Image Text to Speech Conversion Using Azure Cognitive Services

Khan Md. Hasib
ID - 20266015
Department of CSE
BRAC University
khan.md.hasib@g.bracu.ac.bd

Rashik Hasnat
ID - 20266010
Department of CSE
BRAC University
rashik.hasnat@g.bracu.ac.bd

Israt Jahan Ritun
ID - 21166038
Department of CSE
BRAC University
israt.jahan.ritun@g.bracu.ac.bd

Shanuma Afrin Meghla
ID - 21166010
Department of CSE
BRAC University
shanuma.afrin.meghla@g.bracu.ac.bd

***Abstract*—Many news organizations and television outlets have a large number of images in their hands, which they normally keep in some sort of digital repository. Saving a new photo, they must also tag related people in the photo so that the photo can be identified later by searching for the person's name. This process of tagging the photos manually is inconvenient, error-prone, and expensive. Using a straightforward face recognition algorithm also is not feasible because for that we need to train our model with a specific face rectangle which is quite difficult. Moreover, many photos have text written on them which also needs to be searchable. Again manually inputting those texts is not a feasible option. On other side, visual impairment is one of the biggest limitation for humanity, especially in this day and age when information is communicated a lot by text messages rather than voice. In this project, we intend to tackle a problem that no other standard approach has addressed: training a facial recognition model with group images of multiple faces without a clear face rectangle. The captured image undergoes a series of image pre-processing steps to locate only that part of the image that contains the text. Lastly, We convert the text of images into speech. The whole procedure of our project is conducted using Azure Cognitive Services in serverless architecture.**

***Index Terms*—face recognition, text recognition, text to speech recognition, Azure cognitive services**

I. Introduction

The number of visually disabled people has been increasing last year, because of eye disorders, age-related factors, untreated diabetes, injuries and others. For a visually disabled person, reading is one of the most difficulties. Recent advances in cell phones, computers and digital cameras allow the blind to help them by designing camera-based applications. In the world today there are about 285 million visually disabled individuals, 39 million of whom are blind and 246 are visually impaired. Such people have very low scope in their present world to understand precisely what is going on. There is no such device that can be conveniently connected with the environment by such disabled persons. It is really necessary to provide an effective interface for these individuals.

In this area, many improvements have been made to make reading easier for the visually impaired. The technologies used in the current paper have a similar solution but some inconvenience. First, the input pictures taken in past works have nothing intricate, i.e. the test inputs are written on a single sheet of white. It is possible to translate these images to text without pre-processing, but in a real-time environment this technique is not useful [1]. Again, Image processing is a method that uses mathematical sources, including pictures, a sequence, or a video for processing, in the form of mathematical operations. The product of image processing is an image or a series of features or parameters linked to the image. Image processing offers different uses such as digital graphics, scanning, face recognition and text recognition. The identification of various text characteristics, such as its letter size, font size, orientation, history etc. Recognizing the number plate is a good example for extracting text.

OCR executes text extraction from an image. It is a means by which representations of labels, written books, signs, etc. are converted into texts only. OCR contributes to the development of visually disabled reading instruments and telegraph technology. Tesseract library in OCR engine converts the binary image into text that senses contour, incline, pitches, white spaces and joint letters. The accuracy of the accepted text is also verified [2]. Again, Cognitive services are a set of REST APIs that expose the machine learning model to the outer world and help infuse smart, intelligent algorithms to hear, speak, recognize, and interpret user input into the applications, websites, and bots. Azure Cognitive Services reside in the Microsoft public cloud, which guarantees a secure, highly available and smooth performance. That's all we need in order to consume REST APIs in our application and leverage the machine learning capabilities with ease. A few cognitive services do need data to be uploaded by the users, but those services provide the appropriate algorithm.

In our project, We are trying to solve an issue which was not solved by any standard: the training of the multi-face community images in a facial recognition paradigm without a specific face rectangle. This training and tested face by using

Azure Face Api. The image obtained is pre-processed by a sequence of images to only find the portion of the image that comprises the content. This image to text extraction is done by Azure Vision Api. Finally, we are into speech conversion of picture text. Azure Speech Api is used to convert this segment.

The remaining parts of the paper will compose of the following sections:

- A quick overview of the various detection of credit cards is discussed in Section II.
- We have discussed about the workflow diagram of our project at section III.
- In Section IV, descriptions experimental results.
- Section V would include the conditions for conclusion and there is an acknowledgement in section VI.

II. Related Work

Researchers proposed different types of method for image to text, text to speech conversion where main aim is to help the visual impaired people. It is observed that they are still finding it difficult to roll their day today life and it is important to take necessary measure with the emerging technologies to help them.

Benjamin et al. [3] 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. Again, a novel domain adaptation approach for solving cross domain pattern recognition problem was suggested by Yi-Ren [4] where data and features to be processed and recognized are collected for different domains. Shahnawaz et al. [5] introduced a model of image to text conversion for electricity meter reading of units in kilo-watts by capturing its image and sending that image in the form of Multimedia Message Service (MMS) to the server.

In [6], proposed a camera based assistive text reading framework to help visually impaired persons read text labels and product packaging from hand-held objects in daily life. The system proposes a motion based method to define a Region Of Interest (ROI), for isolating the object from untidy backgrounds or other surrounding objects in the camera vision. A mixture of-Gaussians-based background subtraction technique is used to extract moving object region. To acquire text details from the ROI, text localization and recognition are conducted. Then text regions from the object ROI are automatically focused. In an Adaboost model the gradient features of stroke orientations and distributions of edge pixels are carried out by Novel Text Localization algorithm. Text characters in localized text regions are binarized and recognized by off-the-shelf optical character identification software. There is numerous language translators available where we have to manually enter the content and the application do whatever is left of the work of making a translation of the content into the desired language. Some translation platforms even charges for this process. For example the most popular Google translate API charges calculated as per the usage [7].

The proposed system carried out facial identification, image to text and text to speech conversion using Azure Cognitive Services in serverless architecture.

III. Workflow Diagram

In our project we divided our workflow diagram into two phases:

- Phase-I : Facial Identification
- Phase-II: Image Text to Speech Recognition

Fig. 1: Workflow Diagram of Face Identification

The workflow diagram of phase-I is depicted in Fig. 1. Single and multiple photos are taken for training. Training a model simply means learning (determining) good values for all the weights and the bias from labeled examples. In our project, we train the photo with the url link and name of that particular photo. After training, testing the photo. Testing a model means in relation to machine learning models which is primarily used to improve the accuracy.

Train and test model photo is conducted by Azure Face Api. This Api detects faces from a given image. It also compares two faces for similarities, along with identifying any specific face or person in a group of people. As a service response, it provides details to process facial images. It not only determines the human face coordinates, but also the emotion in the face.

Prepare the result text means the information we got from the photo. When we test the photo, if the photo is correctly trained then it is identify the correct name otherwise not. To recognise this, we use confidence level to know the right or wrong result. The position of the photo is also defined here.

Again, we use Azure Speech Api for converting text to speech in the existing photo which we call recognised face. This Api performs the complex task of converting audio to text, text to speech, and speech translation in a much simpler way.

Lastly, in the phase-I, the speech file uploaded in Azure Blob Storage which is a service for storing large amounts of unstructured object data, such as text or binary data. It helps to create data lakes for your analytics needs, and provides storage to build powerful cloud-native.

In the Fig. 2 depicts the phase-II of our workflow diagram. Here, Azure Vision Api used for detecting text from image. This service performs major analyses of the text from images. It can detect the language of the text, detect sentiments, detect key phrases, and list any linked entities. This Api enables to extract rich information from images to categorize and process visual data and perform machine-assisted moderation of images to help curate your services. Basically, it analyzes the image by type, color, etc. It provides lots of information about the image. It can tag the person's face and provide details like age and gender. It also gives score in percentages, for detecting images with adult content or racial issues.

After that, the text is converted to Speech using Azure Speech Api. Thus, we get the conversion what is our main aim for helping the visual impaired people. This whole system is works under serverless architecture. Serverless architecture (also known as serverless computing or function as a service, FaaS) is a software design pattern where applications are hosted by a third-party service, eliminating the need for server software and hardware management by the developer. In our project, we get the benefit using this architecture as it provides better benefit than traditional cloud-based architecture. It offers great flexibility and quicker time to release all at a reduced cost.

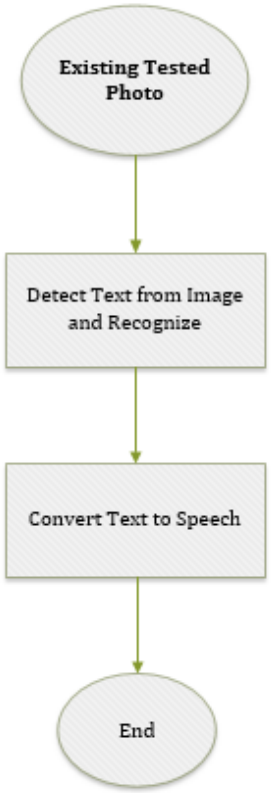

Fig. 2: Workflow Diagram of Image Text to Speech Recognition

IV. Experimental Results

A. Dataset

In this project, we use a lot of photos to train and test. After that, we detect text from that photos and convert those texts to speech. This photos mostly belongs to world famous persons.

B. Tools and Resources

In this process we are going to evaluate our built models using the evaluation metrics provided by the Azure Cognitive Services. For frontend, we used Angular 6 and Bootstrap and for backend we used .net core which an Azure function. For external services, we used Azure Face Api, Azure Vision Api, Azure Speech Api and Azure Blob Storage.

C. Results

Now, we use some single and group photo of Former President "Barak Obama" as our dataset. Train him and test him using Azure Face Api. After a number of attempt, finally our system able to detect him perfectly in the $4th$ iteration. Fig. 3 and 4 denotes the group photo of Obama and Fig. 5 and 6 depicts the single photo. Again, Fig. 7 is the tested photo.

From the following table 1, we get the best confidance level that is 84.90% for the Fig. 6 as when trained multiple times,

Fig. 3: Group Photo of “Barak Obama”

Fig. 4: Group Photo of “Barak Obama”

Fig. 5: Single Photo of “Barak Obama”

Fig. 6: Single Photo of “Barak Obama”

Fig. 7: Tested photo of “Barak Obama”

it correctly identify “Barak Obama”. For Fig. 1, our model detects number of faces and still it is unable to detect the right person, same as for Fig. 2. Again, for Fig. 3, it detects nearly perfect but for perfect we train the model again and get this desire result for Fig. 4. This is the output from our phase-I workflow diagram.

TABLE I: Results of Phase-I: Facial Identification

Image	Detected Face Number	Confidance Level on Test Results
Fig. 3	6	8.38%
Fig. 4	5	**1**7.70%
Fig. 5	1	70.88%
Fig. 6	1	**84.90%**

For phase-II, we use Azure Vision Api for detecting text from image and then use Azure Speech Api for converting text to speech from that image. Fig. 8 depicts the whole scenario of our phase-II workflow diagram.

In the Fig. 8, a picture is taken from our presentation slides’s front page, where our project title name, our name and ID are shown. Azure Vision Api extracts the image into text exact the same thing what it shows in the image. The Text portion recognise the output. Then, when we press the PLAY RESULTS button, then Azure Speech Api converts the text into speech and it pronounces the same text as it is seen in the Fig. 8.

V. Conclusion

As the multimedia system is growing day by day, not understanding a language cannot be a hindrance. The method allows the visually impaired to not feel at a disadvantage when interpreting text that is not written in braille. We would certainly support blind as well as physically disabled persons in our community with our contributions towards this mission. The primary goal of this project to train photos without requiring the user to directly tag the face rectangle the model should be able to benefit from images that have many faces and tags. Then we test the photo and after that detect test from that image and lastly convert the text into speech. Azure Cognitive Services in serverless architecture is used to complete our project. This helps one to make these people communicate with the natural world more easily. Future work could include

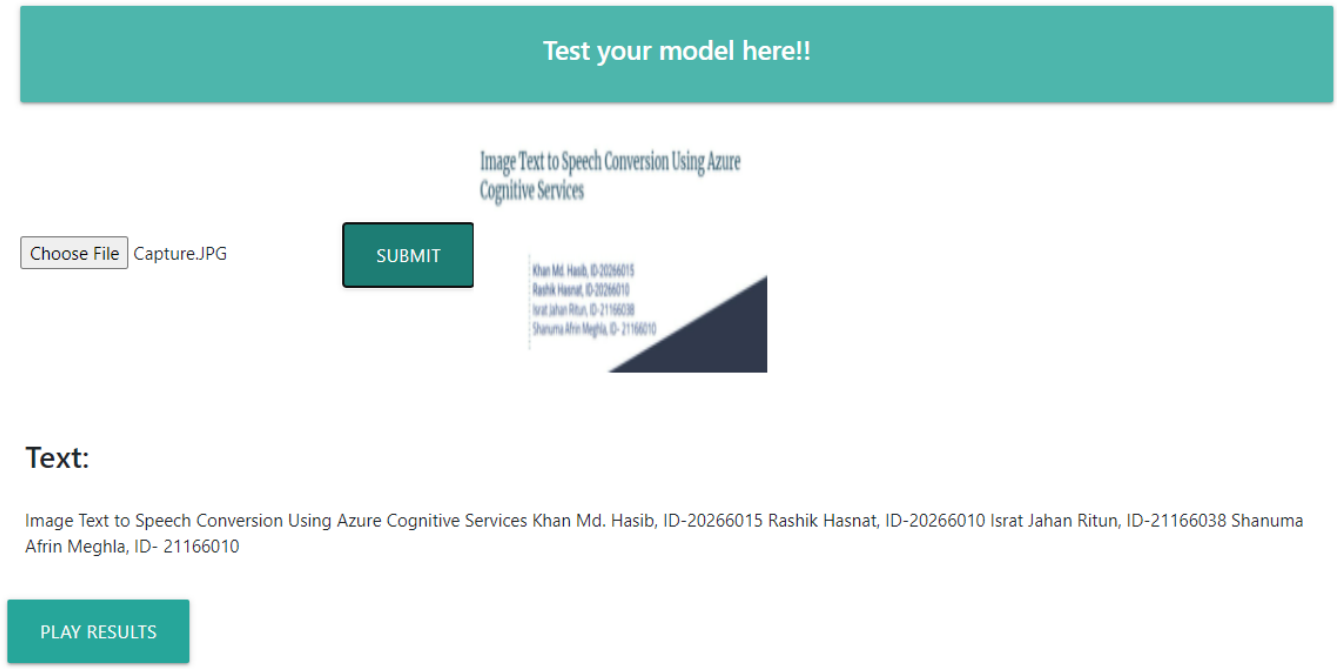

Fig. 8: Output of the Phase-II: Image Text to Speech Recognition

creating systems that track objects and retrieve text from videos rather than static images.

VI. Acknowledgement

We would like to thank **Md. Ashraful Alam Sir** for his time, generosity and critical insights into this project.

References

[1] D.Velmurugan, M.S.Sonam, S.Umamaheswari, S.Parthasarathy, K.R.Arun[2016]. A Smart Reader for Visually Impaired People Using Raspberry PI. International Journal of Engineering Science and Computing IJESC Volume 6 Issue No. 3.

[2] Rajesh, M., Rajan, B. K., Roy, A., Thomas, K. A., Thomas, A., Tharakan, T. B., & Dinesh, C. (2017, April). Text recognition and face detection aid for visually impaired person using Raspberry PI. In 2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT) (pp. 1-5). IEEE.

[3] Benjamin Z. Yao, Xiong Yang, Liang Lin, Mun Wai Lee and Song-Chun Zhu, "I2T: Image Parsing to Text Description" IEEE Conference on Image Processing, 2008 .

[4] Yi-Ren Yeh, Chun-Hao Huang, and Yu-Chiang Frank Wang, "Heterogeneous Domain Adaptation and Classification by Exploiting the Correlation Subspace," IEEE Transactions on Image Processing, vol. 23, no. 5, May 2014.

[5] S. Shahnawaz Ahmed, Shah Muhammed Abid Hussain and Md. Sayeed Salam, "A Novel Substitute for the Meter Readersin a Resource Constrained Electricity Utility" IEEE Trans. On Smart Grid, vol. 4, no. 3, Sept. 2013.

[6] Rajkumar N, Anand M.G, Barathiraja N, "Portable Camera Based Product Label Reading For Blind People.",IJETT, Vol. 10 Number 11 - Apr 2014

[7] Smith, R. (2007, September). An overview of the Tesseract OCR engine. In Ninth international conference on document analysis and recognition (ICDAR 2007) (Vol. 2, pp. 629-633). IEEE.