HIGH SCHOOL EDITION
Journal of Student Research
Volume 14 Issue 1 (2025)

SmartEye: A Machine Learning Approach to Enhance Mobility for the Visually Impaired through Depth Estimation and Object Detection

Daniel Chung[1], Ji Tae Kim[1] and Joyce Pereira[#]

[1]Korean International School Pangyo, Republic of Korea
[#]Advisor

ABSTRACT

Visually impaired individuals encounter significant challenges in their daily lives, particularly when navigating streets and public spaces. Walking in urban environments poses unique difficulties, as they must contend with obstacles, uneven surfaces, and traffic, all of which can create hazardous situations. With the growing prevalence of visual impairment, it is increasingly important to develop effective methods and technologies that can assist these individuals in safely and confidently navigating their surroundings. To address this issue, we propose SmartEye, a machine learning-based mobility assistant system that utilizes depth estimation and object detection. The system features a compact camera module mounted on the user's glasses, which captures the environment in front of them. Through object detection and depth estimation algorithms, SmartEye analyzes the surroundings in real time, identifying obstacles and their distances. The outputs from both the object detection and depth estimation processes are then integrated to provide a comprehensive understanding of the user's environment. This information is communicated to the individual through a speaker attached to the glasses, offering essential guidance and enhancing their mobility and safety while navigating public spaces. The proposed system achieved an absolute relative error of 0.068 and a mean average precision of 57.5 on a public dataset. Additionally, we conducted a real-world study by applying the SmartEye system to real-world street scenarios. The results demonstrated the system's feasibility and effectiveness in assisting visually impaired individuals in navigating complex environments.

Introduction

Blindness encompasses a spectrum of visual impairments, ranging from limited vision to complete lack of sight. Various factors, such as genetics, diseases, injuries, or degenerative conditions affecting the eyes or brain's visual processing centers, can lead to blindness. As of 2020, approximately 43.3 million people worldwide experience limited vision or visual impairments (Pesudovs et al., 2024). To navigate their surroundings, individuals with limited vision typically rely on two main tools: the white cane and guide dogs. The white cane is designed to enhance mobility and independence, allowing users to detect their environment by sweeping the cane from side to side to identify changes in terrain, stairs, curbs, or other obstacles. In contrast, guide dogs are specially trained to assist their handlers in navigating safely, offering additional support in avoiding hazards like uneven stairs or cracked sidewalks.

Despite their usefulness, both white canes and guide dogs have notable limitations that pose significant risks to individuals with visual impairments. White canes are particularly vulnerable in extreme weather conditions, such as strong winds or heavy snowfall, which can obscure obstacles and complicate navigation (BBC, 2015). Additionally, white canes are limited in their ability to detect obstacles at head or knee height or those more than a meter away, such as deep holes or posts (Pyun et al., 2013). Research from Sungkyunkwan University in Seoul, Korea, reveals that over 50% of white cane users have encountered accidents, with 36% experiencing severe incidents, highlighting the safety challenges associated with this tool (Kim et al., 2013).

Similarly, guide dogs, while invaluable, come with challenges. The training and upkeep of guide dogs can be economically burdensome and time-consuming. Approximately 30% of guide dogs fail to complete their training, despite significant financial investments, which average around $30,000 per dog (Bray et al., 2017; Tomkins et al., 2011). Furthermore, the ongoing costs of caring for guide dogs, including medical care, food, and shelter, add to the burden. Guide dogs typically have a working lifespan of only 6 to 8 years, and with a population of around 20,000 guide dogs, their availability is far outstripped by the estimated 43.3 million people living with visual impairments (CLOVERNOOK, 2020; Newen, 2023; Pesudovs et al., 2020).

To overcome these challenges, we propose SmartEye, a machine learning-driven mobility assistant designed to enhance the safety of visually impaired individuals while walking. The SmartEye system consists of wearable glasses equipped with a camera that captures the user's visual perspective. This imagery is processed in real-time to perform object detection and depth estimation, allowing the system to determine the distance between the user and detected objects. The SmartEye system operates on an embedded graphics processing unit and has been evaluated in real-world scenarios, demonstrating its effectiveness and practicality as a mobility aid for the visually impaired.

Background Knowledge

Object Detection

Object detection is a technique rooting from a combination of computer vision and deep learning. It identifies and locates objects from an image or a video using its algorithm. Object detection algorithm classifies objects detected into categories and also determines the location of the objects within the picture. Convolutional Neural Network is often implemented in the algorithm of object detection.

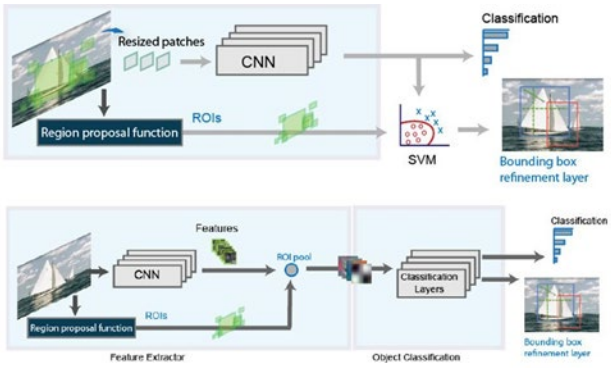


Figure 1. Object Detection Example (MathWorks 2024)

Figure 1 illustrates an example of object detection. The process begins by identifying regions within an image that contain objects, followed by localizing these regions within the inputted image. In this project, this technique will be implemented by processing images captured from the perspective of visually impaired individuals. For each image, the object detection system will identify and localize every object present. The relevant information will then be conveyed to the individual, providing them with a better understanding of their surroundings.

Depth Estimation

Depth Estimation is the process of outputting a depth map which represents distances of pixels from the camera with an inputted image. There are two ways depth estimation can be done, monocular depth estimation and stereo disparity

estimation. This technology is used in various fields such as robotics, 3d modeling, self-driving car, augmented reality and virtual reality, photo enhancement, and etc.

Figure 2. Example of depth estimaton map (Park 2023)

Figure 2 illustrates an example of monocular depth estimation. Monocular depth estimation is extracting 3d information from a 2d image taken from a monocular camera. There have been numerous monocular-based depth estimation approaches utilizing machine learning techniques (Xu et al. 2018; Huang et al. 2021; Xie et al. 2023).

In this research, we combine depth estimation and object detection to predict both the type of objects and their distance from the camera which aims to provide safer walking environments for the visually impaired. A detailed explanation of the proposed method is provided in Chapter 3.

Proposed Method

The proposed method is composed of three modules: the feature extractor, the depth estimator, and the object detection. As shown in Figure 3, first, the feature map is extracted from the input image. Then, the feature map gets used in the depth estimator to operate depth estimation of the input image. With the same feature map, it gets used in object detection for the input image. Therefore, the proposed method is capable of extracting both the depth estimation and the object detection for the input image.

The feature extractor's role is to extract the feature map value from the input image. The input image goes through three sets of Convolutional Neural Network in order to extract the feature map out of it. With the extracted feature map of the input image, the depth estimator runs to depth estimate the input image. Here, the depth estimation is half sized as it allows it to process faster. In our project, speed is a very important factor as we are trying to replace

human eyes with our product. The type of the depth estimator will be monocular depth estimator for it is more efficient and cost saving.

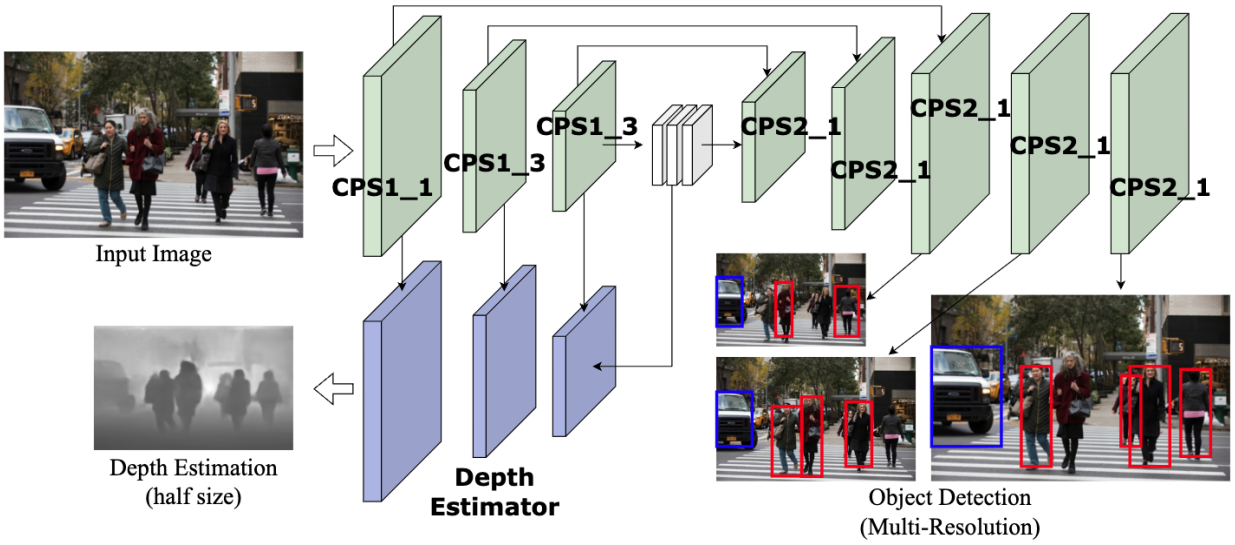


Figure 3. Architecture of the proposed joint training approach of depth estimation and object detection

To train the proposed depth estimation and object detection network, we utilize two types of loss function. For the object detection network, we adapted three loss functions from the original YOLOv5 paper (Jocher et al. 2022): confidence loss, class loss, and regression loss, as we will explain in Equation 1-3. For the depth estimation network, we utilized the mean squared error function which is commonly used in training tasks related to depth estimation.

Equation 1: Confidence loss function

$$L_{conf} = \sum_{i \in S^2} \sum_{j \in B} \mathbb{I}_{i,j}^{obj} (C_i - \hat{C}_i)^2 + \alpha \sum_{i \in S^2} \sum_{j \in B} \mathbb{I}_{i,j}^{noobj} (C_i - \hat{C}_i)^2$$

Here, S and B denote the number of grid cells and the number of predictions of the object detection network, respectively. C and $\hat{C}$ represent the probability of an object being present in the current grid cell and its corresponding ground truth probability. $\boldsymbol{I}^{obj}$ is an indicator function that equals 1 when there is an object in the current grid cell, and 0 otherwise. $\boldsymbol{I}^{noobj}$ functions vice versa.

Equation 2: Class loss function

$$L_{class} = \sum_{i \in S^2} \mathbb{I}_i^{obj} \sum_{c \in classes} (p_i(c) - \hat{p}_i(c))^2$$

The class loss function measures the difference between the predicted class probability and its corresponding ground truth probability. The output variable $\hat{p}$ and $\boldsymbol{p}$ denotes the predicted class probability and its ground truth probability. By subtracting the corresponding ground truth probability to predicted probability, the function captures the disparity or loss value.

Equation 3: Regression loss function

$$L_{reg} = \sum_{i \in S^2} \sum_{j \in B} \mathbb{I}_{i,j}^{obj} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2] + \sum_{i \in S^2} \sum_{j \in B} \mathbb{I}_{i,j}^{obj} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i}\right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i}\right)^2 \right]$$

The regression loss function is utilized to find the accuracy of the predicted bounding boxes. The parameter x and y represent the ground truth bounding box's x and y center coordinates accordingly. $\hat{x}$ and $\hat{y}$ represent the model's predictions on what the bounding box's x and y center coordinate is going to be. Similarly, $\boldsymbol{w}$ and $\boldsymbol{h}$ represent the ground truth bounding box's width and height accordingly. The parameters $\hat{w}$ and $\hat{h}$ represent the model's prediction of the bounding box's width and height. Then with these variables, the regression loss function runs to measure the difference in the ground truths and the predictions of the bounding box in the aspect of x center coordinate, y center coordinate, width, and length.

Equation 4: Mean squared error function

$$L_{mse} = -\frac{1}{XY} \sum_{y \in Y} \sum_{x \in X} |d(x,y) - \hat{d}(x,y)|^2$$

The mean squared error function is utilized to measure the accuracy of the depth estimation. The X and Y each represent the height and width of the input image. The XY, therefore, represents the total number of the grid cells. $d(x,y)$ represents the ground truth depth of the object in the grid cell. $\hat{d}$(x,y) represents the predicted depth of the object in the grid cell. The difference between the ground truth and the prediction is run through as many times as the total number of the grid cells, each value adding up, and finally divided by the total number of grid cells to find the mean of it.

Experimental Results

Depth Estimation and Object Detection Dataset

To train and evaluate the proposed networks, we used three different datasets for depth Estimation and object detection tasks. For the object detection task, we used the Pascal VOC dataset, which includes 20 commonly encountered objects such as person, car, dog, cat, motorbike, and bicycle, all frequently found in street environments. To train and evaluate the depth estimation network, we utilized two datasets: ReDWeb and the KITTI benchmark. The KITTI depth estimation dataset contains 86,000 training images and 7,000 test images. ReDWeb, with 3,600 samples, offers a broader range of lifestyle scenarios, including people, obstacles, and more varied training images.

(a) (b) (c)

Figure 4. Snippet of each dataset
(a): Kitti depth, (b): ReDWeb, and (c): Pascal VOC

Depth Estimation Evaluation

To evaluate the performance of the proposed method, we used the absolute relative error (ARE) function. The calculation process for ARE is detailed in Equation 5.

Equation 5: Absolute relative error function

$$\frac{1}{T}\sum_{p}\frac{|d_p - \hat{d}_p|}{d_p}$$

The absolute relative error function finds the difference between the depth value of the cell and the model's prediction of the depth value of the cell. Then divide the difference by ground truth depth value in order to analyze the difference relatively. Repeat and accumulate the results for every cell and finally divide the value by the number of cells to find the average. The lower the score is, the higher the performance of the method is.

Table 1. Absolute relative error rate comparison with the state-of-the-art depth estimation methods

Model	AbsRel
PAD-Net (Xu et al. 2018)	0.082
H-Net (Huang et al. 2021)	0.076
SwinV2-L (Xie et al. 2023)	0.065
Proposed	0.068

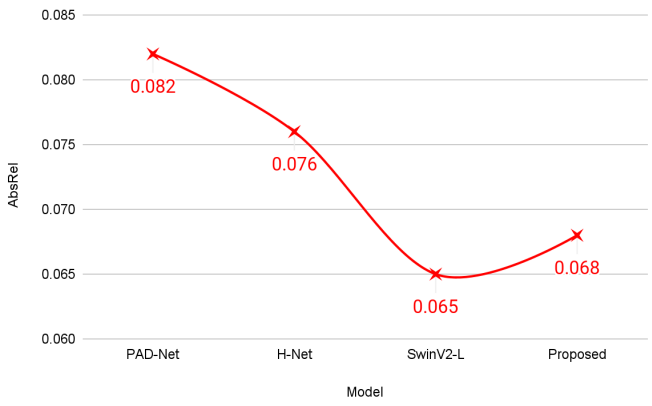


Figure 5. Visual comparison of absolute relative error rate with the state-of-the-art depth estimation methods

Table 1 and Figure 5 present the ARE comparison between the proposed method and state-of-the-art methods. We first trained the proposed method and evaluated its ARE on both the KITTI benchmark and ReDWeb datasets. For comparison, we selected PAD-Net (Xu et al., 2018), H-Net (Huang et al., 2021), and SwinV2-L (Xie et al., 2023), as these methods have demonstrated relevant and competitive results.

From Table 1, it is evident that the proposed method achieves a lower absolute relative error compared to PAD-Net and H-Net, indicating better performance. However, when compared to the SwinV2-L method, the proposed method shows slightly lower performance. Despite this, it is important to note that SwinV2-L operates at a significantly slower speed, making it unsuitable for real-time operation on an embedded board. Considering that the proposed

method only has a 0.003 higher absolute relative error than SwinV2-L and can effectively operate on our embedded board, the proposed method is the more practical choice for this application.

Object Detection Evaluation

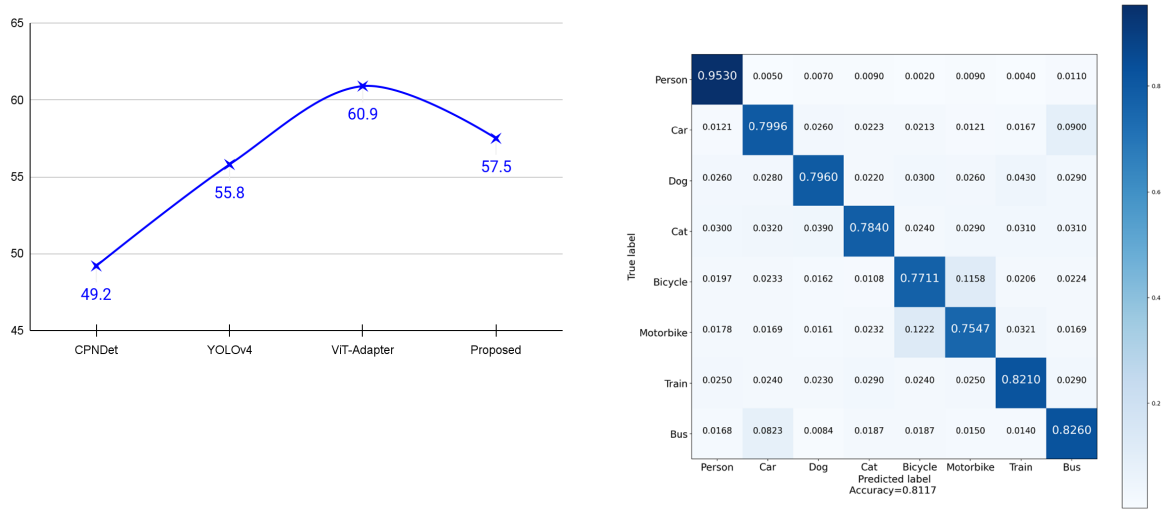
To evaluate the proposed object detection network, we utilized mean average precision (mAP) often used in many object detection tasks. The computation of mAP is explained in Equation 6.

Equation 6: Mean Average Precision

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i$$

Table 2. mAP comparison with the state-of-the-art object detection methods

Model	mAP
CPNDet (Duan et al. 2020)	49.2
YOLOv4 (Wang et al. 2021)	55.8
ViT-Adapter (Chen et al. 2022)	60.9
Proposed	57.5



(a) (b)

Figure 6. Object detection evaluation result and confusion matrix

Table 2 and Figure 6 summarize the evaluation of the proposed object detection network. Similar to Chapter 4.2, we first evaluated the mAP of the proposed method and compared it with the state-of-the-art object detection methods which are CPNDet (Duan et al. 2020), YOLOv4 (Wang et al. 2021) and ViT-Adapter (Chen et al. 2022).

The proposed method achieved a mean Average Precision (mAP) of 60.9 which surpasses both CPNDet and YOLOv4 by a significant margin. However, it did not exceed the performance of the ViT-Adapter. While the ViT-

Adapter demonstrated the highest accuracy among the comparison models, its reliance on transformer architectures results in high memory requirements and substantial computational costs. This makes it less suitable for deployment on embedded computing boards, where resource efficiency is crucial.

Conclusion

In this research, we introduced SmartEye, a machine learning-based mobility assistant designed to enhance navigation for visually impaired individuals through depth estimation and object detection. The proposed system achieved an absolute relative error of 0.068 and a mean Average Precision of 57.5 on a public dataset. Additionally, we conducted a real-world study that implemented the SmartEye system in various street scenarios, demonstrating its feasibility and effectiveness in helping visually impaired individuals navigate complex environments. In the future, we plan to develop a pixel-level segmentation network to further improve the accuracy of the assistance provided by the system.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

BBC. (2015, Nov 18). "*How dangerous are white canes?*": BBC https://www.bbc.com/news/disability-34855311

Bray, E. E., Sammel, M. D., Cheney, D. L., Serpell, J. A., & Seyfarth, R. M. (2017). Effects of maternal investment, temperament, and cognition on guide dog success. Proceedings of the National Academy of Sciences, 114(34), 9128-9133.

CLOVERNOOK. (2020, Sep 18). "*GUIDE DOGS VS. WHITE CANES: THE COMPREHENSIVE COMPARISON*": CLOVERNOOK https://clovernook.org/2020/09/18/guide-dogs-vs-white-canes-the-comprehensive-comparison/

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778). https://doi.org/10.48550/arXiv.1512.03385

Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., Kwon, Y., Michael, K., ... & Jain, M. (2022). ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. Zenodo.

Kim, S. Y., & Cho, K. (2013). Usability and design guidelines of smart canes for users with visual impairments. international Journal of Design, 7(1).

MathWorks (2024, Apr 17). "*Why Object Detection Matters*": MathWorks https://www.mathworks.com/discovery/object-detection.html

Newen, Alice. (2023, Oct 29). "Guide Dog Statistics Australia": Nalzo https://nalzo.com.au/blogs/tips/guide-dog-statistics-australia

Park, Minseo, (2023, Sep 7). “Monocular depth estimation & Stereo disparity estimation”: Jolabokaflod https://velog.io/@jolabokaflod/Monocular-depth-estimation-Stereo-disparity-estimation

Pesudovs, K., Lansingh, V. C., Kempen, J. H., Tapply, I., Fernandes, A. G., Cicinelli, M. V., ... & Bourne, R. (2024). Global estimates on the number of people blind or visually impaired by cataract: a meta-analysis from 2000 to 2020. Eye, 1-15.

Pyun, R., Kim, Y., Wespe, P., Gassert, R., & Schneller, S. (2013, June). Advanced augmented white cane with obstacle height and distance feedback. In 2013 IEEE 13th international conference on rehabilitation robotics (ICORR) (pp. 1-6). IEEE.

Tomkins, L. M., Thomson, P. C., & McGreevy, P. D. (2011). Behavioral and physiological predictors of guide dog success. Journal of veterinary behavior, 6(3), 178-187.