing them to efficiently extract text and notes from various sources including handwritten documents and tutorial videos in both MP4 and text formats. The conversion of scanned handwritten documents into searchable formats not only facilitates accessibility for individuals with visual impairments but also provides a valuable tool for all students seeking to organize and utilize their handwritten notes more effectively. Central to our project is a staunch commitment to data accuracy, ensuring that the extracted information precisely aligns with the source material, thus bolstering the reliability and trustworthiness of the extracted content. In essence, our project endeavors to optimize text recognition to create a more inclusive and time-efficient experience for the student community, empowering them to navigate and utilize educational resources with ease. Furthermore, by extending accessibility options through the conversion of extracted text into MP3 format, we aim to enhance the learning experience for students with diverse needs and preferences, reaffirming our dedication to fostering an inclusive educational environment. Through these concerted efforts, we aspire to contribute to the advancement of educational technology, ultimately striving to make learning more accessible, efficient, and empowering for all students.

FUTURE WORK

Our future work like advancements in text extraction technology from videos will enable seamless conversion directly into required documents, streamlining the process for users. Similarly, handwritten documents will undergo efficient conversion into editable formats such as Word documents or PDFs, enhancing accessibility and usability. These advancements signify a significant leap forward in document processing capabilities, offering users a convenient and versatile solution for managing textual content. By harnessing cutting-edge techniques, such as Optical Character Recognition (OCR) and advanced video processing, these innovations promise to revolutionize how information is extracted, converted, and utilized, ultimately enhancing productivity and efficiency for individuals across various domains.

REFERENCES

[1] Handwritten Text Recognition Using Convolutional Neural Network Atman Mishra,Sharath Ram,Kavyashree arXiv:2307.05396v1 [cs.CV] 11 Jul 2023

[2]Text Recognition and Extraction from Video Kiran Agre ,Ankur Chheda,Sairaj Gaonkar ,Mahendra Patil International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Published by, www.ijert.org ICIATE - 2017 Conference Proceedings

[3] Text to Speech Conversion S.Venkateswarlu, D.B.K.Kamesh,J.K.R.Sastry,Radhika Rani Indian Journal of Science and Technology, Vol 9(38), DOI: 10.17485/ijst/2016/v9i38/102967, October 2016

[4] Text – To – Speech Synthesis (TTS) Nwakanma Ifeanyi,Oluigbo Ikenna,Okpala Izunna IJRIT International Journal of Research in Information Technology, Volume 2, Issue 5, May 2014, Pg: 154-163

[5] A Survey on Neural Speech Synthesis Xu Tan,Tao Qin,Frank Soong ,Tie-Yan Liu arXiv:2106.15561v3 [eess.AS] 23 Jul 2021

[6] Text Extraction from an Image using CNN P. Rajeshwari,D.Vinay Sekhar Reddy ,G.Pranay,Mohd Arbaz Mazharuddin © 2022 JETIR April 2022, Volume 9, Issue 4 www.jetir.org (ISSN-2349-5162)

[7] MAGE TEXT EXTRACTION AND ITS LANGUAGE TRANSLATION Shubham Nagmoti,Kapil Bhoyar,Shantanu Raut,Saransh Jamgade,Nikhil Mangrulkar,Aniket Pathade Article

in Journal of Research in Engineering and Applied Sciences · April 2021 DOI: 10.46565/jreas.2021.v06i02.008

[8] Extraction of Text from an Image and its Language Translation Using OCR G.R.Hemalakshmi,M.Sakthimanimala,J.Salai Ani Muthu International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Volume 4, Issue 4, April 2017 ISSN (Online) 2394-2320

[9] OCR-based Text Extraction from Images Vijayalakshmi V.,Amruta Ashok,Naik Chinmayee M.S International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering ISO 3297:2007 Certified Impact Factor 7.047 Vol. 10, Issue 5, May 2022 DOI: 10.17148/IJIREEICE.2022.10572

[10] EXTRACTING TEXT FROM VIDEO Jayshree Ghorpade,Raviraj Palvankar,Ajinkya Patankar,Snehal Rathi Signal & Image Processing : An International Journal (SIPIJ) Vol.2, No.2, June 2011 DOI : 10.5121/sipij.2011.2209 103

[11] Text Recognition and Extraction from Video Kiran Agre,Ankur Chheda,Sairaj Gaonkar,Mahendra Patil International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Published by, www.ijert.org ICIATE - 2017 Conference Proceedings

[12] Text Extraction in Video Ankur Srivastava,Dhananjay Kumar,Om Prakash Gupta,Amit Maurya,Mr.sanjay kumar Srivastava International Journal of Computational Engineering Research||Vol, 03||Issue, 5||

[13] HANDWRITTEN TEXT RECOGNITION SYSTEM SUSHMITHA,SUSHMITHA B POOJARY,VARSHITHA,VIDYA K C,SHILPA © 2021 IJCRT | Volume 9, Issue 7 July 2021 | ISSN: 2320-2882

[14] Handwritten Text Recognition Using Convolutional Neural Network Atman Mishra,Sharath Ram,Kavyashree arXiv:2307.05396v1 [cs.CV] 11 Jul 2023

[15] Handwritten Text Recognition: with Deep Learning and Android Shubham Sanjay Mor,Shivam Solanki,Saransh Gupta,Sayam Dhingra,Monika Jain,Rahul Saxena International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958 (Online), Volume-8 Issue-3S, February 2019