Navigational Assistance for the Visually Impaired Using Computer Vision

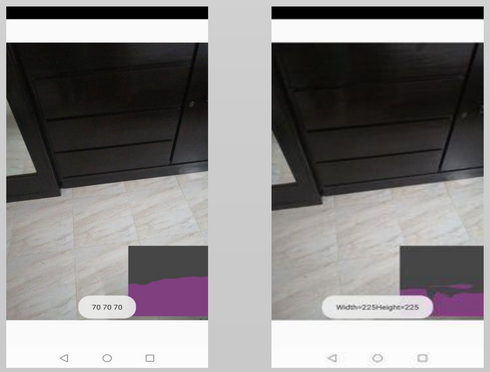
Shadab Hafiz Chowdhury, Ishrat Jahan Ananya, Sarah Suad, Nabiul Hoque Khandakar
Department of Computer Science and Engineering, North South University

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. This project uses deep learning to implement a semantic segmentation algorithm that recognizes walkable areas in an interior environment in real-time, directing users away from obstacles such as furniture or people. We test ShuffleNet and DeepLabv3 and implement the former into an app that can be used on any android phone.

Introduction

Visually isual stimulus, they cannot interact or navigate the world around thimpaired, face many difficulties in daily life. As they are bereft of vem easily. Locating and picking up an object is a difficult task. 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”.
There are currently 285 Million people around the world with varying levels of visual impairment. In Bangladesh, 6 million people exhibit some form of visual impairment, ranging from mild to severe.
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. One of the primary goals of Computer Vision research has been to emulate human sight and everything that entails. 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. Below is a demo of the project.



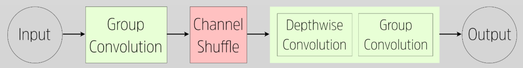
Methodology

The development process had three distinct steps. 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 converted 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.

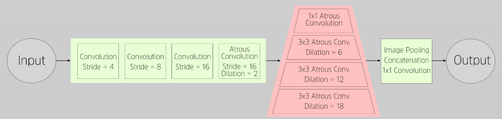
Semantic Segmentation

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. Below we discuss two models chosen for the task.

Shufflenet
Shufflenet is a CNN architecture that was developed from the ground-up to be extremely efficient using limited computing resources. 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 has been further developed since then.



DeepLabV3
Developed by Google, DeepLabV3 is one of the most accurate and powerful architectures for semantic segmentation available today [10]. The DeepLabv3 archtecture 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.



Audio Output Generation
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.

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

Results

The final output is generated based on the grid where 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.

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

Conclusion

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.
Furthermore, we could exercise a collaborative effort to develop astandard benchmark dataset for this and develop a more tactile interface to give audio feedback to the user.

Acknowledgements

Special thanks to our supervisor MS. TANJILA FARAH, Senior Lecturer , ECE Department, NSU