International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 12, Issue 3 [March. 2023] PP: 305-310

A Hybrid Approach to Converting Visuals into Speech for the Visually Impaired

S.YASHWANTH REDDY
INFORMATION TECHNOLOGY
ANURAG GROUP OF INSTITUTIONS

B. HITESH
INFORMATION TECHNOLOGY
ANURAG GROUP OF INSTITUTIONS

MD. SIRAJUDDIN
INFORMATION TECHNOLOGY
ANURAG GROUP OF INSTITUTIONS

K. SANTOSH CHANDRA RAO
INFORMATION TECHNOLOGY
ANURAG GROUP OF INSTITUTIONS

ABSTRACT *The development of image-to-audio conversion technology has significantly improved the lives of blind people by allowing them to access media and information that was previously inaccessible. The deaf community now has a way to interact with various media, works of art, and sources of information thanks to technology that transforms visual data into an auditory format. This article gives a thorough review of image-to-audio conversion technology, covering its background, uses, advantages, drawbacks, and potential future advancements. We look at how important accessibility is for the deaf community as well as how image-to-audio conversion technology encourages freedom, inclusion, and fair access to information. Image-to-audio conversion technology has a wide range of uses in industries like education, journalism, the arts, navigation, medicine, finance, and gaming. By giving blind students access to visual aids in audio format, this technology has completely changed their educational experience for them. Deaf people may now view movies, TV shows, and online content thanks to image-to-audio conversion technology.*

Date of Submission: 15-03-2023 Date of acceptance: 30-03-2023

I. INTRODUCTION

Visual impairment is one of humanity's most significant limitations, especially today when information is increasingly communicated via text messages (both electronic and paper-based) rather than voice. The device we've proposed is intended to assist people who are visually impaired. In our planet of 7.4 billion people, 285 million are visually impaired, with 39 million completely blind (no vision at all) and 246 million having a mild or severe visual impairment (WHO, 2011).

These figures are expected to rise to 75 million blind people and 200 million people with visual impairment by 2020. Because reading is so important in people's daily lives (text is everywhere, from newspapers to commercial products, signboards to digital screens), visually impaired people face many challenges.

Based on current technological advancements in computer vision, digital cameras, and portable computers, it is possible to develop a camera-based technology that combines computer vision technology with other commercial products such as OCR systems. Reading is extremely important in today's society. Because the printed text is everywhere in the form of reports, bank statements, receipts, restaurant menus, and so on, blind users have a difficult time reading these forms. The method Text to Voice Adaption Using a Portable Camera is referred to in order to reduce the frustrating problem. The existing method has a significant disadvantage in terms of size and portability.

II. BACKGROUND

Although there are numerous web-based services that convert text to speech, they are not without limitations.

www.naturalreaders.com:.

1)It is a website where you can upload files in various formats such as doc, txt, and jpeg. However, because OCR is required to extract text from a photo in.png or.jpg format, the user must upgrade.

If a customer does not use it frequently, the monthly subscription costs approximately 1100 rupees, which is simply prohibitively expensive.

Furthermore, the user cannot have a word file for each book they want to listen to.

2)More platforms, such as ttsreader.com, fromtexttospeech.com, and many others, provide text-to-speech services.

The text is converted to speech after it is entered. Again, the problem is that no image is taken, and the user must have a text file or manually enter the text, which is a lot of work.

We have separate platforms that perform separate conversions, such as:

3) A platform called www.onlineocr.net provides a service to convert images to PDF, which takes longer to complete.

After that, the user can copy the text from the pdf or doc file and paste it into one of the text-to-speech conversion websites.

All of these processes take a long time, starting with uploading the.jpg format and then converting it to the docx file, after which the user can copy the text from the output, i.e. the pdf, and paste it into the textbox. However, it is a crisscross method in which the end user must repeat the same process for each page of the book.

The Image Captioning system, for example, generates a textual description of an image using a combination of computer vision and NLP techniques. Text-to-speech synthesis techniques can then be used to convert this textual description into an audio format.
The Image Captioning system analyses the content of an image first, using computer vision algorithms to identify objects, scenes, and other visual features. This data is combined with contextual knowledge and semantic rules to produce a textual description of the image. Natural Language Processing (NLP) techniques are used to analyse the textual description and convert it into a human-readable audio format.
Another system is the Deep-Speaker Embeddings system, which converts an image into an audio representation using deep learning techniques. This system extracts image features with a deep convolutional neural network (CNN) and then generates an audio signal with an NLP algorithm. The resulting audio signal can be used to convey image content information such as color, texture, and shape.

Overall, there are numerous approaches to converting images to audio using NLP algorithms, and the specific approach used will be determined by the application and the system's specific requirements.
Existing systems for image-to-audio conversion using NLP algorithms have several drawbacks. Some of these disadvantages include:
Accuracy: Existing image-to-audio conversion systems may not always produce accurate results. This is because the system's accuracy is determined by the quality of the image analysis and the NLP algorithms used. In some instances, the system may interpret the image incorrectly, resulting in inaccurate audio descriptions.
Limited vocabulary: Existing image-to-audio conversion systems may have a limited vocabulary, limiting the types of descriptions that can be generated. This can make describing complex or abstract images difficult.
Difficulties with abstract concepts: NLP algorithms may struggle with abstract concepts like emotions or ideas. This can make accurately describing images that convey these types of concepts difficult
.
Image-to-audio conversion using NLP algorithms can be time-consuming, especially for large or complex images. This can make using the system in real-time applications difficult.

III. LITERATURE SURVEY

The proposed system is low-cost and allows visually impaired individuals to hear the text. The main idea behind this project is optical character recognition, which converts text characters into audio signals. By segmenting each character, the text is preprocessed before being used for character recognition. Following segmentation, the letter is extracted and the text file is resized. The text file is then used to generate the audio signal. MATLAB16

Our paper describes the development of a real-time system in an outdoor environment based on object detection, classification, and position estimation to provide visually impaired people with voice output-based scene perception. The system is inexpensive, lightweight, straightforward, and simple to use. The module is built into the stick, and the pi-camera is used to take the photo, with the camera moved in the desired direction by a controller. The valuable insights gained from the feedback are then used to modify the system to better meet the needs of the user. The object detection and classification framework employs a multi-modal fusion-based mask RCNN with motion, sharpening, and blurring filters for efficient feature representation. The detected objects and their positions are classified by image recognition.

We proposed a unified system for extracting text from images and converting it into a target-language audio track. This method can assist the visually impaired in determining the attitude and demeanor of the person with whom they are interacting. All of the tasks associated with this application are accomplished by issuing commands. This system is unique in that it reads handwritten papers, analyses them for personality, and provides audio output to the blind. This technology allows people who are blind or visually impaired to read and comprehend First Information Reports (FIRs), business cards, chats, doctor's prescriptions, invoices, addresses, and other documents.

This research helps visually impaired and elderly people detect text in order to identify medicine. The researchers propose developing an app that will help visually impaired people scan images and convert the detected text into voice messages. The Google vision library is used to create an Android application that primarily contains three important functionalities: text recognition, text detection, and text-to-speech conversion. An in-built camera is used to scan the medicine image.

OCR is a mechanism that converts text images that have been typed, handwritten, or printed into machine encoded text. Using the phone's camera, this system will assist you in taking a picture or scanning a document that is present with the user. The image will be scanned, and the application will read the English text and convert it to speech format. To generate the speech output, the Text to Speech Module is used. The goal of delivering the output in the form of voice/speech is to provide the information on the document to the visually impaired.

IV. EXPERIMENTAL

A. Dataset

1) Standard Dataset:

Dataset Description:
In this project, we are using the Inshorts image dataset which is provided by Kaggle. Here you will be provided with different images.

B. Experimental Environment

Windows 10 operating system as test platform,

- CPU is Intel Core I7 6500u, which has dual cores running on 2.4GHZ.
- RAM 16GB.
- GPU is Nvidia GTX 970, CUDA Cores 1664, 4GB GDDR5

C. Methodology:

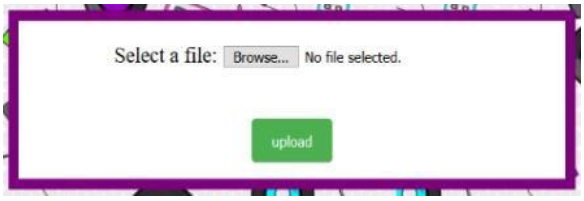
The website is entirely based on the Python framework, with Python 3.6.7 being used for the project. According to the project description, an uploaded image is converted into audible mp3 format.

As a result, the procedure is divided into six stages:

1. Obtaining an image from a website user.
2. Examining the image.
3. Image processing.
4. Convert an image to text.
5. Text-to-speech.
6. Play the audio file in the browser.

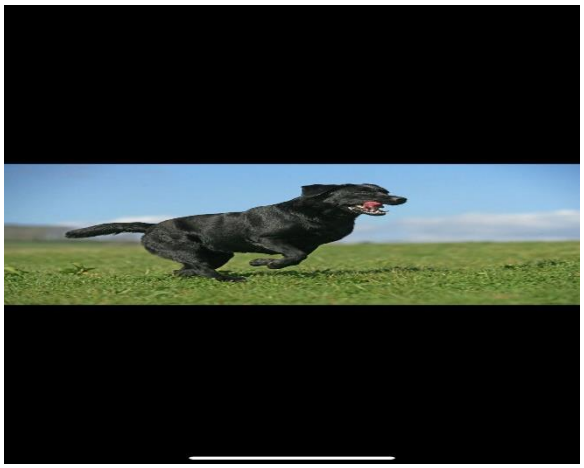
1) Obtaining an image from the user in a web app :-

For users to interact with the Image to Speech API, a web-based platform is provided. The user must upload an image to the web form. To send the image to the Python script, use the 'POST' method.



2) Reading the Image :-

Once the image has been received by the Python script, Open-cv2 CV's class is used to read it and convert it to a greyscale image of 1's and 0's



3) Process Image :-

The image array is further dilated and eroded to remove noise from the image, and the image pixels are increased and converted to greyscale so that they can be passed to the OCR Engine.

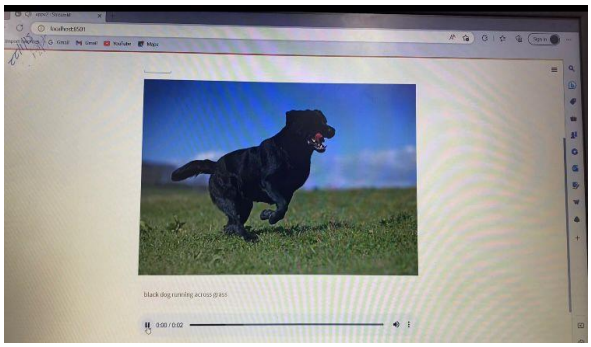
4) Text to Image :-

The clear image with no noise is sent to the tesseract OCR engine. The typed data in the image is converted to string format by the OCR engine.
The string's accuracy is determined by the image's clarity; however, there are errors in the conversion of punctuation symbols such as (",',!) in some cases.

5) Speech to Text :-
The Google Text to Speech API gTTS is used to convert the previous phase's generated string to audio format.
GTTS also provides the option to slow the voice of speech, but for convenience, this is set to false in the program.

6) Return the Audio File :-
a) Play on Browser
The audio is returned using the return statement because the flask environment allows for direct file return and media playback on the browser.

b) Mail the Audio The other or more private way to save the media file is using the mail service. For this Yagmail API of python is used, which provides unlimited emails in a day and allows to integration G-Mail and attachments.



Experiment's Evaluation Factors:
The three algorithms are evaluated in accordance with the proposed model according to the following Factors:
• Accuracy
• Performance
• Execution Time

V. ADVANTAGES

1) The book's mp3 format is useful for visually impaired readers.
2) Because the output is in audible format, the user can listen to it while doing other tasks, maximizing time.
3) Because the book is audio, the user does not need to carry the book.
4) Books are primarily used by students, and many times while reading, students intend to sleep; however, with our audio format, they can listen to books as well as live audio lectures.
5) Several research(s) say that listening improves the power of imagination, so the end users can simulate the scenes occurring in the book faster than reading.

VI. FUTURE SCOPE

1) The product includes an Optical Character Recognizer tool, which can only capture textual data, leaving various images throughout the book uncaptured. As a result, the audiobook does not include any of the book's diagrams or images.
2) Because only one image can be uploaded, if there are more pages, the end user must upload them one by one, which takes more time.
3) The entire text-to-speech conversion is dependent on internet connectivity; with a slow internet connection, the speech conversion may take a long time to generate or may even break the code.
4) To convert the text to audio, GTTS will require an internet connection. As a result, it may be slower than other offline APIs.

Currently, only one image can be uploaded at a time; work is being done to allow users to upload multiple images at once. The audio generated is per page, so if you have multiple pages, you'll get many mp3 format, so depending on the order in which the image is uploaded, the audio files can be combined into a single.mp3, making it easier for users to download.

The programme is currently limited to the English language; however, a user may upload an image in another language and the programme will fail. Tesseract-OCR can recognise a variety of languages. As a result, in the future, it may be a goal to implement various languages in the project.

The string generated by OCR contains various punctuation symbols, including (",',!) Because the Text to Speech service does not recognise them, they are spelled (double inverted comma etc.). (Optimization of strings)

VII. CONCLUSION

To summarise, image-to-audio conversion technology is a valuable tool that can significantly improve the lives of deaf people by providing them with access to information and media that they would not otherwise have. This technology has the potential to transform a variety of fields, including education, media, art, navigation, healthcare, finance, and gaming, by allowing all individuals, regardless of hearing ability, equal access. Despite its difficulties and limitations, such as accuracy and implementation costs, image-to-audio conversion technology has enormous potential benefits. With advancements in machine learning and artificial intelligence, technology is constantly evolving, providing more accurate and efficient results. The ability of image-to-audio conversion

technology to promote independence, inclusion, and equal access to information Better educational, professional, and social outcomes for the blind community. It has the potential to create more jobs.

Opportunities, improved communication, and reduced social isolation are just a few of the advantages. Furthermore, image-to-audio conversion technology has broader societal implications because it promotes accessibility and inclusion for all individuals, regardless of ability. It is important to continue investing in the development of this technology to ensure its effectiveness and accessibility for all. Finally, image-to-audio conversion technology is a game-changing innovation that has the potential to change the way we interact with visual content while also promoting greater accessibility and inclusion for the blind community.

REFERENCES

[1]. Sneha.C. Madre and S.B. Gundre, "OCR Based Image Text to Speech Conversion Using MATLAB," 2019 Second International Conference on Intelligent Computing and Control Systems (ICICCS).

[2]. P Rohit, M S Vinay Prasad, S J Ranganatha Gowda, D R Krishna Raju, and Imran Quadri, "Image Recognition Based Smart Aid For Visually Impaired People," International Conference on Communication and Electronics Systems (ICCES), 2020.

[3]. Sujata Deshmukh, Praditi Rede, Sheetal Sharma, and Sahaana Iyer, "Voice-Enabled Vision For The Visually Disabled," 2022 International Conference on Advances in Computing, Communication, and Control (ICAC3)

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

[5]. Abhishek Mathur, Akshada Pathare, Prerna Sharma, and Sujata Oak, "AI-based Reading System for the Blind Using OCR," 3rd International Conference on Electronics, Communication, and Aerospace Technology (ICECA), 2019.

[6]. "Raspberry Pi-based Intelligent Reader for Visually Impaired Individuals," 2nd International Conference on Innovative Mechanisms for Industrial Applications (ICIMIA), 2020. Vaibhav V. Mainkar, Tejashree U. Bagayatkar, Siddhesh K. Shetye, Hrushikesh R. Tamhankar, Rahul G. Jadhav, and Rahul S. Tendolkar

[7]. "Machine Learning based approach to Picture Description for the Visually Impaired", Asian Conference on Innovation in Technology (ASIANCON), 2021, by Raghunandan Srinath, Anisa Fathima, S. Arpitha, Chaitanya S. Rao, and T. Kavya

[8]. "Vivoice - Reading Assistance for the Blind using OCR and TTS," International Conference on Computer Communication and Informatics (ICCCI), 2022. R. Prabha, M. Razmah, G. Saritha, RM Asha, Senthil G. A, and R. Gayathiri

[9]. "Smart Machine Learning System for Blind Support," International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), 2021, S. Durgadevi, K. Thirupurasundari, C. Komathi, and S. Mithun Balaji

[10]. 2nd International Conference on Inventive Systems and Control (ICISC), Sandeep Musale and Vikram Ghiye, "Smart reader for visually handicapped," 2018.