Multimedia Tools and Applications (2020) 79:29473–29491
https://doi.org/10.1007/s11042-020-09464-7

Development of a wearable guide device based on convolutional neural network for blind or visually impaired persons

Yi-Zeng Hsieh[1] · **Shih-Syun Lin**[1] · **Fu-Xiong Xu**[1]

Received: 23 September 2019 / Revised: 24 June 2020 / Accepted: 28 July 2020 /
Published online: 11 August 2020

Abstract
This study proposes a design for a wearable guide device for blind or visually impaired persons on the basis of video streaming and deep learning. This work mainly aims to provide supplementary assistance to white canes used by visually impaired persons and offer them increased freedom of movement and independence using the proposed wearable device. The considerable amount of environmental information provided by the device also ensures enhanced safety for its users. Computer vision in the proposed device uses an RGB camera instead of the RGBD camera commonly used in computer vision. Deep learning is applied to convert RGB images into depth images and calculate the plane for detecting indoor objects and safe walking routes. A convolutional neural network (CNN) is adopted, and its neural network structure, which is similar to that of the human brain, simulates a neural transmission mechanism similar to that triggered in human learning. Therefore, this system can learn a large number of feature routes and then generate a model from the learning result. The proposed system can help blind or visually impaired persons identify flat and safe walking routes.

Keywords Blind or visually impaired persons · Wearable device · Deep learning · Convolutional neural networks

1 Introduction

Persons with disabilities face a number of daily obstacles and challenges, including daily movement and communicating with other persons, which limit their freedom to interact

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

Yi-Zeng Hsieh
yzhsieh@mail.ntou.edu.tw

Fu-Xiong Xu
ab19696a@gmail.com

[1] National Taiwan Ocean University, Keelung, Taiwan

and engage with the world around them independently. Modern electronic technologies can be developed to address such challenges and improve the lives of individuals with physical and sensory disabilities. The use of appropriate auxiliary equipment can greatly impact an individual's continued quality of life and promote and maintain independent living. At present, advances in high-speed microprocessors lead to various assistive technology systems [3, 9, 11, 15, 20, 26, 32, 36, 37]. These various systems allow people with disabilities to use limited voluntary motions for communication, computer manipulation, and control of household appliances. Each system has its own considerations, applicability, and limitations. The white cane is a common travel aid for blind or visually impaired persons, but it only ensures that a small ground area in front of the person is clear or locate any nearby obstacle on the ground via contact with the cane. Although the white cane has many advantages, such as being lightweight, small when folded, and low cost, its main limitation is its limited sensing range (approximately 1–2 m). Some blind persons may get around with the help of a guide dog; however, only a few blind or visually impaired persons have access to guide dogs because fully trained guide dogs are costly, and the breeding fee for guide dogs is usually high. Therefore, easy-to-use and cost-efficient electronic travel aids are needed to help visually impaired persons by expanding their ability to perceive unfamiliar environments. The Toyota's robot "BLAID" is a wearable device for helping the blind to detect some signs such as traffic lights, toilets, etc., but it didn't detect the obstacle and it still relied on the guide dog to help them walk. However, the "BLAID" did not release the method on how to implement their device and it still described their device scenario. The "BLAID" also did not provide their device experiments and reference papers about technical reports. Our proposed method adopted the semantic segmentation based on the CNN model that can describe the environment information and environment depth information based on TX2 hardware. Our system would be achieved a real implementation system.

At present, the white cane is commonly used as a guide aid. Objects and surfaces, such as metal, plastic, wood, or guide bricks, can be recognized by the different sounds made when such objects are tapped with a cane. However, many problems exist with this guide method. For example, if the cane does not strike an object or surface, then the user will be unaware of its presence. In particular, white canes and guide dogs are not ideal for navigating unfamiliar environments.

Various types of guide robots have been developed to address these shortcomings [32, 36], but most of the early designs used sensing components to avoid obstacles [11]. However, such an approach cannot cope with unexpected situations. At present, machine learning technology is becoming increasingly mature, and high-performance GPU hardware architecture is progressing considerably with the ability to analyze images and run high-dimensional operations in real time. By combining the former and current machine learning application methods, a depth image is generated to allow distance measurements. The combination of these measuring components can improve obstacle avoidance and guidance of guide robots and cope with most emergencies. However, the infrared light of a camera is susceptible to external environmental disturbances due to its depth [9, 37]. In previous years, scholars have proposed the use of a single camera [20, 21] or the addition of a laser [1, 7] to calculate the image depth value and assist with the ranging process. However, the calculation results are not widely used in various scene images. Therefore, computer vision in this study uses a single RGB camera, designs an algorithm for predicting image depth values in RGB images, and establishes depth information through deep learning [22]. Deep learning is developed using a convolutional neural network (CNN)-based deep prediction algorithm, which converts RGB images into grayscale as the CNN input and predicts image depth

training using the depth value of each pixel as the desired value. The obtained depth image is applied to obtain a flat path and provide autonomous obstacle avoidance using CNN's radar ranging [5, 48]. Compared with the use of the white cane or a guide dog, the device can calculate and communicate a safer route for the wearer at a greater distance and in far more detail with higher robustness regarding unexpected obstacles in an unknown environment. Therefore, the proposed device is safer and more reliable and convenient.

The contribution of our proposed system is to use a low-cost device with an RGB camera to predict the obstacles and the walking plane for guiding the blind safely walking. The low-cost device is described as details in section III. The hardware of our system is built-in Nvidia Jason TX2 combined with the RGB camera. The TX2 is an application client to stream video image and the RTSP protocol is adopted to stream the image into the server. The deep learning algorithm is processed by the server, and then, after processing the resulting output returns to the TX2 client. Our software algorithm is illustrated by the five steps. The first step is to predict the environment depth by the four stages CNN method, but the predicted depth environment information is rough. Therefore, the second step is to fine-tune the depth environment information by the scale-invariant mean squared error. Then, the walking plane must be found to guide the blinds walking safely, and hence the floor and plane detection methods are adopted. Moreover, the blind must know the safe road distance to avoid users colliding with the obstacles. Finally, our proposed system will tell the blind the distance of the safe road through the earphones. The flowchart of our system is shown in Fig. 1.

The remainder of this paper is organized as follows. Section 2 briefly reviews the machine learning and scene segmentation. The proposed walking plane detection method is introduced in Section 3. The simulation results are discussed in Section 4. Finally, some concluding remarks are given in Section 5.

2 Related work

The advent of the era of deep learning has made depth imaging is a major research topic to date. Adopting learning mechanisms based on neuropsychology is the first step of machine learning, but machine learning algorithms require many iterations, many calculations, and are too dependent on hardware to produce training results. In the early days, the efficiency of machine learning development is high. However, the rapid development of computer hardware in recent years has remarkably improved the computing power of computers and considerably enhanced neural network technology. Machine learning has been promoted by many scholars, and the system architecture of computer vision has become increasingly sophisticated with the improvement of depth perception to further enhance recognition accuracy, which is the key technology in near-depth imaging. There are many machine learning applications as following. Xu et al. [46] proposed based on the human skeleton map for fall prediction. Xu et al. [45] proposed a novel multi-feature fusion (MFF) CNNs framework for the Drosophila embryo of interest detection. Xu et al. [44] proposed a novel edge-oriented

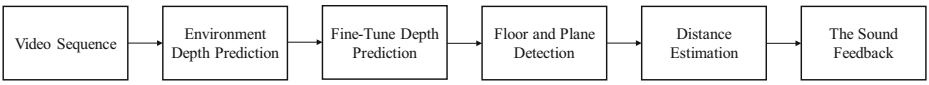


Fig. 1 The algorithm flowchart

framework to improve the performance of existing salient detection methods. Xu et al. [43] considered the cross-modal retrieval task, from the perspective of optimizing ranking model. Xu et al. [42] described the body's adjustment process from the physical point of view. Xu et al. [41] proposed a new algorithm of phase consistency detection based on dimensionality reduction. Xu and Li [39] proposed an algorithm exploiting the transaction data behind the social media stocks. Liu et al. [23] discussed an in-depth exploration for the most popular DCGAN. Liu et al. [24] discussed the relationship between image semantic segmentation and animal image research. Xu et al. [40] proposed a new recommended method using collaborative filtering. Xu [38] proposed a two-dimensional numerical model for machine learning to simulate major U.S. stock market index.

The accurate prediction of environmental depth information is important in predicting geometric relationships within an environment. Straub et al. [34] proposed real-time inference algorithm to evaluate the real environment. There are some blind dog systems developed to help the blind [2, 33, 47]. Knowing the geometric relationship of objects helps to provide rich object features and environments, such as in 3D modeling [13, 29, 30], physical and support models [17, 27], and robotics [6, 31]. Saxena et al. [30] used markov random field (MRF) to infer a set of plane parameters and used supervised learning for training to obtain the image depth values and segmented images after image color segmentation and build a model.

Saxena et al. [29] developed a method of depth prediction by using a single image. Supervised learning was also used to solve the predicted depth information requirement, and a single image (including unstructured outdoor environments, such as forests, trees, and buildings) was collected as the training dataset with a corresponding ground depth map as the desired output. The panoramic depth value information of the image was still needed because local features cannot predict the depth of a single point. In turn, MRF was used to combine multiscale local and global image features and simulate the depth of each point and the relationship between depths at different points through the local and global depth values for predicting depth information.

Silberman et al. [31] used color depth (RGBD) images to predict the major surfaces of an indoor scene and its object environment and related relations. In terms of depth of prediction, the depth of the scene was established via multiscale deep learning, whereby the area was shaded in the environment, and depth information in the complex scene was inferred through fine-tuning.

Liu et al. [22] put forward a model for depth prediction from a single image (RGB) that combined the strength of deep CNN and continuous CRF to predict the depth of new images efficiently. They also proposed a complete convolutional network and an improvement model of a novel superpixel pooling method, which accelerates the overall training speed by using deeper networks to obtain enhanced prediction performance.

In the field of image labeling, Long et al. [25] used a convolutional network as a powerful visual model and proposed a new fully convolutional network (FCN) for segmentation that combines feature structure and improves the output of spatial precision; the fully convolutional classifier can be fine-tuned for segmentation, as shown in Section 4.1. Although the score of the standard metric is high, the output is unsatisfactory.

PSPNet [49] is a network based on the FCN [25] architecture. The traditional FCN uses the input image as the CNN input to deconvolute the same feature as the input image [4, 35]. In the process, global information is not added to the network. This condition leads to a lack of global semantic information in the FCN, which results in errors. Therefore, Zhao et al. [49] suggested the addition of a global-scene-level to multiscale feature ensembling on the

basis of the FCN architecture; this addition allows the network to contain local and global information. Among them, PSPNet optimization and loss function are based on ResNet [12].

He et al. [12] observed that the continuous deepening of the network theory can yield better results; however, the experimental evidence does not necessarily indicate improvement. Deepening the network can be done easily but causes the gradient to disappear and results in decreased accuracy. Deep networks have a degradation problem, which makes them difficult to train. Accordingly, He et al. [12] proposed residual learning to solve such problems. The network structure of ResNet is based on the VGG19 network modification. The residual learning method mentioned above is added to the VGG19 network. When the feature map size is decreased, the number of feature maps will increase to maintain the network complexity and solve the problem of network degradation.

3 Proposed system

The wearable guide device with deep learning for blind or visually impaired persons developed in this study aims to provide environmental information to the wearer while using a white cane and allow them to easily and safely navigate unfamiliar environments. CNN is used in model development to make preliminary predictions on the depth information of RGB images, further reference and improve a multiscale deep network [29], and strengthen the prediction of environmental depth information. This environmental depth information is then used to predict a safe route for the wearer to walk in, complete a fast algorithm to determine flat routes and depth-marking areas on the basis of deep learning, inform the wearer the distance of the safe path, and ensure that every single step taken by the wearer is safe. The main research includes (1) body device design, (2) indoor depth prediction, (3) depth detail adjustment, and (4) plane detection and establishment, as shown in Fig. 2.

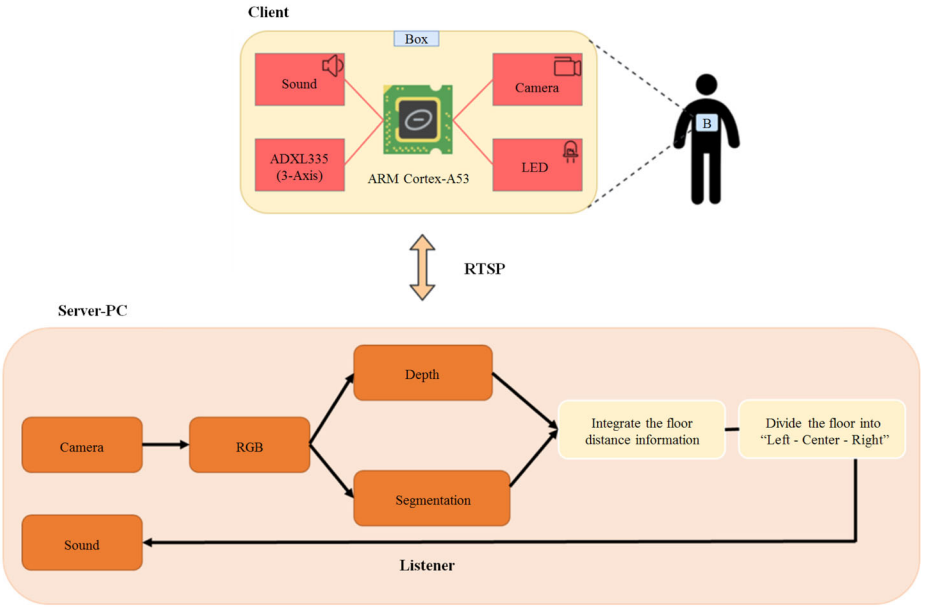


Fig. 2 Architecture of the proposed system

3.1 Hardware design

This study adopted the lightweight computing power of an Nvidia Jetson TX2 as the wearable operating core to reduce the size of the hardware device worn by the user. This system can run a complete operating system (Linux), save power, is lightweight, and has excellent control performance, which make it suitable as the central control server for video streaming in the proposed system. The streaming process uses real-time streaming protocol (RTSP) and the H.264 format to stream images and introduces the images into the trained model enabling the system to determine the range of planes in front of the user that can be walked on and guide the user to a safe path. The magnitude of the sound reminder is used to tell the user whether they should continue walking. The communication protocol is shown in Fig. 3.

3.2 Design of the wearable guide device

The proposed wearable device was 3D printed using a stable ABS to allow the user to hang the device near their chest and listen to the information provided with headphones. The lens in the device provides a video stream at a 20° horizontal downward position. The device is shown in Fig. 4.

The middle of the device has an adjustable elastic buckle belt, which can be used by blind or visually impaired persons to strap the device around their waist. Figure 5 shows how the device is worn.

3.3 Indoor depth prediction

This study refers to the CNN network described by Saxena et al. [29] with four stages, as shown in Fig. 6, wherein the input layer is a 304 × 228 RGB image; the first and second stages are a 9 × 9 convolutional filter, and the moving step filter is course 2 2 × 2 max-pooling; and the third and fourth stages are a 5 × 5 convolutional filter and a 5 × 5 convolution, respectively. The depth value is predicted by using the following equation:

$$\hat{d}_{i,j,k} = w_{lr}^{T} F_{i,j,k} + b_l, \tag{1}$$

where the depth value is $\hat{d}_{i,j,k}$.

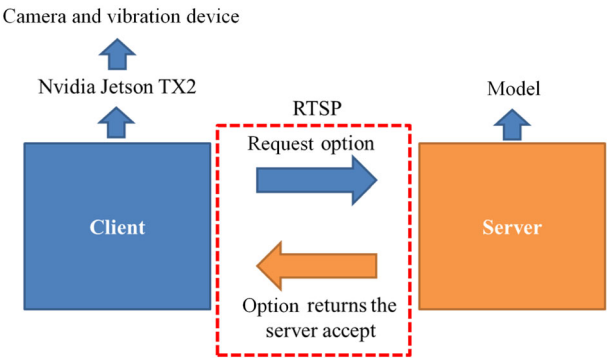


Fig. 3 Communication protocol of the proposed system

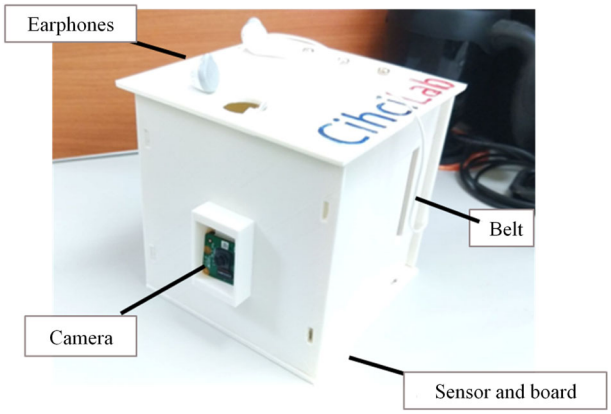


Fig. 4 Proposed device components

The feature vector $F_{i,j,k}$ and baseline (b_l) can be used to compute the pooling parameter H_l as follows:

$$F_{i,j,k} = f(I_{i,j,k}, \theta_f) = W_l H_{l-1}, \tag{2}$$

$$H_l = pool(nonl(W_l H_{l-1} + b_l)). \tag{3}$$

The loss function $L(\theta_f, \theta_{lr})$ is expressed as (4). Then, the stochastic gradient descent algorithm is adopted to compute the CNN weight as follows:

$$L(\theta_f, \theta_{lr}) = \frac{1}{N} \sum_{i,j,k} (d_{i,j,k} - \hat{d}_{i,j,k})^2. \tag{4}$$

Fig. 5 Wearing the device

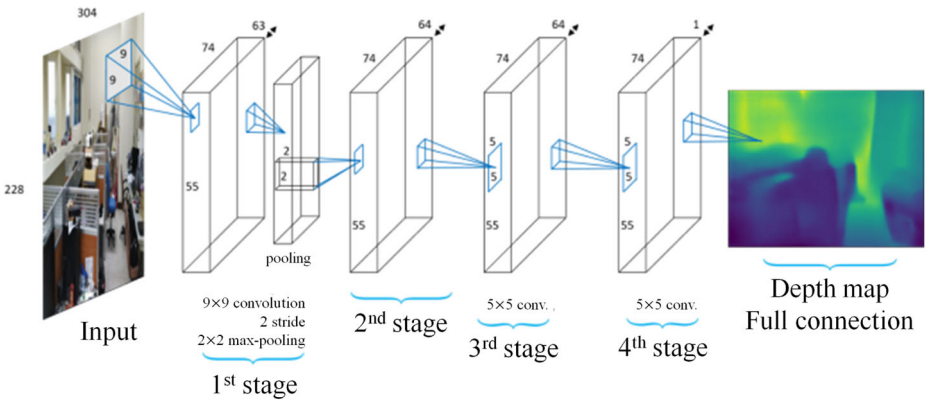


Fig. 6 Depth image generation

3.4 Depth detail adjustment

Given that this study uses only a single RGB camera to achieve computer vision [19], determining the means of adding depth information to the RGB images is the core topic of this article on indoor computer vision. For outdoor prediction, [28] adopted the multiscale deep network method, which is expected to enhance the efficiency of indoor depth information. This neural network is composed of the following subnetworks [8, 14, 16]: (1) coarse-scale network roughly predicts the depth of the panorama, and (2) the value is then input to the fine-scale network for local adjustment to achieve accurate predictions.

3.4.1 Coarse-scale network

This subnetwork has seven layers, including five convolutional and two fully connected layers. The first and second convolutional layers involve downsampling with max-pooling to reduce the feature map dimension and accelerate the operational efficiency and reduce overfitting. The sixth and seventh layers are fully connected and use upsampling to improve the output feature map dimension and achieve the final output image resolution and the 1/4 effect of the input image. Although the conversion between down- and upsampling will result in blurry predictions, the final output will have a better predictive effect than the direct output of the fifth layer, as shown in Fig. 7.

3.4.2 Fine-scale network

The main task of this subnetwork is to fine-tune the output of the coarse-scale network and improve the blurring generated during down- and upsampling. The fine-scale network is divided into four convolutional layers with an original input in the first layer. An input image is also downsampled in a max-pooling manner, and the second layer merges the output of the first layer and the coarse-scale network, as shown in Fig. 8.

The depth information pair is detected by the above-mentioned CNN architectures, and the relationship between the panorama and the point is estimated using a scale-invariant error. The scale-invariant mean squared error (in log space) is defined as

$$D(y, y^*) = \frac{1}{2n}\sum_{i=1}^{n}(\log y_i - \log y_i^* + \alpha(y, y^*))^2. \tag{5}$$

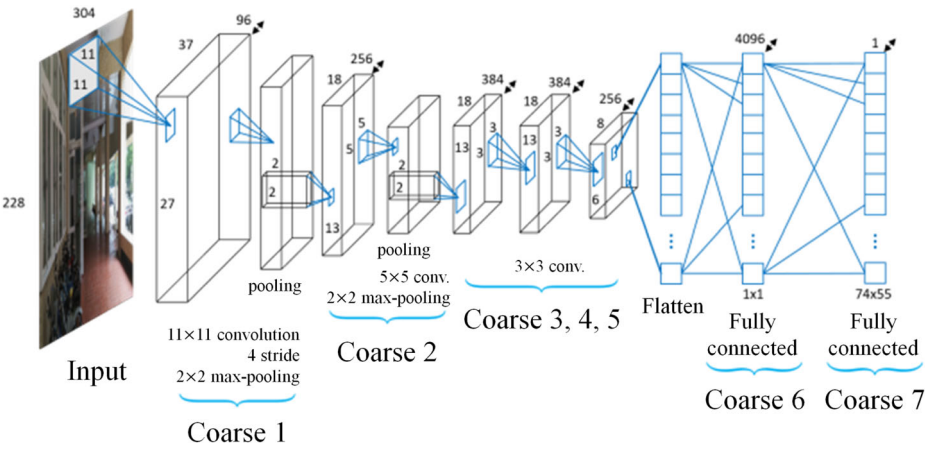


Fig. 7 Architecture of coarse-scale network

where $\alpha(y, y^*) = \frac{1}{n}\sum_i(\log y_i^* - \log y_i)$. For any predicted value y, e^{α} is the parameter that best corresponds to the actual distance. To ensure that the overall error size remains unchanged, the scalar multiplication for all y will have the same scale-invariant.

3.5 Plane detection and development

This study uses the MIT ADE20K scene parsing dataset as the indoor dataset. This dataset is used to separate the images into objects in the environment; distinguish the floors, walls,

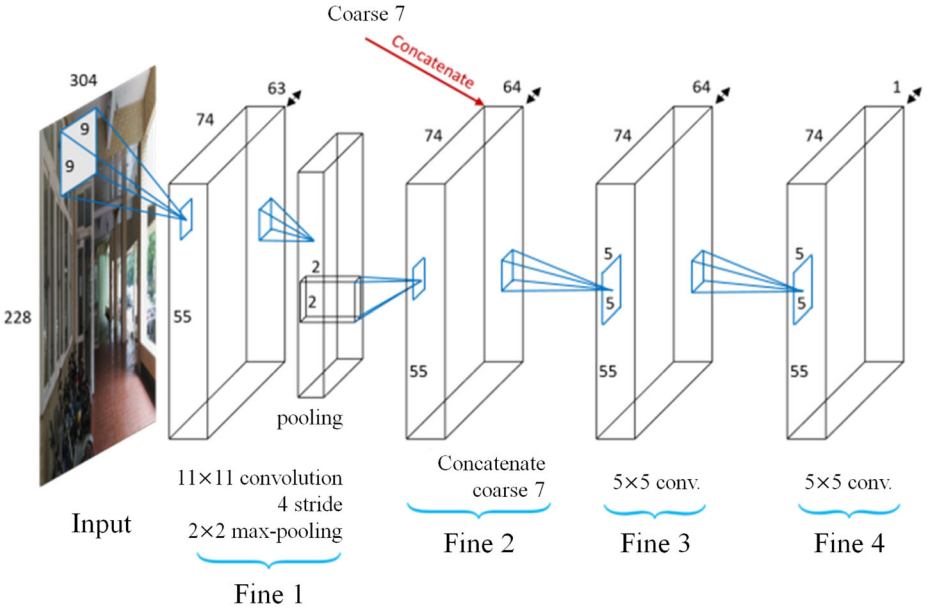


Fig. 8 Architecture of the fine-scale network

and obstacles through the color blocks; and apply the trained parameters. The picture taken by the device wearer while walking is shown in Fig. 9.

Scene segmentation is a basic task in computer vision, and its goal is to classify each pixel in the image, which is potentially used in areas such as autonomous driving and robot perception. The main advantages of the PSPNet [49] segmentation method are as follows.

(1) Based on FCN (Fully Convolutional Network) target segmentation framework, complex background features are embedded.
(2) Based on the deep supervision loss function, an effective optimization strategy is proposed for ResNet.
(3) A state-of-the-art scene parsing and semantic segmentation system have been established, and it contains many practical implementation strategies.
(4) One of the routes is multi-scale feature extraction because the higher-level features in the deeper network contain more semantic information. However, it contains less spatial location information.
(5) Another route is based on structural prediction, for example, by using CRF (Conditional Random Field) as a subsequent step to extract the segmentation results.

The pyramid pooling module proposed in [49], as shown in Fig. 9, is also utilized. In this network architecture, ResNet is adopted to extract features after inputting the images, and the feature map size is 1 for 1/8 input images. The feature map is divided into three blocks of 1×1, 2×2, and 3×3. This three-level pooling kernel fuses the small feature maps into the global information. Then, the original feature map (feature map) is connected with the output of the pyramid pooling module. Auxiliary loss is added to improve network training. The experiment in [49] obtained auxiliary and master branch loss values of 0.4 and 0.6, respectively. Figure 10 shows the auxiliary loss on ResNet101, wherein each brown block is the responsibility block, followed by auxiliary loss.

The depth of the network is crucial to the performance of the model. When the number of network layers is increased, the network can extract more complex feature patterns, so theoretically better results can be obtained when the model is deeper. The deep network has a degradation problem. When the network depth increases, the network accuracy becomes saturated, or even decreases such as that the 56-layer network is even worse than the 20-layer network. This will not be an overfitting problem. We know that deep networks have problems with gradients disappearing or exploding, which makes it difficult to train deep learning models. Therefore, the problem of the degradation of deep networks is very surprising.

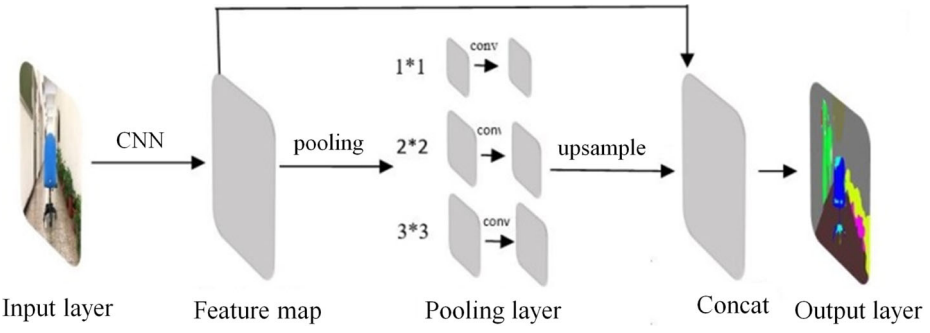


Fig. 9 Architecture of plane detection

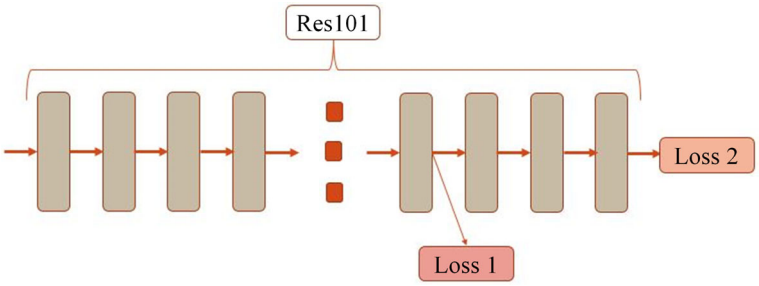


Fig. 10 Architecture of Res101

The degradation problem of deep networks at least shows that deep networks are not easy to train. We build a deep network by stacking up new layers. An extreme case is that these added layers do not learn anything, just copy the features of the shallow network, because the new layer is identity mapping. In this case, the deep network should be at least as effective as the shallow network, and there should be no degradation. The ResNet [12] proposed residual learning to solve the degradation problem. For a stacked layer structure, the learned feature is recorded as the input. ResNet hopes that it can learn the residual as the original learning feature. The residual learning is easier than direct feature learning. When the residual is 0, the accumulation layer only performs identity mapping, the network performance will not decrease. In fact, the residual will not be 0, which will also make the accumulation layer learning based on the input features.

The aforementioned model has obtained the color segmentation map. Thus, this study obtains the floor area and detects the boundary with the plane edge to detect and draw. The Canny algorithm is adopted to detect the edges on the basis of the upper and lower thresholds of the input between 2:1 and 3:1. A threshold is used to determine whether a pixel is an edge. The following criteria are used in detection:

(1) If the pixel gradient intensity is greater than the upper threshold, then the pixel is an edge.
(2) If the pixel gradient intensity is less than the lower threshold, then the pixel is not an edge.
(3) If the pixel gradient intensity is above and below the threshold and a pixel exists with a gradient intensity greater than the upper threshold, then the pixel is an edge; otherwise, the pixel is not an edge. The flow is shown in Fig. 11.

4 Experimental results and discussion

The performance of the proposed device is compared with those of other available methods and devices to provide users with real-time safe and reliable guidance. The main research

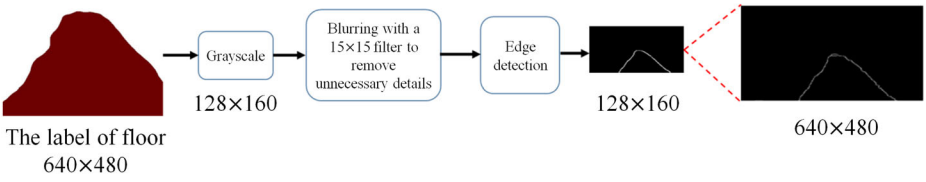


Fig. 11 Plane edge detection

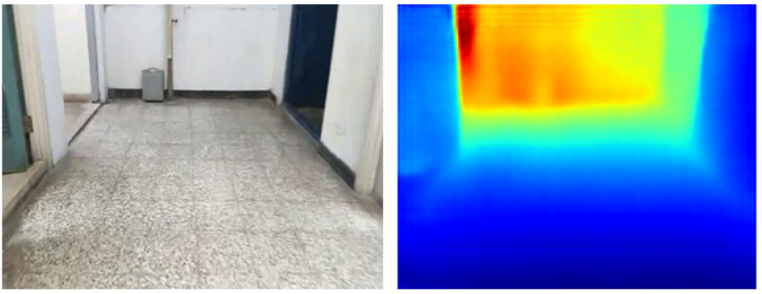


Fig. 12 Depth prediction image

results are divided into (1) depth image combined with label results, (2) comparison of distance results, (3) comparison of plane information and detection.

4.1 Depth image combined with label results

At the beginning of the experiment, the original RGB image was input into the coarse- and fine-scale network layers. The coarse-scale network layers were used to perform global depth prediction on the input image for outputting the prediction results of the coarse-scale network layers to the fine-scale network layers. The first-layer output of the fine-scale network layers was combined with the prediction results of the coarse-scale network layers and output to the second layer. A preliminary predicted depth map was subsequently derived. The fine-scale network layers were then used to fine-tune the prediction results of the coarse-scale network layers. The final output predicted image depth is shown in Fig. 12.

Figure 12 shows the scene depth map of an office and a corridor. This study used the color gradation method to differentiate between the near and distant scenes for ensuring that the system had a single RGB image to determine the far-reaching ability in the scene.

The same image was then labeled with the result of the depth prediction map, and the training model was trained using the MIT ADE20K scene parsing dataset. The image was then subjected to semantic segmentation, and the objects in the environment were segmented by color, as shown in Fig. 13.

Figure 14 shows the result of image labeling. By using various colors to distinguish between items, the flat plane was depicted in brown, and the edges were then used to

Fig. 13 Semantic segmentation image

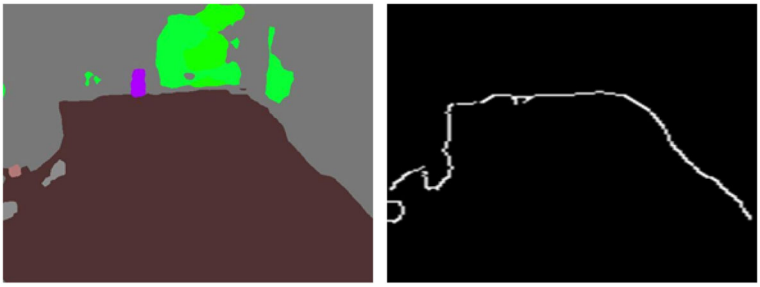


Fig. 14 Plane semantic image

describe the threshold of the brown area of the ground plate to remove unnecessary information. The result is shown in Fig. 14.

With the plane feature map, the label and depth information was used to combine two pixels and derive the combination of plane and depth. This ground depth value was used by the system to measure the user's path surface, length, and other information, as shown in Fig. 15.

This trained predictive model was then applied to the wearable device designed in this study. The wearers can hang the device on their chest, and the system will change the direction in which they can safely walk with the assistance of a white cane and increase mobility and safety for the wearer.

4.2 Comparison of distance results

The results of the proposed method were similar to those achieved by laser-based range finding in the scene with an accuracy of up to 98.52%. Compared with the accuracy of the ultrasonic sensor at 97.19%, the proposed method exhibited an improvement in the accuracy rate of 1.32%. Compared with Kinect's accuracy of 94.76%, the proposed method exhibited an accuracy rate improvement of 3.75%. Ultrasonic methods are only suitable for detection of less than 4 m in general applications, and obstacles more than 4 m away are difficult to detect. Kinect recommends measuring distances from 1.2 m to 3.6 m in official data. The detection range is 57°. According to [18], the distance error at 5 m is 7 cm, and the large distance will result in additional vanishing points because blind spots are likely to appear in plane detection. The proposed method has decreased sensitivity to distance

Fig. 15 Resulting image

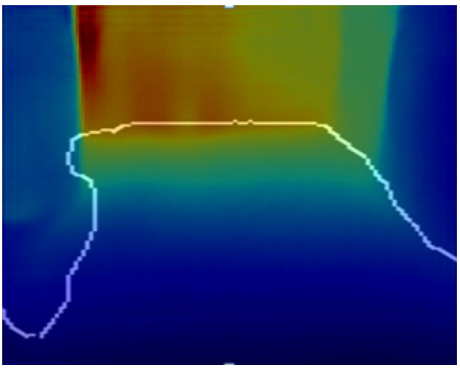


Table 1 Distance comparison

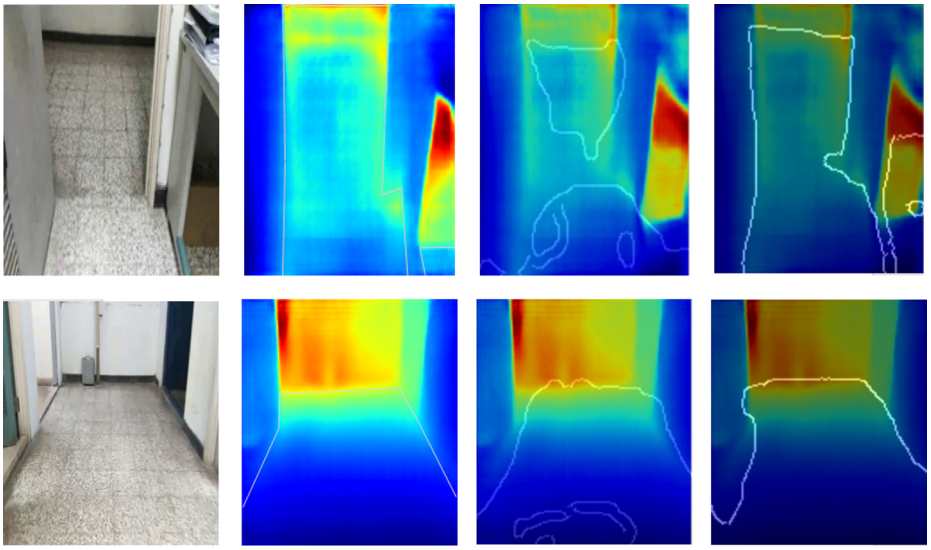
Scenes	A	B	C	D	E	F	G
YDLIDAR X4 (baseline)	1.924	1.721	3.012	3.990	2.732	4.603	4.236
HC-SR04	1.906	1.733	3.005	3.970	2.702	4.132	3.994
Kinect	1.894	1.691	2.913	3.780	2.620	4.102	3.820
Ours	2.06	1.725	3.014	4.000	2.721	4.615	4.211

restrictions. Table 1 presents the distance comparison of the target obstacles measured in different scenarios.

4.3 Flat information detection comparison

At present, common semantic segmentation models, such as FCN8s, do not have sufficient global information, and the local information causes errors when resegmentation is removed. However, this study used PSPNet with global-scene-level and introduced the ResNet optimization method to detect the walkable plane. The comparison results are shown in Fig. 16.

Given that the purpose of this experiment is to provide additional information about accessible planes for blind or visually impaired persons, ensuring that the range enclosed by the model overlaps with the area where the ground truth overlaps is important. This study divided the two selected areas into points, which were categorized into boxes. The overlap rate was calculated for 10 different scenarios. After many experiments, the FCN8s model caused too much ruggedness, too many turns, or too many obstacles. According to



(a) Image (b) Ground truth (c) FCN8s (d) PSPNet

Fig. 16 Comparison of floor label results. **a** Original map, **b** ground truth marked by hand, **c** result plot of the floor framed via FCN8s, and **d** result plot of the floor framed via PSPNet

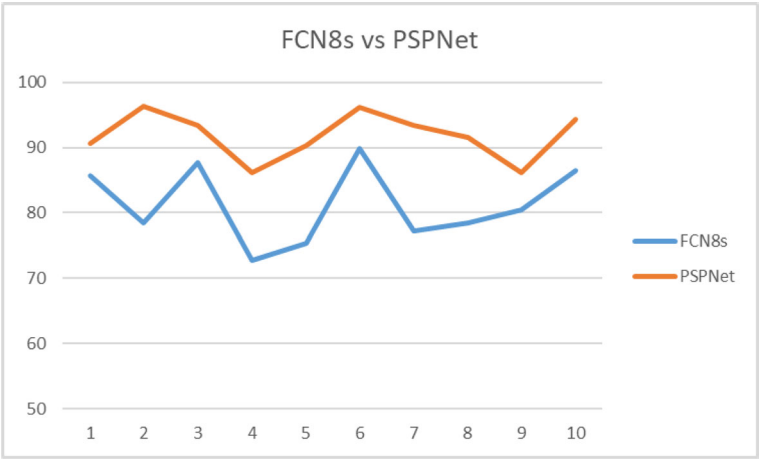


Fig. 17 Correct ratio of plane detection (FCN8s vs. PSPNet)

the statistical results shown in Fig. 17, PSPNet is better than FCN8s because the overlap rate of the average frame selection increased by 10.54%.

This study also lists 15 real experiment environments to ensure that the proposed method can identify the floor plane. From Fig. 18, the found walking plane is shown by the edge boundary. Less noise influence is observed in the different environments, and the method can also avoid obstacles, such as people, chairs, and stairs.

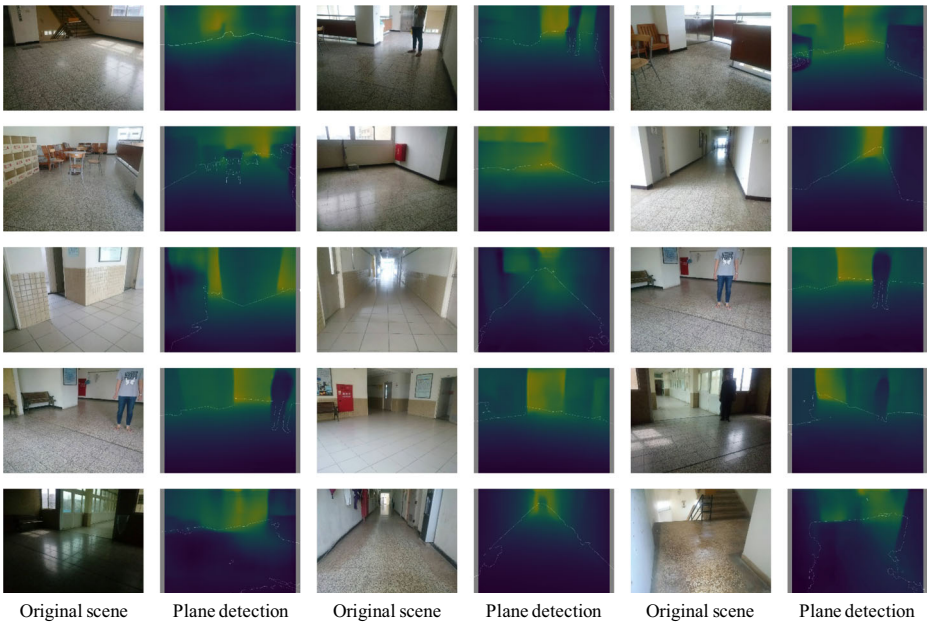


Fig. 18 Plane detection result

Table 2 Compared with 3D-KHT [10]

Case	Algorithms	Precision (%)	Recall (%)
1	Our method	96.32%	94.58%
	3D-KHT	68.51%	62.38%
2	Our method	97.51%	96.83%
	3D-KHT	67.22%	65.14%
3	Our method	98.25%	97.33%
	3D-KHT	60.82%	64.55%
4	Our method	97.38%	98.61%
	3D-KHT	61.85%	62.33%
5	Our method	96.37%	97.58%
	3D-KHT	61.53%	62.74%
6	Our method	96.64%	97.59%
	3D-KHT	62.56%	63.74%
7	Our method	97.68%	96.92%
	3D-KHT	62.57%	61.16%
8	Our method	96.33%	97.82%
	3D-KHT	63.43%	62.79%
9	Our method	98.24%	97.56%
	3D-KHT	62.73%	61.37%
10	Our method	97.77%	96.28%
	3D-KHT	62.85%	61.49%
11	Our method	98.45%	96.25%
	3D-KHT	62.11%	61.54%
12	Our method	97.23%	96.58%
	3D-KHT	63.55%	62.35%
13	Our method	96.74%	97.35%
	3D-KHT	62.61%	61.55%
14	Our method	97.39%	96.53%
	3D-KHT	63.51%	62.81%
15	Our method	97.43%	96.71%
	3D-KHT	61.40%	62.77%

Our proposed plane detection method performance was compared with the 3D-KHT [10]. We adopted the real environment as experiment results and show the precision, recall of our method and 3D-KHT. The estimation result of our performance is better than 3D-KHT. The recall score of our method got higher scores than 3D-KHT. According to compare with the 3D-KHT, our proposed method got better precision and recall than the 3D-KHT shown in Table 2.

5 Conclusions

This study successfully designs a plane detection system for indoor operation. The proposed system is applied to a wearable device for the accurate and safe determination of a walkable plane in the space in front of a visually impaired wearer by means of streaming images. The

wearable device can also determine the length of the plane. The wearer can walk to a destination under the safe guidance of the device, which improves the shortcomings of external infrared interference cameras or laser-assisted ranging from the external environment interference. The use of the proposed wearable device with a white cane allows blind or visually impaired persons to achieve safe and independent movements similar to those provided by a guide dog.

Additional efficient algorithms will be required due to the limitations of the hardware process. Future studies will focus on the use of streamlined hardware devices and efficient algorithms to develop smaller and more lightweight and accurate devices similar to the one proposed in this study; such studies will provide detailed information to their users and offer increased independence and safety.

Acknowledgements This research was supported in part by the Ministry of Science and Technology (contracts MOST-108-2221-E-019-038-MY2, MOST-107-2221-E-019-039-MY2, MOST-109-2634-F-008-007, and MOST-109-2634-F-019-001) of Taiwan. This research was also funded by the University System of Taipei Joint Research Program (contract USTP-NTUT-NTOU-109-01), Taiwan.

References

1. Achar S, Bartels JR, Whittaker WLR, Kutulakos KN, Narasimhan SG (2017) Epipolar time-of-flight imaging. ACM Trans Graph 36(4):37:1–37:8
2. Azenkot S, Feng C, Cakmak M (2016) Enabling building service robots to guide blind people a participatory design approach. In: 2016 11th ACM/IEEE international conference on human-robot interaction (HRI), pp 3–10
3. Bai J, Lian S, Liu Z, Wang K, Liu D (2018) Virtual-blind-road following-based wearable navigation device for blind people. IEEE Trans Consum Electron 64(1):136–143
4. Baig MH, Jagadeesh V, Piramuthu R, Bhardwaj A, Di W, Sundaresan N (2014) Im2depth: scalable exemplar based depth transfer. In: IEEE Winter conference on applications of computer vision, pp 145–152
5. Caltagirone L, Scheidegger S, Svensson L, Wahde M (2017) Fast lidar-based road detection using fully convolutional neural networks. In: 2017 IEEE intelligent vehicles symposium (IV), pp 1019-1024
6. Chin LC, Basah SN, Yaacob S, Din MY, Juan YE (2015) Accuracy and reliability of optimum distance for high performance kinect sensor. In: 2015 2nd international conference on biomedical engineering (ICoBE), pp 1–7
7. Diamantas S, Astaras S, Pnevmatikakis A (2016) Depth estimation in still images and videos using a motionless monocular camera. In: 2016 IEEE international conference on imaging systems and techniques (IST), pp 129–134
8. Eigen D, Puhrsch C, Fergus R (2014) Depth map prediction from a single image using a multi-scale deep network. In: Proceedings of the 27th international conference on neural information processing systems, vol 2, pp 2366–2374
9. Fabrizio F, Luca AD (2017) Real-time computation of distance to dynamic obstacles with multiple depth sensors. IEEE Robot Autom Lett 2(1):56–63
10. Fernandes LA, Oliveira MM (2008) Real-time line detection through an improved hough transform voting scheme. Pattern Recognit 41(1):299–314
11. Forouher D, Besselmann MG, Maehle E (2016) Sensor fusion of depth camera and ultrasound data for obstacle detection and robot navigation. In: 2016 14th international conference on control, automation, robotics and vision (ICARCV), pp 1–6
12. 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
13. Hoiem D, Efros AA, Hebert M (2005) Automatic photo pop-up. ACM Trans Graph 24(3):577–584
14. Islam MA, Bruce N, Wang Y (2016) Dense image labeling using deep convolutional neural networks. In: 2016 13th Conference on computer and robot vision (CRV), pp 16–23
15. Islam MM, Sadi MS, Zamli KZ, Ahmed MM (2019) Developing walking assistants for visually impaired people: a review. IEEE Sens J 19(8):2814–2828

16. Jin Y, Li J, Ma D, Guo X, Yu H (2017) A semi-automatic annotation technology for traffic scene image labeling based on deep learning preprocessing. In: 2017 IEEE international conference on computational science and engineering (CSE) and IEEE international conference on embedded and ubiquitous computing (EUC), pp 315–320
17. Karsch K, Liu C, Kang SB (2014) Depth transfer: depth extraction from video using non-parametric sampling. IEEE Trans Pattern Anal Mach Intell 36(11):2144–2158
18. Khoshelham K (2011) Accuracy analysis of kinect depth data. In: International archives of the photogrammetry, remote sensing and spatial information sciences, pp 133–138
19. Kuznietsov Y, Stückler J, Leibe B (2017) Semi-supervised deep learning for monocular depth map prediction. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 2215–2223
20. Lee HS, Lee KM (2013) Simultaneous super-resolution of depth and images using a single camera. In: 2013 IEEE conference on computer vision and pattern recognition, pp 281–288
21. Liaquat S, Khan US, Ata-Ur-Rehman (2015) Object detection and depth estimation of real world objects using single camera. In: 2015 Fourth international conference on aerospace science and engineering (ICASE), pp 1–4
22. Liu F, Shen C, Lin G, Reid I (2016) Learning depth from single monocular images using deep convolutional neural fields. IEEE Trans Pattern Anal Mach Intell 38(10):2024–2039
23. Liu S, Yu M, Li M, Xu Q (2019) The research of virtual face based on deep convolutional generative adversarial networks using tensorflow. Phys A: Stat Mech Appl 521:667–680
24. Liu S, Li M, Li M, Xu Q (2020) Research of animals image semantic segmentation based on deep learning. Concurr Comput: Pract Exp 31(1):e4892
25. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), pp 3431–3440
26. Maurer M (2012) White cane safety day: a symbol of independence. National Federation of the Blind
27. Michels J, Saxena A, Ng AY (2005) High speed obstacle avoidance using monocular vision and reinforcement learning. In: Proceedings of the 22nd international conference on machine learning, pp 593–600
28. Naseer T, Burgard W (2017) Deep regression for monocular camera-based 6-dof global localization in outdoor environments. In: 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp 1525–1530
29. Saxena A, Chung SH, Ng AY (2005) Learning depth from single monocular images. In: Proceedings of the 18th international conference on neural information processing systems, pp 1161–1168
30. Saxena A, Sun M, Ng AY (2009) Make3d: learning 3d scene structure from a single still image. IEEE Trans Pattern Anal Mach Intell 31(5):824–840
31. Silberman N, Hoiem D, Kohli P, Fergus R (2012) Indoor segmentation and support inference from rgbd images. In: Proceedings of the 12th European conference on computer vision—volume part V, pp 746–760
32. Sokic E, Ferizbegovic M, Zubaca J, Softic K, Ahic-Djokic M (2015) Design of ultrasound-based sensory system for environment inspection robots. In: 2015 57th international symposium ELMAR (ELMAR), pp 141–144
33. Stejskal M, Mrva J, Faigl J (2016) Road following with blind crawling robot. In: 2016 IEEE international conference on robotics and automation (ICRA), pp 3612–3617
34. Straub J, Freifeld O, Rosman G, Leonard JJ, Fisher JW (2018) The manhattan frame model—manhattan world inference in the space of surface normals. IEEE Trans Pattern Anal Mach Intell 40(1):235–249
35. Tian H, Zhuang B, Hua Y, Cai A (2014) Depth inference with convolutional neural network. In: 2014 IEEE visual communications and image processing conference, pp 169–172
36. Toha SF, Yusof HM, Razali MF, Halim AHA (2015) Intelligent path guidance robot for blind person assistance. In: 2015 International conference on informatics, electronics vision (ICIEV), pp 1–5
37. Štrbac M, Marković M, Popović DB (2012) Kinect in neurorehabilitation: computer vision system for real time hand and object detection and distance estimation. In: 11th Symposium on neural network applications in electrical engineering, pp 127–132
38. Xu Q (2013) A novel machine learning strategy based on two-dimensional numerical models in financial engineering. Math Probl Eng 2013:1–6
39. Xu Q, Li M (2019) A new cluster computing technique for social media data analysis. Clust Comput 22:2731–2738
40. Xu Q, Wu J, Chen Q (2014) A novel mobile personalized recommended method based on money flow model for stock exchange. Math Probl Eng 2014:1–9
41. Xu Q, Li M, Li M, Liu S (2018a) Energy spectrum ct image detection based dimensionality reduction with phase congruency. J Med Syst 42(49):1–14

42. Xu Q, Wang Z, Wang F, Li J (2018b) Thermal comfort research on human ct data modeling. Multimed Tools Appl 77(5):6311–6326
43. Xu Q, Li M, Yu M (2019a) Learning to rank with relational graph and pointwise constraint for cross-modal retrieval. Soft Comput 23:9413–9427
44. Xu Q, Wang F, Gong Y, Wang Z, Zeng K, Li Q, Luo X (2019b) A novel edge-oriented framework for saliency detection enhancement. Image Vis Comput 87:1–12
45. Xu Q, Wang Z, Wang F, Gong Y (2019c) Multi-feature fusion cnns for drosophila embryo of interest detection. Phys A: Stat Mech Appl 531:121808
46. Xu Q, Huang G, Yu M, Guo Y (2020) Fall prediction based on key points of human bones. Phys A: Stat Mech Appl 540:123205
47. Yin LS, Sheng YK, Soetedjo A (2008) Developing a blind robot: study on 2d mapping. In: 2008 IEEE conference on innovative technologies in intelligent systems and industrial applications, pp 12–14
48. Žbontar J, LeCun Y (2016) Stereo matching by training a convolutional neural network to compare image patches. J Mach Learn Res 17(1):2287–2318
49. 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), pp 6230–6239