North South University
Department of Electrical and Computer Engineering

CSE499B – Senior Design Project

Navigational Assistance for the Visually Impaired Using Computer Vision

Group Members:

Shadab Hafiz Choudhury	1631335642
Ishrat Jahan Ananya	1631636042
Sarah Suad	1632282642
Nabiul Hoque Khandakar	1631164642

Faculty Advisor:
Tanjila Farah
Department of Electrical and Computer Engineering

Summer 2020

Acknowledgments

We are extremely grateful to all the Faculty and Staff of Electrical and Computer Engineering Department of North South University for providing us with the opportunity and platform to carry out this project. We would also particularly like to acknowledge our supervisor, Ms. Tanjila Farah, Senior Lecturer of the Department of Electrical and Computer Engineering, for her efforts, providing guidance and advice over the course of our senior design project.

Chapter 1

Overview

1. Abstract

Blind people face many difficulties in daily life, one of which is navigation. There are several solutions leveraging the use of computer hardware and artificial intelligence to help guide them. However, most current solutions use complicated hardware and so are not suitable for everyone. Therefore, the goal of this project is to design and develop a functional substitute, or at the minimum, a complement, for the walking stick. Numerous solutions exist for outdoors navigation over longer distances, and the aim for this project is to specialize in indoors navigation instead.

Computer vision technology has advanced a long way. Depending on the type of neural network used, certain tasks can be accomplished at a very high speed even on limited hardware such as on a mobile phone. In a developing country like Bangladesh, there are hundreds of thousands of blind people who have extreme difficulty navigating indoor spaces due to being unable to see the layout of furniture and clutter in a room. The goal of this project is to develop a straightforward software-based solution that will help the visually impaired navigate indoor spaces by using just their mobile phones, avoiding the costs associated with hardware solutions. For this purpose, an android app would be developed that uses a deep learning algorithm to segment the image of a room into different sections that denote the floor, walls, furniture, clutter, etc. The app would then convert the segmented image into an audial description that a visually impaired individual can understand. We test ShuffleNet and DeepLabv3 , two of the most popular and effective architectures for semantic segmentation and implement the former into the app.

2. Table of Contents

1. Abstract 2
2. Table of Contents 3
3. Introduction 4
4. Background and Previous Research 5
    4.1. Use of a Singular Hardware Device such as a Walking Stick 6
    4.2. Use of multiple bodily mounted sensors. 6
    4.3. Software-based solutions for navigational assistance 7
5. Methodology 8
    5.1 Development of Mobile App 8
    5.2. Semantic Segmentation and Models 8
        5.2.1. ShuffleNet 9
        5.2.2. DeepLabv3 10
    5.3. Generation of Audio Output 11
6. Dataset 11
7. Conclusion 12
8. References

3. Introduction

Visually impaired, or blind people, face many difficulties in daily life. As they are bereft of visual stimulus, they cannot interact or navigate the world around them easily. Locating and picking up an object is a difficult task, as they have to feel around to find out where the object is. However, this is possible if they take their time. Similarly, although they cannot read, they can use braille for physical texts and text-to-speech software for computerized texts.

However, there is no easy way for them to navigate through a room full of obstacles, as their perception is limited to the length of a walking stick, which itself has an extremely narrow "field of view", and whatever they can touch. This causes a large number of inadvertent collisions with various obstacles in the room such as furniture and other objects, leading to injuries or damage to the contents of the room.

There are currently 285 Million people around the world with varying levels of visual impairment, ranging from moderate to severe. Out of this 285 million, 40 million are fully blind and have no sense of sight at all.

In Bangladesh, 6 million people exhibit some form of visual impairment, ranging from mild to severe. Estimates for the number of individuals who are fully blind range from 750,000 to 1 million [1, 2].

A lot of work has been done before on helping visually impaired people navigate. However, most prior research has been focused on a hardware-based solution, whether it is a high-tech walking stick equipped with a myriad of devices, or a multitude of sensors mounted all over the body. They mainly used radio or ultrasound waves to detect obstacles.

Despite being effective solutions, one of the major pitfalls of hardware-based solutions is that hardware is expensive to manufacture, import and maintain. In a developing country like Bangladesh, the majority of the 750 thousand blind people are underprivileged and will not be able to bear those expenses easily.

By contrast, software-based solutions have the advantage in that they have no material costs once development is complete. They can be duplicated and distributed for no cost beyond the initial development costs. Therefore, in order to solve the problem of helping visually impaired people navigate, we decided to turn to a software-based solution using computer vision.

One of the primary goals of Computer Vision research has been to emulate human sight and everything that entails. This includes object detection, object classification, motion analysis, etc. Without going into the details of each application of computer vision, the most important aspect for helping blind people navigate is object detection and classification.

The goal of this research is to use purely computer vision to help a blind person gain a rudimentary understanding of an interior area's layout, allowing them to plan out how to proceed. This would be a significant step in making moving around easier for them.

4. Background and Previous Research

A lot of work has been done on providing navigational assistance for visually impaired people. However, most prior research had been focused on a hardware-based solution, whether it is a high-tech walking stick equipped with a myriad of devices, or a multitude of sensors mounted all over the body. They mainly used radio or ultrasound waves to detect obstacles. While this is an effective method, it has the same shortcoming of being unable to help visually impaired people perceive anything outside a very narrow "field of view". As such, they were not a proper substitute for visual navigation.

A computer-vision based solution would come closer to providing the visually impaired with sight. It could act in a similar way as a human guide who directs the blind person away from obstacles. Using computer vision, we can also find out what type of obstacle it is, whether a chair, table, shelf, et cetera, further helping navigation. Previous research has been done on navigating outdoors using computer vision and on developing CNNs that can decipher the general layout of a room [3]. This project will take notes from the outcomes of previous research in similar fields but attempt to work from scratch to develop a functional device for indoors navigation with the intent of guiding a visually impaired person.

In practice, the final output device could act very much like some Self-Driving devices, albeit calibrated to suit a human. Taking inspiration from solutions used for Self-Driving cars, it was determined that techniques for 'Free Space Detection' [4] could be applied to this use case effectively. In vehicular automation, this approach usually utilizes multiple stereo cameras and LIDAR for 3D mapping. The developed project will instead simply make use of the core concept.

A short summary and analysis of the most influential papers, as discussed throughout the course, follows.

4.1. Use of a Singular Hardware Device such as a Walking Stick

A recent approach for a hardware-based solution for helping visually impaired people navigate was Sahoo, Lin and Chang's design of a Walking Stick Aid [5]. This solution integrated a total of eight different hardware components into a single walking stick. A Raspberry Pi and PIC microcontroller were used for control devices, and a vibration motor and buzzer were used for communicating with the user. The device integrated a discrete power supply and three sensors: Ultrasound, Water Level and GPS.

Though the device proved to be highly effective, especially when paired with a mobile app, there were a few shortcomings. The presence of so many different components on the device increases the weight, expense and risk of failure.

4.2. Use of multiple bodily mounted sensors.

A hardware-based alternative to using a walking stick is to use multiple sensors mounted on the body, as evidenced in Bousbia-Salah, Bettayeb and Larbi's work in developing a navigational aid for blind people [6]. This particular system attached two ultrasonic sensors to an individual's shoulders, an accelerometer to their waist, and several more ultrasonic sensors and vibration motors throughout their body and walking stick.

This system proved quite effective. However, it is a relatively older approach, and more modern approaches are able to fulfil the same goal with fewer sensors thanks to advancements in surface modelling. It also has the same disadvantages as the first paper examined.

Both these two hardware-based approaches have similar downsides. Therefore, a software-based solution is needed, which is examined next.

4.3. Software-based solutions for navigational assistance.

Saleh, Saleh, Nazari and Hardt's work on Outdoor Navigation for the Visually Impaired [7] is a fairly comprehensive approach based on deep learning. Here, a smartphone camera is used to get inputs, which is then segmented into different classes using the Google DeepLabV3 model. It is an effective solution that is actually suitable for use in all environments, not just outdoors.

However, the training and testing of the model was done on a dataset that used a large number of outdoors images. Therefore, it is not as efficient or accurate as it could be for interior areas. As a general computer vision-based solution for helping blind people navigate, this project is clearly effective and shows the strength of the approach.

As one of the weaknesses of image classification problems is having a context that is too broad, specialization should be carried out. It should be possible to develop a similar solution that is optimized for interior environments.

5. Methodology

With the background and previous research in mind, the solution developed in this research project involves:

Firstly, a mobile app that passes image frames to a computer vision algorithm continuously at a given rate/frames per second.

Second, a semantic segmentation algorithm that takes the passed frames and converts them to a segmented image where different classes of objects are detected and assigned a pixel colour.

Finally, an output function that 'reads' the segmented image and checks for walkable space or blocked space in areas where the user may walk. It then transmits this information to the user in the form of audio through text-to-speech.

It should be noted that the mobile phone, in this context, will be mounted at the user's shoulder and aimed at a slight angle downwards. This can be done using a simple system of straps and holsters. This position and angle are similar to that of a person with their gaze aimed at the floor a few steps in front of them. From this viewpoint, the bottom third of the image frame will cover the area immediately in front of the user's feet, while the middle third of the image will cover the area the user will reach with a few steps.

5.1. Development of Mobile App

The front-end of the mobile app includes a camera preview function built using Google's Camera2 API. This section is relatively simple, and will simply involve generating a preview using the aforementioned API and then extracting individual frames from the preview. TextureView and ImageView classes are used to display the camera preview and the segmented output respectively. The Handler class is used to assign separate threads for capturing preview frames and for processing the image segmentation. Once a preview frame is captured, it is placed in a Bitmap object. This breaks the image down into a grid and allows color data to be extracted from each square. The segmentation algorithm then classifies each pixel and assigns it a new color based on the class. After each segmented frame is evaluated using the output function, the segmented image is fitted to a bitmap and displayed. The output function also makes use of Android Developer's TextToSpeech class to notify the user of the evaluation. An auditory response will inform the user if the path is 'clear' or has an 'obstacle in front'.

5.2. Semantic Segmentation and Models

One of the most common applications of computer vision and deep learning is image segmentation. Image segmentation is simply the process of breaking down a normal image into different components that can be efficiently analysed by a computer.

There are two categories of image segmentation: instance segmentation and semantic segmentation. In instance segmentation, each individual instance of an object is segmented and labelled separately. In semantic segmentation, the overall image is broken down to different sections.

As the primary goal of the algorithm is to detect 'walkable space' versus 'non-walkable space', semantic segmentation is more suitable. There are many architectures for semantic segmentation available, such as Fast-CNN, R-CNN, U-Net, FCN, etc. Each architecture has its advantages and disadvantages in regards to processing speed and accuracy.

Two models were chosen for further work: Shufflenet and DeepLabV3

5.2.1. ShuffleNet

Shufflenet is a CNN architecture that was developed from the ground-up to be extremely efficient using limited computing resources [8].

One of the goals of the project was to minimize hardware costs. While the assumption can be made that even underprivileged people can afford a basic, low-end smartphone, it cannot be assumed that the device will have a lot of power. Therefore, ShuffleNet sacrifices accuracy for processing speed and limited resource use.

In ShuffleNet, group convolutions are utilized. In grouped convolutions, filters are separated into several different groups. Each group applies filters in a parallel manner, and at the end the final output layer is composed by combining the outputs of each group. Grouped convolutions were first proposed in AlexNet [9] and have been further developed since then.

Most of the computation cost is minimized by taking advantage of grouped convolutions. This method has one weakness – some information can get exaggerated or marginalized due to how grouped convolutions work. Therefore, a 'Channel Shuffle' function is carried out which flattens the output layers and minimizes the effects of information loss.

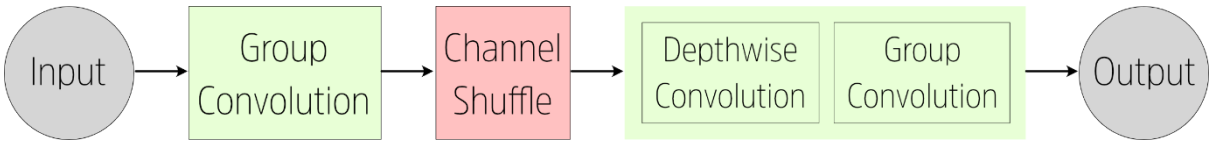

Fig. 1: ShuffleNet Architecture

5.2.2. DeepLabV3

Developed by Google, DeepLabV3 is one of the most accurate and powerful architectures for semantic segmentation available today [10]. The DeepLabv3 architecture makes use of Atrous Convolutions (also called dilated convolutions). In an Atrous convolution, holes are placed between each unit or pixel of the filter. So, each convolution will cover a wider area, but the filter size will remain the same.

For example, a 3x3 filter with a Dilation Rate of 2 will cover the same area as a 5x5 filter. The second and fourth positions for the rows and columns will be discarded. In DeepLabv3, a spatial pyramid is used, using multiple Atrous filters with increasing Dilation Rate in succession.

The architecture of DeepLabv3 is given below.

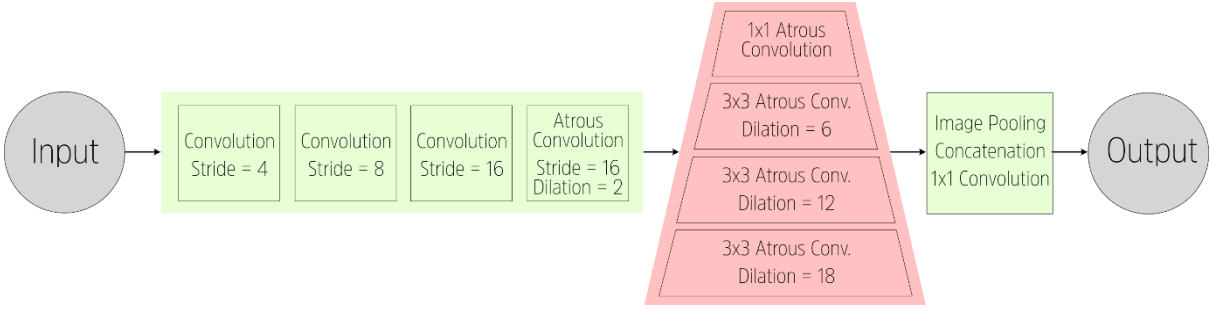

Fig. 2: Deeplabv3 Architecture with Atrous Spatial Pyramid and Pooling

5.3. Generation of Audio Output

Once a segmented image has been received, the final task is to generate an audio output. The output segmented image will be divided into a grid of 9 parts. For guiding a user, the most important part is where their next few steps will be.

Background	Background	Background
Towards the Left	Next 2-4 Steps	Towards the Right
Heading Left	Current Step	Heading Right

Fig.3: Grid Organization

The presence of free walking space or clutter is determined by the colour of the segmented image in that particular section of the grid.

If purple is used to denote the floor in the segmented image, then the function will check for the prevalence of purple in the 'Current Step' position of the grid. This current step is the most important position in our grid, since it is where the user will step next.

If another object appears in the 'Current Step' position, then the segmented output will return a different colour. The function will send a call to the text-to-speech function to tell the user "Obstacle in front!" This message will have the highest priority.

Additional warnings can also be built in in a similar way. If an obstacle is detected in 'Heading Left' or 'Heading Right', the system can tell the user "Obstacle to left/right!" If an obstacle is detected in the "Next 2-4 Steps", the system will tell the user "Obstacle up ahead."

Overall, the expectation is that the user will be aware of obstacles nearby, and stop as necessary when they get too close.

6. Dataset

The primary dataset to use in this project is the MIT ADE20k Dataset for Scene Segmentation [11, 12]. This dataset features 20,120 images taken from a wide variety of scenes both outdoors and indoors. As there is a high incidence of indoor images, this dataset is better suited for the context of this research than other popular image segmentation datasets such as Cityscapes or PASCAL VOC.

The ADE20k Dataset features a total of 150 Classes. However, most of these classes are either superfluous, or too finely detailed, for the task at hand. Therefore, the class labels were consolidated into the primary classes that will be applicable to the process of interior navigation. The consolidated class labels are given below:

1 (wall) <- 9 (window), 15 (door), 33 (fence), 43 (pillar), 44 (sign board), 145 (bulletin board)
4 (floor) <- 7 (road), 14 (ground, 30 (field), 53 (path), 55 (runway))
5 (tree) <- 18 (plant)
8 (furniture) <- 8 (bed), 11 (cabinet), 14 (sofa), 16 (table), 19 (curtain), 20 (chair), 25 (shelf), 34 (desk)
7 (stairs) <- 54 (stairs)
26 (others) <- Class number larger than 26

7. Conclusion & Future Work

Using Deep Learning, interior surfaces can be efficiently and accurately deciphered in a format suitable for the computer. As such, it can be used to develop an extremely cost-effective solution for helping blind people navigate in an environment full of obstacles, such as interior spaces. As the system is in real-time and implements a 'Free Space Detection' approach, any sudden changes in the environment such as a person walking in front will also be detected by the system. Therefore, this research proposes that a semantic segmentation model is developed using one of the mentioned architectures and datasets, and integrated into a mobile app that will handle the inputs and outputs. Such an app would be quite useful to the visually impaired.

With regards to our future work, we will concentrate on 1) developing a more efficient semantic segmentation architecture with an enhanced inference pipeline, 2) building up a more tailored dataset with annotated images taken from the perspective of visually impaired people, 3) providing a more intuitive feedback to the VI by using haptic interfaces, spatial sound etc.

A long-term priority will be to increase the speed & accuracy of the inference (generating outputs from images) process on our proposed system. The simplest way to accomplish this is to delve into implementing better state-of-the-art neural network models that are capable of delivering more accurate inferences on a wider range of inputs. The usage of these such neural networks though comes with a cost of taking up more space on the user's smartphone. These big multi-layered networks can often take up upto 500 Megabytes on the user's smartphone, which is something that we cannot afford if we are trying to put out our solution on a mobile device.

As a resolution to this problem, we can set up the smartphone's camera to stream data to a server running our algorithms. This server will include Graphics Processing Units (GPU's) to further boost the time and accuracy of our inferences and increase the frame rate of our system while freeing up CPU cycles on the smartphone for other processes. With the help of ever advancing mobile internet technology, this can be done seamlessly and without taking up much of the resources of the user's smartphone

When evaluating technical feasibility and performance, there is a lack of standard computer vision benchmark datasets out in the open. Relatively small datasets can be collected from open-sourced resources, making it impossible to compare the performance between various solutions. It may have been caused by the lack of open computer vision datasets tailored to developing solutions for the VI. Generic benchmark datasets, such as ImageNet (classification) and COCO (object detection) could potentially be used to some extent. However, these datasets lack objects of vital importance to the VI, for example, corridors, stairs, elevators.

A collaborative effort for creating a high-quality benchmark dataset for developing and evaluating computer vision solutions for the VI is required. User-based evaluation presents a common challenge in the academic community. The involvement of VI individuals to research projects is challenging and often suffer from selection bias. Moreover, lack of standardized evaluation methods and potential reporting bias limit the representativity of these experiments. To aid the problem in hand, we will be looking to create a custom dataset of our own with the help of VI people by capturing annotated images of the visually impaired people. With this, we are expecting to acquire a healthy amount of samples for training and that too by focusing on objects like stairs, corridors and all other necessary elements that conventional datasets often leave out. We hope to gain valuable insight into specific requirements for subjects who have been coping with visual dysfunction for a long period of time, that might not be obvious from data collected on normal sighted subjects who have had simulated vision loss for only a brief period.

Lastly, we would like to address the development of tactile/acoustic interfaces to provide better feedback to visually impaired people when it comes to alert them regarding obstacles on their path. The app we have implemented performs localization only, but we hope to transform it into a full-featured wayfinding app with an accessible UI that offers turn-by-turn directions to a desired destination as well as optional announcements of nearby points of interest and also provide haptic feedback as an alert when coming across a hazard or caution.

Some visually impaired travelers might prefer navigation directions presented using spatialized (3D) sound, as implemented in the Microsoft Soundscape App, and we will experiment with this type of interface as a possible alternative (or supplement) to verbal directions. We note that travelers with residual vision may prefer a visual UI (e.g., an Augmented Reality interface that superimposes high-contrast arrows on the smartphone screen to guide the user) over an audio one. An extensive evaluation/refinement process is to be carried out with the help of blind users, aimed at improving system performance and usefulness of the system.

8. References:

[1] Sutradhar, I., Gayen, P., Hasan, M. et al. Eye diseases: the neglected health condition among urban slum population of Dhaka, Bangladesh. BMC Ophthalmol 19, 38 (2019). https://doi.org/10.1186/s12886-019-1043-z

[2] Six million Bangladeshis visually impaired: speakers, The Daily Star, URL: https://www.thedailystar.net/city/news/six-million-bangladeshis-visually-impaired-speakers-1856473, fetched on 24/10/2020

[3] V. Hedau, D. Hoiem and D. Forsyth, (2009) "Recovering the spatial layout of cluttered rooms," 2009 IEEE 12th International Conference on Computer Vision, Kyoto, pp. 1849-1856.

[4] Neumann, L., Vanholme, B., Gressmann, M., Bachmann, A., Kahlke, L., & Schule, F. (2015). Free Space Detection: A Corner Stone of Automated Driving. 2015 IEEE 18th International Conference on Intelligent Transportation Systems. doi:10.1109/itsc.2015.210

[5] Sahoo, N., Lin, H.-W., & Chang, Y.-H. (2019). Design and Implementation of a Walking Stick Aid for Visually Challenged People. Sensors, 19(1), 130. doi:10.3390/s19010130

[6] Bousbia-Salah, M., Bettayeb, M., & Larbi, A. (2011). *A Navigation Aid for Blind People. Journal of Intelligent & Robotic Systems, 64(3-4), 387–400.* doi:10.1007/s10846-011-9555-7

[7] Saleh, S., Saleh, H., Nazari, M. A., Hardt, W., (2019). Outdoor Navigation for Visually Impaired based on Deep Learning, Actual Problems of System and Software Engineering (APSSE 2019), 2514.

[8] Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6848-6856).

[9] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

[10] Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

[11] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso and A. Torralba, (2017), "Scene Parsing through ADE20K Dataset," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 5122-5130, doi: 10.1109/CVPR.2017.544.

[12] Zhou, Bolei, et al. (2019), "Semantic Understanding of Scenes Through the ADE20K Dataset." International Journal of Computer Vision 127: 302–321. https://doi.org/10.1007/s11263-018-1140-0