AI-Based Obstacle Detection and Navigation for the Blind using Convolutional Neural Network

Surapol Vorapatratorn
Center of Excellence in Artificial Intelligence and Emerging Technologies
School of Information Technology, Mae Fah Luang University
Chiang Rai, Thailand
Surapol.vor@mfu.ac.th

***Abstract*— Using stereo cameras and Convolutional Neural Networks, we present an AI-based obstacle detection and navigation system for the visually handicapped (CNN). The technology will employ wireless bone conductive headphones to guide users through stereo-directional sound patterns. First, we convert stereo photos to depth images, which may be utilized to determine each obstacle's depth level. Then, using our 2D-based Horizontal Depth Accumulative Information, we isolate obstacle pictures from the depth image (H-DAI image; side view projection). To train the AI model, the obstacle picture will be transformed to our Vertical Depth Accumulative Information (V-DAI image; top view projection). There are 34,325 example photos in the training dataset, with 7 different obstacle kinds. The influence of each picture input type, such as depth image, obstacle image, and V-DAI image, is investigated in our experiment. According to the findings, utilizing a V-DAI picture with CNN achieves the maximum accuracy of classification of 93.61 percent and the quickest prediction speed of 10,169 samples per second.**

Keywords*— visually impaired, the blind, convolutional neural network, obstacle detection, computer vision, machine learning.

I. INTRODUCTION

There are over 285 million visual impairments worldwide, including totally blind 39 million and low vision 246 million[1]. For these, guide dogs and long canes are well-known mobility aids. These solutions, however, are unable to protect their entire body from hazards such as bus poles, tree branches, open windows, street signs, and fences. There are various technologies for these detections including GPS [2], laser range [3-5], ultrasonic sound waves [6-8] have been used in commercial products and research to solve those problems. Nowadays, the RGB-D technique is efficient and extensive for obstacle detection systems [9-10]. However, it does not operate in outdoor settings, particularly when exposed to bright sunshine.

Fig. 1. Our obstacle detection and navigation system.

Having encountered limitations of previous developments, using a stereo camera, this study presents a real-time navigation and obstacle avoidance technology for the blind, as shown in figure 1. Our detection is based on stereo camera which has 10 meters coverage area and 5 meters warning area. Moreover, Using our Depth Accumulative Information (DAI), we can quickly classify obstacles, processing on a computer laptop. Navigate user with seven different sound patterns via bone conductive feedback as shown as figure 2.

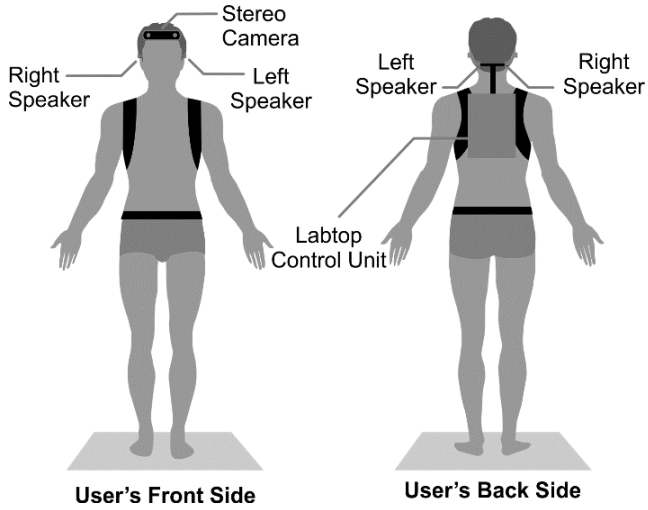

Fig. 2. The hardware installation.

II. STEREO VISION

There are variety of using stereo camera applications. These large variety of the stereo camera make researcher more difficult to choose the best one. Developer has to be concerned a series of features, including the bit-rate, image quality, speed shutter and weight. Therefore, to reach the match of application, we choose the ZED stereo camera [11] in Fig. 3. This stereo camera will be used to investigate because it was the highest performing compared with the various types.

Fig. 3. ZED Stereo camera (left), depth images from ZED (right).

The devices of a stereo vision system are horizontal aligned and split by a baseline length. [12]. Using the stereo camera allows the overlapping two images to extract a disparity image called depth, the distance between the stereo camera and object as shown in Fig. 4.

The stereo camera's optical axes are completely parallel, both picture planes are co-planar, there is no lens distortion. The measurement of the distance between an item and the camera., Z, also known as depth, can be calculated using Equation (1):

$$z = \frac{b \times f}{d}; d = x^l - x^r \quad (1)$$

2021 25th International Computer Science and Engineering Conference (ICSEC) | 978-1-6654-1197-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICSEC53205.2021.9684607

978-1-6654-1197-4/21/$31.00 ©2021 IEEE

where the focal length is specified by f, the baseline is defined by b, and the disparity is defined by $d = x^l - x^r$. The x^l and x^r are positions where the object occurs on the virtual plane of the left and right camera which is difference position, respectively resulting in d which is a positive number.

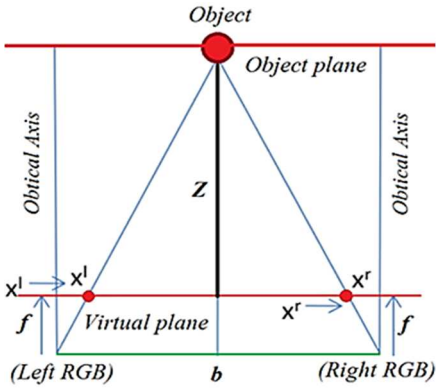

Fig. 4. The relation between focal length, baseline, disparity, depth.

III. PROPOSE METHOD

The depth image from the ZED stereo camera, which is installed on the user's helmet, is the first step in our system. Horizontal-Depth Accumulative Information (H-DAI), which depicts a side-view depth image projected, will be created from the depth picture. In our 2D-based quick ground removal approach, H-DAI is utilized to cut off the ground region of the depth picture, leaving just the obstacle image. The obstacle picture is then transformed to V-DAI (Vertical-Depth Accumulative Information), which is a virtual projected obstacle image in a bird's-eye view. The input data (feature vector) for the classification phase is the outcome of applying V-DAI. Finally, the anticipated obstacle class is the outcome of the categorization. The user's wireless bone conductive headphone converts this data into 3D-tone and speech guidance. Figure 5 depicts a system diagram of the proposed approaches.

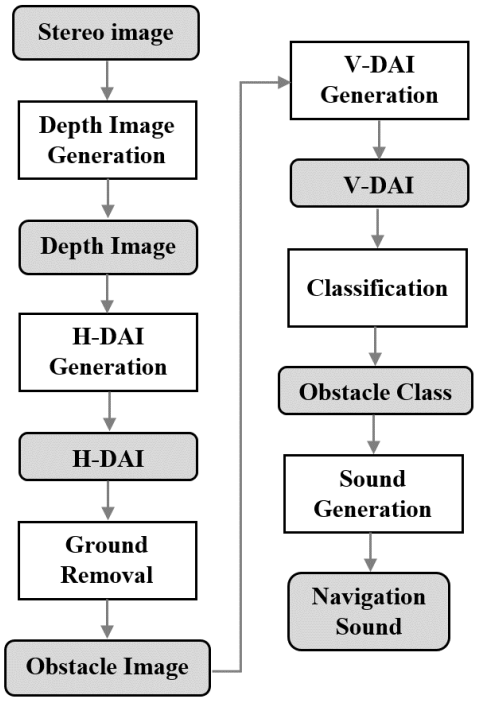

Fig. 5. The proposed approaches are depicted in a system diagram.

A. Ground removal

To get the H-DAI picture and side view depth image projection, the depth image frequency along the Y-axis was measured using the bin count approach in the first stage. The ground curve that indicated the pathway was shown by the greatest value of each H-DAI picture column. The roadway was removed from the depth picture and replaced with the obstacle image using ground curve information. As seen in Figure 6, high depth values indicate items that are close by, whereas low depth values indicate items that are far away.

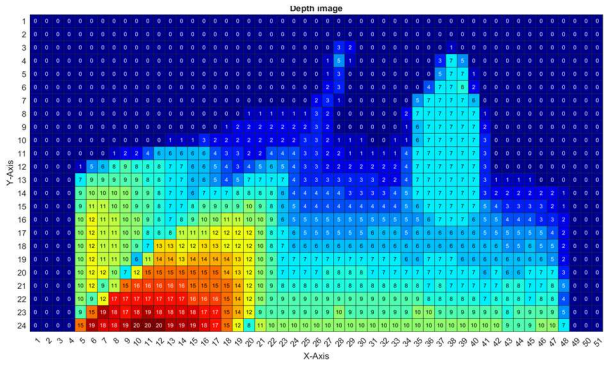

Fig. 6. To convert to H-DAI, take a depth picture.

The frequency of depth values is counted consecutively from the bottom rows of the depth image to the top rows of the depth image to create a horizontal–DAI (H-DAI) picture. The warm hue indicates a high frequency of depth value, whereas the cool color indicates a low frequency of depth value. As illustrated in Figure 7, the H-DAI picture is a virtually projected image from the side utilized to determine the route region.

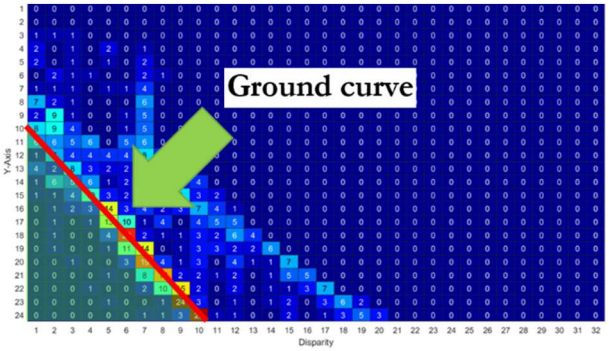

Fig. 7. To convert to H-DAI, generate a depth image.

The ground mark is then created by selecting H-DAI data that is less than the ground curve. This information is kept in binary format. As indicated in Figure 8, the route section is logic 1 and the remainder is logic 0. (left). Then, as illustrated in Figure 8, we erase the data at the ground mark of the depth picture (right).

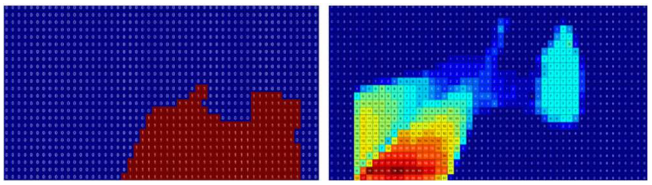

Fig. 8. Floor marking (left), floor depth image deleted (right).

Findings of the ground removal stage from the depth picture, it looks that there is still an obstacle zone indicating a

5 meter distant item. Finally, we must remove the obstruction that is beyond the considered range using the un-region of interest mark, as illustrated in Figure 9 (left). As seen in Figure 9 (right), the outcome is an obstruction picture.

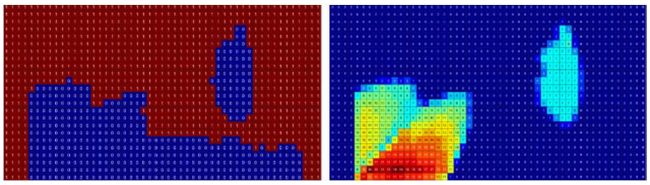

Fig. 9. Uninterested-in-regions-of-interest (left), obstacle image (right)

B. OBSTACLE CLASSIFICATION

We created an obstacle model to guide people across various hurdles so that AI could recognize them. Our second suggestion is to use the V-DAI function to quickly and accurately categorize obstacles. As shown in Figure 8, we built an obstacle model to aid users in overcoming various obstructions so that machine learning could recognize them. Data collection collects a large number of stereo images from the stereo camera as the user walks along the path in various ways. The next phase is data preparation, which entails using the V-DAI feature to transform the obstacle image and building the training obstacle models for expert labeling. Finally, obstacle model training is used to teach our machine learning throughout the obstacle categorization stages. Finding the frequency of depth values that occur in each column or width of the obstacle image (depth image ground removed) and counting from the left column along the width of the obstacle image until the right column is the final step in creating a V-DAI feature for the training set, as shown in Figure 10, is the final step in creating a V-DAI feature for the training set (left). Figure 10 displays the whole V-DAI picture (right).

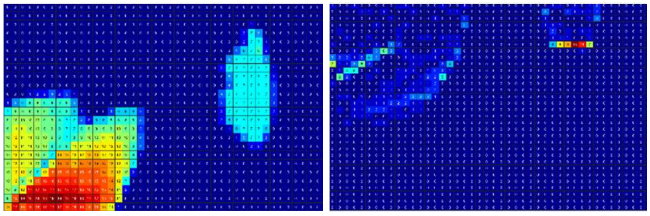

Fig. 10. Obstacle image (left), the V-DAI image (right).

This V-DAI is a top-down or birds-eye view virtual projection used to train obstacle models. We will utilize the categorization learner tool in this stage. [13], import the V-DAI feature vectors produced during the feature extraction stage, as well as the obstacle class label generated during the expert labeling step. The next stage is to partition the entire dataset into three parts: training set, validation set, and test set. A training set is a collection of data that is used to update the data in the neural structure to reduce mistakes. The validation set contains data that is used to verify the capacity to generalize attributes and must also be used to cease training models. Finally, because the Test set contains data that has never been trained previously, it is utilized to evaluate the model's performance. The dataset was divided into three parts for model training: 70 percent training, 15 percent validation, and 15 percent testing. Then, using the Convolutional Neural Network (CNN) learning technique, we trained an obstacle model with a variety of picture types, including depth images, obstacle images, and V-DAI images[14]. As shown in Figure 11, we propose a convolutional neural network topology that includes convolutional, batch normalization, ReLU, max pooling, fully connected, and softmax layers. The final phase in model training is to use CNN learner to assess the efficiency of each picture kind. Various categorization performance parameters are measured. The capacity to categorize barriers accurately is the first key. The second factor is prediction speed, which refers to how quickly machine learning processes the obstacle class during the classification stage. The last performance statistic is training time, or time spent on training models. All three parameters were weighed to determine which image type was best for obstacle categorization.

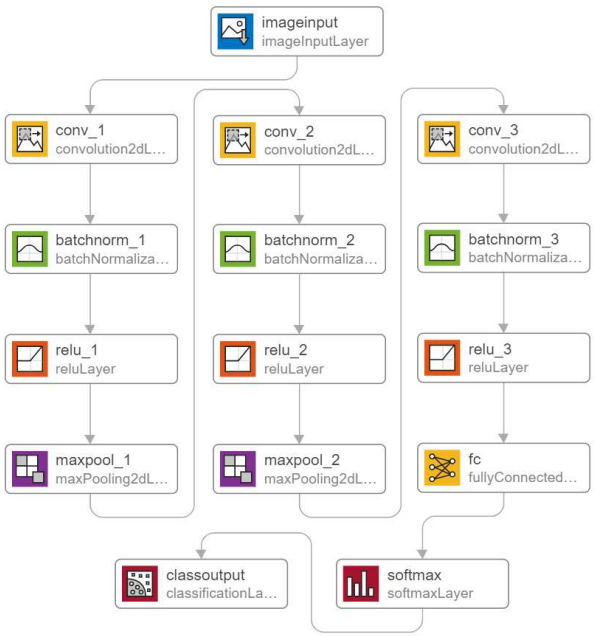

Fig. 11. Our convolutional neural network structure.

C. SOUND FEEDBACK GENERATION

The obstacle class is obtained from the classification phase in the navigation stage. The user's navigation begins with the 7 obstacle classes transformed to a female navigation voice in English. This voice is made up of four sets of voices: four for describing the surrounding area and three for avoiding obstacles. "All clear" (no obstacles), "Left object" (objects on the left side), "Right object" (objects on the right side), "Parallel object" (objects on both sides, no need to avoid), "Keep left" (dodge to the left), "Keep right" (dodge to the right), and "Slow down" (stop walking). When an impediment is recognized, the navigation voice is recorded in wav format and sent to the user through a bone conductive stereo headset. Figure 12 shows a three-dimensional tone pattern with sound spectrum and left-right altitude.

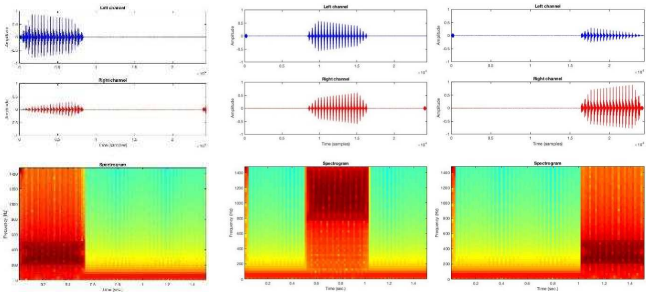

Fig. 12. Sound pattern for left object class (left), stop class (center), right object class (right).

IV. TESTING AND EVALUATION

Training a stereo camera on the data collector's head position to capture a person's perspective view, as illustrated in Figure 13 (left), and then recording a stereo picture in 1,344 376 pixels at a sampling rate of 15 frames per second in mp4 gray video format, is how the training set data is gathered. All stereo images collected from sidewalks with obstructions along the path, such as electrical poles, bushes, road signs, wall, and etc, as shown in Figure 13 (right).

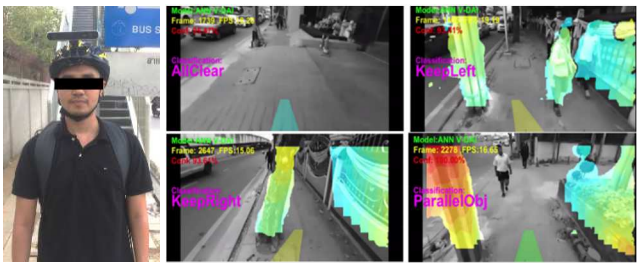

Fig. 13. ZED mounted on user head (left), the example resulting interface (right).

The stroll is 2.3 kilometers long, takes 39.61 minutes to complete, and contains 34,325 stereo pictures. Table I shows the dataset details. For the feature extraction procedure, the stereo photos will be transformed into obstacle images.

TABLE I. INFORMATION FOR EACH DATASET

Dataset	Distance (km.)	Duration (sec.)	Sample (frame)
Chula-1	0.24	208	3,048
Chula-2	0.40	404	5,909
Chula-3	0.16	193	2,756
Chula-4	0.50	440	6,565
Chula-5	0.35	452	6,774
Chula-6	0.65	680	9,273
Total	**2.30**	**2,377**	**34,325**

For a supervised learning approach, expert labeling is processed to provide labels for use in model training phases. Preparing movies for specialists who define the video frame numbers of each stereo picture is the first step in this study's labeling. In addition, we've attached obstacle photos for further detail, and the expert's dataset video is provided in Fig. 14. The expert will specify the type of obstacle class in each video frame. To identify the barrier, we employ seven unique classes: left object, right object, parallel object, keep left, keep right, all clear, and halt. In the feature table, the labeling results will be recorded in CSV format. Table II shows the findings of the expert labeling.

TABLE II. NUMBER OF IMAGE IN EACH CLASS

No.	Class	Number of sample	Percentage (%)
1.	All clear	6,985	20.35
2.	Left object	6,502	18.94
3.	Right object	8,733	25.44
4.	Parallel object	7,519	21.91
5.	Keep left	1,389	4.05
6.	Keep right	2,412	7.03
7.	Stop	785	2.29
Total		**34,325**	**100**

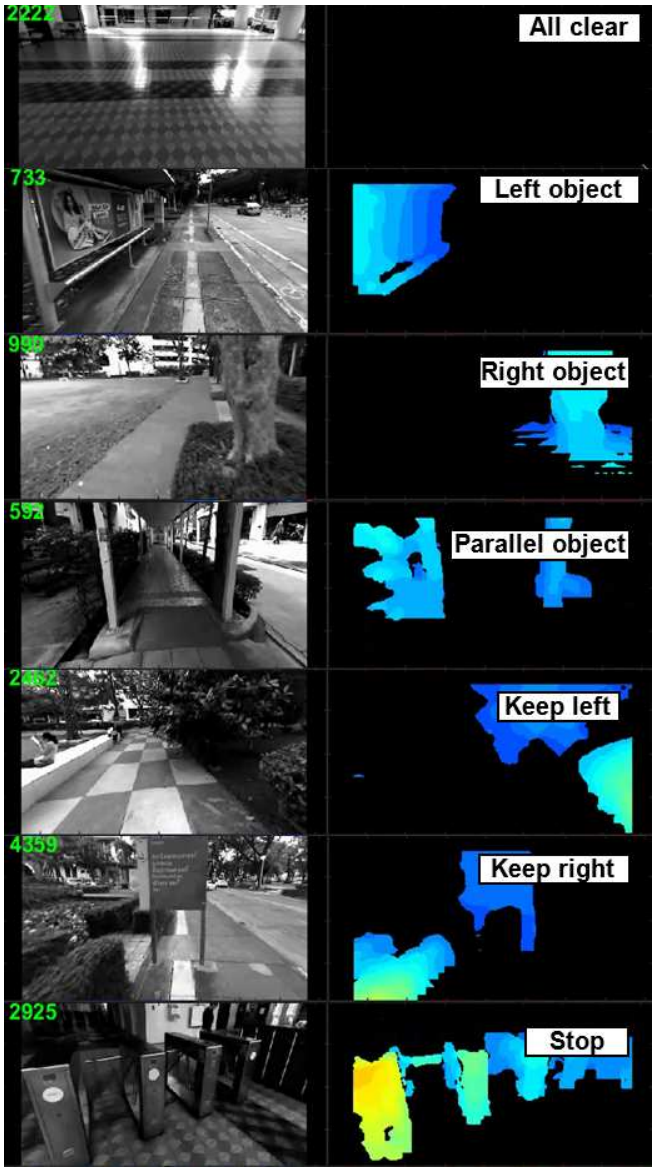

Fig. 14. For expert labeling, use a sample picture and a depth image.

In order to speed up the procedure,, we reduce the size of the depth image, obstacle image and V-DAI image to smaller image of a square 20 × 20 pixels, the accumulative graphic for each image type, including depth image, obstacle image and V-DAI image as shown in Fig. 15-17.

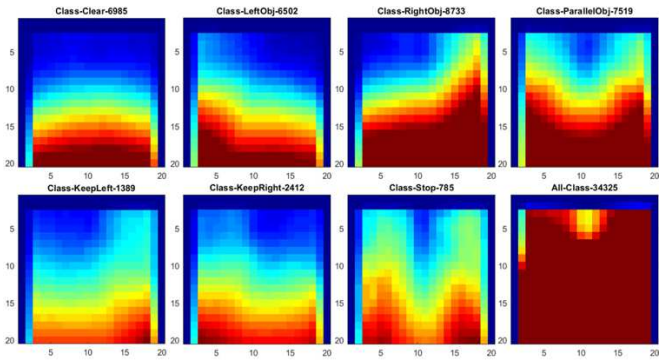

Fig. 15. Accumulative graphic for each depth image class.

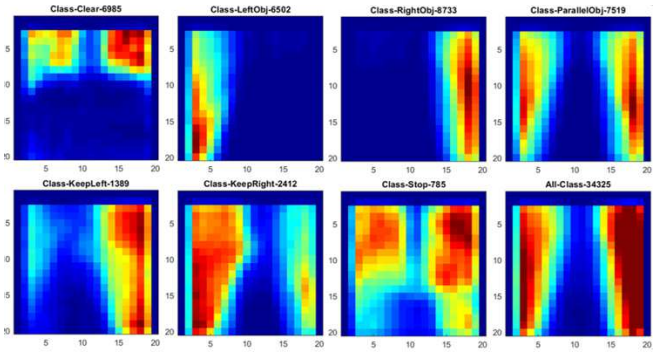

Fig. 16. Accumulative graphic for each obstacle image class.

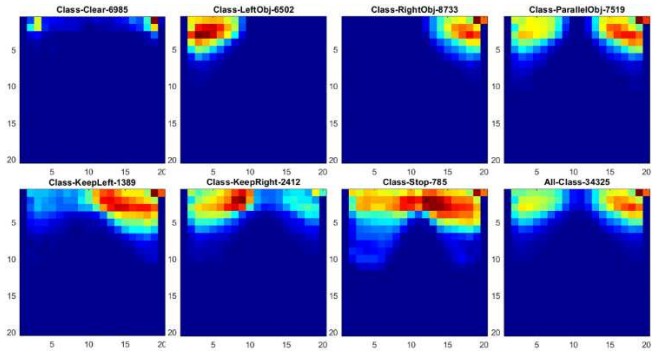

Fig. 17. Accumulative graphic for each V-DAI image class.

We conducted studies to discover the most suitable machine learning in order to deliver the best obstacle model performance in terms of processing speed and accuracy. Using the Convolutional Neural Network (CNN) learning technique, we trained the model with a variety of picture types, including depth images, obstacle images, and V-DAI images. The classification performance is calculated using the averaged value from 10 rounds of training, and the visible features picture of each convolutional layer after training is shown in Fig.18-20, as well as the experiment results in Table III. The V-DAI picture with the highest categorization accuracy. According to the results of the trial, the V-DAI picture with CNN has the greatest classification accuracy of 93.61 percent and the quickest prediction speed of 10,169 samples per second.

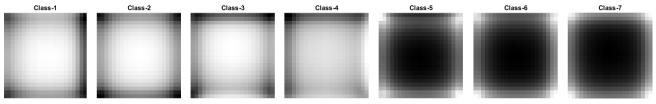

Fig. 18. Visualized features image of first-convolutional layer.

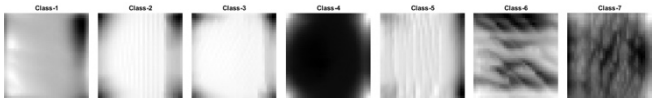

Fig. 19. Visualized features image of second-convolutional layer.

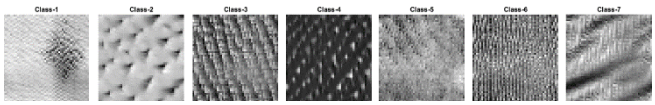

Fig. 20. Visualized features image of third-convolutional layer.

TABLE III. CNN PERFORMANCE FOR EACH IMAGE TYPE COMPARISON

Metrics	Image type		
	Depth	Obstacle	V-DAI
Accuracy (%)	89.3436	84.7369	93.6085
Precision, PPV	87.0746	83.3447	90.8342
Recall, TPR	83.0958	77.8623	88.6538
Fall-out, FPR	1.8854	2.7191	1.1063
F1-score	84.6806	80.0083	89.5340
Training Time (sec)	151.1635	132.5499	88.9037
Prediction Speed (sample/sec)	8,143	7,640	10,169

We tested our system outside of the Chulalongkorn University region, which is used to train the obstacle model, to check that it still works. The researchers picked a 670-meter walk from the researcher's residence to the Phayathai sky-train station in Pathumwan, Bangkok, for the walking route. The researcher picked this experimental walking path because it has a variety of hazards, including people, motorcyclists, automobiles, traffic signs, electric poles, barriers, trees, plants, and other obstructions. Figures 21-22 show the various possibilities.

Fig. 21. Real-world obstacle classification result.

Fig. 22. Real-world obstacle classification result.

There are many situations where our system is malfunctions as shown in Fig. 23-25, such as in a scenario where pedestrians are walking in the opposite direction. This malfunction can lead to a collision which may cause an accident. The solution to this problem is to use the position information of obstacles in the past frame to predict the trajectory of the obstacle's direction. Furthermore, there are dangerous situations, such as when our system navigated the user to a drain, due to our system's inability to detect objects below ground level. We recommend using a white walking stick with our system to scan for holes or drains that are below the level of the walkway. This limitation of our system with drains detection. Unfortunately, our system may lead users toward dangerous objects, such as incoming cars or motorbikes, because our system cannot identify the objects in the walkway. The researcher proposed a future solution of using object recognition based on deep learning techniques.

Fig. 23. The limitation of our system lead to incoming pedestrian.

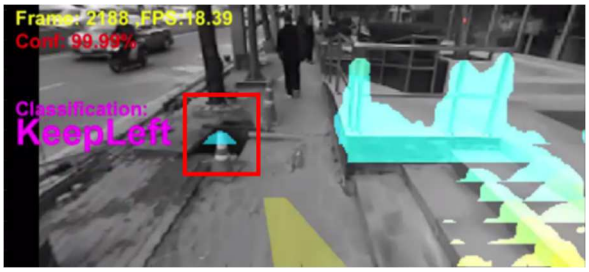

Fig. 24. The limitation of our system lead to drains.

Fig. 25. The limitation of our system lead to dangerous objects detection

V. CONCLUSION AND DISCUSSION

We demonstrate a real-time navigation and obstacle detection system for the blind using a stereo camera. A stereo camera and a machine learning system were used to make our discovery. Our initial contribution is a way for quickly segmenting obstacles using our H-DAI. The second contribution is a quick and accurate obstacle classification approach, with our V-DAI picture with CNN achieving a classification accuracy of 93.61 percent and a classification speed of 10,169 images per second. We ran several tests to assure the greatest degree of obstacle detection accuracy and speed, as faults in detection or excessive processing time might cause significant harm to the user.

The work's limitations, such as its capacity to be used in a dark or foggy setting, might be investigated for future direction. To address this, fusion sensors between the stereo camera and LiDAR, which are available in newer mobile devices, might be employed. In low-vision situations, such sensors can help in detection. Another possible future project would be to improve the accuracy of the results by taking into account the trajectory direction of approaching items. This might be accomplished using object trajectory prediction based on the item's previous and current positions. Object recognition might also be incorporated to improve our navigation system. An arriving object's type would be communicated to the user. This will aid children in being prepared to avoid unsafe or severely deadly items quickly.

REFERENCES

[1] WHO. Draft action plan for the prevention of avoidable blindness and visual impairment 2014-2019: universal eye health: a global action plan 2014-2019. Geneva: World Health Organization; 2013.

[2] JM. Loomis, RG. Golledge and RL. Klatzky, “GPS-Based navigation systems for the visually impaired”, In: Barfield W, Caudell T, editors. Fundamentals of wearable computers and augmented reality, New Jersey, Lawrence Erlbaum Associates, 2001, p. 429-46.

[3] M. Ritz, L. Konig, “Laser technique improves safety for the blind”, MST NEWS. 2005;5:39.

[4] Inc NRI. The laser cane, model N-2000, [Internet]. 2021 [cited 2021 Jul 20], Available from: http://www.nurion.net.

[5] I. Ulrich, J. Borenstein, “The guide cane-applying mobile robot technologies to assist the visually impaired”, IEEE Trans Syst Man Cybern Syst "Hum. 2001;31(2):131-6.

[6] G. Research, “The miniguide mobility aid”, [Internet]. 2021 [cited 2021 Jul 20], Available from: http://www.gdp-research.com.au/minig_1.htm.

[7] K. Ito, M. Okamoto, J. Akita, T. Ono, I. Gyobu, T. Takagi, et al, “CyARM: an alternative aid device for blind persons”, Extended abstracts proceedings of the 2005 conference on human factors in computing systems, 2005, Apr 2-7, Portland, USA, New York, Association for Computing Machinery, 2005, p. 1483-8.

[8] S. Vorapatratorn and K. Nambunmee, “iSonar: An obstacle warning device for the totally blind”, Journal of Assistive, Rehabilitative & Therapeutic Technologies, 2(1), 23114.

[9] R. Damaschini, R. Legras, R. Leroux and R. Farcy, “Electronic travel aid for blind people”, In Pruski A, Knops H, editors. Assistive technology: from virtuality to reality, Amsterdam:, IOS Press, 2005, p. 251-5.

[10] S. Vorapatratorn, A. Suchato and P. Punyabukkana, “Real-time obstacle detection in outdoor environment for visually impaired using RGB-D and disparity map”, Proceedings of the international convention on rehabilitation engineering & assistive technology, 2016, p. 1-4.

[11] Stereolabs, “ZED stereo camera”, [Internet]. 2021 [cited 2021 Jul 20]. Available from: https://www.stereolabs.com/zed/.

[12] SD. Cochran and G. Medioni, “3-D surface description from binocular stereo”, IEEE Trans Pattern Anal Mach Intell, 1992,(10): p. 981-94.

[13] The MathWorks, “Classification Learner” [Internet]. 2021 [cited 2021 Jul 20]. Available from: https://www.mathworks.com/help/stats/classificationlearner-app.html.

[14] A. Krizhevsky, I. Sutskever, GE. Hinton, “Imagenet classification with deep convolutional neural networks”, Adv neural inform process syst, 2017,60(6), p. 84-90.