IMAGE TO AUDIO CONVERSION USING MACHINE LEARNING

Mahesh Timmanna Hegde[1]**,Dr.Shashidhar Kini K**[2]

MCA, Srinivas Institute of Technology, India

E-mail:maheshthegde@gmail.com[1]

E-mail:skinimca@sitmng.ac.in[2]

Abstract

Image text-to-audio conversion is an innovative research area that aims to convert textual information present in images into audible speech. This project proposes a comprehensive approach to tackle this challenge using machine learning techniques. By leveraging advancements in computer vision and speech synthesis, the system can automatically extract and transform text from images into high-quality audio output.

1. INTRODUCTION

In today's digital era, images are a prevalent form of visual content, often containing valuable textual information. However, for individuals with visual impairments or those who prefer audio-based information consumption, accessing the textual content within images can be a significant challenge. Image text-to-audio conversion aims to bridge this accessibility gap by automatically converting text in images into audible speech**.**

The conversion of image text to audio poses several complex challenges, requiring the integration of computer vision, natural language processing, and speech synthesis techniques. By leveraging advancements in machine learning and deep learning, it is now possible to develop sophisticated systems capable of extracting text from images and generating natural-sounding speech.

2. PROBLEM STATEMENT

The Visually impaired people will face difficulties in daily life such as access to information, knowledge and opportunities and there is a need for an assisting aid which will allow them to live life like a common man by forgetting their vision related problems. Visually impaired tend to be suited into special classes and they are treated in special ways which in many cases have resulted in isolating them from the society.

This project helps to widen the scope of the person by giving the description of the text or written document by speech.

3. LITERATURE SURVEY

Raju, et al. made a project on text extraction from video images. In this article author suggested a method for extracting words from photos. Using features from the frequency and spatial domains, they retrieved 13 different elements from photos to classify whether or not they include text. Simple Logistic, J48, and Random Forest classification techniques all have a high success rate of 98%. A dataset with changing text attributes was created using frames from downloaded videos, taking one frame every three seconds into account. The frame image was used to build blocks that were 102*26 in size. The approach involved extracting frames from videos in Mat lab, saving them as JPEG images, and manually selecting frames for further processing. 13 distinct features were extracted from the cropped photos, and Otsu's global thresholding approach was used to convert them into binary representation. Challenges arose with complex video frame backgrounds, making feature selection difficult due to variations in text, font, size, stances, colors, and shapes [1].

Jacobs, et al. worked on Text recognition of low-resolution document images. In this article outlines a method for OCR on images with a constant foreground and background color .By simultaneously tackling segmentation and recognition problems with a global optimization framework, they overcome the difficulties presented by fuzzy and low-quality camera images. A trustworthy system is developed using a machine learning method based on convolution neural networks that is independent of the environment, including the lighting, fonts, sizes, cameras, angles, and focus. The OCR system comprises of two word recognizers—one for words and one for characters—that recognise words using neural networks and dynamic programming. The system is effective at selecting the right term up to 95% of the time and is appropriate for document retrieval [2].

Aman Raj, et al. made a work on An Integrated Model for Text to Text, Image to Text and Audio to Text Linguistic Conversion using Machine Learning Approach. In this article outlines a study proposes an integrated model that does text-to-text, image-to-text, and audio-to-text conversions using machine learning approaches, with a focus on Indian languages. The suggested model, which can translate text, image, and voice, has been tested on sizable datasets of several Indian languages and makes use of cutting-edge methods like speech recognition, machine learning, and computer vision to precisely transcribe and translate the input data. By accurately converting text, images, and audio to text, the experiment results show the model's efficacy. Our proposed model may be used for a variety of purposes, including language learning, accessibility for people who are deaf or nonverbal, and interlanguage communication. The proposed strategy aims to close the communication gap between people who speak various languages and backgrounds [3].

Dave, et al. worked on OCR Text Detector and Audio Convertor. In this article states that one of the most extensively researched areas in computer vision, image processing, and optical character recognition in recent years is text recognition and extraction from various picture documents and their conversion into audio. To infer text from a photograph, many sophisticated text detection methods are employed, such as the FAST and East algorithms. In this article, a combination of simple systems of filters and detectors are employed. Additionally, we provide a framework for geometrical rectification that is essential for restoring the frontal-flat perspective of a document from a single camera-captured image. Our method of geometric rectification is based on an assessment of the homograph matrix of the image. To increase the precision of our OCR system, Tesseract, a deep learning-based recognition engine, has been added. The recognition engine being used by Tesseract is an extended Short-Term Memory (LSTM)-based Recurrent Neural Network (RNN). Then, we used GTTS, a screen reader programme created to turn text into speech for our OCR system, to transform our OCR text output into an audio output. This method is really quick. It is employed to recognize multilingual handwriting and translate it into speech. This technique has an accuracy rate of 85% or better. Finally, using this method, text may be accurately and simply extracted from document images and converted to speech [4].

4. METHODOLOGY

The project describes the pre-processing steps done on the image before the character recognition is done. When the image is captured it is first converted to grayscale as it is easy to detect the text on a gray format image. There might be some noise present in the image due to some external factor or the quality of image. Noise removal algorithms are used to reduce that noise and filter the image for better results.

The inverse Thresholding is done on the image for detection and extraction of region of interest in the image. We detect the boundaries that will be occupied by the text. The squared region is considered and only that part of the image is having some form of text. From the information of the contours the image is cropped to the size of the region of interest.

The cropped image is again filtered to reduce the grainy noise which might occur due to cropping. This image is applied to the thresholding function which will convert the image into a binary format with zeros and ones. This final image is considered for the further operations.

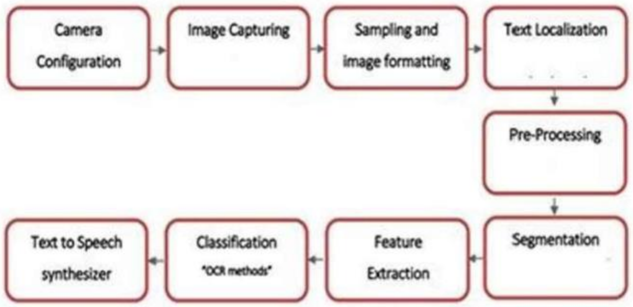

Fig 4.1: Block diagram

The basic setup consists of a computer and a web-camera which is configured to run on the system. The reading mode starts when the user press of the corresponding key in the keyboard. The video starts recording when the program is active. When a corresponding key on the keyboard is pressed that image frame of the video is captured. This is the image used for the process. Sampling and image formatting involves sampling of the captured image and resizing of the image for a standard format. Having a standard format of image makes it easy to run the processing algorithms on the image. The text localization is used in order detect if any text is present in the image. If no text is detected then there is no reason to do the further processing, this step helps to reduce the processing time for some cases. In the pre-processing stage the Image frame is converted into gray scale in order to detect the contours which will be necessary for text recognition. If the position and angle are confirmed to be correct, the image will be cropped according to the intersection points found. Some noise might be present, noise reduction algorithm reduces the noise from images taken from the camera there by improving the image. The main program then runs the OCR algorithm to recognize the text. The required features from the image are extracted and the text that is detected is then recognized and stored in a string. For every word recognized if the probability is less than 45% then that is not considered as it might be some random text. This string of text is given as input to GTTS which will generate an audio file for the text in the string. The Pygame module is used to initialize the speakers and to play the audio. This method has an advantage that it does not have to save the audio file which helps in increasing the speed. In the final step the result is given as output through the earphone or speaker.

5. RESULT

The project's intended result is the creation of a machine learning system that can accurately convert visual text into speech. The system will be able to extract text from images and use natural language processing techniques to further process the text to provide high-quality voice output. This system's main objective is to support people who have trouble reading or visualising written material. Users will be able to access and understand information that was previously inaccessible to them by providing accurate and comprehensible audio renditions of the text. The end goal of this project is to provide people with better access to information and to help them interpret written material by using the system that has been created.

6. CONCLUSION

The creation of a machine learning system that can accurately translate visual text into speech is the project's anticipated outcome. With the aid of natural language processing techniques, this system will be able to extract text from photos and then process it to produce high-quality voice. An effective system that can help those who have trouble reading or seeing written materialis what is anticipated as a result. Users should be able to access and understand information that was previously inaccessible to them thanks to the system's accurate and understandable audio representations of the text.

REFERENCES

[1] - Raju, Nidhin, and H. B. Anita. "Text extraction from video images." Int J Appl Eng Res 12, no. 24 (2017): 14750-14754.

[2] - Jacobs, Charles, Patrice Y. Simard, Paul Viola, and James Rinker. "Text recognition of low-resolution document images." In Eighth International Conference on Document Analysis and Recognition (ICDAR'05), pp. 695-699. IEEE, 2005.

[3] - Singh, Aman Raj, et al. "An Integrated Model for Text to Text, Image to Text and Audio to Text Linguistic Conversion using Machine Learning Approach." *2023 6th International Conference on Information Systems and Computer Networks (ISCON)*. IEEE, 2023.

[4] - Dave, Himank, et al. "OCR Text Detector and Audio Convertor." *Int. J. Res. Appl. Sci. Eng. Technol* 8 (2020): 991-999.