electronics

Article

Design and Implementation of an Intelligent Assistive Cane for Visually Impaired People Based on an Edge-Cloud Collaboration Scheme

Yuqi Ma [1], Yanqing Shi [1], Moyu Zhang [1], Wei Li [1], Chen Ma [1] and Yu Guo [1,2,*]

[1] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China; myq_cs@163.com (Y.M.); kruase@163.com (Y.S.); moyuzhang_ustb@126.com (M.Z.); lw925396172@163.com (W.L.); chen.ma.2001@foxmail.com (C.M.)
[2] Shunde Graduate School, University of Science and Technology Beijing, Foshan 528399, China
* Correspondence: guoyu@ustb.edu.cn

Abstract: Visually impaired people face many inconveniences in daily life, and there are problems such as high prices and single functions in the market of assistance tools for visually impaired people. In this work, we designed and implemented a low-cost intelligent assistance cane, particularly for visually impaired individuals, based on computer vision, sensors, and an edge-cloud collaboration scheme. Obstacle detection, fall detection, and traffic light detection functions have been designed and integrated for the convenience of moving for visually impaired people. We have also designed an image captioning function and object detection function with high-speed processing capability based on an edge-cloud collaboration scheme to improve the user experience. Experiments show that the performance metrics have an aerial obstacle detection accuracy of 92.5%, fall detection accuracy of 90%, and average image retrieval period of 1.124 s. It proves the characteristics of low power consumption, strong real-time performance, adaptability to multiple scenarios, and convenience, which can ensure the safety of visually impaired people when moving and can help them better perceive and understand the surrounding environment.

Keywords: intelligent cane; visually impaired; edge-cloud collaboration; image captioning

Citation: Ma, Y.; Shi, Y.; Zhang, M.; Li, W.; Ma, C.; Guo, Y. Design and Implementation of an Intelligent Assistive Cane for Visually Impaired People Based on an Edge-Cloud Collaboration Scheme. *Electronics* **2022**, *11*, 2266. https://doi.org/10.3390/electronics11142266

Academic Editor: Byung Cheol Song

Received: 3 June 2022
Accepted: 18 July 2022
Published: 20 July 2022

1. Introduction

According to [1], carried out by the World Health Organization (WHO) in 2011, 285 million people are visually impaired, 39 million of whom are completely blind, and 246 million of whom have weak vision/sight. The WHO report predicts that this ratio of blind people will increase. According to a study in the Lancet Global Health, the number of visually impaired people (VIP) could rise to 550 million by 2050. Visual defects make people unable to perceive external information through the visual system, resulting in various inconvenient factors in travel and life.

For example, they are more likely to hit obstacles and fall. In particular, aerial obstacles, such as awnings, tree branches, and similar objects [2], cannot be detected by a general white cane. Moreover, it is hard for them to stand up without any help after falling. They cannot perceive the surrounding situation and recognize traffic lights, which increases their travel risks. With the rapid development of computer hardware, software, and artificial intelligence, some auxiliary facilities and wearable devices have brought convenience to VIPs. However, the existing assistance tools for VIPs are more or less expensive, with limited auxiliary functions, poor interaction, and other shortcomings. We designed a low-cost portable assistance cane based on an edge-cloud collaboration environment to overcome the above problems.

Figure 1 illustrates the interactions between different components of the framework. The intelligent cane can detect aerial obstacles and alert users to avoid them. When a VIP

falls, the intelligent cane can notice the change in the user's state and provide sound alerts. If the fall alarm has not been canceled within the specified time, the person's emergency contacts (such as guardians or family members) will receive their current location via Short Message Service (SMS). Moreover, to help visually impaired people perceive more information about their surroundings and provide powerful real-time performance, all complex image-processed algorithms, such as image captioning, object detection, and traffic light detection, are executed in the cloud server. When the VIP takes an image, the intelligent cane uploads the image to the cloud server and receives the image processing result in the form of a JSON data string, and then feeds it back to the VIP.

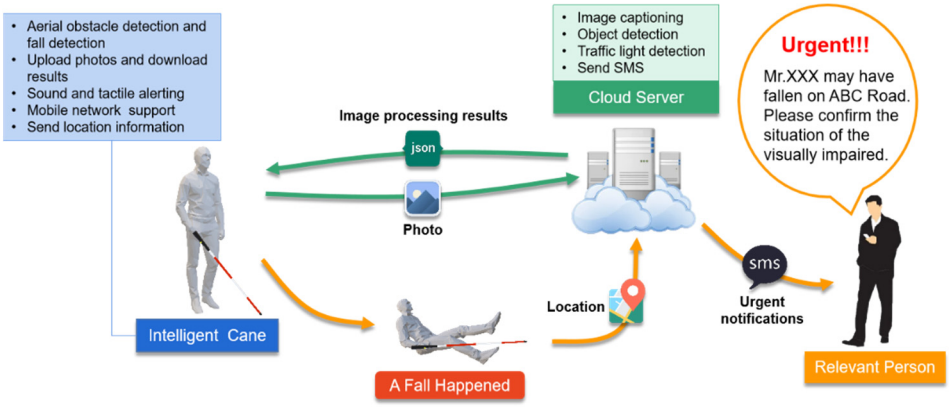

Figure 1. Interactions of the components of the framework.

The main contributions of our work can be summarized as follows:

1. The system has multiple functions and meets the needs of VIPs for assistive products from different perspectives, which facilitates the lives of the visually impaired.
2. For the convenience of moving for VIPs, we designed and integrated aerial obstacle detection, fall detection, and traffic light detection functions. Experiments show that these functions can ensure the safety of the VIP efficiently.
3. To help VIPs perceive more information about their surroundings, we have also designed an image captioning function and object detection function with high-speed processing capability based on an edge-cloud collaboration scheme.
4. The system is low cost, low power, and simple to operate. All functions are implemented using only Raspberry Pi and Arduino with some sensors. All interactive operations can be proceeded with one button.

The intelligent assistive cane for VIPs can ensure the safety of moving visually impaired people and can help them better perceive and understand the surrounding environment.

The rest of the paper is organized as follows. In Section 2, related works are presented. Section 3 introduces the proposed assistive cane. The related experiment and analysis of results are provided in Section 4. Finally, Section 5 concludes this work and discusses the direction of future work.

2. Related Work

Many assistance systems have been proposed, which can be classified into three groups according to the technology: non-vision-based, vision-based, and hybrid technologies. Non-vision-based tools use various sensors to analyze the environment and represent it to the user. Vision-based systems use live video streams or images to provide information about the surrounding environment. Hybrid systems combine vision-based approaches with sensor technologies [3]. This paper classifies them as vision-based systems.

2.1. Non-Vision-Based Systems

There is a lot of research on navigation systems, usually using five technologies: radio frequency identification (RFID), ultra-wideband (UWB), near-field communication (NFC),

Bluetooth Low Energy (BLE), and infrared (IR) [3–11]. G. A. M. Madrigal et al. [4] and C. M. Yang et al. [7] adopted the RFID system to provide information about the user's location in indoor spaces. The SUGAR system [8] uses UWB for positioning, which performs with a high level of accuracy. While these technologies can be used to solve the localization problem, it is limited to a specific location, as it must include a map of navigational tags. They are not independent systems and have many limitations, which greatly reduce the universality, flexibility, and mobility of the system.

Detecting obstacles is an essential function in assistance tools for VIPs. IR sensors and ultrasonic sensors are usually used to detect obstacles. However, IR sensors depend on the reflectance properties of the object [12]. In contrast, ultrasonic sensors are less susceptible to interference and more suitable for use in life scenarios. The traditional white cane allows the user to sense low obstacles, but the perceived distance is very limited, and it cannot sense the aerial obstacles that might pose a danger to VIPs. Unfortunately, few existing intelligent tools take this into account. For example, C. M. Yang et al. [7] and A. Khan et al. [13] used three ultrasonic sensors installed in three different directions on a white cane to detect low obstacles without considering aerial obstacles.

Non-vision-based systems can also have the following functions: voice and vibrate alert, GPS positioning, obstacle detection, navigation [3–15], etc. The existing intelligent tools almost stay at the sensor level, and their functions are relatively singular. Some limitations can be overcome using vision-based systems. This paper combines several computer vision-related technologies to optimize existing solutions.

2.2. Vision-Based Systems

At present, most traffic light detection is used for traffic, with not much for crosswalk. The characteristics of the two traffic lights have a lot of differences. P. S. Swami et al. [16] proposed a traffic light detection system for low-vision or visually impaired people. VIPs capture the image with a camera fixed on their head. Using canny edge detection and circular Hough transformation, the system identifies the shape of the traffic light and further identifies the color of it. It can also calculate the time before lights change, and send a notification to the VIP by voice. However, when the street scene is complex, the background in the image is more chaotic and includes vehicles, trees, houses, pedestrians, and other unrelated targets, resulting in an inability to effectively distinguish traffic lights from the background. With only traffic lights in the picture, the highest accuracy of the traffic light recognition system is only 90.90%, and the recognition time is approximately 7–20 s, which indicates low accuracy and poor real-time performance. P. N. Karthikayan et al. [17] developed a software for smart glasses to detect traffic lights using You Only Look Once (YOLO), an object detection algorithm based on deep learning. Then, the colors (red, green) in the image are detected using OpenCV. They also designed currency recognition by using Convolution Neutral Network (CNN), Oriented FAST, and rotated BRIEF (ORB) algorithm. However, the accuracy and processing speed of this method are not described explicitly.

Yang et al. [18] showed that the use of two RGB-D cameras on different body levels (one on head, second on the waist) allows users to take the navigable direction and pay attention to the potential hazards near their feet. They describe the RGB-D images as traversable lines and categorize obstacles into two levels (floor and low-lying ones), which are transferred to stereo sound and semantic warnings. This navigation method has a large learning cost for VIPs. Refs. [19,20] also used RGB-D cameras to collect and process depth information for navigation. However, RGB-D cameras do not detect the depth of transparent or smooth objects very well, and are not suitable for outdoor use.

M. A. Khan et al. [21] used a raspberry PI as the processor. The design incorporates a camera and sensors for obstacle detection and advanced image processing algorithms for object detection. They used OCR to identify textual information in images, and used TensorFlow object detection API and related libraries to realize facial appearance detection and distance measurement. By using SSD Lite-MobileNet, they can recognize at least 80% objects accurately. A limitation of the model is that there is an upper limit for the number

of objects identified (4–5 objects from a single video frame), which means that many objects will be neglected in a complex environment. In addition, complex environment testing and functional development is insufficient for this design.

M. G. Sarwar et al. [22] developed an LBPH-based face recognition system for VIPs which could help them distinguish acquaintances from strangers. S. Zaman et al. [23] presented a deep learning model that automatically generates highly descriptive image captions and converts the descriptive statement into braille for blind-deaf people. The whole frame of the descriptive image captions is based on the encoder–decoder model.

Most of the vision-based functions used in the assistance systems are single, the real-time performance of the calculation is not considered, and some systems are not actually applied to the assistance tool. This paper designs a mechanism that can perform multiple computer vision tasks well, which comprehensively meets the needs of the visually impaired for visual assistance.

3. The Proposed Assistance Cane

3.1. Overall Architecture and Functional Design

Considering VIPs generally have the psychological characteristics of loneliness, an inferiority complex, and a lack of sense of security, the demand for products of assistance tools for VIPs must provide a sense of security, pleasure, comfort, and familiarity. Presently, a white cane is the most commonly used auxiliary tool in the daily life of VIPs. So, in this paper, we combine artificial intelligence technology with white cane to provide VIPs with great convenience for their lives.

In consideration of the various situations when VIPs are traveling, we designed a movement module which includes aerial obstacle detection, fall detection, and traffic light detection functions to protect their safety. In terms of perception, we have used computer vision-related technologies to achieve image captioning and object detection. We have also adopted an edge-cloud collaboration scheme to guarantee high-speed computing and an emergency messaging system, thereby improving the experience of our product.

Figure 2 shows the IoT architecture of the system. The perception layer includes various sensors. Among them, the MPU6050, ultrasonic model and the button model are used to collect information. After processing the data, the Arduino controls the vibration motor and the speech synthesis broadcast module to respond or send signals to the Raspberry Pi. The Raspberry Pi is the core of the perception layer collecting images and GPS data, and transmits data through the internet layer to the application layer. 4G or 5G networks from mini portable Wi-Fi can support the Internet layer. We have used a cloud server in the application layer, which is responsible for processing images with image captioning algorithms, object detection algorithms, traffic light detection algorithms, or sending an SMS to family members' smartphones. The image processing result is sent to the Raspberry Pi via the internet layer and fed back to the VIP through a Bluetooth headset or sent to the Arduino to control the sensors module.

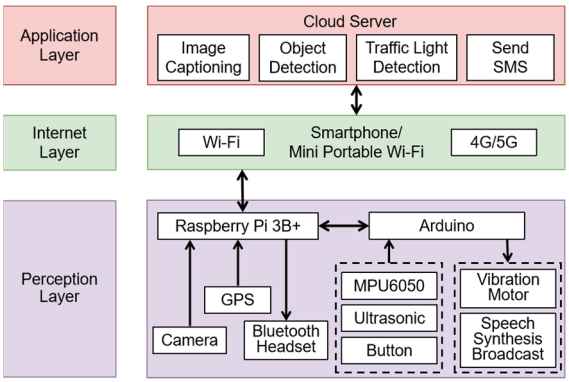

Figure 2. IoT architecture of the intelligent assistance cane.

Figure 3a shows the detailed drawings of the assistive devices. The power bank supplies power to the whole device. The Raspberry Pi connects to the Arduino and a camera by the USB ports and a GPS module by I/O port, while sensor modules are connected to the Arduino through I/O ports. Users can press the power switch to start the system. Figure 3b shows a schematic diagram of a person wearing the device.

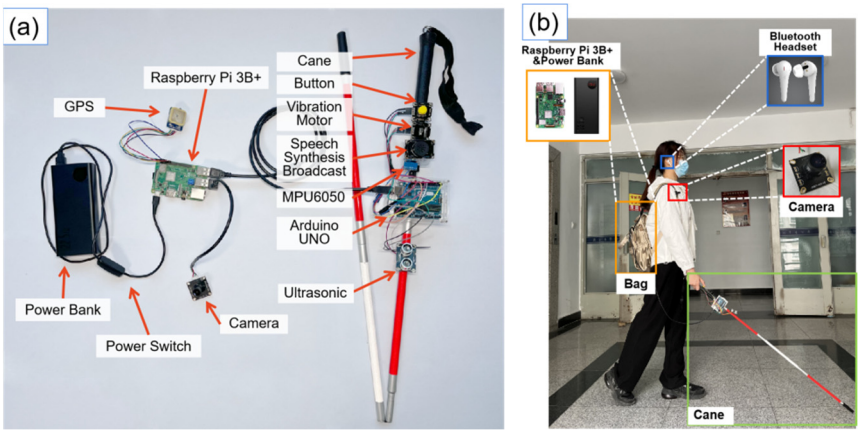

Figure 3. Instructions for the use of the assistance tool for VIPs. (**a**) Detailed drawings of assistive devices for VIPs. (**b**) Schematic diagram of a person wearing the device.

Combined with the functions needed for the assistance tool for VIPs, a schematic diagram of the system workflow has been designed, as shown in Figure 4.

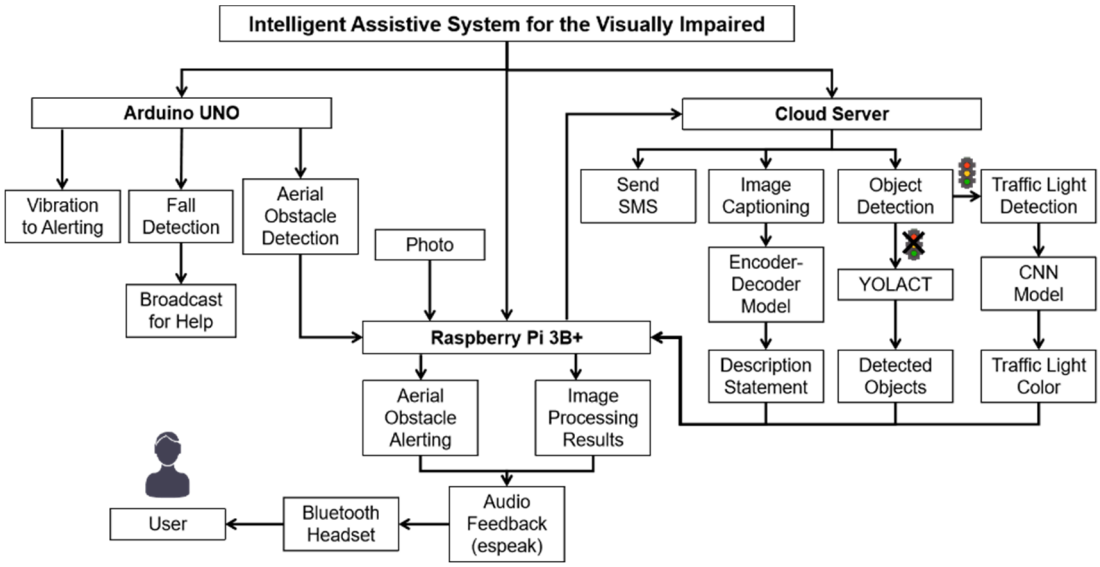

Figure 4. Schematic diagram of the system workflow.

The implementation of each function is detailed below. Section 3.2 presents the functional design of assistive mobility and Section 3.3 presents the functional design of auxiliary visual perception.

3.2. Functional Design of Assistive Mobility

3.2.1. Aerial Obstacle Detection

As illustrated in Figure 5, we use the triangulation method to detect the distance between aerial obstacles and user. When the ultrasonic wave emitted by the transmitter encounters obstacles, the receiver will receive it. By calculating the elapsed time (ET)

difference between the signal sent and received, the length of the reference side of the triangle (RST) [13] can be obtained by (1).

$$\mathrm{RST} = \frac{\mathrm{ET} * v_{sound}}{2} \tag{1}$$

where v_{sound} = 340 m/s.

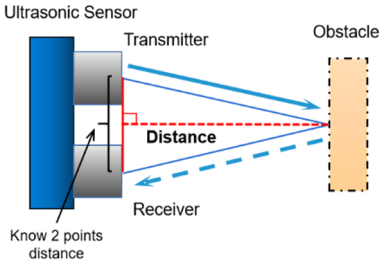

Figure 5. Ranging schematic diagram of the ultrasonic module.

With the distance between the receiver and the transmitter, the triangle's height, which is the distance between the obstacle and the sensor, can be calculated by the Pythagorean theorem.

The ultrasonic module is fixed on the cane. The angle with the ground is approximately 20°, and the height is approximately 55 cm from the ground. Considering the difference in the ultrasonic module relative position when people of different heights use the cane, the detection distance of the ultrasonic is set to 5–150 cm after calculation. The VIP will be told to walk slowly and carefully by a Bluetooth headset when an obstacle is detected. At the same time, the cane will vibrate to alert the VIP.

3.2.2. Fall Detection

Under particular circumstances, such as the rapid flow of people, VIPs may get knocked down, which can also lead to changes in acceleration and the position of the cane. Therefore, we can determine whether a VIP falls or not according to whether the acceleration changes.

The MPU6050 gyroscope sensor can detect changes in the triaxial acceleration and the triaxial angle of the cane in real time. These changes in the cane characteristics are strong evidence to determine whether the VIP falls.

The total sum acceleration vector of a cane, Acc, is an important parameter to distinguish the cane motion state. A small Acc indicates soft movement, while Acc tends to increase with drastic movement. The Acc calculation method [24] is shown in (2):

$$\mathrm{Acc} = \sqrt{\mathrm{A_x}^2 + \mathrm{A_y}^2 + \mathrm{A_z}^2} \tag{2}$$

Figure 6 illustrates the cane posture in a three-dimensional space. $\mathrm{A_x}$, $\mathrm{A_y}$, and $\mathrm{A_z}$ are the accelerations in the x, y, and z axes, respectively. The sensitivity of the MPU6050 module is 2048. The three-dimensional acceleration values are calculated from the sampled data, which are indicated in (3) [25], where $\mathrm{A_k}$ is the acceleration value in a specific direction, X is the sampled data, and g is the acceleration unit. The acceleration of gravity is defined as 1 g, equal to 9.8 m/s^2.

$$\mathrm{A_k} = \frac{\mathrm{X}}{2048}(\mathrm{g}) \tag{3}$$

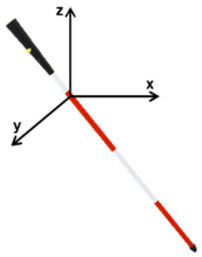

Figure 6. Cane posture in three-dimensional space.

When the cane falls, the cane has an impact phase in contact with the ground. At this time, the acceleration will reach a peak, which is more evident than the general behavior. Therefore, the threshold method can be used to determine when a fall occurs. This paper determines the minimum value MIN of Acc when the user falls through experiments. When the calculated Acc $\geq$ MIN, the user can be determined to have fallen.

When the above situation happens, the speech synthesis broadcast module on the cane will send out a voice for help—"I am blind, please help me". If the user does not cancel the broadcast within 30 s, the cloud server will inform family members of the user's location by SMS.

3.2.3. Traffic Light Detection

Traffic light detection is a sub-content of object detection. This paper adopts the YOLACT object detection algorithm described in detail in Section 3.3.1. Figure 7 illustrates the schematic diagram of the object detection function and the traffic light detection function.

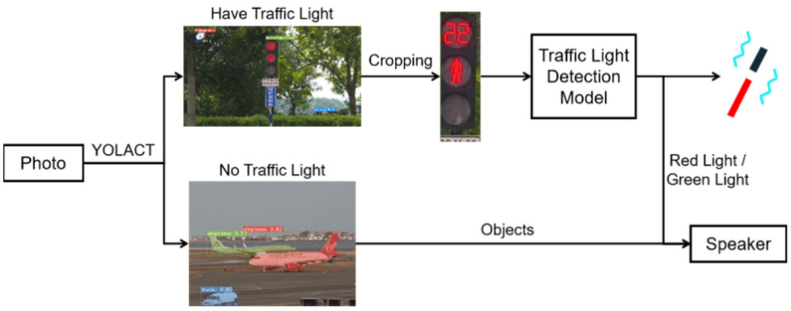

Figure 7. Schematic diagram of the object detection function and the traffic light detection function.

In this paper, the CNN model training for traffic light detection is based on the Python TensorFlow deep learning framework and the MIT (Massachusetts Institute of Technology) open-source traffic light dataset. Figure 8 shows some pictures of the traffic light dataset. The model architecture is shown in Figure 9.

Figure 8. Sample pictures from the dataset.

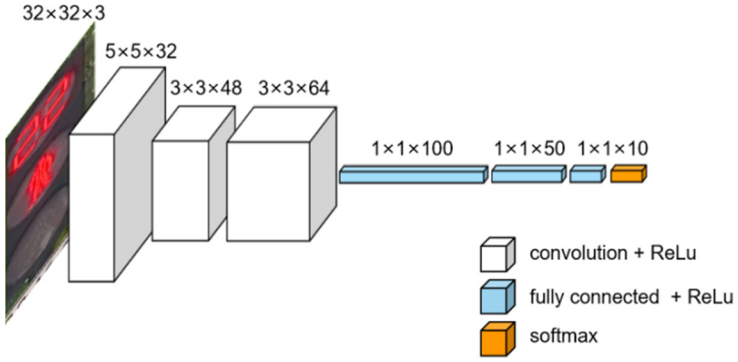

Figure 9. Architecture of the traffic light detection model.

The appearing traffic light will be marked by YOLACT with a bounding box and cropped by obtaining the coordinates of the bounding box during object detection. Considering that the traffic lights on the sidewalk will appear in the central part of the picture when the VIP is crossing the road, we select the traffic light nearest to the center of the photograph for cropping when an image contains multiple traffic lights. After that, traffic light detection is carried out through the pre-trained model, and the recognized traffic light situation is reported through a Bluetooth headset. At the same time, the cane vibrates to remind the blind to pay attention to the report information.

3.3. Functional Design of Auxiliary Visual Perception

3.3.1. Object Detection

According to [26,27], compared with other methods (such as FCIS, Mask R-CNN, RetinaMask, PA-Net, MSR-CNN), YOLACT is the fastest method with the highest frame per second (fps). This base model achieves a mean average precision (mAP) of 32.2 at 33.0 fps when generating a bounding box. Therefore, in practical applications, YOLACT can meet the requirements of high accuracy and real-time performance. Figure 10 shows some examples of YOLACT object detection results.

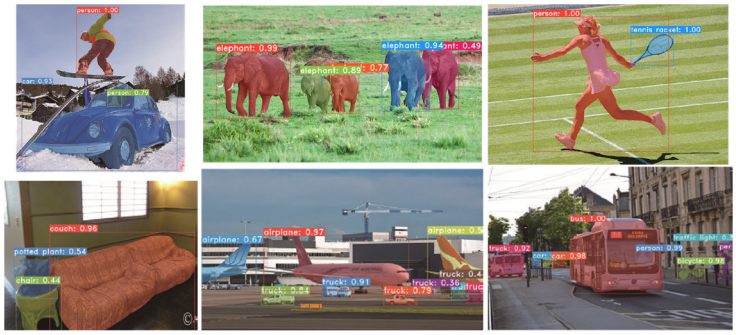

Figure 10. Some examples of object detection results [27].

The object detection model YOLACT is used in this paper to help the VIP perceive the surrounding environment. This paper adopts YOLACT-550 [27], which is built on the feature pyramid network (FPN) and the ResNet101 as the backbone network. The provided pre-trained weights for the MS COCO [28] have also been adopted for convenience.

In an image, multiple identical objects may be detected. Therefore, we can feed back the detected objects and their numbers to the VIP. For example, in the first picture of Figure 10, the results fed back to the VIP would be, "In the current field of view, there are two people and one car".

The VIP users can trigger the function by pressing the button on the cane once, which will be executed by the edge-cloud collaboration scheme described in Section 3.1.

3.3.2. Image Captioning

Although the object detection algorithm can identify all the objects in the image, it cannot identify the relationship between them. Compared with the object detection algorithm, image captioning is a higher-level and more complex perception task. It enables the VIP to perceive the most important information and relationships between objects in an image, which makes it easier for them to understand their surroundings. The two algorithms complement each other to provide a complete sense of the environment.

Image captioning has attracted considerable attention in recent years. It is an algorithm that describes the image's content through language, helping the VIP obtain information about their surrounding environment. The model of this function is based on the image captioning model mentioned in [29]. The image captioning model uses the encoder–decoder framework combined with a soft attention mechanism. An image is fed into the encoder (a CNN) to extract features, and then the encoded image is used as an input of long short-term memory (LSTM) for decoding. The current output of LSTM is used as the input again so that LSTM can predict the next word. Soft attention is deterministic. It can change the attention to the local part of the image with the progress of the decoder and then generate more reasonable words, which makes the whole mode smooth and differentiable.

Figure 11 shows the image processing procedure. The image captioning result is the visualization effect of the soft attention concentration position. The white area indicates the regions which the model roughly attends to. As the model generates each word, its attention changes to reflect the relevant parts of the image and finally concatenate all the words.

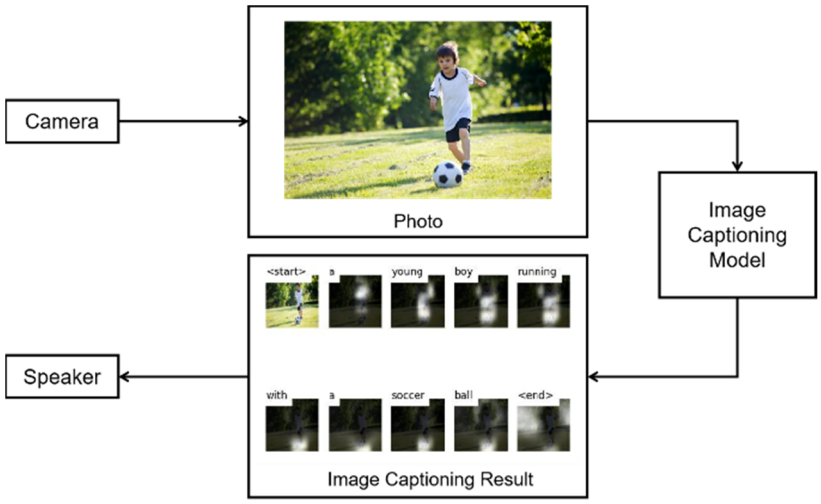

Figure 11. Image captioning function realization process.

The function can be triggered by continuously pressing the button on the cane twice, and then the edge-cloud collaboration scheme described in Section 3.1 will execute this function.

4. Experiment and Result Analysis

4.1. Experimental Environment

Aerial obstacle detection and fall detection experiments were carried out by fixing the Arduino UNO on a cane.

This paper adopts the YOLACT model in [28] and trains the image captioning model on the cloud server. The Flickr8k dataset [30] was used to train and test the model, which consists of 8092 images and comes with five reference sentences per image. Each descriptive sentence has a score on a scale of 0 to 1, representing the least accurate sentence description to the most accurate. Figure 12 illustrates some examples of the Flickr8k dataset. In the image captioning model training, the images are put through the VGG-16 model for feature

extraction by encoding, during which 70% of the dataset is used for model training and 30% is used for testing. The LSTM network model is used for training during decoding.

Figure 12. Examples of the Flickr8k dataset [30].

The image description algorithm, object detection algorithm, and traffic light detection algorithm are all executed on the cloud server. The server configuration is 64 G RAM, 32 core 2.10 GHz CPU, and 2080 Ti GPU. The operation system is Ubuntu20.04. We also deployed these image processing algorithms to a mini-computer for offline testing and compared the image processing time of offline and edge-cloud collaboration to verify the effectiveness of using edge-cloud collaboration. The configuration parameters of the mini-computer are 4 G RAM and 4 core 1.44 GHz CPU; the system is Ubuntu18.04, and both software environments are Python 3.6.

4.2. Result Analysis

4.2.1. Aerial Obstacle Detection

When the ultrasonic detection distance is set to 5–150 cm, the device can detect aerial obstacles approximately 60 cm to 190 cm from the ground. Figure 13 illustrates three kinds of aerial obstacles: branches, fire hydrant doors, and eaves of low buildings.

Figure 13. Example of aerial obstacle detection.

As shown in Table 1, we performed 120 experiments on six aerial obstacles of different sizes and heights. Experiments show that the intelligent cane can effectively detect obvious aerial obstacles. Nevertheless, it is difficult to detect some very tiny aerial obstacles, such as strings. The overall accuracy is 92.5%. We also asked ten volunteers between 150 and 188 cm in height to test this function. The designed cane can accurately detect aerial obstacles from the knee to the head. It can effectively avoid risks for the visually impaired.

Table 1. Accuracy rate of aerial obstacle detection.

Aerial Obstacles	Height Range from the Ground	Number of Experiments	Accuracy Rate
Cabinet door	51–163 cm	20	100%
Eaves of low buildings	104–110 cm	20	100%
Fire hydrant doors	104–176 cm	20	95%
Strings	126–126.5 cm	20	65%
Sticks	136–138 cm	20	95%
Branches	155–185 cm	20	100%
Total	51–185 cm	120	92.5%

4.2.2. Fall Detection

Figure 14 illustrates the variation curve of the Acc when a fall is detected. Through experiments, this paper found that the Acc value is always less than 2.35 g during daily activities and is greater than or equal to 2.35 g during falls. Therefore, the threshold of Acc for determining falls should be set to 2.35 g, which means the user is determined as falling when the calculated Acc $\geq$ 2.35 g.

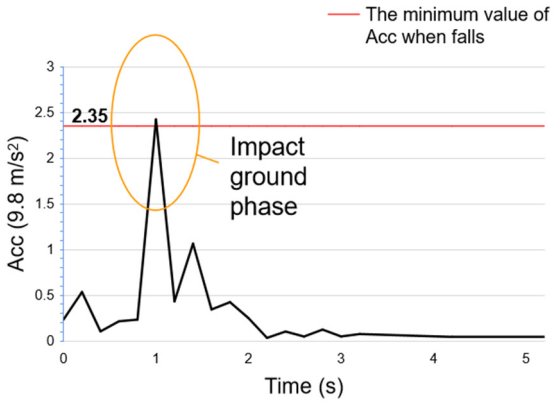

Figure 14. Total acceleration curve of falls.

The experiment was set for falling and daily activities to verify the algorithm's effectiveness. Experiments show that the detection accuracy of fall events is 90%, as shown in Table 2.

Table 2. Accuracy rate of fall detection.

Fall Type	Number of Experiments	Correct Detection Times	Accuracy Rate
Fall to the ground	50	45	90%

Since the MPU6050 sensor is installed on the cane, we tested the false-positive rate of the behaviors that may affect the acceleration value of the cane in the daily activities of a VIP, including walking normally, touching tactile paving, hitting objects with the cane, walking up and downstairs, standing up and sitting down, and bending. We deemed that misjudgment would only occur when objects are hit with a mighty force in the experiment and would operate normally when exploring the path with average strength or other daily behaviors. The false-positive rate of daily activity events is 2.5%, as shown in Table 3.

Table 3. False-Positive Rate of daily activities.

Type of Daily Activities	Number of Experiment	Error Detection Times	False-Positive Rate
Walk normally	20	0	0%
Touch tactile paving by cane	20	0	0%
Hit object by cane	20	3	15%
Walk up and downstairs	20	0	0%
Stand up and sit down	20	0	0%
Bend down	20	0	0%
TOTAL	120	3	2.5%

If the fall alarm is not canceled within 30 s, the assistance device will send the user's location information to the smartphone of the family member or guardian. Figure 15a shows the notification message of the fall occurrence for the visually impaired user. Figure 15b shows the location of the fall occurrence for the visually impaired user sent automatically by the cane.

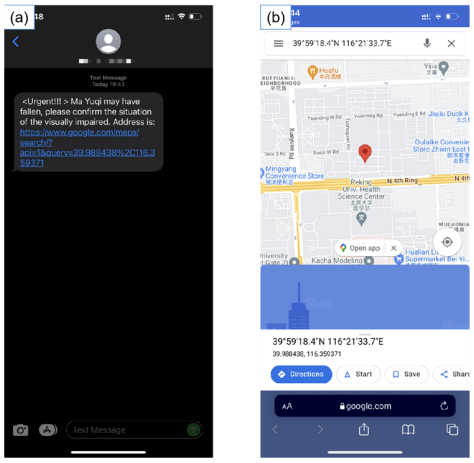

Figure 15. Screen of visually impaired user fall information received by emergency contacts. (**a**) Notification message. (**b**) Location information.

4.2.3. Object Detection and Traffic Light Detection

Figure 16 illustrates a comparison of the results of object detection and image captioning. Image captioning can produce descriptive words for whole scene, such as "in a white shirt" in Figure 16c. In addition, it can identify the gender of characters and the position relationship between objects, such as "girl" and "in front of". In contrast, object detection results are all nouns, which are more comprehensive in the image. For example, words such as "book", "cell phone", "mouse", and "laptop" were detected in object detection but did not appear in the image captioning results. However, object detection cannot distinguish differences between people such as age and gender, whilst image captioning can. At this point, it is less specific than the processing results of the image captioning. Therefore, the two functions can complement each other to provide better environmental information for VIPs.

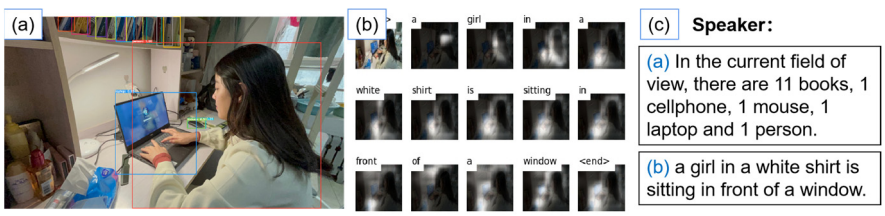

Figure 16. Comparison of the results of object detection and image captioning. (**a**) Object detection processed result. (**b**) Image captioning processed result. (**c**) Comparison of the two results.

The YOLACT-based model has 32.3 box mAP [27] and it can detect 80 objects, including "toilet", "bus", "TV", "toothbrush", "umbrella", etc. In this way, it can meet the daily needs of VIP. Experiments show that the result of the object detection function is comprehensive and accurate.

The dataset used by the traffic light detection model is the MIT open-source traffic light dataset. The images were resized to a dimension of 32 by 32 by 3 pixels in a red-green-blue (RGB) three-channel color image to reduce the time spent in network training.

The convolution layer extracts CNN features and inputs the features into the base classifier. The data structure of the feature data is an array. The label is one-hot encoding to adapt to the model training structure. The model is built by functions in the TensorFlow deep learning framework. The batch size is set to 128 and epoch to 50.

The loss rate of the traffic light detection model is 6.43%. After the training, 1000 pictures containing only traffic lights were used to test the accuracy of the traffic light detection model. The average accuracy is 98.13%.

Figure 17 shows the processing procedure and the results of the traffic light detection. Figure 17a shows the detection of two traffic lights marked by blue and green bounding boxes. As the center point of the green box is closer to the center point of the picture, the traffic light in the green box is cropped. Then, the traffic light detection model is used to recognize the traffic light picture in Figure 17b. If the traffic light is detected as green, the user is told, "Green light ahead, you can pass" through the Bluetooth headset. If it is red, the user will be told, "Red light ahead, do not pass". In this work, we used 200 photos similar to Figure 17a to test the accuracy. The recognition accuracy of this function reached 88.20%.

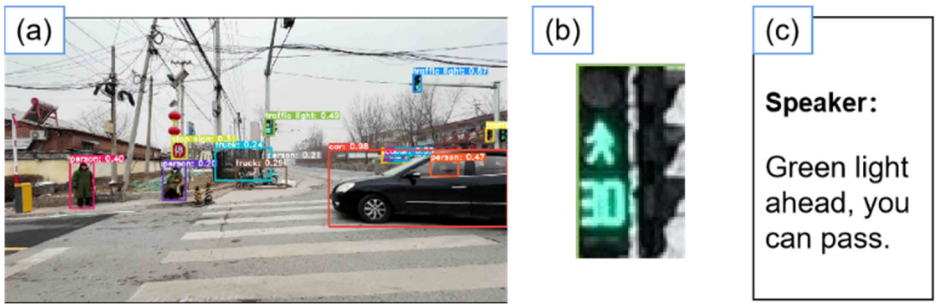

Figure 17. Traffic light detection processing and results. (**a**) Object detection processing results. (**b**) Traffic light image cropping results. (**c**) Report statement.

4.2.4. Image Captioning

We tested the accuracy of the image captioning function in a variety of life scenes. The main elements of each picture can be described and reported correctly. Figures 18–20 show the processing results of some actual-life scenes.

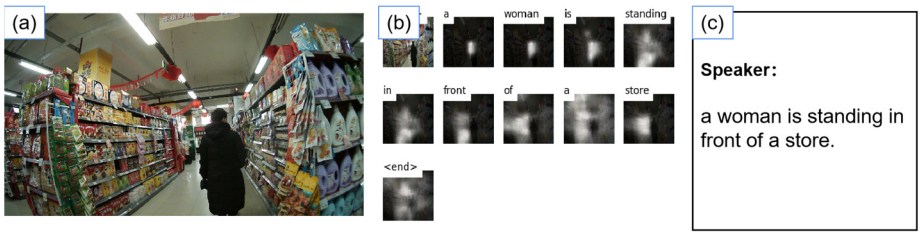

Figure 18. Image processing results of a store scene. (**a**) Processed images. (**b**) Processed results. (**c**) Report contents.

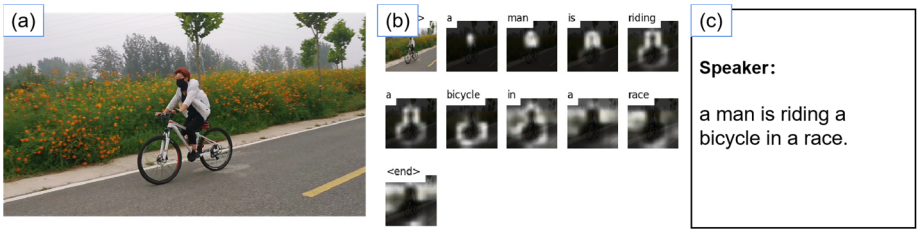

Figure 19. Image processing results of a roadside scene. (**a**) Processed images. (**b**) Processed results. (**c**) Report contents.

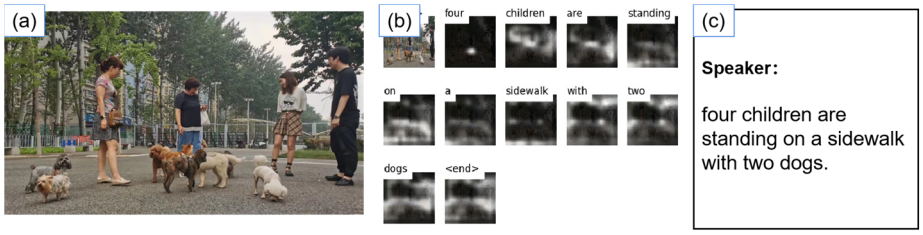

Figure 20. Image processing results of a park scene. (**a**) Processed images. (**b**) Processed results. (**c**) Report contents.

Figure 21 shows the processing results of two similar images. The results are different, but the sentences all have the subject "people" and the verb "standing". The difference is that (a) recognizes the background as "snow", while (b) recognizes it as "wall". As shown in Figure 22, when the elements in the image are not very obvious, it may produce wrong results. (a) misidentifies the bicycle as a "skateboard" and (b) misidentifies the scene as a "waterfall". Therefore, users can take multiple photos at the same location to avoid such a mistake and obtain more accurate and comprehensive information on the surrounding environment.

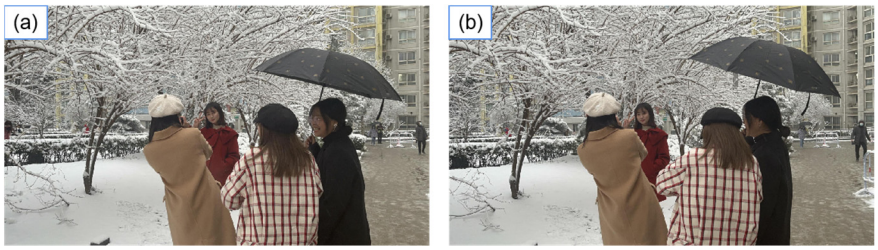

Figure 21. Generated caption: (**a**) "a group of people standing in the snow"; (**b**) "three people are standing in front of a brick wall".

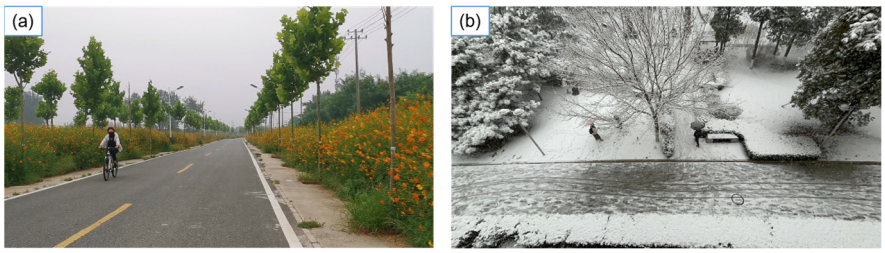

Figure 22. Generated caption: (**a**) "a man is riding a skateboard down a paved road"; (**b**) "a person is in the middle of a waterfall".

To verify the effectiveness of this function, we asked 10 volunteers to test it and score it. Each volunteer rated the 10 processed photos and gave an overall score based on whether the images and their corresponding processed results were consistent. Scores ranged from 1 to 10, with 1 being the lowest and 10 being the highest. The test results are shown in

Table 4. The average score is 8.3, which proves that this function can effectively help VIPs perceive the surrounding environment.

Table 4. Score for image captioning function accuracy.

Volunteer Number	Score
1	8
2	8
3	9
4	9
5	7
6	10
7	8
8	6
9	9
10	9
Average	8.3

4.2.5. Overall Device Performance

The assistance tool is sensitive in detecting aerial obstacles, with an accuracy of 92.5%. Since the cane itself can explore low obstacles, it can help VIPs avoid obstacles in all directions. The fall detection feature of the assistance tool is 90% accurate and can notify family members in the case of an emergency. It enables the visually impaired to receive help as soon as possible after falling. The image captioning score is 8.3 on a scale of 10, and the mAP of object detection is 32.3. The combination of these two functions can provide VIPs with accurate and comprehensive environmental information.

In terms of image processing speed, we compared the processing time of offline and edge-cloud collaboration systems, as shown in Figure 23. When processing images in the offline system, visually impaired users need 3.959 s for the object detection function, 4.632 s for the traffic light detection function, and 4.049 s for the image captioning function from taking a photo to receiving the result. The average image processing time is 4.213 s. When processing in the edge-cloud collaboration system, these data are **1.032 s**, **1.265 s**, and **1.076 s**, respectively. The average processing time is **1.124 s**, improved by 73.32%. Experiments show that the use of edge-cloud collaboration for the intelligent cane can enable VIPs to obtain environmental information promptly under various circumstances.

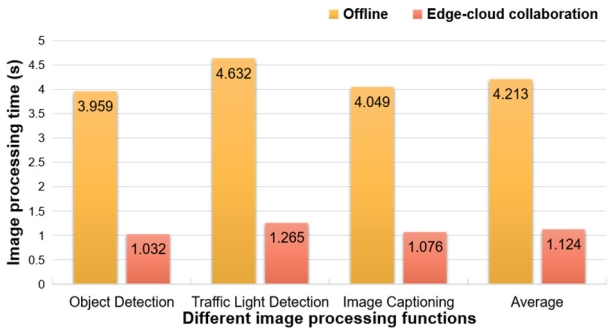

Figure 23. Comparison of the image processing time of offline and edge-cloud collaboration.

According to simulation experiments, blind people usually hesitate in front of a traffic light for approximately 1 min before determining the traffic light situation. However, the device designed in this paper only needs 1.265 s on average to detect the traffic light situation, saving approximately **97.89%** of the time, and can also guarantee that VIPs cross the road safely.

The average power consumption for processing a single image is 2.028 mAh. The capacity of the power bank is 20,000 mAh, which can process approximately **9862** images. Thus, the long battery life of the tool can provide convenient service for VIPs.

5. Conclusions and Future Work

In this paper, an intelligent assistive cane has been proposed for visually impaired people, based on an edge-cloud collaboration scheme. The cane includes five functions: aerial obstacle detection, fall detection, traffic light detection, object detection, and image captioning.

The experimental results show that the proposed intelligent cane can efficiently detect aerial obstacles from the knee to the head with an accuracy of 92.5%. The average accuracy of fall detection reached up to 90%. Furthermore, when VIPs experience a fall event, their emergency contacts will receive their current location by (SMS). When detecting aerial obstacles and traffic lights, the cane vibrates to alert VIP users and ensure their safety. Two functions, namely image captioning and object detection, can help VIPs better understand their environment. Since all image processing is executed in the cloud server, the results can be obtained in an average of 1.124 s for each image, with very high real-time performance.

When using this intelligent assistance cane, VIPs can release their anxiety and gradually integrate into society, which not only makes their lives more wonderful but also lays a solid foundation for building a harmonious society.

In the future, as mentioned in [31], we can offload some tasks from the cloud to the edge to maintain the balance of the network, which can avoid communication delay and insecure transmission, and improve the overall performance of the device. In addition, we can expand more functions based on the IoT, such as real-time monitoring of the location and surrounding environment of the VIP.

Author Contributions: Conceptualization, Y.M., Y.S., W.L., M.Z., C.M. and Y.G.; methodology, Y.M., Y.S., M.Z. and W.L.; software, Y.M.; validation, Y.M.; formal analysis, Y.M., C.M. and Y.G.; writing—original draft preparation, Y.M.; writing—review and editing, Y.G.; visualization, Y.M., Y.S. and M.Z.; supervision, Y.G.; project administration, Y.M. and Y.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the Guangdong Basic and Applied Basic Research Foundation (No. 2020A1515110463), partially funded by the National Natural Science Foundation of China under grant 61772068, and partially funded by the Fundamental Research Funds for the Central Universities under grant FRF-TP-20-063A1Z and FRF-IDRY-20-018.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: "Flickr8K dataset" at https://academictorrents.com/details/9dea07ba660a722ae1008c4c8afdd303b6f6e53b (accessed on 20 June 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bourne, R.R.A.; Flaxman, S.R.; Braithwaite, T.; Cicinelli, M.V.; Das, A.; Vision Loss Expert Group. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: A systematic review and meta-analysis. *Lancet Glob. Health* **2017**, *5*, e888–e897. [CrossRef]
2. Sáez, J.M.; Escolano, F.; Lozano, M.A. Aerial Obstacle Detection With 3-D Mobile Devices. *IEEE J. Biomed. Health Inform.* **2015**, *19*, 74–80. [CrossRef]
3. Plikynas, D.; Žvironas, A.; Gudauskis, M.; Budrionis, A.; Daniušis, P.; Sliesoraitytė, I. Research advances of indoor navigation for blind people: A brief review of technological instrumentation. *IEEE Instrum. Meas. Mag.* **2020**, *23*, 22–32. [CrossRef]
4. Madrigal, G.A.M.; Boncolmo, M.L.M.; Santos, M.J.C.D.; Ortiz, S.M.G.; Santos, F.O.; Venezuela, D.L.; Velasco, J. Voice Controlled Navigational Aid With RFID-based Indoor Positioning System for the Visually Impaired. In Proceedings of the 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), Baguio City, Philippines, 29 November–2 December 2018; pp. 1–5. [CrossRef]
5. Zvironas, A.; Gudauskis, M.; Plikynas, D. Indoor Electronic Traveling Aids for Visually Impaired: Systemic Review. In Proceedings of the 2019 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 5–7 December 2019; pp. 936–942. [CrossRef]
6. Plikynas, D.; Žvironas, A.; Budrionis, A.; Gudauskis, M. Indoor Navigation Systems for Visually Impaired Persons: Mapping the Features of Existing Technologies to User Needs. *Sensors* **2020**, *20*, 636. [CrossRef]

7. Yang, C.M.; Jung, J.Y.; Kim, J.J. Development of Walking Assistive Cane for Obstacle Detection and Location Recognition for Visually Impaired People. *Sens. Mater.* **2021**, *33*, 3623–3633. [CrossRef]
8. Martinez-Sala, A.S.; Losilla, F.; Sánchez-Aarnoutse, J.C.; García-Haro, J. Design, Implementation and Evaluation of an Indoor Navigation System for Visually Impaired People. *Sensors* **2015**, *15*, 32168–32187. [CrossRef]
9. Khan, S.; Nazir, S.; Khan, H.U. Analysis of Navigation Assistants for Blind and Visually Impaired People: A Systematic Review. *IEEE Access* **2021**, *9*, 26712–26734. [CrossRef]
10. Al-Fahoum, A.S.; Al-Hmoud, H.B.; Al-Fraihat, A.A. A smart infrared microcontroller-based blind guidance system. *Act. Passiv. Electron. Compon.* **2013**, *2013*, 726480. [CrossRef]
11. Wahab, M.H.A.; Talib, A.A.; Kadir, H.A.; Johari, A.; Noraziah, A.; Sidek, R.M.; Mutalib, A.A. Smart cane: Assistive cane for visually-impaired people. *arXiv* **2011**, arXiv:1110.5156.
12. Mustapha, B.; Zayegh, A.; Begg, R.K. Ultrasonic and Infrared Sensors Performance in a Wireless Obstacle Detection System. In Proceedings of the 2013 1st International Conference on Artificial Intelligence, Modelling and Simulation, Kota Kinabalu, Malaysia, 3–5 December 2013; pp. 487–492. [CrossRef]
13. Khan, A.; Ashraf, M.A.; Javeed, M.A.; Sarfraz, M.S.; Ullah, A.; Khan, M.M.A. Electronic Guidance Cane for Users Having Partial Vision Loss Disability. *Wirel. Commun. Mobile Comput.* **2021**, *2021*, 1628996. [CrossRef]
14. Zhangaskanov, D.; Zhumatay, N.; Ali, M.H. Audio-based Smart White Cane for Visually Impaired People. In Proceedings of the 2019 5th International Conference on Control, Automation and Robotics (ICCAR), Beijing, China, 19–22 April 2019; pp. 889–893. [CrossRef]
15. Singh, B.; Kapoor, M. Assistive cane for visually impaired persons for uneven surface detection with orientation restraint sensing. *Sens. Rev.* **2020**, *40*, 687–698. [CrossRef]
16. Swami, P.S.; Futane, P. Traffic Light Detection System for Low Vision or Visually Impaired Person Through Voice. In Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 16–18 August 2018; pp. 1–5. [CrossRef]
17. Karthikayan, P.N.; Pushpakumar, R. Smart Glasses for Visually Impaired Using Image Processing Techniques. In Proceedings of the 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 11–13 November 2021; pp. 1322–1327. [CrossRef]
18. Yang, K.; Wang, K.; Lin, S.; Bai, J.; Bergasa, L.M.; Arroyo, R. Long-Range Traversability Awareness and Low-Lying Obstacle Negotiation with RealSense for the Visually Impaired. In Proceedings of the 2018 International Conference on Information Science and System (ICISS '18). Association for Computing Machinery, New York, NY, USA, 27–29 April 2018; pp. 137–141. [CrossRef]
19. Hakim, H.; Fadhil, A. Navigation system for visually impaired people based on RGB-D camera and ultrasonic sensor. In Proceedings of the International Conference on Information and Communication Technology (ICICT '19), Association for Computing Machinery, New York, NY, USA, 15–16 April 2019; pp. 172–177. [CrossRef]
20. Chen, H.; Wang, K.; Yang, K. Improving RealSense by Fusing Color Stereo Vision and Infrared Stereo Vision for the Visually Impaired. In Proceedings of the 2018 International Conference on Information Science and System (ICISS '18), Association for Computing Machinery, New York, NY, USA, 27–29 April 2018; pp. 142–146. [CrossRef]
21. Khan, M.A.; Paul, P.; Rashid, M.; Hossain, M.; Ahad, M.A.R. An AI-Based Visual Aid With Integrated Reading Assistant for the Completely Blind. *IEEE Trans. Hum. Mach. Syst.* **2020**, *50*, 507–517. [CrossRef]
22. Sarwar, M.G.; Dey, A.; Das, A. Developing a LBPH-based Face Recognition System for Visually Impaired People. In Proceedings of the 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), Riyadh, Saudi Arabia, 6–7 April 2021; pp. 286–289. [CrossRef]
23. Zaman, S.; Abrar, M.A.; Hassan, M.M.; Islam, A.N.M.N. A Recurrent Neural Network Approach to Image Captioning in Braille for Blind-Deaf People. In Proceedings of the 2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON), Dhaka, Bangladesh, 28–30 November 2019; pp. 49–53. [CrossRef]
24. Al-Dahan, Z.T.; Bachache, N.K.; Bachache, L.N. Design and implementation of fall detection system using MPU6050 Arduino. In Proceedings of the International Conference on Smart Homes and Health Telematics, Wuhan, China, 25–27 May 2016; Springer: Cham, Switzerland, 2016; pp. 180–187.
25. Kosobutskyy, P.; Ferens, R. Statistical analysis of noise measurement system based on accelerometer-gyroscope GY-521 and Arduino platform. In Proceedings of the 2017 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), Lviv, Ukraine, 21–25 February 2017; pp. 405–407. [CrossRef]
26. Heng, S.S.; Shamsudin, A.U.b.; Mohamed, T.M.M.S. Road Sign Instance Segmentation By Using YOLACT For Semi-Autonomous Vehicle In Malaysia. In Proceedings of the 2021 8th International Conference on Computer and Communication Engineering (ICCCE), Kuala Lumpur, Malaysia, 22–23 June 2021; pp. 406–410. [CrossRef]
27. Bolya, D.; Zhou, C.; Xiao, F.; Lee, Y.J. YOLACT: Real-Time Instance Segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 9156–9165. [CrossRef]
28. YOLACT Model. Available online: https://github.com/dbolya/yolact (accessed on 8 July 2021).
29. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 7–9 July 2015; pp. 2048–2057.

30. Flickr8K Dataset. Available online: https://academictorrents.com/details/9dea07ba660a722ae1008c4c8afdd303b6f6e53b (accessed on 20 June 2021).
31. Guo, Y.; Mi, Z.; Yang, Y.; Obaidat, M.S. An Energy Sensitive Computation Offloading Strategy in Cloud Robotic Network Based on GA. *IEEE Syst. J.* **2019**, *13*, 3513–3523. [CrossRef]