Using Computer Vision and Text-to-Speech for the Visually Impaired

Atanas Trpcheski[1] and Ivan Chorbev[1]

[1] Faculty of Computer Science and Engineering, University Ss. Cyril and Methodius, Skopje, Republic of North Macedonia
atanas.trpcheski@students.finki.ukim.mk

Abstract. Globally there are 49 million people living with blindness and an additional 295 million living with moderate to severe visual impairment. Sight loss of-ten has a drastic effect on the independence and well-being of individuals. From avoiding obstacles on their way to catching a bus, tasks that were once trivial become very challenging, increasing the risk of falls and collisions and the time and effort needed for daily life independence. Computer vision and artificial intelligence could let the blind and visually impaired independently perceive what is around them. Text-to-speech and object recognition AI is improving the lives of more than roughly 40 million people in the U.S. alone with eyesight and speech problems. This paper shows the designs of an intelligent camera app (for Android/iOS/Windows devices) and a separate Web App that can provide its users with narrative information about who and what is around them. Together with the app, we present the architecture be-hind it. This will show how it utilizes leading Computer Vision and Text-to-speech services and how they work together for a seamless and real-time experience for the user. This is particularly useful for someone with visual impairment. Just hold up your phone and hear a description of what's in the camera's field of view (just like asking a friend for help).

Keywords: Computer vision, Text-to-speech, AI, Azure.

1 Introduction

According to a report published by the World Health Organization [19], globally, at least 2.2 billion people have a near or distance vision impairment. In at least 1 billion – or almost half – of these cases, vision impairment could have been prevented or has yet to be addressed. This 1 billion people include those with moderate or severe distance vision impairment or blindness due to unaddressed refractive error (88.4 million), cataract (94 million), glaucoma (7.7 million), corneal opacities (4.2 million), diabetic retinopathy (3.9 million), and trachoma (2 million), as well as near vision impairment caused by unaddressed presbyopia (826 million) [1]. In terms of regional differences, the prevalence of distance vision impairment in low- and middle-income regions is estimated to be four times higher than in high-income regions. With regards to near vision, rates of unaddressed near vision impairment are estimated to be greater than 80%

in western, eastern and central sub-Saharan Africa, while comparative rates in high-income regions of North America, Australasia, Western Europe, and of Asia-Pacific are reported to be lower than 10%.

Population growth and ageing are expected to increase the risk that more people acquire vision impairment. Globally, the leading causes of vision impairment are [2][3][4]: uncorrected refractive errors, cataract, age-related macular degeneration, glaucoma, diabetic retinopathy, corneal opacity, and trachoma. Among children, the causes of vision impairment vary considerably across countries. For example, in low-income countries congenital cataract is a leading cause, whereas in middle-income countries it is more likely to be retinopathy of prematurity. As in adult populations, uncorrected refractive error remains a leading cause of vision impairment in all countries amongst children.

In this paper, we seamlessly utilize and combine into one-two leading services for object recognition (Computer Vision) and Text-to-speech.

A. Microsoft AI Computer Vision [10]. Computer vision is a branch of computer science that aims to help computers recognize and comprehend objects and people in photos and videos. Computer vision, like other types of AI, aims to execute and automate tasks that mimic human abilities. In this situation, computer vision attempts to recreate both how humans perceive and how they interpret what they see.
    - Object classification: The system categorizes the objects in an image into one of several categories. For example, a computer may use object classification to differentiate people from items in a photo and determine how many people are there.
    - Object identification: The system identifies a particular object in a photo, video, or image. For example, with object identification, the system would be able to distinguish people in a photo and analyze their appearance to determine the identity or traits of those people.
    - Object tracking: The system analyzes a video to process the location of a moving object over time. For example, with object tracking, a parking lot surveillance camera could identify cars in a parking lot and provide information about the location and movements of those cars over time.
    - Optical character recognition: The system identifies letters and numbers in images and converts that text into machine-encoded text that can be read by other computer applications or edited by users.

B. Microsoft Text To Speech Engine [11]. Text-to-speech enables applications, tools, or devices to convert text into human-like synthesized speech. The text-to-speech capability is also known as speech synthesis. It uses human-like prebuilt neural voices out of the box, or it is capable of creating a custom neural voice that's unique to a product or brand. This engine uses deep neural networks to make the voices of computers nearly indistinguishable from the recordings of people.

We are turning the visual world into an audible experience by using these leading services by pointing your phone. The users can complete multiple tasks with just one app:

- Short Text: Speaks text as soon as it appears in front of the camera
- Products: Gives audio beeps to help locate barcodes and then scans them to identify products
- Person: Recognizes friends and describe people around you, including their emotions, and get an estimate of their age and gender.
- Scene: Hear an overall description of the scene captured.
- Currency: Identify currency bills when paying with cash

2 Related Work

There are two apps that more or less share the same features we did for this paper. This section introduces them together with the pros and cons of using them related to our app.

A. Google Lookout [12] - Assisted vision: Lookout uses computer vision to assist people with low vision or blindness to get things done faster and more efficiently. Using the phone's camera, Lookout makes it easier to get more information about the world around and do daily tasks more efficiently like sorting mail, putting away groceries, and more. -
    - Built with guidance from the blind and low-vision community, Lookout supports Google's mission to make the world's information universally accessible to everyone.
    - Use Food Labels mode to quickly identify packaged foods by their label
    - Read a whole page of text with Documents mod
    - Use Currency mode to identify banknotes quickly and reliably, with support for US Dollars, Euros, and Indian Rupees.
    - Explore mode offers information about objects in your surroundings.
    - Lookout is available in more than 20 languages.

B. Microsoft Seeing AI [13]. Seeing AI is a free app that narrates the world around you. Designed for the blind and low vision community, this ongoing research project harnesses the power of AI to open up the visual world and describe nearby people, text, and objects. The app enables people to recognize:
    - Short Text - Speaks text as soon as it appears in front of the camera.
    - Documents - Provides audio guidance to capture a printed page and recognizes the text and its original formatting.
    - Products - Scans barcodes, using audio beeps to guide you; hear the name, and package information when available.
    - People - Saves people's faces so you can recognize them and get an estimate of their age, gender, and emotions.
    - Scenes - Describe the world around you.
    - Currency - Recognizes currency notes.

Here we describe some of the pros and cons of using these apps compared to ours.

Pros

- No internet connection is required.

- Big companies that have big teams investing in the overall look and feel.
- Support more than 20 languages

Cons

- The models they use inside the apps are pre-trained, which means the app needs to be constantly updated for fresh models. And they always have to be optimized only to support a specific set of functions as the apps will become huge in storage space. And this is their biggest downside. Our app has a real-time connection to the cloud AI, which means it always pulls the latest and greatest with every image taken (as the cloud AI is constantly being trained with new images coming all around the world). However, one downside of our approach is the latency between the picture being uploaded to the cloud AI and the audio response the user received. If there is a slow internet connection, the app can be unusable.

They are native apps. "Google Lookout" works only on Android devices, and "Microsoft Seeing AI" works only on Windows devices. Our app, using React Native, is cross-platform, which means we support both.

3 System Architecture and Information Flow

In Fig.1 we present the overall project architecture with details on each of the components used. The users interact with either the mobile app or a dedicated website (web app).

A. The mobile app is built using React Native [14]. Thanks to react-native, we write one code, but we can ship it to both Android and iOS devices. The mobile app is simple to use; the user points his device to a scene or object and takes a picture. After a slight delay, he will listen to a voice describing what he is looking at.
B. The website is an alternative to the mobile app and is built with .Net Core [15]. The reason behind using .Net Core is that developed web apps can be published on all operating systems (which will reduce costs). The most significant difference between the mobile and web app is that the users need to manually upload the photos for analysis, and once the analysis is done, he is presented with the description and audio prompt.

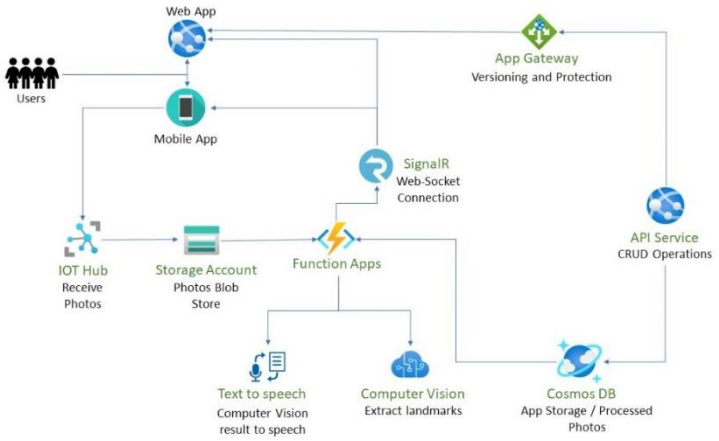

Fig. 1. Project architecture

The web app is used to show a list of all images that have been processed (together with the resulting analysis). So, it is mainly used to showcase the while architecture and capabilities. The mobile app does not have a way to show past activities as it will be used by the end-users who have some visual impairment and do not need this feature.

In the following section, we describe the Azure [16] Cloud Services [17] used in the backend of the app, the reason for using them, and how they are interconnected with each other for a seamless experience.

Once the user captures an image, the image needs to be sent to the cloud for processing. Because we can have multiple users who are uploading images at the same time, we need a service that can handle this load (as most services have a rate limit and can even see multiple requests like a DDoS attack). For this reason, Microsoft Azure is offering a service called Azure IoT Hub. It's a service that allows you to establish bi-directional communication with billions of IoT devices at the same time. So, it's built to handle high loads and will solve our problem with scaling the app to a lot of users.

IoT Hub temporarily uses Blob Storage to temporarily store the images for further processing in our pipeline.

We need a way for the architecture to pick up that a new image is uploaded and send this image for further processing to the AI. For this reason, we are using Azure Functions. They are basically listeners to events and have an additional benefit that they can also be programmed to talk to every service that Azure supports. So, they are used as listeners and also for intra-service communication. In our scenario, they will upload an image and send it to Computer Vision. This Azure function is named "blobToVision()".

Computer Vision is used to extract rich information from images and videos. It boosts content discoverability, automates text extraction, and can analyze videos in real-time. It is just one part of Azure Cognitive Services. We use it for visual data pro-

cessing to label content with objects and concepts, extract text, generate image descriptions, moderate content, and understand people's movements in physical spaces. No machine learning expertise is required.

Once the Computer Vision AI is done, another Azure Function will trigger and store the resulting analyzed data in the Azure Cosmos Database. Azure Cosmos DB is a fully managed, serverless NoSQL database for high-performance applications of any size or scale and 99.999-percent availability. This Azure function is named "visionToCosmos()"

At the moment the resulting Computer Vision data is stored in Azure Cosmos, a new Azure function will listen and pick it up to send to Azure Text-to-Speech for voice generation and creating a final '*.wav' audio file. This *.wav file is played back to the user but is also stored in the Azure Cosmos Database (as the web app can show all stored images and playback the audio files as well). Azure Text-to-speech allows us to build apps and services that speak naturally, with a customized, realistic voice generator that accesses voices with different speaking styles and emotional tones. This Azure function that sends the data for voice generation is named "cosmosToSpeech()." And the one that stores the resulting Speech data back to Cosmos is called "speechToComsos()"

At this point, we have all the data we need, and we just need to push it back to the user. For this, we use Azure SignalR. The Azure SignalR Service simplifies the process of adding real-time web functionality to applications over HTTP. This real-time functionality allows the service to push content updates to clients. As a result, clients are updated without polling the server or submitting new HTTP requests for updates. We feed the data from Cosmos to SignalR via an Azure Function called "cosmosToSignalR()". SignalR will then push the data to both Mobile and Web Clients.

The Mobile/Web users at this point will receive the message "Id" and will pull the REST API Service to get the data they need and play the audio.

In front of the API Service, we put an App Gateway service to protect the whole API (and with it the architecture) by adding auth and DDoS protection.

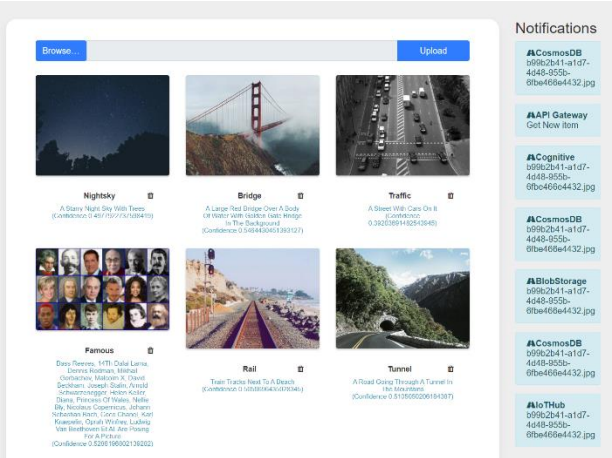

Fig. 2. Web application - Main screen

Some images from the web application together with the resulting analysis are given in figures below. Fig. 2 shows the web application main screen. It consists of a file upload control (that the user can use to upload new images), a grid with all the processed images in the database, and a side notification panel which is used to show the flow of data between all the services. The user can click on an image in the grid to get more details of objects detected and how the final summary was extracted. Fig. 3 shows an example upload of a picture with 'Train Tracks'. We can see that is successfully extracted the correct tags (objects) and did a precise summary ('Train tracks next to a beach'). The same is done in Fig. 4 but now with famous people from around the world. We can see that it was also able to pull their names and years.

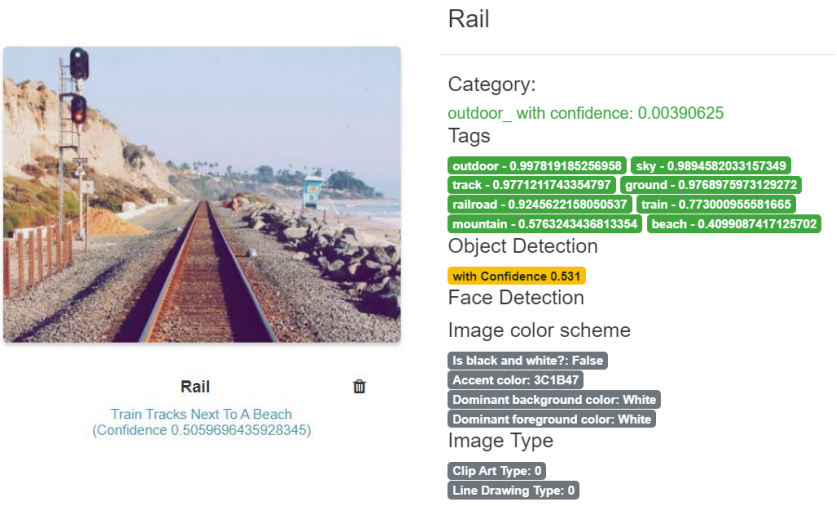

Fig. 3. Scene recognition, processed image "Train Track"

The web application is done mostly to showcase the architecture, how all the services are connected and talk to each other. It also shows all the details that can be extracted from the images in greater detail (as seen in Fig 2). It uses a File Upload control to upload pictures. It is not attended to be used by the target audience (visually impaired people). The mobile application is what will be used by the visually impaired; the user points his device to a scene or object and takes a picture (as it is shown in Fig. 5). After a slight delay, he will listen to a voice describing what he is looking at.

Fig. 4. People recognition, processed image “Famous People”

Fig. 5. Mobile App

4 Pros and Cons of the Proposed Solution

As described above, thanks to the Azure Functions, we have an architecture that does all of the processing, movement of data, and service inter-connectivity automatically. Thanks to the SignalR service and our underlying API, the users will receive the results automatically.

Pros:

- As we are using the Azure Services in real-time, we will always have the best predictions.
- We are using services that scale perfectly with a huge load and can offer 99.9% uptime.

Cons:

- The solution is highly dependent on the internet connection of the user. High speed and reliable internet will offer excellent results. In case of poor internet connection, we can expect more considerable latency.
- The solution is using cloud services, and they are all paid services.

5 Cost Estimation

These estimates, presented in Table 1, are done for real-world and high load/demand scenarios. We need to note that all these services also have Free Tiers (no charge) for up to 8000 simultaneous users.

Table 1. Microsoft Azure Cost Estimate

Service Type	Estimated Monthly cost	Estimated upfront cost
Azure IoT Hub	$10.00	$0.00
Azure Functions	$0.00	$0.00
SignalR	$10.00	$0.00
Storage Account	$61.40	$0.00
Azure Cosmos DB	$143.44	$0.00
App Service	*$146.00*	$0.00
Application Gateway	*$352.15*	$0.00
API Management	*$0.00*	$0.00
Azure Cognitive Services	*$10.00*	$0.00
Support	*$0.00*	$0.00
	$733.00	**$0.00**

6 Conclusion

Our solution will help people with visual impairment, and it will solve some of the everyday day-to-day tasks. But there are still a lot of gaps to fill. We still need to find solutions for some of the problems like:

- Navigation and wayfinding
    - Where to walk to transfer from one but to another?
    - How to find a doctor's office in a medical building?
    - Finding a place nearby to sit down
- Access to media, documents, and physical objects
    - Reading documents with complex layout (OCR limitation)
    - How to use touch panels, like from the microwave oven?
    - How to enjoy watching a movie?
    - How to understand a physical 3D object, like a biological model?

One future step would be to use the phone GPS and Google Maps POIs [18] (point of interest) to add new features to the app. Features like navigating by phone to a POI. The user needs to press the capture button on his phone to start the analysis. We can alternatively take automatic pictures every 2 seconds for real-time analysis. We also need to find a way to help the user hold the camera while walking. That can be something like a Lanyard, Shirt Pocket, or a wearable camera that we will pair with the phone. We will also need to think of a way to prompt to avoid obstacles if the user is moving while using the app. For now, the app is only set up for Text-to-speech in English, and we can easily support other languages as a future step. We only use Microsoft AI services for this paper and the app as described. Potentially, we can scale the app to use Google's AI and unity for even better predictions.

Our app is designed so visually impaired people can get details about objects and text around them with high precision. It's a smart app that uses the camera and sensors on the mobile device to recognize objects and text and tells the user what it finds. The is also a web app that users can manually upload images and see the algorithm results in more detail. The app uses a real-time connection with cloud AI services (it does not use any pre-trained models) which means it will always get the latest and best predictions. As time progresses, the cloud algorithms will just get better and better, and since we connect to the real-time, so will our app.

There are still a lot of gaps to fill, but hopefully, this app is one of the ways to help and get there.

References

1. Global Data on visual impairments 2010, WHO, Feb. 2021. https://www.who.int/blindness/GLOBALDATAFINALfor web.pdf (accessed May. 25, 2022).
2. P. Ackland, S. Resnikoff, and R. Bourne. World blindness and visual impairment: despite many successes, the problem is growing. In Community Eye Health, vol. 30, no. 100, Art. no. 100, 2017.
3. R. R. Bourne et al. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. In Lancet Glob. Health, vol. 5, no. 9, Art. no. 9, 2017.

4. R. R. Bourne et al. Global Prevalence of Blindness and Distance and Near Vision Impairment in 2020: progress towards the Vision 2020 targets and what the future holds.. In Invest. Ophthalmol. Vis. Sci., vol. 61, no. 7, Art. no. 7, 2020.
5. E. Munemo and T. Tom. Problems of unemployment faced by visually impaired people. In Greener J. Soc. Sci., vol. 3, no. 4, Art. no. 4, 2013.
6. Daily Life Problems Faced by Blind People. In Daily Life Problems Faced by Blind People, Feb. 25, 2021. https://wecapable.com/problems-faced-by-blind-people/ (accessed Feb. 25, 2021).
7. Challenges blind people face when living life. In Challenges blind people face when living life, Apr. 15, 2019. https://www.letsenvision.com/blog/challenges-blindpeople-face-when-living-life (accessed Feb. 25, 2021).
8. H. T. V. Vu, J. E. Keeffe, C. A. McCarty, and H. R. Taylor. Impact of unilateral and bilateral vision loss on quality of life. In Br. J. Ophthalmol., vol. 89, no. 3, Art. no. 3, 2005.
9. L. Zhi-Han, Y. Hui-Yin, and M. Makmor-Bakry. Medication handling Challenges among Visually Impaired Population. In Arch. Pharm. Pract., vol. 8, no. 1, Art. no. 1, 2017.
10. Computer Vision. An AI service that analyzes content in images and video. https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/ (accessed May. 25, 2022).
11. Text to speech. A Speech service feature that converts text to lifelike speech. https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech/ (accessed May. 25, 2022).
12. Google Lookout. Lookout uses computer vision to assist people with low vision or blindness to get things done faster and more easily. Using your phone's camera. https://play.google.com/store/apps/details?id=com.google.android.apps.accessibility.reveal&hl=en&gl=US (accessed May. 25, 2022)
13. Microsoft Seeing AI. Seeing AI is an artificial intelligence application developed by Microsoft for iOS. Seeing AI uses the device camera to identify people and objects, and then the app audibly describes those objects for people with visual impairment. https://www.microsoft.com/en-us/ai/seeing-ai (accessed May. 25, 2022)
14. React Native. Open-source UI software framework created by Meta Platforms, Inc. It is used to develop applications for Android, Android TV, iOS, macOS, tvOS, Web, Windows, and UWP by enabling developers to use the React framework along with native platform capabilities. https://reactnative.dev/ (accessed May. 25, 2022).
15. Net Core. .NET is a free, cross-platform, open-source developer platform for building many different types of applications. https://dotnet.microsoft.com/en-us/ (accessed May. 25, 2022).
16. Microsoft Azure. Microsoft Azure often referred to as Azure, is a cloud computing service operated by Microsoft for application management via Microsoft-managed data centers. https://azure.microsoft.com/en-us/ (accessed May. 25, 2022).
17. List of Azure Services. https://azure.microsoft.com/en-us/services/ (accessed May. 25, 2022).
18. Google POIs. Google Points of interest. https://www.google.com/maps (accessed May. 25, 2022)
19. Computer Vision-based Assistance System for the Visually Impaired Using Mobile Edge Artificial Intelligence, https://bit.ly/3BuVumb