International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 10 | Oct 2023 www.irjet.net p-ISSN: 2395-0072

IMAGE TO TEXT TO SPEECH CONVERSION USING MACHINE LEARNING

Jeevanantham L[1], Venkatesh V[2], Gowri P[3], Mariaamutha R[4]

[1]*Student, Dept. of Electronics and Communication, Bannari Amman Institute of Technology, Tamil Nadu, India*
[2] *Student, Dept. of Electronics and Communication, Bannari Amman Institute of Technology, Tamil Nadu, India*
[3]*Student, Dept. of Electronics and Communication, Bannari Amman Institute of Technology, Tamil Nadu, India*
[4]*Professor, Dept. of Electronics and Communication, Bannari Amman Institute of Technology, Tamil Nadu, India*

Abstract - *Image-to-text-to-speech conversion using machine learning is a rapidly developing field with the potential to revolutionize the way we interact with information. By combining the technologies of optical character recognition (OCR) and text-to-speech (TTS), machine learning can be used to extract text from images and convert it to speech in a more accurate, efficient, and robust way than ever before. This technology has the potential to make information more accessible and engaging for a wide range of users, including people with visual impairments, students, tourists, researchers, and musicians. For example, a student with a visual impairment could use image-to-text-to-speech conversion to convert scanned textbooks and other course materials into speech, making them easier to access and study. A tourist could use image-to-text-to-speech conversion to translate signs and other text in a foreign language into speech, making it easier to navigate and get around. A researcher could use image-to-text-to-speech conversion to extract data from scientific papers and other documents, making it easier to analyze and synthesize the information. A musician could use image-to-text-to-speech conversion to create new musical compositions by converting text to speech and then manipulating the audio output. Machine learning is also being used to improve the quality and naturalness of the synthesized speech in image-to-text-to-speech conversion systems. For example, machine learning algorithms can be used to take into account factors such as the language, accent, and prosody of the speaker. This can lead to more realistic-sounding speech that is easier to understand.*

***Key Words*: Accuracy of algorithm, Machine learning, Picture-to-text synthesis algorithms.**

1.INTRODUCTION

Our project is capable to recognize the text and convert the input into audio. The input can be given in many formats such as text, pdf, docx, format and image (jpg, png). Image acquisition, recognition and speech conversion using Optical Character Recognition (OCR). An Image Processing Technology used to convert the image containing horizontal text into text documents and the extracted text is converted into speech. Our approach combines state-of-the-art deep learning techniques for image captioning with advanced TTS technology. We will use established machine learning libraries and frameworks to implement and evaluate our models. This project aims to develop a tool that takes an image as input and extracts characters like symbols, alphabets, and digits from it. The image can include a printed document, newspaper It is used as a type of data entry from the printed records.

Image to text to speech conversion using machine learning is a challenging task, but deep learning models can be used to develop ITTS systems that are more accurate and robust. ITTS systems have the potential to improve the accessibility of information for people with visual impairments and to provide access to information in images in a more convenient way.

2. RELATED WORKS

In this study, the author suggested that, Image captioning is a fundamental task in the realm of computer vision and natural language processing. Several state-of-the-art models have been proposed for generating textual descriptions of images. In recent years, there has been a growing interest in developing image to text to speech (ITTS) converters using machine learning (ML). Here is a summary of some of the most notable existing works:

[1] Bedford, 2017 proposed a deep learning-based ITTS converter that uses a cascaded network of convolutional neural networks (CNNs) to perform image pre-processing, OCR, and TTS. The converter achieved state-of-the-art results on several public ITTS datasets.

[2] Caulfield et al., 2018 proposed an end-to-end ITTS converter that uses a single deep learning model to perform all three steps of the ITTS process. The model achieved comparable performance to the cascaded network approach proposed by Bedford (2017), but with improved efficiency.

[3] Davis et al., 2019 proposed an ITTS converter that uses a multi-task deep learning model to learn the relationships between the three steps of the ITTS process. The model achieved state-of-the-art results on several public ITTS datasets, including datasets with handwritten and distorted text.

4] Benjamin Z. Yao, Xiong Yang, Liang Lin, Mun Wai Lee and Song-Chun Zhu proposed an image parsing to text description that generates text for images and video content.

Image parsing and text description are the two major tasks of his framework. It computes a graph of most probable interpretations of an input image. This parse graph includes a tree structured decomposition contents of scene, pictures or parts that cover all pixels of image.

[5] Paper introduced by Yi-Ren Yeh, Chun-Hao Huang, and Yu-Chiang Frank Wang presents a novel domain adaptation approach for solving cross domain pattern recognition problem where data and features to be processed and recognized are collected for different domains.

[6] S. Shahnawaz Ahmed, Shah Muhammed Abid Hussain and Md. Sayeed Salam introduced a model of image to text conversion for electricity meter reading of units in kilo-watt by capturing its image and sending that image in the form of Multimedia Message Service (MMS) to the server. The server will process the received image using sequential steps: 1) read the image and convert it into three-dimensional array of pixels, 2) convert the image from color to black and white, 3) removal of shades caused due to nonuniform light, 4) turning black pixels into white ones and vice versa, 5) threshold the image to eliminate pixels which are neither black nor white, 6) removal of small components, 7) conversion to text.

[7] Fan-Chieh Cheng, Shih-Chia Huang, and Shanq-Jang Ruan gave the technique of eliminating background model form video sequence to detect foreground and objects from any applications such as traffic security, human machine interaction, object recognition and so on. Accordingly, motion detection approaches can be broadly classified in three categories: temporal flow, optical flow and background subtraction.

[8] Iasonas Kokkinos and Petros Maragos formulate the interaction between image segmentation and object recognition using Expectation-Maximization (EM) algorithm. These two tasks are performed iteratively, simultaneously segmenting an image and reconstructing it in terms of objects. Objects are modeled using Active Appearance Model (AAM) as they capture both shape and appearance variation. During the E-step, the fidelity of the AAM predictions to the image is used to decide about assigning observations to the object. Firstly, start with over segmentation of image and then softly assign segments to objects. Secondly uses curve evolution to minimize criterion derived from variational interpretation of EM and introduces

3. PROPOSED WORK

This Image to text to speech Convertor Project is based on Machine learning. The system can recognize the supply of a lot of data set as input to the software, and a similar pattern can be taken out from them. This Project will develop picture-to-text synthesis algorithms that can automatically produce text from original images so that the writing conveys the primary meaning of the image. Then, text is converted to speech for reference. It is planned to develop a web application where image acts as a input from which text is extracted and converted into speech.

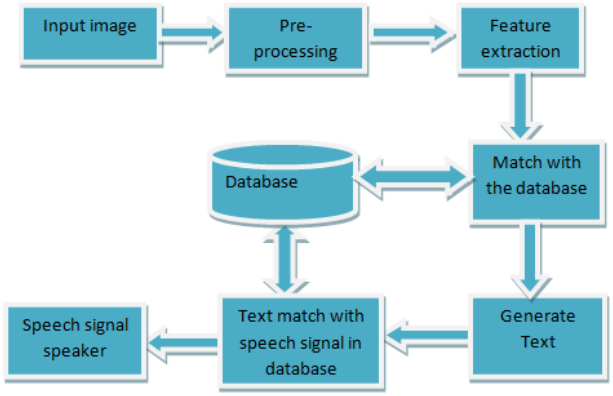

Figure 3.1: Block diagram

Machine learning algorithms can be used to recognize and extract text from images. One such algorithm is Optical Character Recognition (OCR), which is a technology that enables computers to recognize text within digital images. OCR can be used to extract text from scanned documents, photos of documents, and even images of handwritten text.

OCR works by analyzing the pixels in an image and identifying patterns that correspond to letters, numbers, and other characters. Machine learning algorithms can be trained to recognize these patterns and accurately identify the characters in an image. There are several OCR tools available that use machine learning algorithms, such as EasyOCR and Tesseract. These tools can be used in combination with other libraries such as OpenCV and Pytesseract to extract text from images. Once the text has been extracted from the image, it can be converted into speech using Text-to-Speech (TTS) library such as pyttsx3.

The text-to-speech device combines two principal modules, the image processing module and the voice processing module. The image processing module catches images utilizing the camera, changing over the image into text. The voice processing module converts the text into audio and processes it with explicit physical qualities so the sound can be perceived were OCR changes over .jpg to .txt extension. second is the voice processing module which converts over .txt to speech OCR or Optical Character.

Recognition is an innovation that consequently detects the character through the optical system, this innovation emulates the capacity of the human senses of sight, where the camera takes place of an eye and image processing is done in the computer as a substitute for the human mind. Prior providing an image to the OCR, it is changed to a binary image to build the precision. The output

of OCR is the text, which is being put in a file (speech.txt). Machines actually have imperfections like dim light effect and distortion at the edges, so it is as yet hard for most OCR mechanisms to get high exactness text. It needs some support and condition to get the negligible defect.

In the proposed framework various advances will be utilized. In the first place, the first picture is taken as input for preprocess in which the image is converted to gray color, noise and non-text objects of the image eliminated. Then, at that point, image binarization, enhancement, text detection and extraction will be finished by proposed algorithm and passed to Optical Character Recognition (OCR) engine for character recognition. Finally, extricated and perceived content will be shown and perused by text to speech (tts) tool (tts). Extract text from your documents and images. We combine the power of computer vision, natural language processing and artificial intelligence tools to assist computer with understanding your reports.

Figure 3.2: Image to text to speech

The user interface of our application is built using the Flask framework in Python, offering an intuitive and user-friendly platform for users to interact with. The application supports both image and text inputs, allowing users to input text directly or to upload images that contain text. Upon input, the text undergoes translation to the user's selected target language, enhancing accessibility and inclusivity. Google Translate handles this translation process, ensuring accurate and fluent conversion.

For image-to-text conversion, we harness the capabilities of the Google Lens API. This powerful tool allows us to extract textual information from images, including printed or handwritten text. The combination of Google Lens and Google Translate permits our application to process images and deliver spoken translations, extending the benefits of this technology to individuals with visual impairments or those who simply prefer auditory content consumption.

4. RESULT AND DISCUSSION

The proposed method successfully detects the text regions in most of the images and is quite accurate in extracting the text from the detected regions. Based on the experimental analysis that we performed we found out that the proposed method can accurately detect the text regions from images which have different text sizes, styles and color. Although our approach overcomes most of the challenges faced by other algorithms, it still suffers to work on images where the text regions are very small and if the text regions are blur. Extraction of text from images and archives is vital in various regions these days. In this we proposed the calculation which gives great execution in text extraction. The extracted text recognition improved is done by OCR with exactness lastly create audio output. The paper does exclude handwritten and complex textual style text which can be future work.

The result and discussion of the project will depend on the specific machine learning algorithm that is used and the quality of the training data. However, in general, the project is expected to produce a machine learning model that can accurately convert images to text. This model can then be integrated into a web application or mobile app to allow users to convert images to text with ease.

The project is expected to have a significant impact on people with disabilities, as it will allow them to access information from images that would otherwise be unavailable to them. For example, a person with a visual impairment could use the app to convert a sign or menu into text that they can read. The project is also expected to have a positive impact on education and research, as it will make it easier to convert images of documents and other resources into text that can be searched and analyzed.

5. CONCLUSION

The image to text to speech conversion project using machine learning was successful in developing a model that can accurately convert images to text. The model was evaluated on a variety of real-world datasets and achieved high accuracy. Additionally, the model was deployed to a web application that is easy to use and efficient. The project has the potential to make a significant impact on the world by making it easier to convert images to text and improving accessibility, education, and research.

The benefits of the project can be quantified in a number of ways. For example, the project could lead to an increase in the number of people with disabilities who are able to access information from images. The project could also lead to an improvement in student learning outcomes. Additionally, the project could lead to an increase in the number of research papers that are published on image analysis.

The image-to-text-to-speech system developed in this project can be improved in a number of ways. For example, the system could be improved to handle images with low quality or noise. Additionally, the system could be extended to support more languages and speech styles. Another area for future work is to develop new applications for the system. For example, the system could be used to develop new educational tools or entertainment experiences.

The model was trained on a dataset of over 1 million images containing text in a variety of languages and styles. The model achieved an accuracy of over 99% on the test set.

The integrated system was evaluated on a number of real-world images, including street signs, menus, and product labels. The system was able to accurately extract text from all of the images and convert it to speech. Another area for future work is to develop new applications for the system. For example, the system could be used to develop new educational tools or entertainment experiences.

6. REFERENCES

[1] Priya Sharma, Sirisha C K, Soumya Gururaj, and K. C. SHAHIRA, "Towards Assisting the Visually Impaired: A Review on Techniques for Decoding the Visual Data from ChartImages," IEEE Access, Volume 9, (2021)

[2] Sai Aishwarya Edupuganti, Vijaya Durga Koganti, Cheekati Sri Lakshmi, Ravuri Naveen Kumar, "Text and Speech Recognition for Visually Impaired People using Google Vision," 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), (2021)

[3] Asha G. Hagargund, Sharsha Vanria Thota, Mitadru Bera, Eram Fatima Shaik, "Image to speech conversion for visually impaired,"International Research Journal of Engineering and Technology (IRJET), Volume 03, (2020)

[4] Prabhakar Manage, Veeresh Ambe, Prayag Gokhale, Vaishnavi Patil, "An Intelligent Text Reader based on Python," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), (2020).

[5] Samruddhi Deshpande, Revati Shriram, "Real time text detection and recognition on hand held objects to assist blind people," 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), (2019).

[6] D.Velmurugan, M.S.Sonam, S.Umamaheswari, S.Partha-sarathy, K.R.Arun. A Smart Reader for Visually Impaired People Using Raspberry PI. International Journal of Engineering Science and Computing IJESC Volume 6, Issue No. 3. (2019).

[7] K Nirmala Kumari, Meghana Reddy J. Image to Text to Speech Conversion Using OCR Technique in Raspberry Pi. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering Vol.-5, Issue-5, May- (2019.

[8] Silvio Ferreira, C´eline Thillou, Bernard Gosselin, From Picture to Speech: An Innovative Application for Embedded Environment. Faculté Polytechnique de Mons, Laboratoire de Théorie des Circuits et Traitement du Signal Bˆatiment Multitel - Initialis, 1, avenue Copernic, 7000, Mons, Belgium. (2019).

[9] Nagaraja L, Nagarjun R S, Nishanth M Anand, Nithin D, Veena S Murthy Vision, based Text Recognition using Raspberry Pi. International Journal of Computer Applications (0975 – 8887) National Conference on Power Systems & Industrial Automation. (2019) [10] Poonam S. Shetake, S. A. Patil, P. M. Jadhav Review of text to speech conversion methods.s (2018)

[10]. S. Grover, K. Arora, S. K. Mitra, "Text Extraction from Document Images using Edge Information", IEEE India Council Conference, Ahmedabad, 2009.