A Research Proposal for PhD Advisory Committee

Intelligent Navigation System using Deep Learning for Visually Impaired

By

Viral Parmar

Assistant Professor

Information Technology Dept.

Shantilal Shah Engineering College

INDIA

Table of Contents

List of Figures

INTRODUCTION

Visual impairment has been a serious problem affecting every walk of life. Statistics from World Blind Union shows that there are about 253 million persons visually impaired (VI) or blind. A visually impaired person can only sense a few modalities like light, shadow and hearing, and not be able to visually perceive the objects in front of them. They move around and sense the environment based on their experiences using cans which guide them to detect stationary objects and to avoid collision with any moving object. But this isn't enough as guide cans don't guarantee safety because they can't perceive types of objects, dimensions of objects and how they are aligned or oriented.

Evolution in technology, advancement in computing power, memory size and sensor technologies, have contributed to the use and development of technology solutions to help people with disabilities. In today's era where technologies like Amazon Echo, Alexa, Google Home and smartphones for Visually Impaired are paramount and some systems describe images for people with visual impairment. Moreover, technology could enhance the way to experience and sense the surroundings and provide an exact description of the environment.

The main goal of this proposal is to develop a system that can understand the way objects are placed in the environment allowing visually impaired persons to navigate more precisely in an outdoor environment. Our system will be using a deep learning approach to understand the 3D position and approximate dimensions of the object in space. The Navigation guidelines will be provided to the user in the form of audio signals.

BACKGROUND AND RELATED WORK

Existing Computer Vision Systems for VI-person can be broadly classified into mainly four categories: Object Detection, Image Segmentation, Image Classification and Object Recognition. In the past few years, many systems have been developed to assist Visually Impaired Persons using ensemble frameworks encompassing earlier mentioned categories in computer vision.

The state of art assistive model for obstacle detection based on deep learning by [1] was developed for VI People. The main idea behind this research work was to use YOLO v3 for object detection with Darknet-53 as a backbone network and to implement this system on smartphones. This remained as a state of art model as this system was lacking the ability to calculate the distance between obstacles for VI people. At a well-known International Conference on Contemporary Computing Application under IEEE chapter Dhruv Dahiya [2] and his team presented an approach based on Deep Learning to support VI people by helping them in identifying some basic amenities like Restrooms, ATM, Metro, Stations and Pharmacies. The technical grounds were laid on a combination of resnet-50 and Faster-RCNN for object detection. An Accurate and efficient deep learning-

based real-time framework has been proposed. It can be concluded that this method is cheap as it was only made for only 4 types of public amenities.

Despite significant recent developments, currently available visual assistance system cannot adequately perform complex computer vision task that entails Deep Learning. In [3] author proposed an efficient and faster way by employing AI accelerator hardware such as Neural Compute Stick-2 (NCS2) along with model optimization technique (Open VINO) and Smart Depth Sensors such as Open CV AI kit-depth (OAK-D). The Proposed system was able to assess the traffic conditions and detection and localization of hanging objects, crosswalks, moving objects. However, running semantic image segmentation models simultaneously along with other models is nontrivial. None of the image segmentation models was able to run on a single NCS2. The performance of the whole system was suffering due to the usage of few segmentation models. The training process of the neural network is difficult and complex because it requires large resources with an important test phase process. Therefore, depending on AI Accelerator hardware such as NCS2 is not sufficient and effective either. Further, Saleh Shadi [4] and his group, which was showing a comparative study of various Deep Learning Models such as DeepLab[5], Fast R-CNN[6], Muiltibox[7], SSD[8] and DeepLab V3+[9] for Object detection and classification. Among these models DeepLab V3+outperformed other models in terms of size and inference time. DeepLab V3+ is considered one of the most powerful techniques for semantic images segmentation with Deep Learning. This model was used by the authors to develop an Outdoor Navigation system for VI people. It overcomes the limitations of a Navigation system mentioned earlier using hardware accelerators.

Many Challenging issues are still unresolved due to the complex nature of indoor/outdoor scenes and the need to recognize many types of objects in a short period for the system to be useful to a blind person. One way to reduce the complexity of the problem is not to care about providing the exact location of objects in the scene to the blind person. Instead, only detect the mere presence of the objects in the scene. Then this problem becomes similar to what is known as image multi-labelling. The Implementation for the same was carried out in [10] using light-weight pre-trained CNN called SqueezeNet. The architecture of SqueezeNet was improved by resetting the last convolution layer to free weights, replacing the Activation function from RelU to LeakyReLU and adding a batch normalization layer.

So, up till now, there exist some vision-based assistive systems which aim to improve the perception experience of VI. Still, it is extremely difficult for VI to well understand their surroundings. OrCAM Provides the means to get some special information such as reading tests and detecting traffic lights. However, this information is not enough for VI to understand their surroundings. Lately, a growing amount of success has been reported in the field of vision navigation and semantic segmentation tasks. Inspired by these works, in [11] a group developed a data-driven learning approach to predict the walkable instruction with supervised learning over training images collected from an RGBD sensor. Besides RGB and depth data, a semantic label was employed to train the network to detect low lying objects and also to tell the environment information around a VI. But

this came with a limitation that depth sensors always come across failure at glass walls or doors, which is common to see in indoor places.

Mobile application has also been developed for VI people for visual assistance among them the few well-known applications like Seeing AI[12] and Be My eyes[13] are widely accepted. Seeing AI narrates the scene around you. In addition to that it can read the printed documents, scan the barcodes, currency identification and describe the perceived colour. On the other side Be My eyes is a very distinctive idea where the VI peoples are assisted by people having clear vision. But with emerging advancement in AI, standalone systems/Applications for VI person are the future. Amplifying the current Mobile applications for VI with more functionality such as finding the orientation of objects, navigational guidance is the key way to overcome the problems that are their in the existing system.

The enhancement in the navigation system is obvious and therefore the proposed system is now focusing on finding the orientation of an object in 3D space so that the VI could get the instruction very accurately to overcome the obstacle.

RESEARCH GAP

- Most computer vision-based Assistance Systems of VI are developed to detect only the presence of an object in the scene regardless of its size dimension or orientation.

- The hardware unit made to be mounted on VI for visual assistance is inefficient in detecting more than 4-5 classes and in addition to that the hardware unit is not able to detect low lying object.

- Most of the systems are sensing obstacles/objects in front of VI. Which does not give 360-degree experience to a VI

- Standalone systems with limited hardware recourse are difficult to update and it's difficult to accommodate a well-trained deep learning model to detect various classes of objects.

- Using Capsule Network as a new approach to finding the Orientation of an object in a scene.

SCOPE OF WORK

- Developing a navigation system for VI using Deep Learning to capture various modalities like the 3D orientation of an object (i.e., Vehicles) and approximating the dimension of the object in a 3D Space. These modalities will help the navigation system to guide a Visually impaired person more precisely by guiding his movement concerning size and shape and orientation of the object to effectively overcome it. For example: If a car is parked in front of a VI, then the navigation system could be able to tell the orientation (Horizontally parked, vertically parked or at some angle). And how much VI must move to overcome the car.

- It will also detect the change in the elevation in advance while walking on roads.

- The system will be interacting with the Data Cloud for fast inference of objects.

- At the initial level, we will be working with object detection for Cars and Buses.

- Implementing this concept in LiDAR to sense the surrounding environment.

METHODOLOGY AND APPROACH

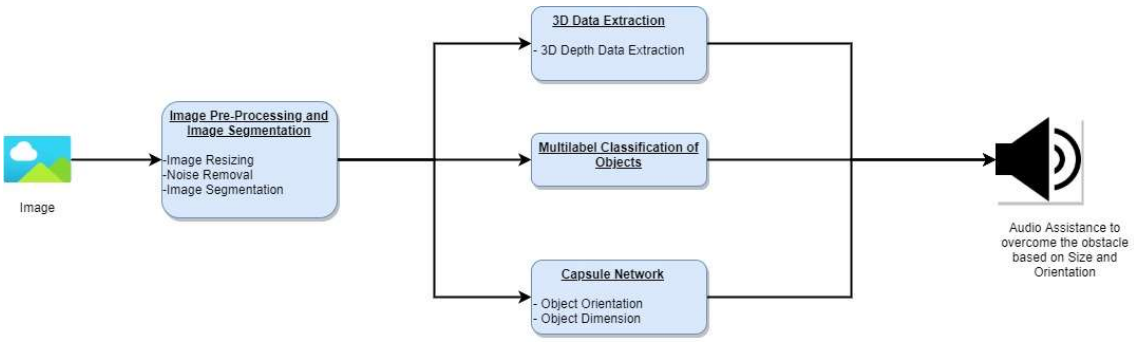


Figure 1: Proposed Work Flow Chart

Image Pre-Processing and Semantic Segmentation

Mostly Image pre-processing consist of Pixel brightness transformation, Geometric transformation, Image filtering and smoothing. Roboflow is one click pre-processing option as well as it could be used to handle the annotation work. It's free, and we can use it with our models whether they're written in Tensorflow, PyTorch, Keras, or some other tool. This will make our pre-processing step fast. And the image pre-processing functionalities offered by Roboflow are enough for my research work to carry on with.

Semantic Segmentation will be another additional layer that will be helping our AI model to recognise various objects in a scene. The most popular semantic segmentation architecture developed for biomedical images segmentation was **I-Net**[14], it's an extended version of U-Net, now applied to a variety of use cases like Self Driving Car to capture different segments or Different classes in real-time, I-Net is more successful than conventional models, in terms of architecture and terms pixel-based image segmentation formed from convolutional neural network layers. It's even effective with limited dataset images. We could extend the application of I-Net for segmenting Vehicles images.

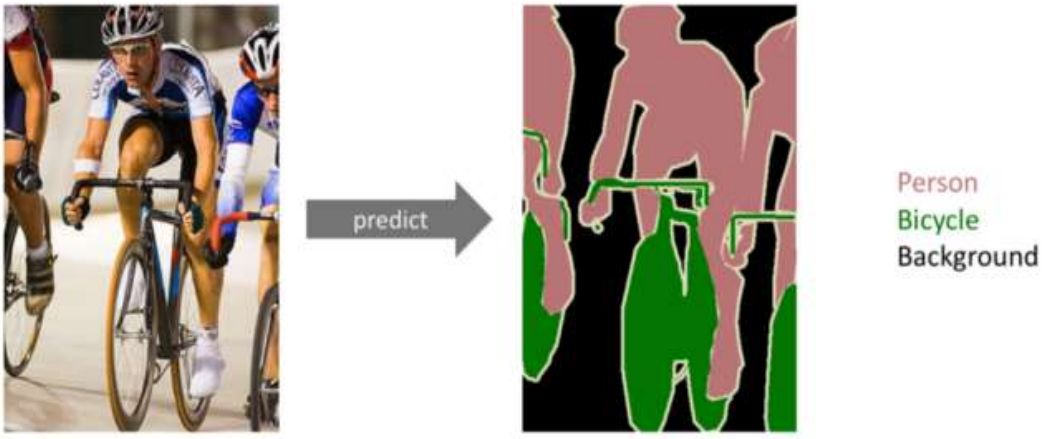


Figure 2: Sample of Semantic Image Segmentation[15]

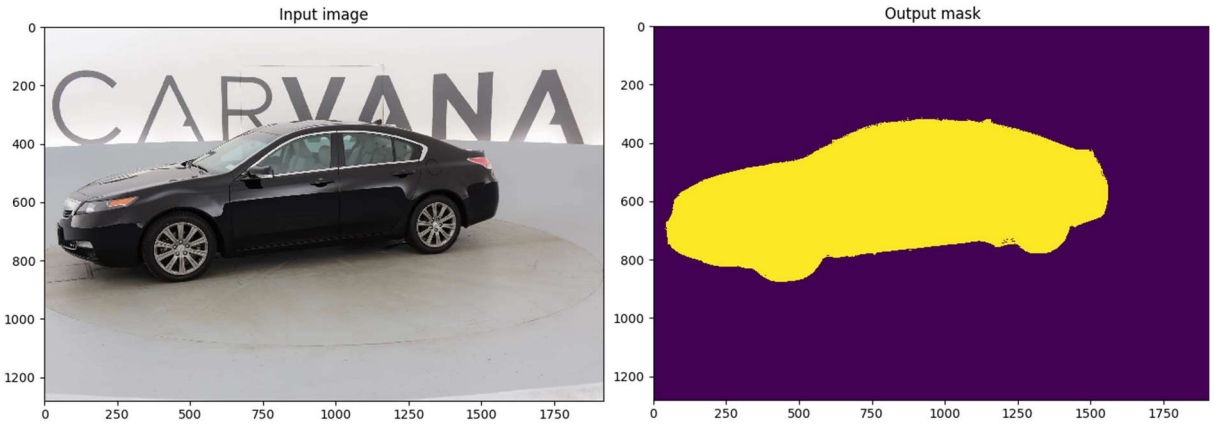


Figure 3: Image Segmentation of Car Image[16]

3D Depth Data Extraction

Predicting depth from a single image is an important problem for understanding the 3-D geometry of a scene. The Perception of the environment is the crucial task for a VI to solve the problems like localization or navigation. The Microsoft Kinect Sensor

provides highly detailed depth images of the environment and is therefore well suited for this task.

Multilabel Classification of Objects

To do Multi-label Classification, we need a good feature extractor, we need a good CNN network as a first part to perform feature extraction. EfficientNet[17] is simple and efficient at performing feature extraction and then we will be designing a simple Multi-label Classifier using CNN. It will predict a vehicles body colour, body direction and body type. Combining these two modules, you can do vehicle detection and multi-label recognition at the same time. Based on this info, some structured info's in outdoor traffic scenes can be extracted.

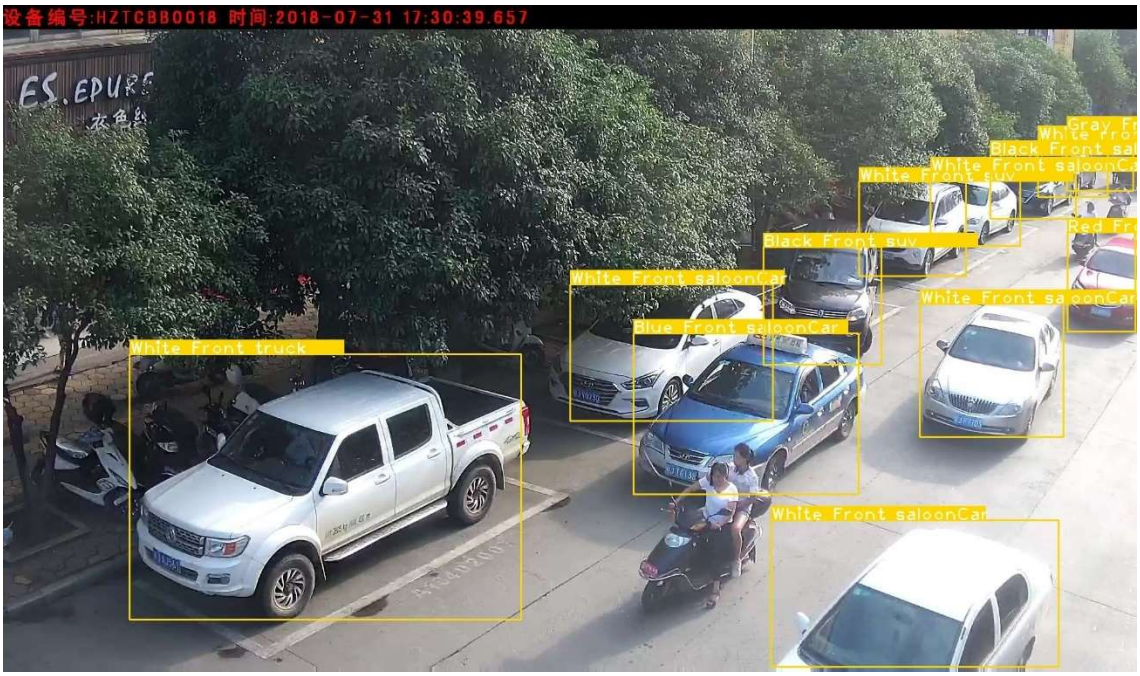


Figure 4: Multi-Label Image Classification[18]

Spatial Information

The convolutional neural networks and its variants are promising for the image classification and object detection tasks, it suffers from some limitations, too. The first limitation is the spatial information of objects is not to be considered due to the pooling layer. The second limitation is a minor modification in the image can change the output results, which means performs poor with adversarial attacks sometimes. To overcome such limitations, Capsule networks are formed. Capsule networks preserve spatial information of the objects and use them for detection. Plus, capsule networks are equivariant, which means modification in the image such as rotation, translation cannot vary the output most.

The research work is related to analyse the working of the new architecture of deep learning called capsule network. Capsule networks work with capsules that generate

output with instantiation parameters like orientation, thickness, skew etc. By analysing the capsule networks, orientation details could be found from the output vector.

The proposed research tries to get the orientation of the object from the output vector of the capsule network. Proper orientation detection of objects can lead to novel development in the field of artificial intelligence and deep learning. Object orientation addresses the major prevalent issue of detecting the way the object is placed on space in the form of horizontal or vertical orientation. VI people could be benefited by knowing the orientation of an object to overcome it properly without any accident.

Audio Assistance

We will be using Bluetooth enabled wireless earbuds for audio assistance in this project. For Text-to-speech an offline package Festival speech synthesis system will be used, which will be giving environment description to VI. Few of the commands that are to be given by the VI will be recognized by Vosk Speech Recognition Library which is having higher accuracy in speech recognition. Triggers words will be used to start the system such as (trigger word: "start") to start the system and in the same way trigger words like "describe the front", "describe the left" "describe the right" will describe the view in front, left and in right directions. At initial stage the description we will be starting will description of car, bike, and heavy vehicles. Once the system will calibrate the distance from the object and the orientation of the object, it will guide the VI person saying "Go to your left or right to overcome the obstacle. At during that course of time the system will be going to warn user if there is something on left while he is moving left to over come obstacle and vice versa.

NAVIGATION SYSTEM EVALUATION DATASETS

A key step in validating a proposed idea or system is to evaluate over a suitable dataset. Below is the list of the datasets that will be used to validate the proposed system:

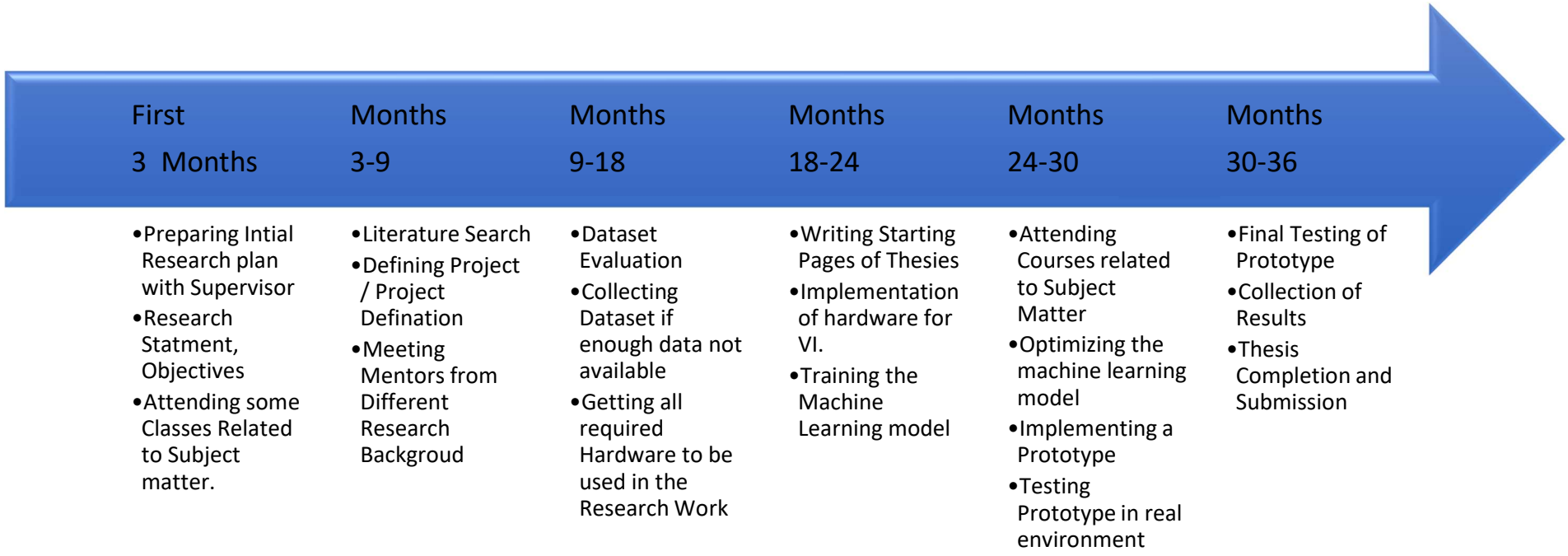
- **ApolloCar3D**[19]**:** It contains 5,277 driving images and over 60K car instances, where each car is fitted with an industry-grade 3D CAD model with the absolute model size and semantically labelled key points. This dataset is above 20 times larger than PASCAL3D+ and KITTI, the current state-of-the-art.
    - **Usage/Applications**
        1. **GSNet**[20]: Geometric and Scene-aware Network an end-to-end multi-task network that can estimate 6DoF car pose and reconstruct dense 3D shape simultaneously. Vehicle pose estimation and shape reconstruction are tightly integrated into the

system and benefit from each other, where 3D reconstruction delivers geometric scene context and greatly helps improve pose estimation precision

2. **Optimal Pose and Shape Estimation for Category-level 3D Object Perception**[21]: a category-level perception problem, where one is given 3D sensor data picturing an object of a given category (e.g., a car), and has to reconstruct the pose and shape of the object despite intra-class variability (i.e., different car models have different shapes).

- **NUSCENES**[22]**:** The strength of nuScenes is in 1000 carefully curated scenes with 3d annotations, which cover many challenging driving situations. nuImages complement this offering by providing 93,000 2d annotated images from a much larger pool of data. We also provide past and future camera images, resulting in a total of 1,200,000 camera images.
    - **Usage/Applications**
        1. **Cityscapes 3D**[23]: Dataset and Benchmark for 9 DoF Vehicle Detection. Detecting vehicles and representing their position and orientation in the three-dimensional space.
        2. **MonoLoco**[24]: Monocular 3D Pedestrian Localization and Uncertainty Estimation.

WORK PLAN


First 3 Months
•Preparing Intial Research plan with Supervisor
•Research Statment, Objectives
•Attending some Classes Related to Subject matter.
Months 3-9
•Literature Search
•Defining Project / Project Defination
•Meeting Mentors from Different Research Backgroud
Months 9-18
•Dataset Evaluation
•Collecting Dataset if enough data not available
•Getting all required Hardware to be used in the Research Work
Months 18-24
•Writing Starting Pages of Thesies
•Implementation of hardware for VI.
•Training the Machine Learning model
Months 24-30
•Attending Courses related to Subject Matter
•Optimizing the machine learning model
•Implementing a Prototype
•Testing Prototype in real environment
Months 30-36
•Final Testing of Prototype
•Collection of Results
•Thesis Completion and Submission

HARDWARE AND SOFTWARE

Hardware

Microsoft Kinect

The Kinect sensor returns the depth stream data as a succession of the depth image frame. The Kinect sensor returns raw depth with a 16-bit greyscale format with a viewable range of 43 degrees vertical and 57 degrees horizontal. Well, this is not just an image; behind the scenes, the Kinect sensor runs a series of algorithms on the captured data to give you more than an image, which tells you how far each pixel in that frame is. The depth pixel contains the distance between the Kinect device and the objects in front of the device, in millimetres. The data is represented based on the X and Y coordinates in the depth sensor view.

Jetson Xavier NX Developer Kit

The NVIDIA® Jetson Xavier NX™ Developer Kit brings supercomputer performance to the edge. It includes a Jetson Xavier NX module for developing multi-modal AI applications with the NVIDIA software stack in as little as 10 W. We can also now take advantage of cloud-native support to more easily develop and deploy AI software to edge devices.

Software

TensorFlow Library

TensorFlow is an end-to-end open-source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets researchers push the state-of-the-art in ML and developers easily build and deploy ML-powered applications.

Jupyter Lab

JupyterLab is a web-based interactive development environment for Jupyter notebooks, code, and data. JupyterLab is flexible: configure and arrange the user interface to support a wide range of workflows in data science, scientific computing, and machine learning. JupyterLab is extensible and modular: write plugins that add new components and integrate with existing ones.

OpenCV

OpenCV (Open-Source Computer Vision Library) is an open-source computer vision and machine learning software library. OpenCV was built to provide a common infrastructure for computer vision applications and to accelerate the use of machine perception in commercial products. Being a BSD-licensed product, OpenCV makes it easy for businesses to utilize and modify the code.

Python

Python is an interpreted high-level general-purpose programming language. Its design philosophy emphasizes code readability with its use of significant indentation. Its language constructs ,as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects

BIBLIOGRAPHY

[1] N. Rachburee and W. Punlumjeak, "An assistive model of obstacle detection based on deep learning: YOLOv3 for visually impaired people," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 4, pp. 3434–3442, 2021, doi: 10.11591/ijece.v11i4.pp3434-3442.

[2] D. Dahiya, H. Gupta, and M. K. Dutta, "A Deep Learning based Real Time Assistive Framework for Visually Impaired," *2020 Int. Conf. Contemp. Comput. Appl. IC3A 2020*, pp. 106–109, 2020, doi: 10.1109/IC3A48958.2020.233280.

[3] J. K. Mahendran, D. T. Barry, A. K. Nivedha, and S. M. Bhandarkar, "Computer Vision-Based Assistance System for the Visually Impaired Using Mobile Edge Artificial Intelligence," *Proc. IEEECVF Conf. Comput. Vis. Pattern Recognit. Work.*, pp. 2418–2427, 2021.

[4] S. Shadi, S. Hadi, M. A. Nazari, and W. Hardt, "Outdoor navigation for visually impaired based on deep learning," *CEUR Workshop Proc.*, vol. 2514, pp. 397–406, 2019.

[5] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2018, doi: 10.1109/TPAMI.2017.2699184.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017, doi: 10.1109/TPAMI.2016.2577031.

[7] C. Szegedy, S. Reed, D. Erhan, D. Anguelov, and S. Ioffe, "Scalable, High-Quality Object Detection," 2014, [Online]. Available: http://arxiv.org/abs/1412.1441.

[8] W. Liu *et al.*, "SSD: Single shot multibox detector," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9905 LNCS, pp. 21–37, 2016, doi: 10.1007/978-3-319-46448-0_2.

[9] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11211 LNCS, pp. 833–851, 2018, doi: 10.1007/978-3-030-01234-2_49.

[10] H. Alhichri, Y. Bazi, N. Alajlan, and B. Bin Jdira, "Helping the visually impaired see via image multi-labeling based on SqueezeNet CNN," *Appl. Sci.*, vol. 9, no. 21, 2019, doi: 10.3390/app9214656.

[11] Y. Lin, K. Wang, W. Yi, and S. Lian, "Deep learning based wearable assistive system for visually impaired people," *Proc. - 2019 Int. Conf. Comput. Vis. Work. ICCVW 2019*, pp. 2549–2557, 2019, doi: 10.1109/ICCVW.2019.00312.

[12] "Seeing AI." https://www.microsoft.com/en-us/ai/seeing-ai.

[13] “BE MY EYE.” https://www.bemyeyes.com/.

[14] W. Weng and X. Zhu, “INet: Convolutional Networks for Biomedical Image Segmentation,” *IEEE Access*, vol. 9, pp. 16591–16603, 2021, doi: 10.1109/ACCESS.2021.3053408.

[15] A. S. Identi and C. Study, “Understanding Semantic Segmentation with UNET 3 . What is Semantic Segmentation ?,” *Towards Datascience*, 2019. https://towardsdatascience.com/understanding-semantic-segmentation-with-unet-6be4f42d4b47.

[16] Milesial, “U-Net: Semantic segmentation with PyTorch,” *2020*. https://github.com/milesial/Pytorch-UNet.

[17] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 10691–10700, 2019.

[18] Even, “Vehicle car detection and multi-label classification,” 2020. https://github.com/CaptainEven/Vehicle-Car-detection-and-multilabel-classification.

[19] X. Song *et al.*, “APOLLOCAR3D: A large 3D car instance understanding benchmark for autonomous driving,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 5447–5457, 2019, doi: 10.1109/CVPR.2019.00560.

[20] L. Ke, S. Li, Y. Sun, Y. W. Tai, and C. K. Tang, “GSNet: Joint Vehicle Pose and Shape Reconstruction with Geometrical and Scene-Aware Supervision,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12360 LNCS, pp. 515–532, 2020, doi: 10.1007/978-3-030-58555-6_31.

[21] J. Shi, H. Yang, and L. Carlone, “Optimal Pose and Shape Estimation for Category-level 3D Object Perception,” 2021, doi: 10.15607/rss.2021.xvii.025.

[22] H. Caesar *et al.*, “Nuscenes: A multimodal dataset for autonomous driving,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, no. March, pp. 11618–11628, 2020, doi: 10.1109/CVPR42600.2020.01164.

[23] N. Jourdan, M. Ag, and M. Cordts, “Cityscapes 3D: Dataset and Benchmark for 9 DoF Vehicle Detection.”

[24] L. Bertoni, S. Kreiss, and A. Alahi, “MonoLoco: Monocular 3D Pedestrian Localization and Uncertainty Estimation.”