Multimedia Tools and Applications (2024) 83:6555–6578
https://doi.org/10.1007/s11042-023-15644-y

The development of assisted- visually impaired people robot in the indoor environment based on deep learning

Yi-Zeng Hsieh[1] · **Xiang-Long Ku**[1] · **Shih-Syun Lin**[2]

Received: 31 May 2021 / Revised: 7 June 2022 / Accepted: 22 April 2023 /
Published online: 10 June 2023

Abstract
The indoor positioning for visually impaired people has influence on their daily life in unknown indoor environment. This study designs the robot that can assist the blind walking safety and navigate in indoor environment by a single camera. The sense classification is proposed to position the blind in indoor environment by proposed convolutional neural network framework and integrate the semantic segmentation to find the road surface through a depth camera to guide the blind walking. The proposed vision-based sense classification method is compared with the traditional WiFi triangular-positioning method, and the average error of x-y coordinate position result as (9.25,3.65) is better. From the experiment, the designed robot can help the visually impaired people to indoor navigation in unknown indoor environment.

Keywords Deep learning · Convolutional neural network · Indoor positioning · Visually impaired people

1 Introduction

According to the definition of the World Health Organization (WHO), a person with visual acuity less than 0.05 is visually impaired. Visually impaired people encounter many obstacles in daily life, and need white canes or guide dogs for assistance (https://www.tfb.org.tw/web/news/about_qa.jsp). Guide dogs can help the visually impaired people to avoid danger when walking but the guide dogs cannot lead the visually impaired people to a designated place. The cost of training a guide dog is high, and the training process is very long. White canes are preferred tools for the visually impaired, but there are many limitations about the safety of white cane (http://www.guidedog.org.tw/aboutguidedog/about-1.html). With the development of science and technology,

✉ Shih-Syun Lin
linss@mail.ntou.edu.tw

[1] Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei City 106, Taiwan

[2] Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung City, Taiwan

intelligence robots which will be more efficient than the training of guide dogs also can assist the visually impaired people to avoid obstacles. Visually impaired people often need to ask others to lead the way to a designated place in an indoor environment. This situation causes a major problem for the visually impaired people in an unfamiliar indoor environment. The motivation of this study is owing to that indoor environments such as office space or school space are fixed, therefore an indoor map is built that the robot can understand to indoor navigation by adopting the deep learning methods to calculate the distances and to identify the direction of obstacles. It is hoped that the proposed robot can help the visually impaired people to avoid obstacles, and the destinations in the indoor environment are arrived.

In order to locate the robot indoor position, this study intends to use convolutional neural network (CNN) to perform indoor location classification by scenes image to achieve indoor positioning. The semantic segmentation method is also adopted so as to find the flat road surface to safety navigation. The purpose of semantic segmentation is to predict the category corresponding to each pixel in the environment image, and then based on the segmentation results the robot can know whether the road ahead is walkable. On the other hand, to detect the obstacle is detected through YOLOv3 to allow the robot to complete the obstacle avoidance and measure the distance between robot and obstacle by the Intel RealSense depth camera in this study.

This study is to develop the robot by classifying the indoor scenes in order to indoor positioning algorithm. The algorithm adopts RGB images by proposed convolutional neural network framework to calculate the indoor positioning which the indoor space is spited off 35 categories regions without other sensors. The robot location is estimated by its scene classification, and to go a step further the robot can assist the blind to indoor navigation by path planning and using semantic segmentation method to calculate the flat road distance by transforming the region area using depth camera to make sure the walking road surface is safe. From the above the motivations, the contributions of this study are summarized as following:

1) The sense classification method for indoor positioning is adopted through a single RGB camera. 2)The semantic segmentation is used to find walking safety road and the road region is transformed to real distance coordinate through depth camera to help the blind walking.
2) The proposed robot can provide the walking and navigation information for the blind who are in unknown indoor environment. The designed robot is also integrated with the YOLOv3 to detect the objects to help the blind avoid obstacles.
3) The experiments show the x-y coordinate average error of proposed image-based sense classification indoor positioning method compared with the traditional triangular-wifi positioning are (9.25,3.65) and (19.45,8.65), separately. From the experiment result, the proposed indoor positioning method can be more reliable.

The following is the organization of the paper: Section 1 introduces the research motivation and goals. Section 2 presents the application of the robot and the related application of the algorithm used in this article. Section 3 discusses the robot architecture and algorithm. The experimental results are discussed in Section 4. Section 5 offers conclusions and suggestions.

2 Related works

Most of the early related researches focused on the improvement of the white cane by adding sensors such as ultrasonic, infrared, laser rangefinder, etc. Some researchers would use small cameras for image processing. These are collectively referred to as electronic travel aids. (ETA) [3] proposed a wearable obstacle avoidance system to assist the visually impaired to navigate safely when wearing this cane and glasses. The ultrasonic sensor is placed in the cane and the cane can detect obstacles on the ground. Glasses embedded with ultrasonic sensors and small cameras are used to detect whether there are obstacles at a specific angle above the head. The information obtained is fed back to the user through sound and vibrational feedback.

The advancement of computer vision uses two cameras to form a stereo vision system[8] used the Tyflos navigation system which integrates the data of each sensor and two cameras and provides feedback to the visually impaired through sound. The stereo method performed two 32*32 diagram analyses [4] and then analyzed the location of obstacles.

Early robots used sensors to avoid obstacles [27]. With the advancement in hardware and computing power, many studies built the sensor values to transform maps and integrate more environmental information [10, 26] proposed an ultrasonic environment detection robot that uses different ultrasonic reflection rates to distinguish object materials. It can provide the robot with more information and help the robot be more efficient in avoiding obstacles.

The use of depth cameras in robots has gained popularity. Microsoft Kinect (A Depth Camera) used for mobile robots can be effective for navigation. However, this kind of depth camera is easily affected by sunlight. It cannot detect transparent objects like glass. This is a big problem for obstacle detection. Therefore, the depth camera is used for obstacle avoidance, and a sensor is used for data integration and to strengthen the recognition ability of obstacles. To assist the robot, [28] proposed to use Kinectv2 a laser rangefinder, and an ultrasound sensor to perform data integration. This method can make up for the shortcomings of a sensor and reduce the blind spot of the robot. After integrating environmental information, the AStar algorithm is used for path planning.

In the indoor environment, the Global Positioning System (GPS) cannot be used. Therefore, the most widely used indoor positioning technology is to deploy wireless local area network (Wireless LAN, WLAN) equipment. It is mainly due to its low cost and high utilization rate [7] used the most common mobile device, by turning on the WiFi hotspot functions on the mobile devices will increase the accuracy of WLAN. This research uses mobile hotspots and WLAN devices to build a WiFi node map and confirms the current location through the nodes of the map corresponding to the Dynamic Access Points and Fingerprints (DAFs).

With the development of deep learning, a large amount of data is collected for deep learning calculation and analysis [32] suggested that WLAN indoor positioning is a low-cost solution and satisfies the location-based services (LBS). Therefore, a localization algorithm is proposed for extracting indoor environment features. The information of each access point is planned in information integration format to establish the relationship between the RSS signal and the location coordinates. Then, it combines the selected RSS information to train the neural network for indoor positioning. This proposed algorithm has a higher positioning accuracy than traditional positioning algorithms. To predict the indoor location, [17] proposed a solution using an indoor twisted magnetic field. Due to the ambiguity in the magnetic field data, it is difficult for the existing magnetic field positioning method to locate in a wide space. To improve accuracy, it is necessary to integrate multiple

magnetic pole sensors and isolate devices that interfere with the magnetic field. Therefore, to solve this problem, the deep learning method is used for indoor positioning. First, extract the features from the magnetic field sequence and generate the magnetic force map as the input of CNN.

From 1949 to the present, the learning mechanism based on neuropsychology started the first step of machine learning. The machine learning algorithm requires a large number of iterations with a lot of calculations, therefore it is hardware dependent. With the rapid growth of computer hardware technology, computing power has been greatly improved. This has helped neural network technology to flourish. The architecture of computer vision systems has become sophisticated. Deep learning is used to improve image recognition.

Object recognition is a classic technology used to understand the content of pictures especially in the field of computer vision. It can use local feature detection methods to recognize more complex tasks. The accuracy has been greatly improved, i.e. from the traditional use of the Scale-Invariant Feature Transform (SIFT) algorithm [29] to the current use of the deep learning model [13] proposed the use of a four-layers CNN for object recognition.

In deep learning algorithms, the classic algorithms used to find objects are YOLO [20, 22] and RCNN [11, 23], both of which are target detection algorithms. Then the border of the object is adjusted through Non-Maximum Suppression (NMS) for detection and positioning. Subsequently, the improved fast-RCNN has been improved in speed. The basic process is to first generate a large number of region proposals through CNN, and then use another CNN to extract the characteristics of the potential target regions and perform category judgments.

In object classification, there is a neural network that frames the target as described above. Another neural network classifies various objects and adds color labels. Semantic segmentation [2, 18, 33] can effectively distinguish indoor environment objects by color. The categories of indoor environment are chairs, tables, people, floor, windows, doors, walls, etc., The most classic network of semantic segmentation is fully convolutional networks (FCN) [25]. Classic CNN uses a fully connected layer to obtain a fixed-length feature vector for classification after the convolution layer. FCN classifies the picture at the pixel level to complete the semantic segmentation problem. It can accept an input of any size and uses a deconvolution layer to perform the feature map of the last convolution layer. Upsampling restores the size of the original image so that a prediction can be generated for each pixel while retaining the spatial information in the original input image.

FCN can effectively complete semantic segmentation, but in complex environments or similar objects have misjudgments [33] proposed an improved algorithm Pyramid Scene Parsing Network (PSPNet) based on the FCN architecture [9] effectively improved the accuracy of object detection by concatenating deep and shallow features to increase the amount of information of shallow features, and giving enough context when segmenting at the shallow layer. ResNet is also used to extract features and add auxiliary loss to the back of the network to strengthen network training and accelerate network convergence [9, 28] presents a robust probabilistic deep Q network algorithm that would assist in robotic surgery for censorious surgeries in real-time [19] through a self-learning network segmentation process requiring minimal training and can learn by itself from the previous experimental outcomes.

As mentioned above, many sensors are used as the main electronic travel aids (ETA) and some methods combine computer vision with artificial intelligence. Wearable devices have been developed to improve the navigation of blind and visually impaired people [19, 24, 30] proposed the vibration-only modality, compared with audio-kinesthetic or

multimodal vibro-audio solutions and touch-screen for blind users [15] developed a wearable blind guide device based on image streaming and deep learning. Its main goal is to assist the white walking stick used by the visually impaired. The wearable device for the blind uses a single RGB camera and uses a CNN network to predict depth information. Compared with a depth camera, it reduces the weight when worn and can obtain distance information. The calculation is divided into two parts: (1) use CNN to predict the depth information of RGB images and pass the depth of the environment, and the information includes the walkable distance; (2) use PSPNet to semantically cut the RGB image, and show the road surface in black. Integrating the two data can get the depth value corresponding to each road surface. The wearable device shakes and blurs the picture when taking pictures, and for the complex calculations, the data needs to be streamed back to the server. Therefore, this paper expects to use a wheel-shaped robot to improve this problem. The robot is loaded with high computing equipment and can operate without the need for a network environment. Therefore, it solves the problem of delay caused by the communication with the server.

3 Methodology of the designed assisted-blind robot system

According to [15], semantic segmentation can assist the visually impaired by telling whether the road ahead is safe, and then knowing the farthest distance through depth information. However, the use of wearable devices may cause poor photo quality. There will be misjudgments when performing deep learning. The wearable device has a size limit for convenience so its performance is limited. It needs to be uploaded to the server for calculation through a network. The above problems are expected to be solved by wheeled robots. The solutions will be explained in detail below.

The wheeled robot developed in this paper guides the visually impaired to a designated place indoors and leads the visually impaired to take a safe path. The operation process of this robot is shown in Fig. 1. When the robot is started, it rotates on the spot first. It

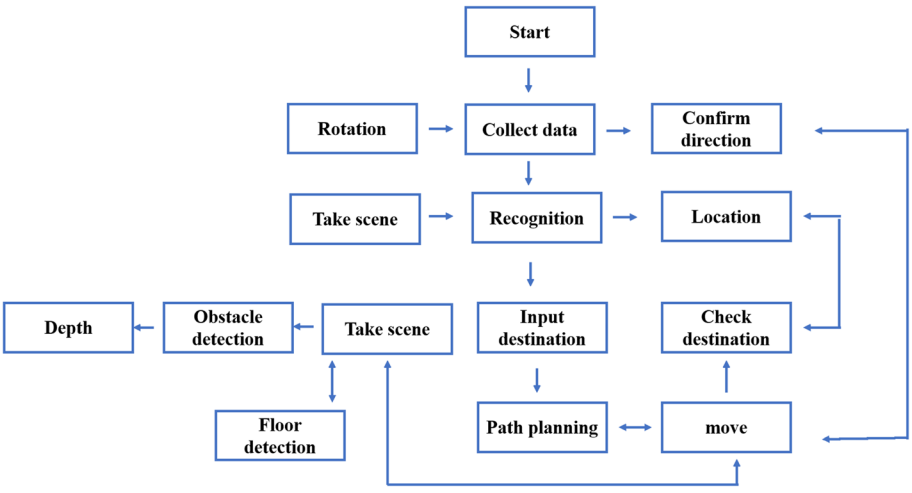

Fig. 1 The flowchart of proposed robot indoor navigation system

performs a magnetic pole correction and confirm the orientation of the robot. It then takes a picture of the scene to confirm its current position and then wait for the visually impaired to input the target position. After the robot obtains the target position it performs path planning, obstacle recognition, and movable range recognition. Finally, the robot confirms whether it has reached the destination through scene recognition. For detailed research methods, refer to the following chapters.

The hardware architecture used in this research and the detailed specifications of the computer, depth camera, motor control board, and motor are shown in Fig. 2, which is the wheeled robot and structure design for this research. The depth camera uses Intel RealScene camera and provides the environment depth information. The gyroscope sensor is applied for robot direction. The notebook is computing kernel of object detection and semantic segmentation deep learning algorithm and sending instructions to control robot motors.

The deep learning algorithm of this research is applied to the robot in an indoor environment. It helps in obstacle recognition and indoor positioning. It also ensures safety while walking in front of the robot. Through a large amount of indoor data collection, the CNN scene recognition is performed. The semantic segmentation is completed through the pyramid scene parsing network and YOLOv3 is used.

3.1 The proposed convolutional neural network architecture for sense recognition

This study uses CNN as sense recognition method architecture and the input layer is 64*64 image. The CNN architecture is described as the following. The first and second layers are convolutional and the kernel size is 3*3. The 32 feature maps are generated and the activation function is linear rectification function (ReLU) followed by 2*2 max-pooling. The third and fourth layers are convolutional layers in which the kernel size is

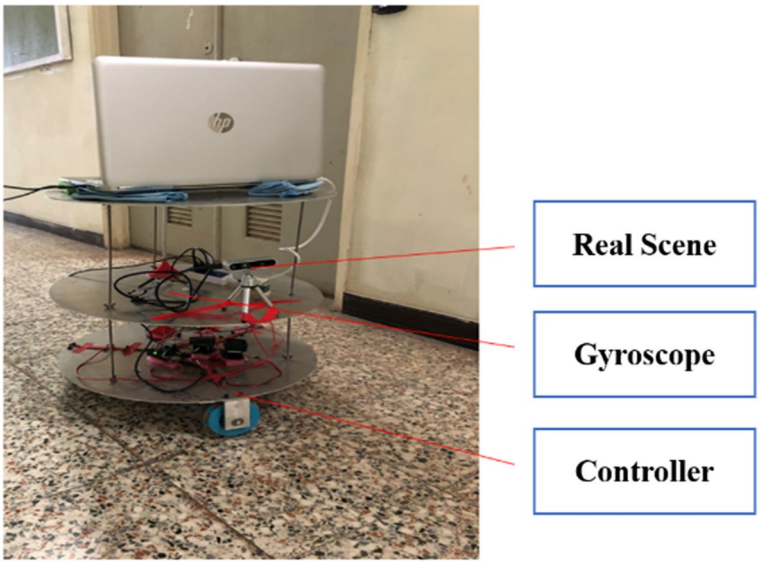

Fig. 2 The designed robot consists of RealScene camera, gyroscope and controller

3*3 and 64 feature maps are generated. The activation function is also ReLU followed by 2*2 max-pooling. The fifth and sixth layers are convolutional with a kernel size of 3*3 and 128 feature maps are generated. The activation function is ReLU followed by 2*2 max-pooling. There are two connected hidden layers (1024, 512). Finally, the output layer is 35 and uses the softmax function as shown in Fig. 3. The following section introduces the convolutional layer, pooling layer, and activation function methods used in this study.

- **Convolutional Layer**

The Convolutional method is to generate 32, 64, and 128 feature maps through a 3*3 kernel with one stride. The purpose is to extract the image features and the size of the image remains unchanged. The kernel uses the sliding window method to calculate. A large number of weight values and feature maps are used to continuously modify the weight values to find the most suitable parameters for the image.

- **Pooling Layer**

The method of pooling is to retain more important features and filter out the unimportant information of the image. This study uses 2*2 max-pooling to reduce the dimensionality of the picture and the amount of calculation.

- **Activation function**

The activation function is an important part of the neural network. Choosing the wrong activation function may cause the network to fail to train and cause the gradient to disappear during backpropagation. The selection of ReLU in this paper can effectively avoid this problem.

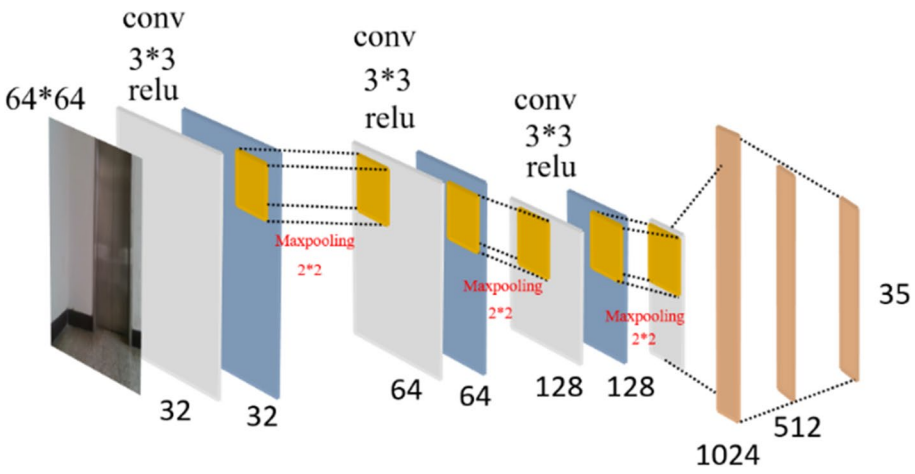

Fig. 3 The convolutional neural network architecture for sense-based classification on indoor positioning method in this study

3.2 Pyramid scene parsing network

This paper uses Pyramid Scene Parsing Network (PSPNet) [33] as the deep learning algorithm of semantic segmentation, which allows the robot to detect whether the road ahead is walkable. There is 224*224 input image. First, features are extracted through ResNet [14] which produces a feature map that is one-eighth of the original image. It then uses spatial pyramid pooling [5] to divide it into 1*1, 2*2, and 4*4, and generate 1, 4, and 16 feature maps, respectively. There are a total of 21 blocks, and CNN and average pooling are performed on these 21 blocks, respectively. After that, upsampling of these 21 blocks is performed and the method of dilated convolution [31] is used to restore to eight points of the original image. Because it needs to be concatenated with the feature map generated by the previous ResNet, the network uses the auxiliary loss function [9] method to train the network.

The following part describes the methods of dilated convolution, auxiliary loss function, and spatial pyramid pooling. The robot will inform the visually impaired of the results. The distance value will be calculated as the number of steps, and each step will be 40 cm. The visually impaired will be notified by voice, such as the safety distance ahead is 9 steps, or the safety distance on the right is 3 steps, or the safety distance on the left is 3 steps.

- **Dilated convolution**

[31] proposed the importance of the use of dilated convolution for semantic segmentation. The purpose of this method is to fill 0 between the kernels. The dilated rate can be used to determine how many 0 s are filled between the kernels. The calculation can increase the amount of information on the network but too many zeros will cause gridding, which affects the loss of correlation between the feature map and the kernel generated by dilated convolution.

- **Auxiliary loss function**

PSPNet uses auxiliary loss function, that combines with ResNet's res4b22 using Loss1 function and ResNet's res5c using Loss2 function. It can be learned from [9] that if the auxiliary loss is used for multi-target detection tasks and it can effectively accelerate the convergence speed. PSPNet also can be seen from the image that the convergence of loss is higher than that of the twice loss function in multi-objective detection. Multiplying the coefficient before loss function can also effectively speed up the convergence. In this paper, the Loss1 function is the fourth residue block of ResNet and cross-entropy as Eq. 1 is used. Loss2 function is the fifth residue block of ResNet and cross-entropy is also used. The final total loss is shown in Eq. 2.

$$\mathrm{L}\left(y_{pred}, y_{label}\right) = -y_{label} \log y_{pred} \tag{1}$$

$$Loss_{\text{total}} = 0.4 * Loss1 + Loss2 \tag{2}$$

- **Spatial Pyramid Pooling [5]**

After generating the feature map from the original image through ResNet50, perform spatial pyramid pooling to cut the feature map into 1*1, 2*2, 4*4 blocks (spatial bins),

a total of 21 blocks, and then perform CNN on these 21 blocks respectively. 2*2 max-pooling of this method is to effectively extract features from images of various sizes and the dimensionality is reduced the training speed is accelerated.

3.3 You only look once v3

The obstacle avoidance system of the robot uses YOLOv3 (You Only Look Once) [20–22]. The network is based on detecting objects for training. YOLOv3 takes images by the robot as neural network input to filter out the best bounding box with the highest confidence score.

The method cuts the feature maps of the image into 13×13, 26×26, 52×52 units and sets each unit with 3 anchor boxes as the candidate area of the bounding box. It can find the most suitable four vertex coordinates t_x, t_y, t_w, t_h as the final box of the object. The X and Y of the object can be obtained and the center point of the bounding box can be calculated. The distance of the object can be obtained by corresponding the center coordinate to the depth map generated by Intel RealSense D435.

The distance value is calculated as the number of steps and each step is 40 cm. The visually impaired is notified by voice, such as the distance to the obstacle in the front is 5 steps, the obstacle on the right is 3 steps away, and the obstacle on the left is 4 steps away. The process of YOLO detection is as follows:

- Resize the picture to 416 * 416 as the input of the neural network.
- Run the neural network to get some bounding box coordinate positions, object confidence, and probability of each category. YOLO uses the Eqs. (3)–(6) to predict b_x、b_y、b_w、b_h values, separately.

$$b_x = \sigma(t_x) + c_x \tag{3}$$

$$b_y = \sigma(t_y) + c_x \tag{4}$$

$$b_w = P_w \, e^{t_w} \tag{5}$$

$$b_h = P_h \, e^{t_h} \tag{6}$$

- **Perform non-maximum suppression (NMS) and filter boxes.**

[21] proposed a 53-layer convolutional architecture and named it DarkNet-53. The network is mixed with ResNet to solve the gradient problem. The network performs upsampling and fusion concatenation. Multiple sizes are used for prediction. The sizes are 13×13, 26×26, 52×52, respectively. Training with feature maps of multiple sizes can improve the detection effect of small targets. In this paper, the COCO data set [6] is used for target prediction. The result of the integration of YOLO results and depth map is shown in Fig. 4.

- **Indoor environment scene dataset establishment**

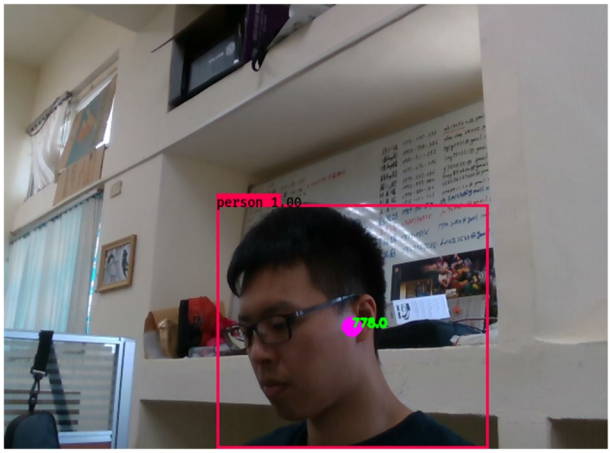

Fig. 4 YOLOv3 combined with depth map result

The indoor positioning method of this paper needs to be built on a digital map. Therefore, it is necessary to build an indoor map of the experimental environment at first. The fourth floor of school building is used as an experimental environment and is drawn according to the pixel scale relationship. From a 2D image map, the actual length is 240 cm which occupies a total of 340 pixels in the 2D image map. Therefore, each pixel is 0.705 cm. A coordinate system belonging to the robot can be established. It is similar to the center pixel coordinates in the 2D image map. This center coordinate is a fixed value. As there are 35 blocks in the total image map, there are also 35 center coordinates. The Y pixel is the sum of the pixel coordinates and the 2D image map. It is converted from the depth value of the robot and the scene. When the robot rotates +90 degrees to take an image, it is calculated by Eq. 7, and when the robot rotates −90 degrees to take the image, it is calculated by Eq. 8.

$$Y_{pixel} = 240 + (d * 0.705) \tag{7}$$

$$Y_{pixel} = 580 - (d * 0.705) \tag{8}$$

The fourth floor of the experimental environment is divided into 35 blocks. The test speed of the robot on the road is about 45 (cm/s). Considering the robot width is 50 cm, the interval of blocks is 100–120 cm. The scene positioning results are obtained from the results of scene recognition. The robot rotates 90 degrees left and right to take pictures to position location in 35 blocks and uses a depth camera for the depth value of the scene image. The distance between the robot and the scene can be obtained from the depth camera. The indoor positioning dataset is collected by the above Eqs. 7 and 8. The training indoor data set is a total of 39,176. The verification indoor data set is 19,588. The testing dataset number is 9000. The training and testing dataset are divided into 35 categories and each category is about 1118 and 257 separately.

- **WiFi positioning during robot movement**

When the robot moves, it needs to keep positioning itself according to the planned route. The robot must stop and do the scene positioning. The research method in this

paper under this condition uses WiFi triangle positioning. When the robot encounters an obstacle, it is avoided. It uses triangle indoor WiFi positioning to correct its position. The following introduces the 1D-CNN architecture of WiFi triangle indoor positioning. The input layer comprises three groups of WiFi signal values from t to t + n shown in Fig. 5 and n is 256.

The first layer is a convolutional layer with a kernel size of 3*3. It generates 32 feature maps and uses ReLU as the activation function. The second layer is a convolutional layer. The size of the layer and kernel is 3*3. It generates 64 sets of feature maps by using ReLU as the activation function. Finally, there are 2 connected hidden layers (60, 10). The final output layer is (X_{pixel}, Y_{pixel}) as shown in Fig. 6.

- WiFi positioning based on sense recognition

The actual application in robots is found that the WiFi signal has several problems which include such as weak signal strength due to distance from AP, and base station is blocked by buildings and the signal is easily obscured, and interference with the WiFi signal causes the value to fluctuate significantly. These factors lead to WiFi positioning. Therefore, this paper hopes to improve the shortcomings of WiFi positioning through scene recognition shown in Fig. 7, and WiFi triangular-positioning is used in the robot walking, and perform CNN scene positioning to correct the WiFi triangular-positioning error after every 60 seconds.

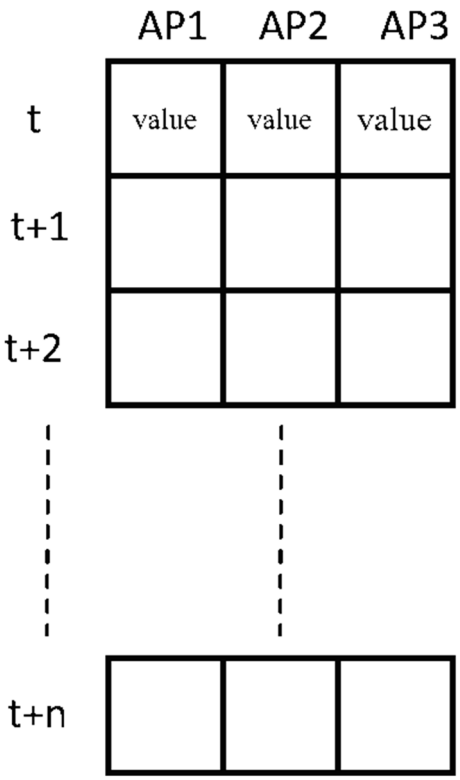

Fig. 5 The input of three AP WiFi signal values arrangement for 1D-CNN

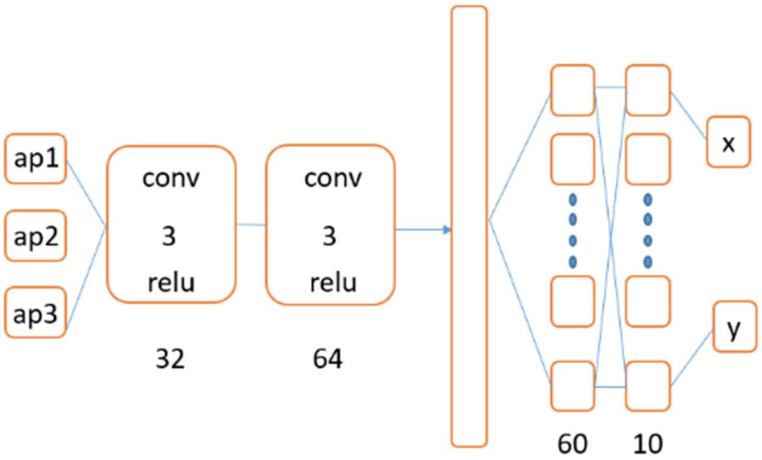

Fig. 6 Convolutional neural network architecture for WiFi triangle indoor positioning

- **Robot path planning**

The path planning in this study uses the AStar algorithm [12]. Through the establishment of a cost function, g(n), in Eq. 9, the cost is from the starting point to node n. According to the alpha value to determine the difficulty of Astar's work. An alpha=0, only judge whether the node can pass. At alpha=1, Astar will carry out standard path planning. The alpha value can be from 0 to 1. This research uses alpha=0.9 to update $g(n)$ to $g^{'}$(n). Then through the cost evaluation function as heuristic estimated cost, h(n), in Eqs. (10–12), get the node n to the destination. The h(n) is as also octile distance [16]. The Eqs. 10 and 11, the parameter $n.x$ and $n.y$ is a robot x-y position in time n. The $goal.x$ and $goal.y$ is a destination x-y position. There are many types of AStar's cost evaluation functions. $h_{diagonal}$(n) is the number of steps that can be moved along the diagonal line, and $h_{straight}$(n) is the Manhattan distance. AStar obtains the best path by Eq. 13 to confirm that f(n) is the minimum value, and the path keeps getting closer to the smaller f(n) as shown in Fig. 8.

$$g'(\mathrm{n}) = 1 + alpha * (g(n) - 1) \tag{9}$$

$$\mathrm{h}_{\mathrm{diagonal}}(\mathrm{n}) = \min\Big(abs(n.x - goal.x), abs(n.y - goal.y) \tag{10}$$

$$\mathrm{h}_{\mathrm{straight}}(\mathrm{n}) = \Big(abs(n.x - goal.x) + abs(n.y - goal.y) \tag{11}$$

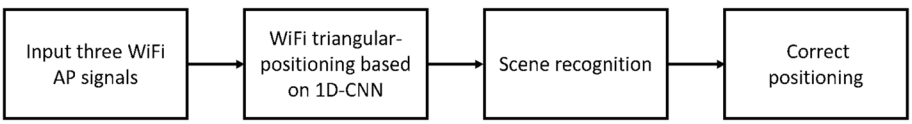

Fig. 7 The proposed WiFi positioning based on sense recognition method

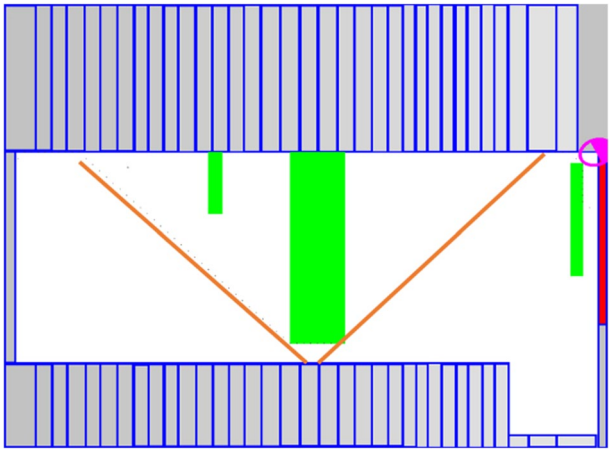

Fig. 8 The result of path planning (Green color is an obstacle. The orange color represents the planned path. The red color is the destination. The purple color is the robot)

$$h(n) = 1.414 * h_{diagonal(n)} + h_{straight}(n) \tag{12}$$

$$f(n) = g(n) + h(n) \tag{13}$$

4 Experiment

This paper uses ADE20K indoor semantic segmentation data set [1]. It has a total of 150 categories with 638 training data sets, 400 verification data sets, 350 testing data sets. The input image size is 224*224 pixels. The training parameters are as following

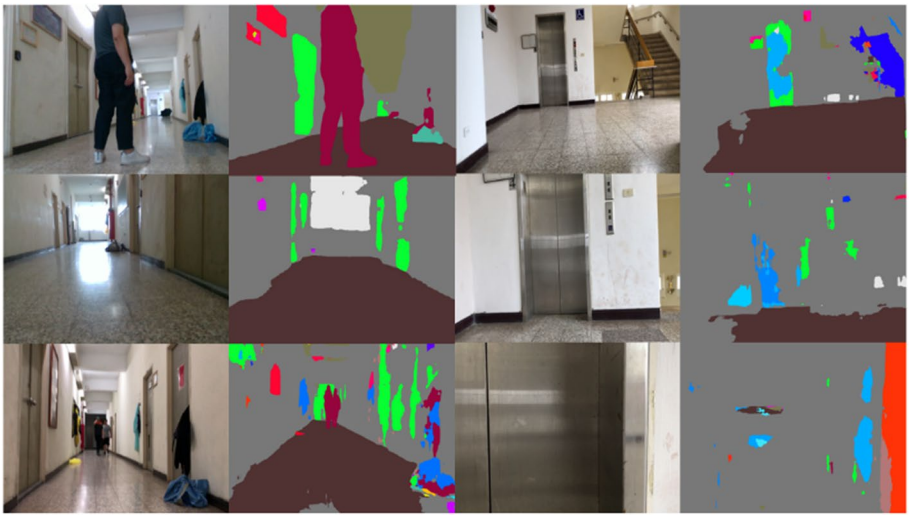

Fig. 9 The semantic segmentation of PSPNet result

Table 1 Training parameters

Mean IoU	42.04
Pixel Accuracy (%)	80
Inference Speed (GTX1050)	1/0.6(s)
Training Time(hour)(GTX1080ti)	25
Max Epoch	2000
Optimizer	Adam
Training learning rate	0.001

Table 1, and it can be seen that the network does present the road surface in brown. Although some pixels predictions have some errors, the visually impaired can still be used in a simple and organized environment (Fig. 9). Figures 10, 11 and 12 show the semantic segmentation results of the robot in continuous obstacle avoidance in an indoor environment. It can be known that fully convolutional networks (FCN) and pyramid scene parsing Network (PSPNet) have a certain degree of accuracy for semantic segmentation. In [33], FCN is used as the baseline for judging semantic segmentation. The semantic segmentation method adopts PSPNet. Therefore, this research discusses the differences as shown in Fig. 13. The left side is the resulting map of FCN. The right side is the resulting map of PSPNet. The dotted line is the error place. Because the FCN network architecture ignores the detailed information of the original image, it results in an unclear composition structure. For robot control, the road information will be wrong causing a risk for the visually impaired. In PSPNet, the pre-feature extraction and the post-feature extraction are concatenated, so this problem can be effectively reduced.

- **Scene recognition indoor positioning based on deep learning experiment result**

This paper uses proposed CNN architecture for scene recognition. The fourth floor of school building is used as the research data. Inputting a 64*64 pixels image is as

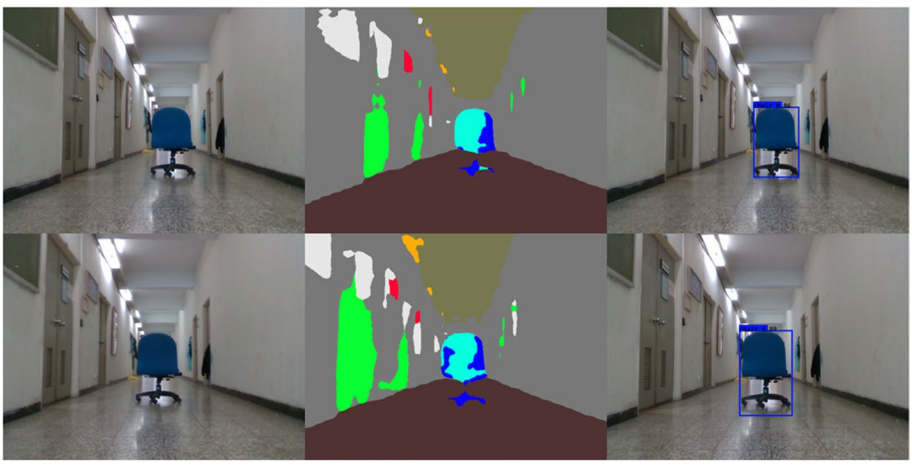

Fig. 10 The real avoiding obstacle sequential images result (left: original image, center: semantic segmentation, right: YOLO)

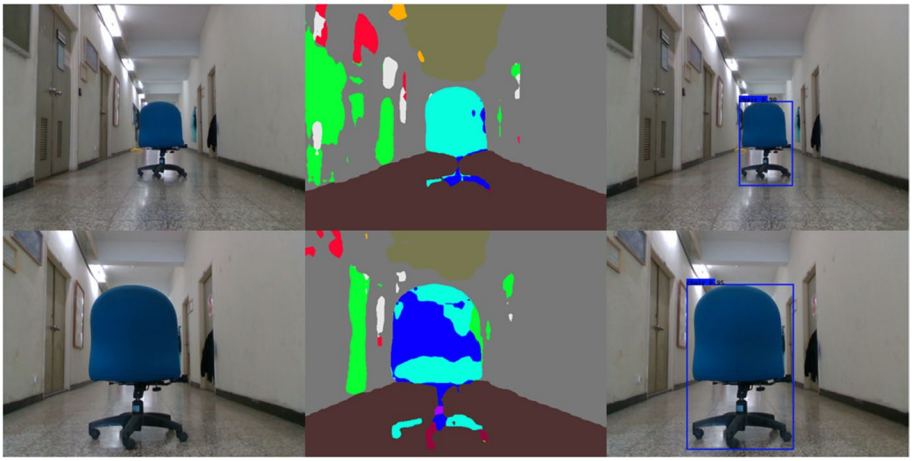

Fig. 11 The real avoiding obstacle sequential images result (left: original image, center: semantic segmentation, right: YOLO)

indoor scene. The training data set collects a total of 39,176 pieces of indoor data. The verification data set is 19,588 in total. The total number of testing data sets is 9000. The training dataset is divided into 35 categories and each category is about 1118. The training parameters of this paper are shown in Table 2. The confusion matrix of testing data result is shown in Table 3. Scene recognition is easily affected by the camera angle and light. A large amount of data can be collected to compensate for this problem. The network loss function is cross-entropy, and the total epoch is 100 times. The training accuracy is 99% and the test accuracy is 94%. The training accurate-epoch relationship is shown in Fig. 14. Figure 15 shows a random image from each type of testing

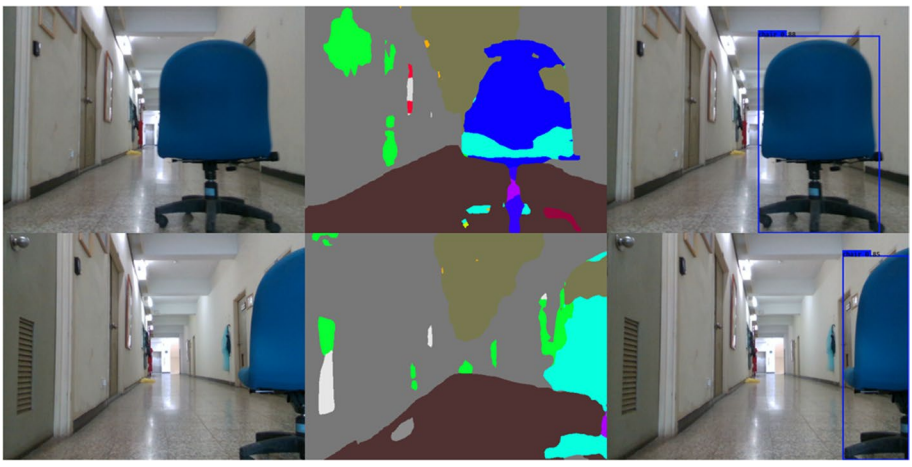

Fig. 12 The real avoiding obstacle sequential images result (left: original image, center: semantic segmentation, right: YOLO)

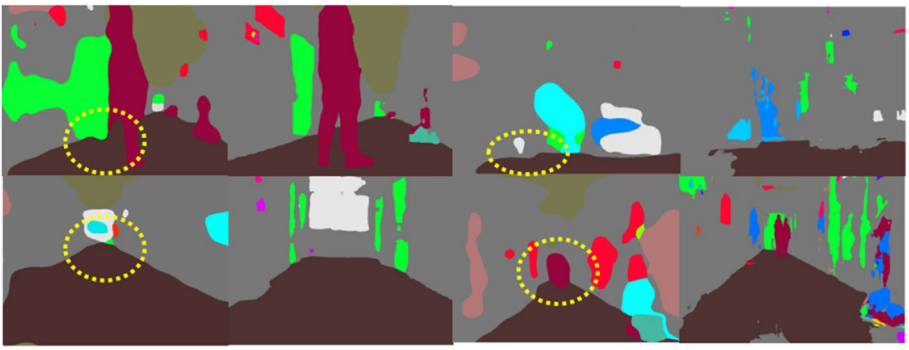

Fig. 13 The semantic segmentation comparison with FCN and PSPNet

data to view the classification results and there are 3 classification errors. Figure 16 shows the positioning of the scene in a 2D image map with the depth camera after scene recognition.

- **Sense recognition combined with WiFi triangle positioning method experiment result**

The sense recognition is used for indoor positioning to fine-tune triangle WiFi positioning. For triangle WiFi indoor positioning, WLAN and AP sharing devices are used to collect WiFi signal strength. This paper uses a 1D-CNN for training and three WiFi AP signals are collected as 1D-CNN input and output is the robot X-Y position. The architecture of 1D-CNN is that the first hidden convolutional layer is 32 feature maps, and the second convolutional layer is 64 feature maps, and finally two fully connected-layers are connected by 60 neurons and 10 neurons. The training parameters are shown in Table 4 and the loss function is mean square error, and the output is a 2D image view. There are a total of 169 training data sets, 160 verification data sets, and 110 testing data sets. Figure 17 shows the relationship between loss and accuracy. Figures 18 and 19 show the corresponding prediction results of the X and Y coordinates of the verification data. From top to bottom, curves are such as training accuracy, verification accuracy, training loss, and verification loss separately.

Table 2 Training parameters and Training/Testing result of the proposed method

Max epoch	100
Loss function	Cross-entropy
Optimizer	Adam
Activation function	Relu
Training learning rate	0.001
Training loss	$1.235X10^{-3}$
Testing Accuracy (%)	99%
Validation Accuracy (%)	97%
Testing loss	$1.953\ X10^{-2}$
Test Accuracy (%)	94%

Table 3 The proposed method is presented by the confusion matrix of testing data result in 35 category

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
1	242			3																															
2		2442				3				1		1			1						1							1							1
3	1		244		0					1		1		1	1						1		1					1				1		1
4	1			242	0			1		1		1		1	1				1		1		1					1			1	1		1
5	1				240			1		1				1		1			1				1					1			1	1		
6	1					242		1		1		1			1	1			1	1				1				1			1		1	1
7	1	1	1	1			242			1		1		1	1	1				1				1				1					1	1
8	1	1		1				242		1		1		1	1					1			1	1			1	1				1	1	1
9		1		1					242					1	1						1		1				1				1	1		
10	1			1	1			1		246			1		1		1			1	1		1	1			1	1		1	1	1		1
11	1			1	1			1			242		1		1		1		1	1	1	1		1			1			1	1	1		1
12	1	1		1	1			1				257	1		1	1	1		1	1		1		1				1	1	1	1			1
13	1	1		1			1		1				242		1	1	1	1	1			1		1			1		1	1	1	1		1	
14	1	1		1	1	1	1		1					242	1	1	1	1				1		1			1		1	1	1	1		1	
15	1			1	1	1	1		1	1		1			238		1	1		1					1	1			1	1		1	1	
16	1		1		1	1				1		1		1		242		1	1	1	1	1		1	1	1					1		1	
17			1		1			1		1		1	1	1	1		240	1	1	1	1	1			1	1			1		1		1	
18			1		1	3		1		1			1	1			1	244	1	1	1				1	1	1		1	1	1		1	1	
19	1	1			1		1	1	1	1	1		1		1		1	1	242	1		1				1	1		1	1			1	1	
20		1	1				1		1		1			1	1	1	1	1	1	236		1		1	1	1	1			1		1	1	1	1
21	1		1				1	1	1		1			1	1	1	1	1	1	1	242	1		1	1										1
22			1					1						1		1	1		1	1		242		1	1							1			1
23					1		1	1			1		1					1	1		1	1	248				1			1				1	
24					1	1	1				1		1					1	1		1	1		242			1							1	
25					1	1	1		1		1	1	1					1	1	1	1	1			242		1		1	1				1	
26						1			1		1	1	1		1					1	1					242			1	1	1				
27				1					1		1	1	1		1		1	1			1	1			1		242		1	1	1		1		
28		1	1	1					1		1		1		1	1	1	1			1	1			1	1	1	242		1	1		1	1	
29		1	1	1					1		1			1		1	1	1		1		1		1	1	1	1		242	1			1	1	
30		1	1			1			1		1	1								1				1		1	1	1	1	238				1	1

Table 3 (continued)

31					1	1	1			1	1	1	1		1				1				1				1	1		242				1
32	1				1	1	1	1		1		1	1		1	1	1							1	1		1	1	1		246			1
33	1	1		1		1	1	1		1	1				1	1	1		1					1	1		1	1	1	1		242		
34	1	1		1				1		1					1	1	1		1					1	1		1	1	1	1			242	
35		1		1		3						1							1					3	3		1	1	1	1				242

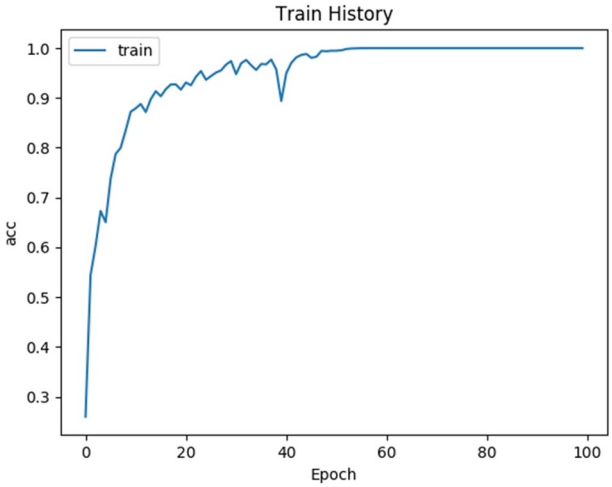

Fig. 14 The training accuracy for 100 learning epochs

Comparison of the proposed method and traditional WiFi triangular-positioning as shown in Fig. 20 and Table 5, is a straight-line comparison between the proposed scene recognition method (purple) and the WiFi positioning method (blue). The robot collects WiFi signals for positioning when it stops to perform scene positioning. Both output (X_{pixel}, Y_{pixel}) on the 2D plane, where Y is the camera depth value, and the X is the center pixel coordinate to identify the scene. The drift of the WiFi causes some errors.

Table 5 shows the numerical comparison of WiFi and the proposed positioning method on the 2D floor plan. From the experimental data, it can be found that WiFi has a large error in the prediction coordinates. The average error of (X_{pixel}, Y_{pixel}) in these 20 data is (19.45, 8.65), and the proposed positioning method is (9.25, 3.65).

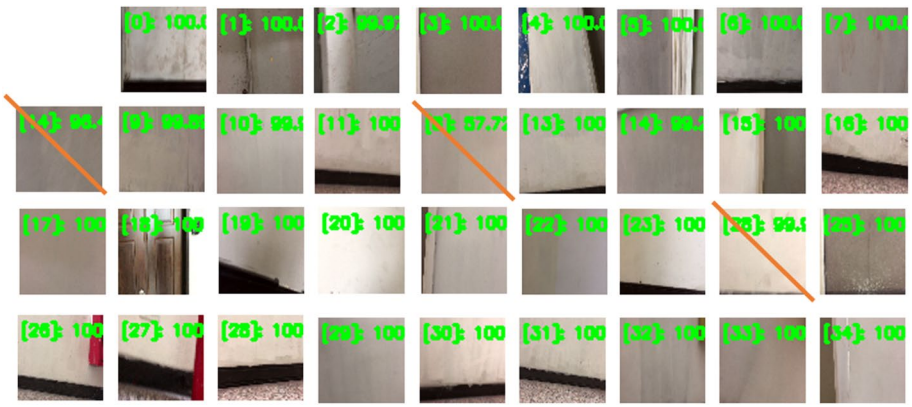

Fig. 15 The proposed image-based sense classification result of testing data

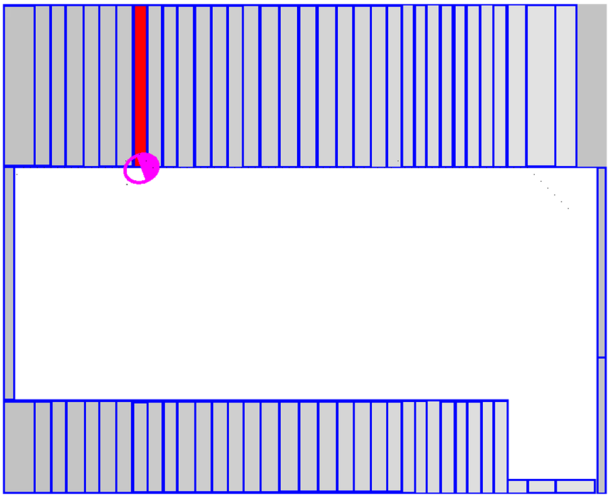

Fig. 16 The indoor positioning location of the 2D image

Table 4 The WiFi positioning of training parameters and training/testing results

Training loss (RMSE)	0.253
Loss function	Root mean square error
Max epoch	10,000
1D-CNN architecture	3*32*64*60*10*2
Training learning rate	0.001
Optimizer	Adam
Activation function	Relu
Accuracy (%)	95
Validation Accuracy (%)	85
Testing loss (RMSE)	0.298
Test Accuracy (%)	84

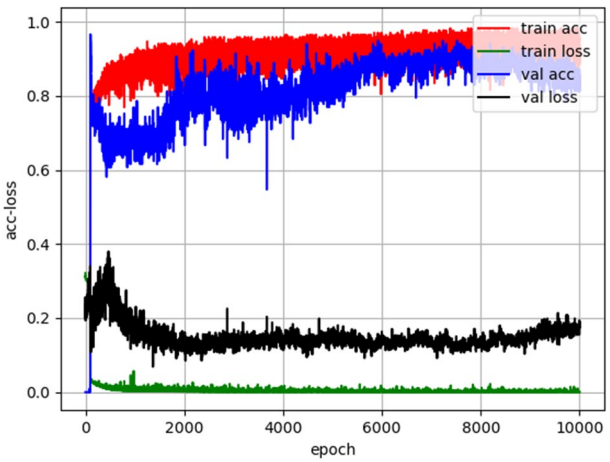

Fig. 17 Loss result (Top-bottom training accuracy, validation accuracy, training loss and validation loss)

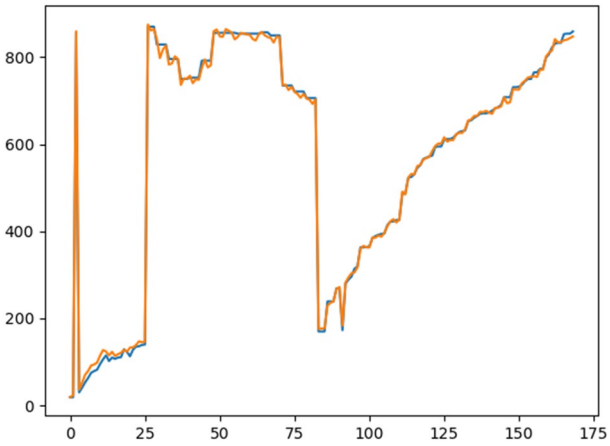

Fig. 18 The indoor positioning result of coordinate X output

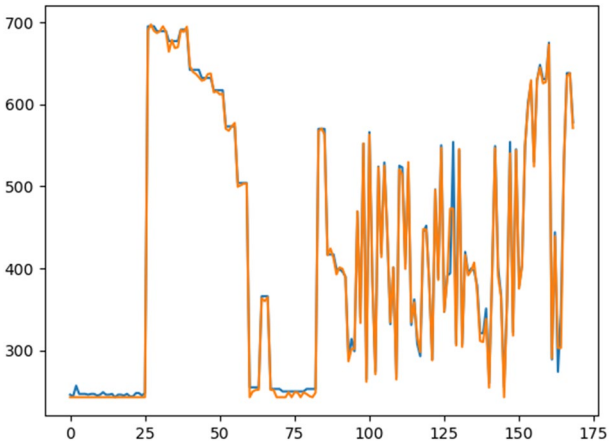

Fig. 19 The indoor positioning result of coordinate Y output

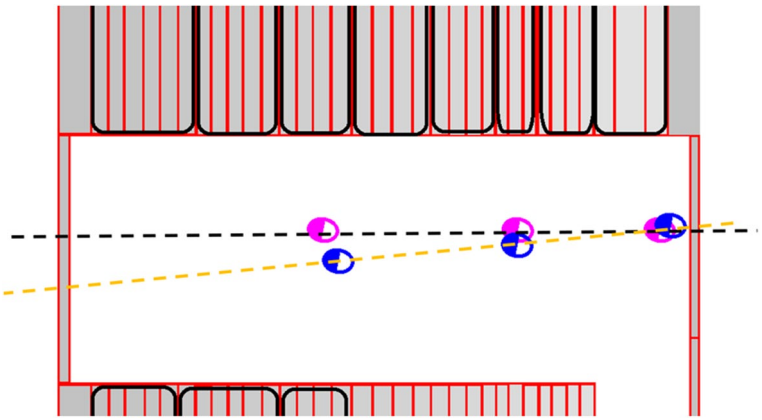

Fig. 20 The proposed positioning method and traditional triangle WiFi positioning comparison result shown in 2D coordinate graph

Table 5 The proposed positioning method and traditional triangle WiFi positioning comparison results

	Traditional WiFi triangular-positioning	The proposed positioning method	Ground truth
1	(845,596)	(870,610)	(856,612)
2	(755,628)	(755,650)	(759,646)
3	(838,321)	(831,325)	(848,323)
4	(763,264)	(755,265)	(768,263)
5	(685,253)	(690,258)	(684,251)
6	(625,395)	(615,400)	(620,401)
7	(578,400)	(570,409)	(570,406)
8	(471,410)	(450,420)	(455,416)
9	(379,394)	(366,400)	(357,399)
10	(265,381)	(225,388)	(235,386)
11	(158,338)	(133,340)	(120,340)
12	(71,259)	(35,260)	(27,257)
13	(78,407)	(35,410)	(35,413)
14	(73,557)	(35,571)	(29,571)
15	(155,545)	(105,562)	(117,559)
16	(269,541)	(247,541)	(239,554)
17	(381,541)	(366,565)	(359,569)
18	(737,536)	(755,555)	(740,549)
19	(798,619)	(831,640)	(805,637)
20	(790,413)	(790,420)	(797,419)
error	(19.45,8.65)	(9.25,3.65)	0

5 Conclusion

The lack of global positioning system assistance for indoor positioning makes indoor positioning a challenging task. In this paper, the proposed CNN based sense recognition method is used for indoor positioning. The robot completely relies on the camera to complete the positioning task. The CNN scene positioning method in this paper is regular which makes path planning more efficient. The WiFi signal is easily subject to physical limitations such as the problem of WiFi signal being blocked by buildings, difficulty to receive the signal and the signal drift, etc. It is hoped that this study can improve the indoor positioning by deep learning method. When the system detects the road that is feasible to walk in 0.6 seconds using the semantic segmentation method, and the farthest judgment distance is close to 10 m. This is a task that a white cane cannot complete. The identification of obstacles and their distances are calculated using YOLOv3 with depth camera. The development of the robot greatly compensated for the shortcomings of wearable devices and greatly solved the problem of the computing capacity of wearable devices. The robots equipped with computing equipment can inform the visually impaired in time. However, the wearable device needs to transmit the data to the server for calculation through the network. These wearable devices cannot function when the network is not ideal.

From the experimental results, it is clear that the scene recognition accuracy is 90% and it has good performance in road judgment and obstacle detection. This blind guide robot starts to work when it receives the destination instruction. Firstly, it performs indoor positioning and path planning. When an obstacle is encountered and the obstacle avoidance is completed, indoor positioning and path planning are performed and the visually impaired will be lead to the destination. In the process, it will continue to determine whether the road is safe.

From the experiment results, the designed robot can definitely assist the blind walking in indoor environment, but there are some issues that will be overcome in the future. It is hoped to improve the semantic segmentation method because the indoor environment road is always affected by the light resulting in road reflection. Another, the sense indoor positioning combined with the simultaneous localization and mapping (SLAM) will be considered to establish the indoor environment effetely and quickly. Besides, multipath interference affecting WiFi signal noise must be filtered to get position more preciously to reduce the computation error. For better user experience, the system can use the wearable device or smart phone that will help the blind to interact with the designed robot and more easily do indoor navigation through the smart device/phone speech recognition.

Acknowledgments This paper was partly supported by Ministry of Science and Technology, Taiwan, under MOST 110-2221-E-019 -051 -, 109-2221-E-019 -057 -, 110-2634-F-019 -001 – and 110-2634-F-008 -005 -.

Data availability The datasets generated during and/or analyzed during the current study are available in the COCO and ADE20K repository, http://cocodataset.org/#home and https://groups.csail.mit.edu/vision/datasets/ADE20K/, separately.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. ADE20K (n.d.) https://groups.csail.mit.edu/vision/datasets/ADE20K/. Accessed 28 Jun 2019
2. Badrinarayanan V, Kendall A, Cipolla R (2017) SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 39(12):2481–2495
3. Bharathi S, Ramesh A, Vivek S (2012) Effective navigation for visually impaired by wearable obstacle avoidance system. In: 2012 International Conference on Computing, Electronics and Electrical Technologies (ICCEET), pp 956–958. https://doi.org/10.1109/ICCEET.2012.6203916
4. Bourbakis N, Kavraki D (2005) A 2D vibration array for sensing dynamic changes and 3D space for blinds' navigation. In: Fifth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'05), pp 222–226. https://doi.org/10.1109/BIBE.2005.1
5. Chen L, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2018) DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848
6. COCO (n.d.) http://cocodataset.org/#home. Accessed 28 Jun 2019
7. Costilla-Reyes O, Namuduri K (2014) Dynamic Wi-Fi fingerprinting indoor positioning system. In: 2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN), pp 271–280. https://doi.org/10.1109/IPIN.2014.7275493
8. Dakopoulos D, Boddhu SK, Bourbakis N (2007) A 2D Vibration array as an assistive device for visually impaired. In: 2007 IEEE 7th International Symposium on BioInformatics and BioEngineering, pp 930–937. https://doi.org/10.1109/BIBE.2007.4375670
9. Du Y, Czarnecki WM, Jayakumar SM, Farajtabar M, Pascanu R, Lakshminarayanan B (2020) Adapting auxiliary losses using gradient similarity. https://doi.org/10.48550/arXiv.1812.02224
10. El Lahib M, Tekli J, Issa YB (2018) Evaluating Fitts' law on vibrating touch-screen to improve visual data accessibility for blind users. Int J Human-Comput Stud 112:16–27, ISSN 1071-5819,. https://doi.org/10.1016/j.ijhcs.2018.01.005
11. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp 580–587. https://doi.org/10.1109/CVPR.2014.81
12. Hart P, Nilsson NJ, Raphael B (1968) A formal basis for the heuristic determination of minimum cost paths. IEEE Trans Syst Sci Cybern 4:100–107

13. Hayat S, Kun S, Tengtao Z, Yu Y, Tu T, Du Y (2018) "A deep learning framework using convolutional neural network for multi-class object recognition," 2018 IEEE 3rd international conference on image, vision and computing (ICIVC), Chongqing, pp. 194–198
14. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770–778. https://doi.org/10.1109/CVPR.2016.90
15. Hsieh YZ, Lin SS, Xu FX (2020) Development of a wearable guide device based on convolutional neural network for blind or visually impaired persons. Multimed Tools Appl 79:29473–29491. https://doi.org/10.1007/s11042-020-09464-7
16. Kumar N, Vámossy Z, Szabó-Resch ZM (2016) Heuristic approaches in robot navigation. In: 2016 IEEE 20th Jubilee International Conference on Intelligent Engineering Systems (INES), pp 219–222. https://doi.org/10.1109/INES.2016.7555123
17. Lee N, Han D (2017) "Magnetic indoor positioning system using deep neural network," 2017 international conference on indoor positioning and indoor navigation (IPIN), Sapporo, pp. 1–8
18. Lin G, Milan A, Shen C, Reid I (2017) RefineNet: multi-path refinement networks for high-resolution semantic segmentation. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), Honolulu, HI, pp 5168–5177
19. Naga Srinivasu P, Balas VE (2021) Self-learning network-based segmentation for real-time brain M.R. images through HARIS. PeerJ Comput Sci 2(7):e654. https://doi.org/10.7717/peerj-cs.654 PMID: 34435099; PMCID: PMC8356652
20. Redmon J, Farhadi A (2017) YOLO9000: Better, Faster, Stronger. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), Honolulu, HI, pp 6517–6525
21. Redmon J, Farhadi A (2018) "Yolov3: An incremental improvement", CoRR, vol. abs/1804.02767
22. Redmon J, Divvala S, Girshick R, Farhadi A (2016) "You only look once: unified, real-time object detection," 2016 IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, pp. 779–788
23. Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149
24. Santos ADPD, Suzuki AHG, Medola FO, Vaezipour A (2021) A systematic review of wearable devices for orientation and mobility of adults with visual impairment and blindness. IEEE Access 9:162306–162324. https://doi.org/10.1109/ACCESS.2021.3132887
25. Shelhamer E, Long J, Darrell T (2017) Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 39(4):640–651
26. Sokic E,Ferizbegovic M,Zubaca J,Softic K, Ahic-Djokic M (2015) "Design of Ultrasound-based Sensory System for environment inspection robots", International symposium ELMAR (ELMAR), Zadar,Croatia ,28–30 September
27. Sonali KK, Dharmesh HS, Nishant MR (2010) Obstacle avoidance for a mobile exploration robot using a single ultrasonic range sensor. INTERACT-2010, pp 8–11. https://doi.org/10.1109/INTERACT.2010.5706156
28. Srinivasu PN, Bhoi AK, Jhaveri RH et al (2021) Probabilistic deep Q network for real-time path planning in censorious robotic procedures using force sensors. J Real-Time Image Proc 18:1773–1785. https://doi.org/10.1007/s11554-021-01122-x
29. Swaminathan R, Nischt M, Kuhnel C (2008) Localization based object recognition for smart home environments. In: 2008 IEEE International Conference on Multimedia and Expo, pp 921–924. https://doi.org/10.1109/ICME.2008.4607586
30. Tekli J, Issa YB, Chbeir R (2018) Evaluating touch-screen vibration modality for blind users to access simple shapes and graphics. Int J Human-Comput Stud 110:115–133, ISSN 1071-5819,. https://doi.org/10.1016/j.ijhcs.2017.10.009
31. Wang P et al. (2018) "Understanding convolution for semantic segmentation," 2018 IEEE winter conference on applications of computer vision (WACV), Lake Tahoe, NV, pp. 1451–1460
32. Xu Y, Wang Y, Ma L (2010) "A Novel WLAN Indoor Positioning Algorithm Based on Positioning Characteristics Extraction," 2010 Fourth international conference on genetic and evolutionary computing, Shenzhen, pp. 134–137
33. Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), Honolulu, HI, pp 6230–6239