AlterEye: Footpath and Obstacles Detector System using Mask R-CNN

Jan Miller F.Jaro
BS Computer Engineering
Batangas State University - Alangilan
Batangas City, Philippines
Janmiller.jaro@g.batstate-u.edu.ph

***Abstract*— According to WHO (World Health Organization), near or far vision impairment affects at least 2.2 billion people globally, making their lives more difficult than those of persons who have a sense of sight, especially while traveling from one place to another. Furthermore, having an alternate sight is essential for them to be safe while traveling, therefore the researcher developed a system called "AlterEye" that can identify footpaths and obstacles along the path. This system used Mask R-CNN and COCO pre-trained model to create the said system, and after many tests and analysis of results, it was proven that Mask R-CNN is effective to use in this study because it produced a model with 90.53% accuracy. Using the 15th model in image testing shows that it can detect footpaths and obstacles with an accuracy of 100% and 96%, respectively, while in video testing, the accuracy peaks at 99% and plays around between 99% and 85%.**

Keywords—Mask R-CNN, labelMe, PixelLib, Inference

I. Introduction

Traveling from one place to another remains a challenge for blind and visually impaired people. Because for them to go to other places it require proper guidance to protect them from harm or accidents. That is why footpaths are very important for them as it separate vehicles and pedestrian, [1] it keep them safer while they are walking towards specific locations. Living in a world that is inaccessible for people who have disabilities is saddening, that is why the of need a system that can guide people with disabilities is a must.

According to data from WHO (World Health Organization) Near or far vision, impairment affects at least 2.2 billion individuals globally. At least one billion, or roughly half of these cases, of vision impairment, may have been averted or corrected. This 1 billion includes people with moderate or severe distance vision impairment or blindness due to uncorrected refractive error (88.4 million), cataracts (94 million), age-related macular degeneration (8 million), glaucoma (7.7 million), diabetic retinopathy (3.9 million), and near vision impairment due to uncorrected presbyopia (3.9 million) (826 million) [2]. There are so many tools like trained dogs and white canes that can help blind and visually impaired people, but it is not enough because they cannot provide them with all the information and features for safe mobility that individuals with sight have [3]. People with this kind of disability are prone to pedestrian accidents, particularly in low and middle-income countries.

A lot of existing research was made to help people with the said disabilities to have accessibility while navigating, this research [4] created an artificial navigation system with customizable sensitivity help from an ultrasonic proximity sensor and a GPS module for visually impaired people to move confidently and independently in both indoor and outdoor settings. This system can detect any potential difficulties and potholes with the Ultrasound reflection qualities The system's attachment to the clothing, footwear, body area, and walking stick makes its use more versatile and dependable.

Moreover, in this paper [5] researchers made a “SmartCane” system, which consists of a robotic white cane and mobile device software, that assists blind and visually impaired people in navigating indoors. The mobile device's software may communicate with the robotic white cane to plan a route and guide a BVI user to a location within an interior area. The study of [7] made a "smart eye system", this technology is created to be portable. With the help of GPS, the device continuously monitors the user's present location. When obstacles are spotted on the way, the device gives a warning. It also helps when it comes to the identification of people based on previously saved photographs. Because all data is supplied into the system before it is used, it does not require Internet connectivity to function. This is especially useful if Internet access is not available throughout the city. Furthermore, the device does not employ any Android or other touchscreen technologies, making it very simple and easy to operate.

In this study [8] researchers produced a clever and efficient path direction robot to help vision-impaired persons navigate This is an innovative device and alternative to using guide dogs. The robot can walk along many paths and then recall and retrace all of them, making it a suitable substitute for a guide dog, which is often a luxury for 90% of blind individuals who live in low-income settings. The robot may direct the user to locations that cannot be tracked via GPS; because most navigation systems for the blind that have been developed use a complex combination of positioning systems, video cameras, location mapping, and image processing algorithms, they have designed an affordable low-cost prototype navigation system that serves the purpose of guiding visually impaired people both indoors and outdoors. With the amount of research that shows the struggles of blind and visually impaired people in navigating.

This research proposed a system that can analyze footpaths and obstacles along its path to provide guidance for people in need, to give them accessibility that people with no disabilities have and for them to live a life safer from harm. Furthermore, this paper aims to develop a model that can detects footpaths and obstacles on them. Additionally, labelMe was used to annotate the footpath dataset, and PixelLib to process the instance segmentation for the footpath to be classified from an image, and COCO pre-trained model to train and validate the dataset; more in-depth explanation will be given in methodologies.

II. Methodology

A. *Processing of Dataset to Make a Model*

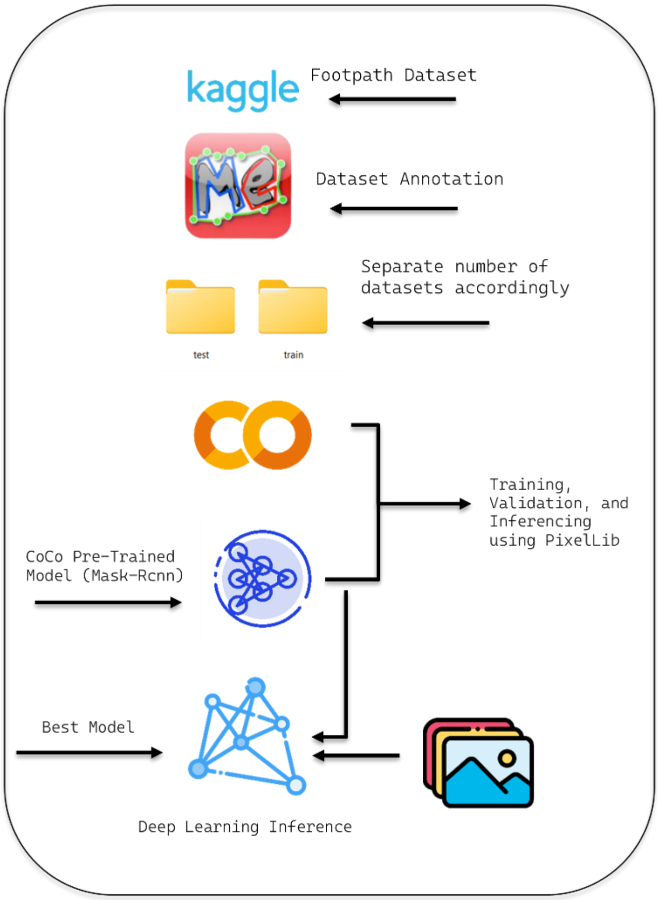

kaggle
Footpath Dataset
Dataset Annotation
Separate number of datasets accordingly
test
train
Training, Validation, and Inferencing using PixelLib
CoCo Pre-Trained Model (Mask-Rcnn)
Best Model
Deep Learning Inference

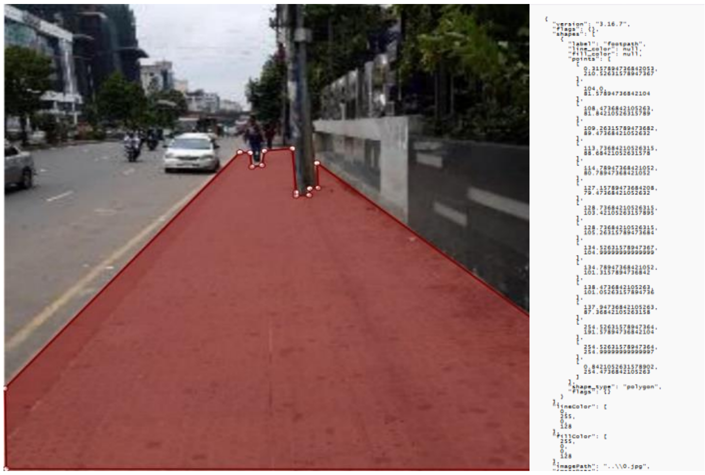
Figure 1. Creating the System

Figure 1 shows the steps it took to create a system with the best model, from gathering the data that is needed in this research, from annotating using labelMe, to separating the amount of data to be able to train them properly, then using the Google Colab as the IDE, using the pre-trained coco model so it can create the best model that the system needs, and lastly, the inferencing, to test the system if it accurately identity the objects that it needs.

Figure 2. Footpath Dataset

Figure 2 is an example of the dataset that was used to make the model for the system it will be annotated using labelMe

B. *Annotation of Dataset*



```
{
  "version": "3.16.7",
  "flags": {},
  "shapes": [
    {
      "label": "footpath",
      "line_color": null,
      "fill_color": null,
      "points": [
        [
          0.3157894736842053,
          210.52631578947367
        ],
        [
          104.0,
          81.57894736842104
        ],
        [
          108.4736842105263,
          81.84210526315789
        ],
        [
          109.26315789473682,
          89.47368421052632
        ],
        [
          113.73684210526315,
          88.68421052631578
        ],
        [
          114.78947368421052,
          80.78947368421052
        ],
        [
          127.15789473684208,
          79.47368421052632
        ],
        [
          128.73684210526315,
          103.42105263157895
        ],
        [
          128.73684210526315,
          105.26315789473684
        ],
        [
          134.52631578947367,
          104.99999999999999
        ],
        [
          134.78947368421052,
          101.3157894736842
        ],
        [
          138.4736842105263,
          101.05263157894736
        ],
        [
          137.94736842105263,
          87.36842105263158
        ],
        [
          254.52631578947364,
          191.57894736842104
        ],
        [
          254.52631578947364,
          254.99999999999997
        ],
        [
          0.8421052631578902,
          254.4736842105263
        ]
      ],
      "shape_type": "polygon",
      "flags": {}
    }
  ],
  "lineColor": [
    0,
    255,
    0,
    128
  ],
  "fillColor": [
    255,
    0,
    0,
    128
  ],
  "imagePath": "..\\0.jpg",
```

Figure 3. Annotation of the footpath's dataset using labelME

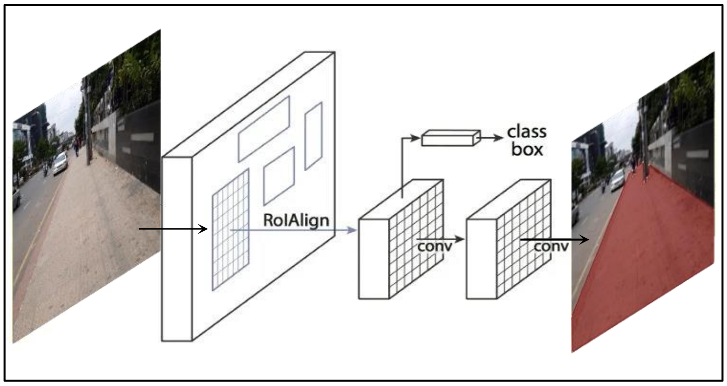


Figure 4. Mask R-CNN framework for instance segmentation

C. *Mask RCNN for Instance Segmentation*

Mask R-CNN is an enhanced network that develops on Faster R-CNN. Faster R-CNN gives the RPN network the last layer of the feature map extracted by CNN. When the target is too small in the original input image, the target is more prominent in the feature map [9]. The convolution layer is likely to filter out small targets as noise in the middle of the hidden layer, even if some simple features can be extracted when they are input into the RPN network. Mask R-CNN applied the Feature Pyramid Network (FPN); the idea of FPN is to apply the same feature map extracted by CNN to the RPN network, which avoids the filtering of small target features. By increasing the residual structure, the feature map of “top-down” is fused with the original convolution to enhance the feature of the target. [9].

The Mask R-CNN process starts with the first stage consisting of two networks, backbone (ResNet, VGG, Inception, etc..) and region proposal network. These networks run once per image to give a set of region proposals. Region proposals are regions in the feature map which contain the object. In the second stage, the network predicts bounding boxes and object classes for each of the proposed regions obtained in stage 1. Each proposed region can be of different sizes whereas fully connected layers in the networks always require fixed-size vectors to make predictions. The size of these proposed regions is fixed by using either the RoI pool (which is very similar to MaxPooling) or the RoIAlign method. Faster R-CNN is a single, unified network for object detection Faster R-CNN predicts object class and bounding boxes.

Mask R-CNN is an extension of Faster R-CNN with an additional branch for predicting segmentation masks on each Region of Interest (RoI). The Mask R-CNN framework for instance segmentation is the second stage of Faster R-CNN, RoI pool is replaced by RoIAlign which helps to preserve spatial information which gets misaligned in the case of the RoI pool. RoIAlign uses binary interpolation to create a feature map that is of fixed size e.g., 7 x 7. The output from the RoIAlign layer is then fed into the Mask head, which consists of two convolution layers. It generates a mask for each RoI, thus segmenting an image in a pixel-to-pixel manner.[10]

In this paper, the model configuration for the training used Mask R-CNN, with a batch size of 4 while using the resnet101 as the network backbone and with only a class of 1, and it also used the CoCo pre-trained model h5 for the training of the dataset.

D. *Evaluation of the Model*

To determine the best model that has been created to begin the process of inferencing. The researcher uses the mAP(mean Average Precision) so the author can compare the accuracy of each model that was created. The accuracy of one model improves when the mAP has increased. In the process of Average Precision (AP), the recall phases are continuous, which interpolates the accuracy, however it stops if the AP detects the reduction of the impact of the curve wiggles.

The area under the interpolated curve is defined as an AP that computes using the following method below, interpolated precision P_{interp} (2) distinguishes a value of the degree of recall r as the maximum standard for the accuracy that has been discovered in all stages of the recalls. Furthermore, the Mean Average Precision (mAP) is the average precision (AP) throughout all the tests also known as mean average precision (mAP) (3), where O is the number of queries included in the set and AP i is the average precision (AP) for a specific query, O.

$$AP = \Sigma_{i=1}^{n-1}(r_i + 1 - r_i)P_{interp}(r_i + 1) \qquad (1)$$

$$P_{interp}(r_i + 1) = \max_{r' \geq r_i + 1} p(r') \qquad (2)$$

$$mAP = \frac{\Sigma_{i=1}^{0} AP_i}{O} \qquad (3)$$

E. *Model Inferencing and Testing*

The study is utilizing Google Collab, Mask R-CNN, Python, PixelLib and the Image AI detection library; It has two essential functions: to detect an image and video. While for model inferencing, the chosen deep learning COCO model h5 file and its accompanying JSON configuration file were used. Furthermore,0 for the testing procedure, the researcher gathered images and videos from the internet that is not part of the training and validation dataset to avoid biases in the accuracy of the model. These images were implemented to prevent biases in the testing accuracy (4) findings.

$$Accuracy = \frac{\#\ of\ Detected\ Object}{Total\ \#\ of\ Objects} X\ 100 \qquad (4)$$

III. Results and Discussions

In this section, the training, validation, and testing findings are discussed.

A. Training and Validation Results

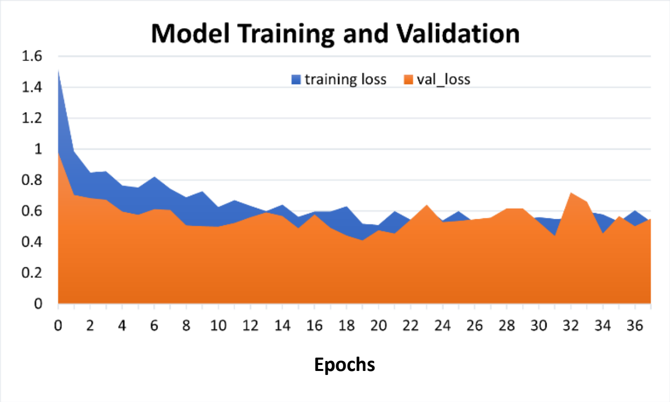


Figure 5. Training Loss and Validation Loss

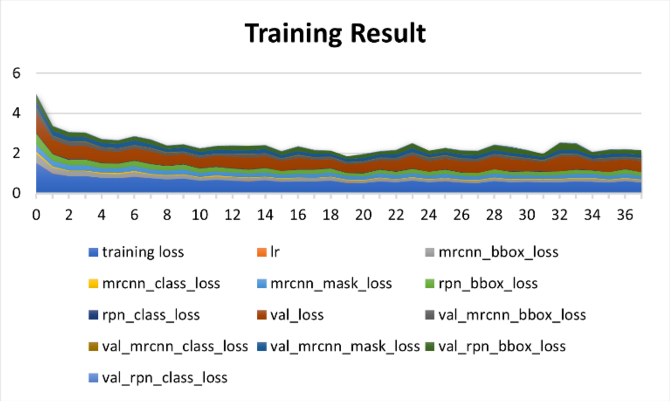


Figure 6. Detailed Training Results

Figure 5 shows that the peak of the training loss and validation loss is from epoch 0 with the percentage of training loss at 151% and for the validation loss is 98%, then it starts to decrease per epoch and it ends with a training loss of 53% and validation loss of 55%, while in the process it was not consistent on decreasing because it is randomly increasing in every epoch but it is not passing the peak and it was still low from what was the peak of losses.

In Figure 6, their result starts decreasing from its peak every epoch, with this ,it can easily understand how the layers of Mask R-CNN because of the added information. Figure 6 shows that every training decreases the losses that is needed to create the best-trained model for the system.

B. Model Evaluation

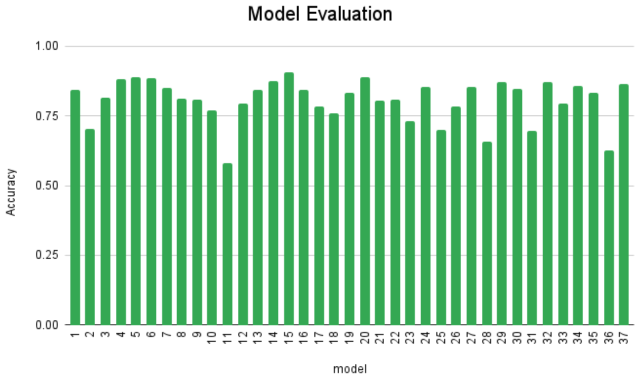


Figure 7. Evaluation of the Model

Figure 7 shows, that the first epoch produce an accuracy of 84% then it falls to 70% then it plays between the accuracy of 88% -70%, then the 11th epoch produced 58%, then after that, it starts increasing again then the 15th epoch produces the peak with an accuracy of 91% after that, it starts to play again with 88% - 65% accuracy. Figure 7 demonstrated that the number of training or epochs does not guarantee a model with the highest accuracy, because sometimes it is random.

C. Image and Video Inferencing

The 15th model was applied to the image and video to determine its accuracy in the real-life scenario

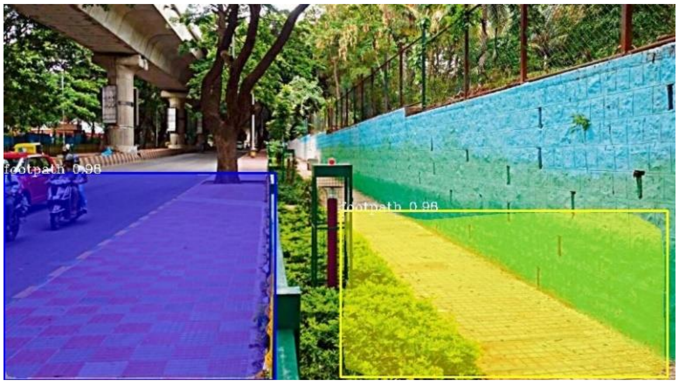


Figure 8. Detection of Two Footpaths and the obstacles in Image

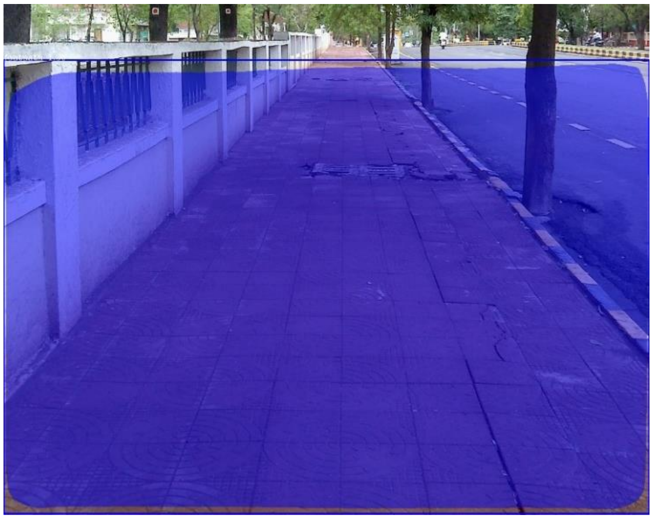


Figure 9. Inferenced image with 100% Accuracy

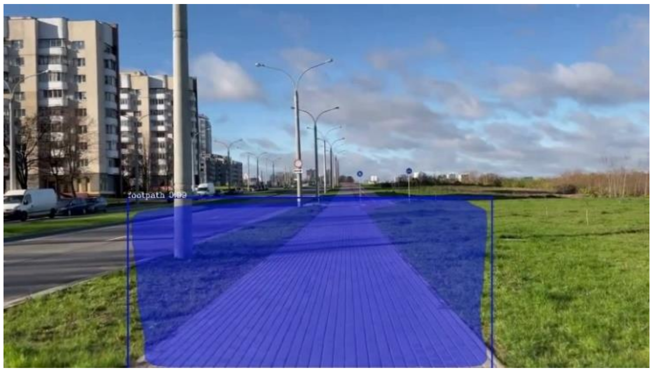


Figure 10. Model Testing using video

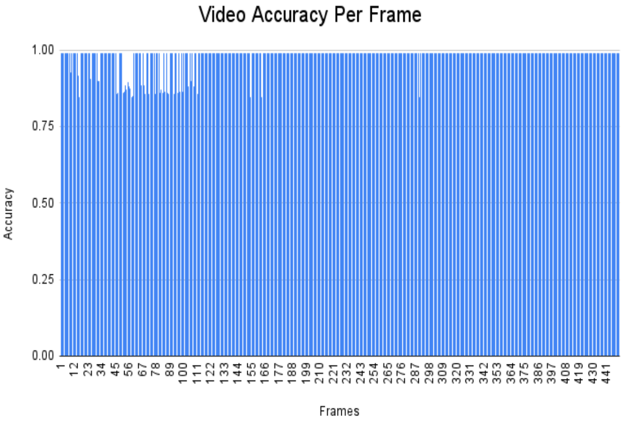


Figure 11. Result of the 15th model in video per frame

Figure 7 shows that to remove the biases, the researcher search for an image from the internet and downloaded it, then test it while using the best model that has been created, the image produces a result of 96% accuracy for both detected footpaths, the model also detected the obstacles along the way which means it is an effective guide for the blind or visually impaired. In Figure 9, the accuracy reached 100% in detecting the footpath, the model has this kind of reliability to its purpose.

Furthermore, in Figure 10, to test it in a video, the researcher downloaded a video from the internet of a person that is walking on a footpath, to test if the model can accurately determine a path in real life scenario while moving in motion. As what the Figure 11 shows, the lowest accuracy falls to 85% while its peak was 99%, and the accuracy per frame was almost consistently stable at 99% from frame 166 to 441, the obstacle along the way affects the accuracy when detecting the path.

IV. Conclusion

The sense of sight is very important to humans because it helps people to navigate to certain areas they want to go to, and it keeps people safe since it allows them to avoid numerous things that can hurt them and endanger their lives, such as automobiles and obstacles in their daily lives. Losing it is an unimaginable idea, which is why developing a system that can serve as an alternate sight for humans is essential, as it can guide them from one place to another more securely. The accuracy of the pathway detector was trustworthy because it demonstrated excellent accuracy results in the testing for both image and video tests.

The researcher used Mask R-CNN because it can produce a reliable model for the system; the 15th model was the best-trained model out of 37 models, with a peak at 0.905388 mAP. The model that was created can detect a footpath and obstacles along its way it has the assurance to help blind and visually impaired people, as evidenced by the image tests which achieved 100% accuracy.

Acknowledgment

The author would like to express his heartfelt gratitude to his advisor for guiding him throughout the research process; this research would not have been possible without his comments and recommendations. The author wants to thank God for being with him, uplifting him, and allowing him to overcome the difficulties he faced.

REFERENCES

[1] "Sight Distance (Obstruction Removal) - Road Safety Toolkit - IRAP." Road Safety Toolkit. https://toolkit.irap.org/safer-road-treatments/sight-distance-obstruction-removal/ (accessed: Dec. 04, 2022).

[2] "Blindness and vision impairment." World Health Organization. Oct. 22, 2022. https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment (accessed: Dec. 04, 2022).

[3] W. Elmannai and K. Elleithy. "Sensor-Based Assistive Devices for Visually-Impaired People .." National Library of Medicine. 2017. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5375851/ (accessed: Dec. 04, 2022).

[4] A. Sen, K. Sen and J. Das, "Ultrasonic Blind Stick for Completely Blind People to Avoid Any Kind of Obstacles," 2018 IEEE SENSORS, 2018, pp. 1-4, doi: 10.1109/ICSENS.2018.8589680.

[5] Q. Chen et al., "CCNY Smart Cane," 2017 IEEE 7th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER), 2017, pp. 1246-1251, doi: 10.1109/CYBER.2017.8446303.

[6] M. Avila, M. Funk, and N. Henze, "Dronenavigator," *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility - ASSETS '15*, 2015.

[7] I. Joe Louis Paul, S. Sasirekha, S. Mohanavalli, C. Jayashree, P. Moohana Priya and K. Monika, "Smart Eye for Visually Impaired-An aid to help the blind people," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), 2019, pp. 1-5, doi: 10.1109/ICCIDS.2019.8862066.

[8] R. K. Megalingam, S. Vishnu, V. Sasikumar, and S. Sreekumar, "Autonomous Path Guiding robot for visually impaired people," *Cognitive Informatics and Soft Computing*, pp. 257–266, 2018.

[9] J. Shi, Y. Zhou and W. X. Q. Zhang, "Target Detection Based on Improved Mask Rcnn in Service Robot," 2019 Chinese Control Conference (CCC), 2019, pp. 8519-8524, doi: 10.23919/ChiCC.2019.8866278.

[10] "How Mask R-CNN Works? | ArcGIS API for Python." https://developers.arcgis.com/python/guide/how-maskrcnn-works/ (accessed: Dec. 04, 2022).