A Visually Impaired Assistant Using Neural Network and Image Recognition with Physical Navigation

On-Chun Arthur Liu, Shun-Ki Li, Li-Qi Yan, Sin-Chun Ng[✉],
and Chok-Pang Kwok[✉]

School of Science and Technology, The Open University of Hong Kong,
30 Good Shepherd Street, Ho Man Tin, Kowloon, Hong Kong
scng@ouhk.edu.hk, cpkwok@study.ouhk.edu.hk

Abstract. In Hong Kong, over 2.4% of the total population suffered from visual impairment. They are facing many difficulties in their daily lives, such as shopping and travelling from places to places within the city. For outdoor activities, they usually need to have an assistant to guide their ways to reach the destinations. In this paper, a mobile application assisting visually impaired people for outdoor navigation is proposed. The application consists of navigation, obstacle detection and scene description functions. The navigation function assists the user to travel to the destination with the Global Positioning System (GPS) and sound guidance. The obstacle detection function alerts the visually impaired people for any obstacles ahead that may be avoided for collision. The scene description function describes the scene in front of the users with voice. In general, the mobile application can assist the people with low vision to walk on the streets safely, reliably and efficiently.

Keywords: Visually impaired · Neural network · Image recognition · Navigation

1 Introduction

According to the statistics in the year 2015, Vos stated in [1] that over 940 million people around the world are classified as visually impaired or blind. In Hong Kong, over 174,800 are considered as visually impaired, which is 2.4% of the Hong Kong population [2]. They are facing many difficulties in their daily lives, such as daily consumption and travelling. Due to the disability problem, visually impaired people are seldom to have outdoor activities. The main reason is the unfamiliar routes and the traffic conditions that lead to the destinations. Although there is a cane to assist visually impaired people for outdoor activities, they need to tap everywhere to detect any obstacles around them. Hence, it has chances to hit other pedestrians accidentally and may in turn annoy other peoples. For people who are with low vision, they may not have a cane to assist

M. Han et al. (Eds.): ISNN 2020, LNCS 12557, pp. 270–281, 2020.
https://doi.org/10.1007/978-3-030-64221-1_24

them for outdoor activities. As they are hard to see, they may hit obstacles like temperate traffic signs or fire hydrants easily. For this group of disabilities, they like to stay at home all the time because of insufficient support to them. Moreover, visually impaired people are unable to have "sightseeing" activities because they need a tour guide to "describe" the actual landscape or scenes to them.

To solve the above issues, this paper proposes an intelligent application installed on mobile devices for visually impaired people to assist their outdoor activities. By applying neural network and image recognition technologies, an obstacle detector has been developed to alert the user for any obstacles close to them. Furthermore, the application can describe the conditions around the user so that they can make their right decisions easier. With this function, the application can act as a "tour guide" to describe the actual scenes in front of the users. Also, navigation is an important function for visually impaired people. With the use of navigation guidance, visually impaired people can reach the destinations safely, reliably and more efficiently.

2 The Existing Solutions

To provide a safe outdoor travelling assistance for visually impaired people, the navigation and detection functions are essential to satisfy their specific needs. Due to the development of artificial intelligence (AI) and derived technologies, there are several solutions in the market [7–10]. Although those applications are performing well, those applications contain different scopes of weaknesses, especially in costs, language support, and compatibilities that they are not applicable in Hong Kong [3–6]. There are two main types of programs currently available in the market to assist the visually impaired. One is a simple mobile phone application, and the other one is to combine hardware such as a blind cane to interact with the users.

Eye-D. Eye-D[1] is a mobile application. Its main function is to assist visually impaired people to find where they are, search nearby facilities, and describe their surroundings. The advantage of this application is that it gives users a clear idea of where they are and allows the visually impaired know what is in front of them. But the disadvantage is that there is no voice indication, relying upon an on-screen text, and only describes the object ahead, does not explain the action, making it difficult for the visually impaired to know what is happening ahead.

WeVoice. WeVoice[2] is a mobile application proposed by InnoTech Association[3]. The main function is to read the text aloud. Users need to take a picture with their phone, and the software will then read the text in the photo. The disadvantage is that it is difficult for the visually impaired to target objects accurately.

[1] https://eye-d.in/.

[2] https://play.google.com/store/apps/details?id=hk.com.redso.read4u&hl=zh_HK.

[3] https://www.facebook.com/InnoTechAssociation/.

Smart City Walk. Smart City Walk[4] is proposed by Hong Kong Blind Union[5]. The main function is to display the current location, search nearby facilities with voice input, and navigate the user to the destination via voice and text. Although it can be accurately positioned indoors, it can only be used in several buildings with iBeacon [11] installed.

WeWalk. WeWalk[6] is a mobile application with a blind cane. It has a built-in ultrasonic detector that alerts the user to obstacles in front of them. Besides, it has voice navigation to guide visually impaired people to their destinations. But its disadvantages are expensive, and only in English and Turkish, and language navigation is limited to Sweden and parts of Europe.

Although many enterprises developed different types of applications for the visually impaired with the advanced technologies for travelling, the users are still inconvenient to perform outdoor activities with no application describing the environment ahead, prompting with obstacles, and providing navigation at once. The existing applications have many functions, but they are not fulfilling the needs of the visually impaired in Hong Kong. For example, some of the applications require the users to take photographs, and it is difficult for the visually impaired to know the exact location of the objects. These applications also cannot remind the visually impaired before there are obstacles. The applications with these functions do not support Chinese or cater for the Hong Kong market.

According to the insufficient supports to visually impaired people, this paper proposes an application with low-cost hardware that uses voice to communicate with visually impaired people by describing the scene in front of them and providing obstacles prompt with navigation.

3 The System Design

The system has three main parts: image processing, navigation, and voice interaction. The image processing part is done on a Raspberry Pi 3 Model B+ and a machine learning server. The image comes from the Raspberry Pi 3 Model B+, with the camera attached which is put on by the user to allow hand-free video capturing feature. The image processing part includes scene description system and obstacle prompt system based on the latest convolutional neural networks (CNN), which will identify the objects from the camera and output as the description in a sentence, the object names, and the object location on the image. The navigation part is processed on a smartphone, which will provide a way to get to the destination based on the existing global positioning system (GPS) on the smartphone. The voice interaction part is processed on a smartphone and using an existing speaker and microphone, which is using offline voice recognition and text-to-speech technology to have interaction for the visually impaired user.

[4] https://play.google.com/store/apps/details?id=com.hkblindunion.smartcitywalk.android&hl=zh_HK.

[5] https://www.hkbu.org.hk/.

[6] https://wewalk.io/en/.

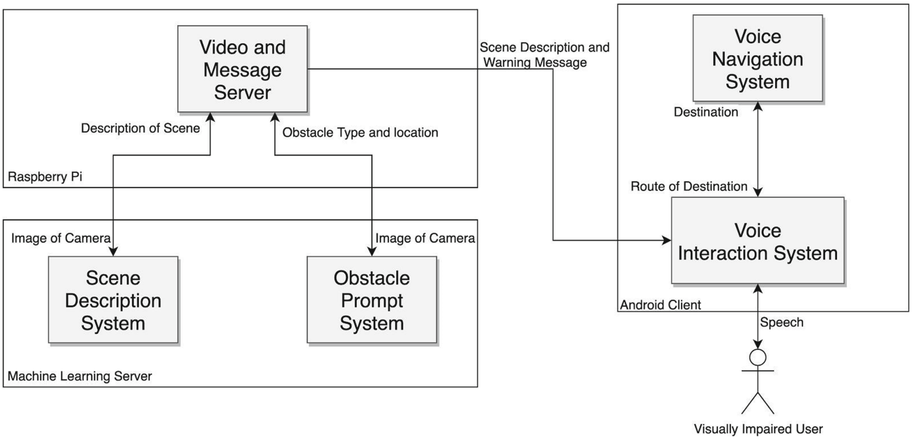

Fig. 1. Components of the system.

The application is supported by client and server devices as shown in Fig. 1 and Fig. 2. The operations of each component are listed below.

1. Mobile device (Android client)
   - Perform voice interaction with the user.
   - Return navigation information upon user request.
   - Use MQTT[7] clients to connect the Raspberry Pi and prompt the user when dangerous information or scene description is received.
   - Open a WIFI hotspot for Raspberry Pi connection.
2. Raspberry Pi (Operation component)
   - Connected to Raspberry Pi WIFI hotspot.
   - Host an MQTT server for data transferring between Android client.
   - Be a hands-free camera device that can easily be installed on the walking stick or worn on the neck since the user may have to pick up the walking stick.
   - Host a video streaming server which will do motion detection and serve images to clients by MJPEG.
3. Machine Learning Server (Server)
   - Serve the request through Flask[8] framework.
   - Process the request of object detection and image caption separately due to the inference time of deep neural networks (DNN) is relatively large.
   - Response the object name and relative direction from the object detection server when the confidence of the object in the image is high enough and the object may be dangerous to users.
   - Return the description in understanding sentences from the image caption server.

[7] http://mqtt.org/.

[8] https://flask.palletsprojects.com/en/1.1.x/.

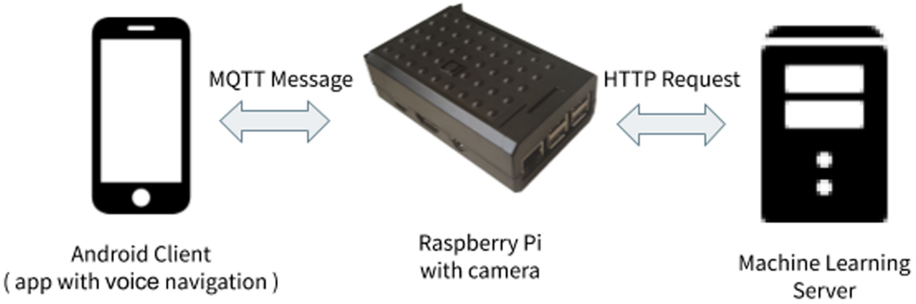

Fig. 2. Hardware components of the application system.

4 The System Implementation

4.1 Overview of the Mobile Application System

The application uses both hardware and software to achieve the goal. The device can be mounted into the walking stick or mounted into the neck ring. In terms of software, the user interface will be a mobile application that connects to the hardware (Raspberry Pi) via WIFI and searches Raspberry Pi using multicast DNS. This program is mainly responsible for interacting with the user. After the program starts, it will connect with the hardware and then start obstacle recognition, environment description, and map functions at the same time. The user can activate the speech recognition function through buttons to give instructions to the system, such as navigation. The system will convert the results of the three functions into speech, as shown in Fig. 3 based on the priority order as shown in Fig. 4.

Figure 4 shows the judgment process when the text-to-speech function receives instruction. This process ensures the system reads one type of information only to the user at the same time without overlapping. The system will handle the obstacle detection at first because it will directly affect the safety of the user if the user is close to the obstacle. After that, the user will receive the navigation instruction when the user arrives at the intersection, and the navigation system will suitably prompt the user. In general, the system will describe the surrounding environment for the user, but it will only describe the environment when the message is received within five seconds, avoiding to describe the environment that has passed.

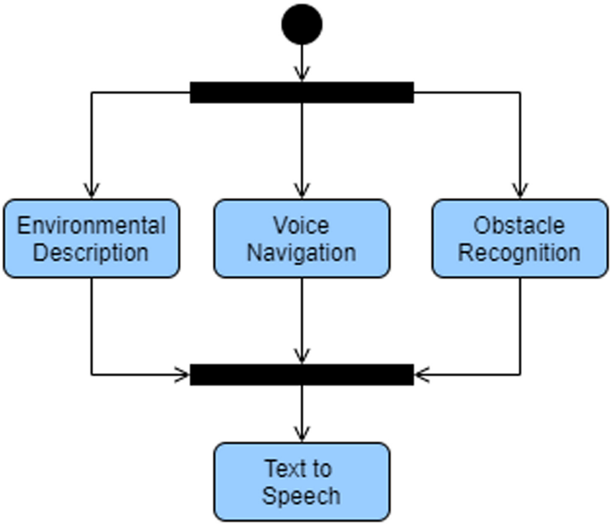

Fig. 3. Active diagram of the application.

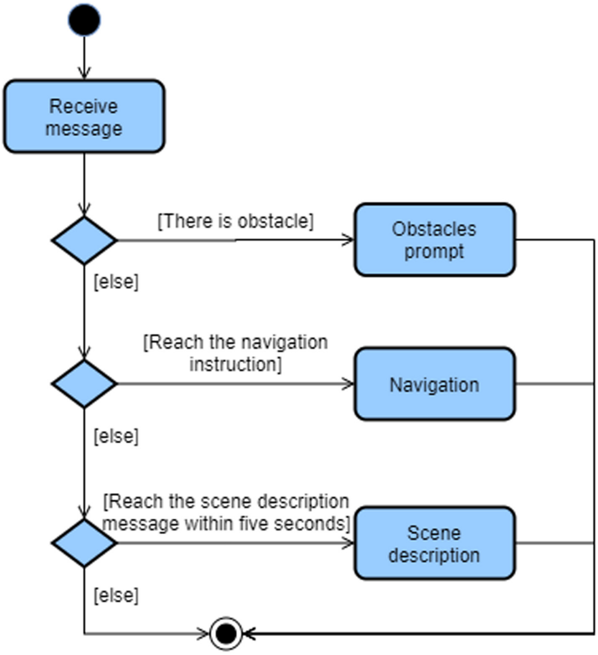

Fig. 4. Priority order for the text to speech function.

4.2 Voice Navigation

Voice navigation aims to make map navigation easier for the visually impaired people through the voice recognition function that allows users to confirm the destination by voice instead of typing through the keyboard. To achieve the goal, the cloud speech recognition service provided by Google[9] is applied, which is built into all Android devices by default. Through the cloud technology, it can convert the speech from the user into text within two seconds, and it has an automatic debugging function to revise the possible errors in speech recognition.

This feature uses dialog flow to analyze and filter the voice input by the user, such as the user saying "Take me to the Open University of Hong Kong" or "Go to the Open University of Hong Kong" and dialog flow responds the action "navigation" and the destination "The Open University of Hong Kong".

Three-Dimensional (3D) Direction Prompt. To let the visually impaired users know the correct direction, rather than the turn left or turn right of a traditional navigation service like Google Map, the 3D direction prompt was implemented. 3D direction prompt applies the latitude and longitude of the user and the next navigation point to calculate the correct direction. Then, it determines whether the user is oriented in that direction by using a gyroscope on the mobile phone. Figure 5 shows that the Mapbox[10] API will return the route and the next navigation point on the route (redpoint). The yellow-green-blue circle represents the user, where blue and the red line represents the user, where blue and the red line represents the direction the user is oriented in heading to the place.

Figure 6 shows that the user is facing the wrong direction since the right (or left) direction is the yellow line, but the face direction of the user is the red line. When the yellow line falls into the green (or yellow) area, the right (or left) channel of the earpiece will alert the user that turns back to the correct direction.

The volume of the left or right channels of the 3D direction prompt is linear as shown in Fig. 7, a reminder and different levels of "beep" sound in left-right channels to navigate the user back to the correct path. To avoid excessive volume, the maximum volume is set to 50, as shown in Fig. 8. Since it is impossible for the user exactly facing the correct direction, there is a buffer of 50° in the correct direction to prevent the user from still hearing the beep when facing the correct direction. To keep the user facing the correct direction, the system uses beep sound on the left-right channels and the small-large volume to inform the user.

[9] https://cloud.google.com/speech-to-text.

[10] https://www.mapbox.com/.

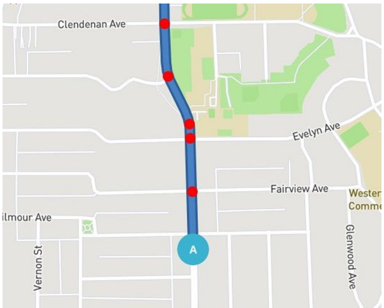

Fig. 5. A sample output of the Mapbox route API service. (Color figure online)

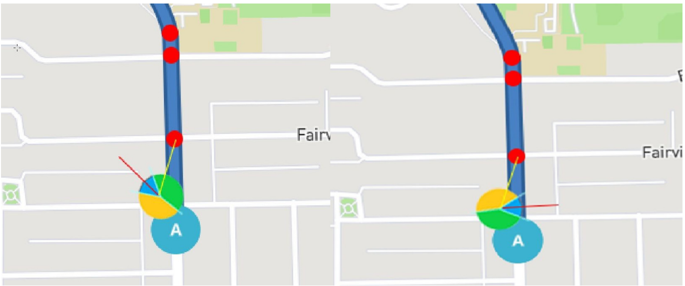

Fig. 6. A sample output of the user facing the wrong direction. (Color figure online)

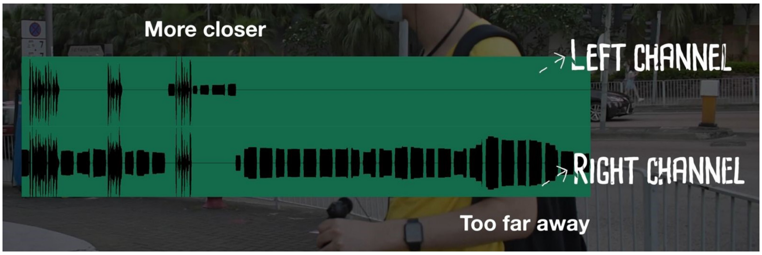

Fig. 7. Volume track for the direction indicator.

4.3 Obstacles Prompt and Scene Description

The design of the obstacles prompt and scene description system as shown in Fig. 9. Fast detection of obstacles is a definite criterion for the users to avoid any dangerous situation. To provide the accurate information to the users, the captioned function includes the following procedures:

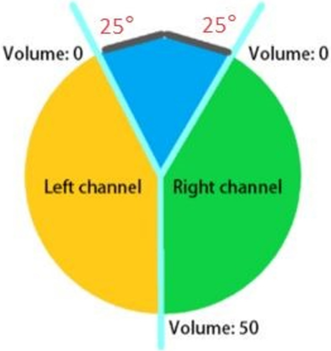

Fig. 8. The 3D direction prompt design.

Step 1. Capture the scene of the road.
Step 2. Filter irrelevant information and adjust the confidence of objects.
Step 3. Sort out the obstacles based on depth estimation algorithm.
Step 4. Illustrate the results to the user with voice.
Step 5. Repeat Step 1 to Step 4 to provide the updated information.

Figure 10 shows a sample image after depth estimation. For example, if an obstacle is detected as shown in Fig. 11, the obstacle prompt will give an immediate warning to the user with the scene description in voice. By post-processing the result of object detection, users will be prompted with the upfront obstacles. Scene description using image caption technology can produce a humanized text description, which can help the users know the scenes ahead. For example, a possible scene description can be "a group of people standing around a bus stop." The processes are similar to obstacle detection, while the application is kept responding to the scene captured from time to time.

Scene description using image caption technology can produce a humanized text description, which can help the visually impaired to know what the world is. For example, a possible scene description can be "a group of people standing around a bus stop." The processes are similar to obstacle detection, while the application is kept responding to the scene captured from time to time.

When the user faces the wrong direction at first, the user is prompted by the voice "You are in the wrong direction, please adjust the facing direction according to the beep sound". And when the user is facing the right direction, the system will also use the voice prompt "You are now in the right direction".

To describe the obstacles ahead, the system will detect the objects on the captured images from the camera as fast as possible.

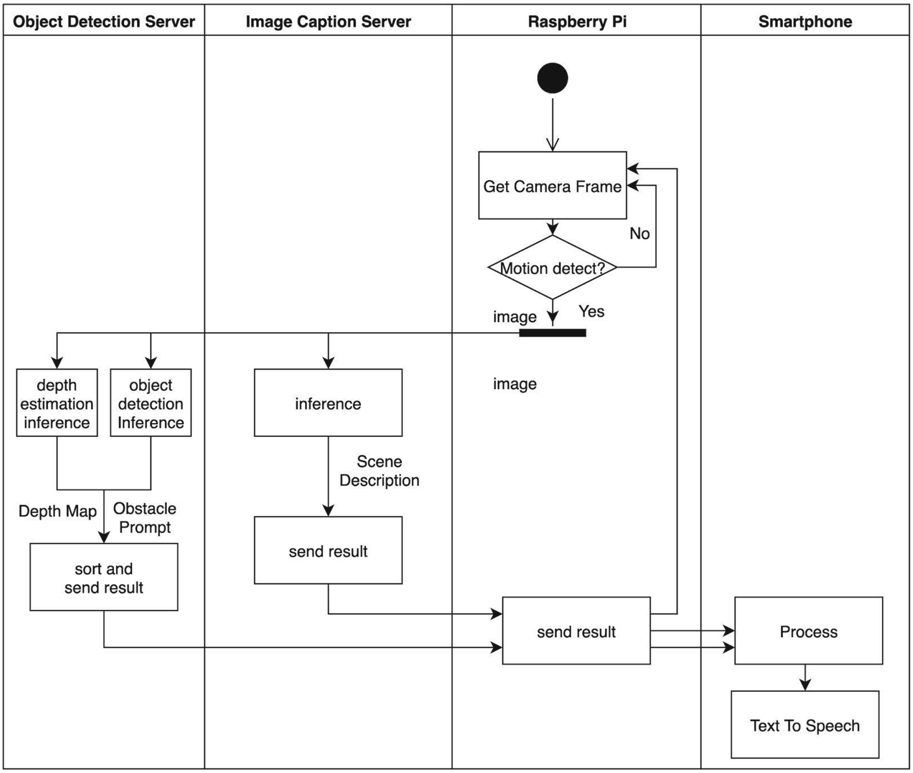

Fig. 9. Obstacles prompt and scene description design.

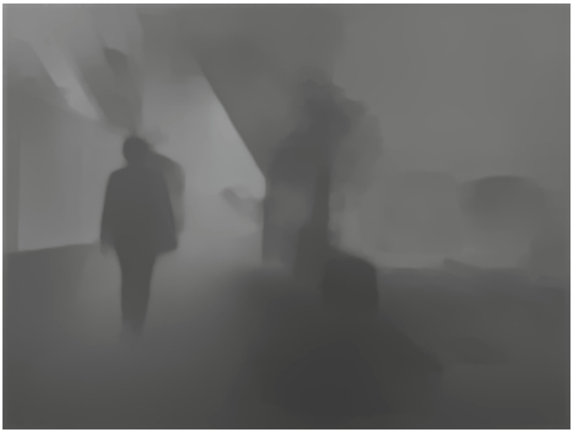

Fig. 10. Image after depth estimation.

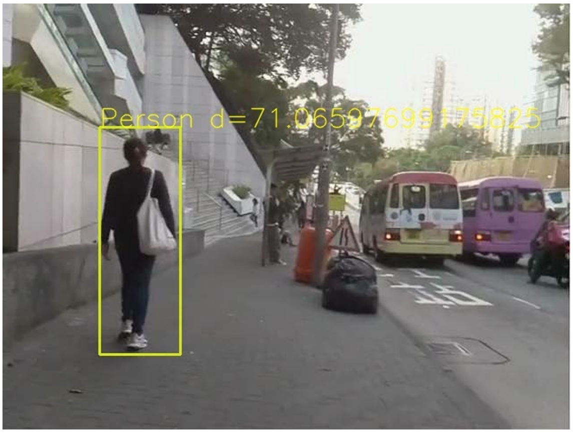

Fig. 11. Image after obstacle detection.

5 Evaluation Results

Observation tests will be used to assess whether the solution helps the users go out and reach the destinations efficiently and safely. A number of the visually impaired people are invited to the observation test by using the usual way, and the application proposed in this paper to walk from Fat Kwong Street (at Homantin) to The Open University of Hong Kong. It will be evaluated by counting the number of objects and people that collided with the users, and the time taken for the users reach the destination. Table 1 shows the average results of user evaluation.

Table 1. Average result of user evaluation.

	Using the usual way	Using our solution
Average number of times colliding with objects	23.5	15.5
Average number of times colliding with people	5	0.5
Time to find the correct direction (in seconds)	8.5	4
Time taken to reach the destination (in minutes)	3.4	2.5

According to the results, the number of object collisions has been reduced by about one-third (from 23.5 to 15.5), and the number of human collisions has dropped significantly to one-tenth (from 5 to 0.5). The data indicated that the use of the new application can reduce the chance of accidents and thus improve the user safety during outdoor activities. Besides, the users can reach the destination more efficiently (from 3.4 minutes to 2.5 minutes) when the proposed application with 3D direction prompts is used.

6 Conclusions

This paper introduces a mobile application as an outdoor assistant to help people with low vision go outside efficiently and safely. On this basis, a real-time environment description system has been developed to let the users know the surrounding in front of them through the voice description. An obstacle prompt system has been implemented for avoiding the users to collide with any object in front of them during the journeys. The 3D direction prompts of the navigation system have been designed to allow the users to recognize the correct direction by using sound from the left-right channels. From the evaluation results, it is shown that the use of the proposed mobile application can assist the people in need with visual disability to travel outside in a safe and efficient way.

References

1. Vos, T., et al.: Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet **388**(10053), 1545–1602 (2016)
2. Statistics on People with Visual Impairment. https://www.hkbu.org.hk/en/knowledge/statistics/index
3. Image Description-Computer Vision - Azure Cognitive Services. https://docs.microsoft.com/zh-tw/azure/cognitive-services/computer-vision/concept-describing-images
4. Pricing-Computer Vision API. https://azure.microsoft.com/en-gb/pricing/details/cognitive-services/computer-vision/
5. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)
6. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural imagecaption generator. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015). https://doi.org/10.1109/cvpr.2015.7298935
7. Graves, A., Navdeep, J.: Towards end-to-end speech recognition with recurrent neural networks. In: International Conference on Machine Learning, pp. 1764–1772 (2014)
8. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
9. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
10. Alhashim, I., Wonka, P.: High quality monocular depth estimation via transfer learning. arXiv preprint arXiv:1812.11941 (2018)
11. Newman, N.: Apple iBeacon technology briefing. J. Direct Data Digit. Mark. Pract. **15**(3), 222–225 (2014). https://doi.org/10.1057/dddmp.2014.7