A Wearable Device for Indoor Imminent Danger Detection and Avoidance With Region-Based Ground Segmentation

ZHONGEN LI, FANGHAO SONG, (Graduate Student Member, IEEE), BRIAN C. CLARK, DUSTIN R. GROOMS, AND CHANG LIU, (Senior Member, IEEE)

1School of Electrical and Computer Engineering, Ohio University, Athens, OH 45701, USA
2Ohio Musculoskeletal and Neurological Institute, Ohio University, Athens, OH 45701, USA
3Department of Biomedical Sciences, Ohio University, Athens, OH 45701, USA
4College of Health Sciences and Professions, Ohio University, Athens, OH 45701, USA

Corresponding author: Chang Liu (liuc@ohio.edu)

ABSTRACT

Avoiding objects independently in indoor environments for individuals with severe visual impairment is one of the significant challenges in daily life. This paper presents a wearable application to help visually impaired people quickly build situational awareness and traverse safely. The system utilizes Red, Green, Blue, and Depth (RGB-D) camera and an Inertial Measurement Unit (IMU) to detect objects and the collision-free path in real-time. A region proposal module is presented to decide where to identify the ground from 3D point clouds. The segmented ground area can act as the traversable path, and its corresponding region in the image is removed to prevent detecting painted objects. The system can provide information about the category, distance, and direction of the detected objects by fusing the depth image and the neural network results. A 3D acoustic feedback mechanism is designed to improve the situational awareness for visually impaired people, and guild them traverse safely. The advantage of this system is that our 3D region proposal module can robustly propose the potential ground region and greatly reduce the computation cost of the ground segmentation. Besides, a typical machine-learning-based approach may miss objects because they could not be recognized, though they still may pose a danger. Another advantage of our approach is that the imminent danger detector can detect such unrecognizable objects to help users avoid a collision. Finally, experimental results demonstrate that the proposed system can be a useful indoor assistant tool to help blind individuals with collision avoidance and wayfinding.

INDEX TERMS

Convolutional neural network, ground segmentation, object detection, point cloud, region of interest, visual impairment, wearable assistive system.

I. INTRODUCTION

According to the World Health Organization (WHO) report [1], there are no less than 2.2 billion people with visual impairment or blindness worldwide. Severely visually impaired individuals (i.e., those with best-corrected visual acuity worse than 20/200 in the best-seeing eye or those with less than 20 degrees of the functional visual field) face many challenges and risks in navigating their environments. These challenges include performing autonomous navigation and environmental perception in unfamiliar indoor environments. When entering complex buildings such as large shopping malls, university buildings, or hospitals, one must perceive the surrounding environment and avoid obstacles to reach the destination safely. For the blind and visually impaired, accurate object detection and timely feedback are required to help them complete indoor activities. Therefore, these challenges add to the classic problems of spatial navigation and obstacle avoidance.

Traditionally, visually impaired people usually rely on navigational aids, namely guild dogs [2] and canes [3] for help when detecting obstacles. However, they may still suffer injuries from hanging obstacles due to the limited detecting range of white canes and guide dogs. To improve mobility and safety for visually impaired people, advanced sensing techniques have adopted (e.g., ultrasonic sensors [4], [5],
visual sensors [6], [7], and laser emitters [8]). However, ultrasonic sensors are poor at target identification or localization. Besides, laser emitters are expensive, heavy, and require high power consumption, which makes them unsuitable for wearable and affordable applications. Vision-based assistants, such as monocular cameras, stereoscopic cameras, and Red, Green, Blue, and Depth (RGB-D) cameras, have been widely used for target finding, localization, navigation, etc. For example, an important function of vision-based applications is to help visually impaired users find specific objects to use, such as chairs [9], sign patterns in road environments, markers posted on podiums and classroom doors [10], up/down stairways [11]–[14], and elevators [15]. For other applications, like visual simultaneous localization and mapping (VSLAM) for blind navigation has become a popular research field [16]–[19] as well. Because these applications assume object avoidance technology has been adopted, the visually impaired can reach their destinations independently and safely. Therefore, a wearable real-time object avoidance system is needed to assist the visually impaired in navigation and environmental perception.

With the rapid and widespread development of sensor technology, RGB-D cameras become affordable and portable. They can directly provide three-dimensional environmental information without the need for computational depth calculations required by RGB cameras. For the RGB-D based object detection system, numerous studies have been conducted in the past years [20]–[23]. However, indoor assistive object detection still faces many challenges: low-cost and effective obstacle recognition and avoidance, real-time and rich feedback, and the complexity of a holistic system on compact portable devices for visually impaired users. This paper proposes a wearable device for indoor object detection and avoidance system with region-based ground segmentation to improve mobility for visually impaired people. The proposed system utilizes an RGB-D camera as the perceptual sensors to extract information about the surroundings, such as the available free path, and the types and directions of identified obstacles. An inertial measurement unit (IMU) is combined to help track the real-time orientation of the camera in use. Through complementary fusion, more accurate and richer environmental information can be obtained from point clouds and images.

All calculations are completed locally in real-time. Hence no internet connection is needed, which might have privacy and access issues. The solution provides 3D sound feedback through speakers to offer accurate situational awareness and imminent danger alerts for the visually impaired. With a miniaturized design, as shown in Fig. 1, the system can reduce the risk of catastrophic injury, thereby improving indoor mobility and quality of life.

The main contributions of this paper are:
- A region-based ground segmentation can robustly segment the ground and reduce calculation time;
- Complementary data fusion from point clouds and images to obtain accurate and informative detection results;
- A 3D acoustic feedback mechanism optimized for different priorities to improve the situational awareness for blind people.

The rest of the paper is structured as follows. Section II reviews various RGB-D based object recognition systems, ground segmentation approaches, and obstacle avoidance methods. Section III elaborates on the principle components of the proposed framework. After that, Section IV shows the experimental evaluations and discussions. Finally, the conclusions, along with future work, are discussed in the end.

II. RELATED WORK

The usage of robotic and autonomous driving techniques has been widely approached to aid safe navigation for visually impaired people. This research builds on important related work to object recognition systems, ground segmentation approaches, and local obstacle detection solutions.

A. RGB-D BASED OBJECT RECOGNITION SYSTEMS

Since RGB-D based devices have advantages over other sensing devices as discussed above, in this section, we focus on object recognition systems with RGB-D sensors.

Kayukawa et al. [20] developed a vision-sensing system, BBEEP, for clearing the path in front of a blind user. This work was able to detect pedestrians and predict their future positions based on consecutive frames. By using 3D odometry API, the system could remove the influence of the suitcase rotation in real-time. However, this system is not suitable to carry out for blind users because it was implemented based on a suitcase.

Pradeep et al. [24] proposed a wearable navigation system with obstacle detection and avoidance functionalities. Obstacles were detected by scanning 2D grids neighborhood.
quantized from the 3D point cloud. The system used haptic feedback to alert users when obstacles were detected in front. However, the system lacks obstacle modeling to recognize the category of objects. While object detection is only considered during the path planning phase, the system does not support real-time obstacle avoidance.

Li et al. [21] proposed an intelligent situational awareness and navigation aid, which was based on a Google Project Tango device to help visually impaired people traverse indoor area independently. Dynamic planning was performed by real-time perception with an RGB-D camera. Based on a time-stamped occupancy map, real-time object detection was implemented through two 2D projections. This method has a lower cost but less information. For example, the type and shape of objects cannot be determined. The detection of long-range ground and obstacles for visually impaired people was presented in [25]. Although the proposed system improves the pathfinding task for the blind and visually impaired, it also does not provide any category information of the detected objects in the user’s surroundings.

B. GROUND SEGMENTATION FROM POINT CLOUDS

Ground segmentation is the first step in object detection because it aims to distinguish between viable ground areas and obstacles. Besides, visually impaired people can rely on the detected ground as the traversable road areas and search the optimal route [31].

Imai et al. [26] proposed an approach to identify the walkable path by using an RGB-D camera and an accelerometer. Their work estimated the initial orientation by decomposing the gravity on the tri-axis. Besides, the ground height and normal vectors were taken into consideration when detecting ground. However, the ground height calculation by one depth column is error-prone, if an obstacle happened to exist in there.

The Random Sample Consensus (RANSAC) is an iterative method and can be used for robustly estimating plane parameters from a set of data points. Takizawa et al. [9] used the RANSAC algorithm to extract the wall and ground for helping blind people find an available seat. Bai et al. [27] applied the RANSAC for coarse ground detection. Then a time-dependent adaptive ground height segmentation algorithm among adjacent frames was adopted to refine the ground result. In [28], the traversable area preliminarily was obtained using the RANSAC segmentation. Then, the preliminary traversable area was further enlarged by a growing region algorithm. However, these works treat the camera as static, and real-time orientation tracking is not considered.

3D Hough transform (HT) for plane detection from point clouds were proposed in [32]. HT describes every plane by its slope. The points in the parameter space will correspond to planes detected in the object space. After processing is finished, the accumulator in the parameter space will search the satisfying point, which corresponds to the shape of the interested plane in the original space. Although HT is less accurate than the RANSAC, it needs less time to detect planes [33].

There are other methods, like using graph descriptor, surface normal, or color information, to detect ground from the point cloud. Graph-based ground segmentation was proposed using a spatial mesh [34]. Aladren et al. [22] used both depth and color images to detect the ground. Although their system achieves high ground detection accuracy, it is too computationally expensive to use in real-time. Holz et al. [35] calculated the local surface normal on the integral image. Then, these points were clustered and segmented in the normal space.

C. OBSTACLE AVOIDANCE APPROACHES

Lee et al. [23] created a 2D probabilistic occupancy map to avoid obstacles in real-time during assistive navigation. The system built a 3D voxel map of the environment to analyze the traversability. Then, a 2D traversable grid map was used to support dynamic path planning. Point cloud clustering for Object detection were presented in [21], [29]. First, the detected ground was removed, and then the Euclidean segmentation was performed on the remaining points to cluster the objects. If the distance between two points was less than the threshold, they were merged. Finally, the algorithm’s output was a set of object clusters, where each cluster was a collection of points of the same object. However, these methods are highly dependent on the size threshold of the merged objects and cannot provide category information to which the objects belong.

Object recognition based on machine learning methods has become increasingly popular. Tapu et al. [30] proposed a DEEP-SEE system to detect both dynamic and static objects using the You-Only-Look-Once (YOLO) object recognition method [36]. The system can provide the obstacle category and distance information. However, this neural network needs heavy computation overhead, making it hard to fulfill real-time detection on mobile devices. Lightweight image-based neural networks have proposed to increase the accessibility of real-time object detection to mobile devices, such as MobileNet [37], YOLO-LITE [38], ShuffleNet [39] etc. However, these lightweight algorithms do not consider distance information directly when performing object recognition. So, the painted object on the ground might be detected to mislead the visually impaired people.

To overcome the shortcoming of lightweight image-based algorithms, Bai et al. [27] proposed a 2.5-D object detection, which combined MobileNet object detection and depth image based object detection. The intersection between their mapped areas was used to improve the detection result. However, refinement by intersection may miss objects because they could not be recognized by a machine-learning-based approach even though they still may pose a danger.

Table. 1 summarizes the functions and limitations of different assistive systems for visually impaired people. Compared with these works, our system is versatile. It can not only dynamically track the orientation and altitude of the camera...
TABLE 1. Comparison of different object detection and avoidance systems. ✓ := supported, × := unsupported, ∼ := unclear from the publication.

| Sensing Devices          | Wearable | Real Time | Orientation Tracking | Height Estimation | Ground Segmentation | Object Recognition | Feedback     |
|--------------------------|----------|-----------|----------------------|-------------------|---------------------|-------------------|--------------|
| Kayukawa et al. [20]     | RGB-D    | ×         | ×                    | ×                 | ×                   | ×                 | Acoustic     |
| Pradeep et al. [24]      | Stereo Camera & IMU | ✓         | ✓                    | ×                 | ×                   | ×                 | Acoustic     |
| Li et al. [21]           | RGB-D & IMU | ✓         | ✓                    | ✓                 | ×                   | ×                 | Acoustic     |
| Yang et al. [25]         | RGB-D & IMU | ✓         | ✓                    | ×                 | ×                   | ×                 | Bone-Conduction |
| Imai et al. [26]         | RGB-D & Accelerometer | ~         | ~                    | ✓                 | ✓                   | ×                 | ×            |
| Takizawa et al. [9]      | RGB-D    | ×         | ~                    | ×                 | ×                   | ✓                 | Acoustic & Haptic |
| Bai et al. [27]          | RGB-D & IMU | ✓         | ✓                    | ×                 | ×                   | ✓                 | Acoustic     |
| Yang et al. [28]         | RGB-D & IMU | ✓         | ~                    | ✓                 | ✓                   | ×                 | Bone-Conduction |
| Aladren et al. [22]      | RGB-D    | ✓         | ~                    | ×                 | ✓                   | ×                 | Acoustic Tactile |
| Lee et al. [23]          | RGB-D & IMU | ✓         | x                    | ✓                 | ✓                   | ×                 | & Acoustic Braille |
| Wang et al. [29]         | RGB-D    | ✓         | ✓                    | ×                 | ×                   | ✓                 | & Acoustic Bone-Conduction |
| Tapu et al. [30]         | Smart-phone | ✓         | ✓                    | ×                 | ×                   | ×                 | ✓            |
| Ours                     | RGB-D & IMU | ✓         | ✓                    | ✓                 | ✓                   | ✓                 | Acoustic     |

FIGURE 2. The workflow of the proposed system.

to segment the ground surface, but also provide real-time avoidance of obstacles using a portable device.

III. THE PROPOSED FRAMEWORK

As shown in Fig. 2, the proposed system first obtains RGB and depth images from an RGB-D camera and gets acceleration and angular velocity from the IMU. Then, the real-time orientation and height of the camera are estimated by the enhanced depth image and IMU data. The system lifts RGB images and depth maps to 3D point clouds and reconstructs the actual 3D scene based on the orientation and height tracker. Then the potential ground area’s region of interest (ROI) is proposed, thus the ground plane can be segmented effectively to search for available paths and objects are detected over the ground. To further increase the stability, a feedback calibration for orientation and height estimation is presented. The final objects with type, distance, and direction are calculated by fusing the output of the real-time CNN object detector and the corresponding depth image. To enhance system reliability, an imminent danger detector is implemented to capture the unrecognizable objects. The region corresponding to the ground area from the point cloud is removed from images before feeding into the CNN object detector, which reduces false-positive results such as painted objects on the ground. The system provides informative 3D acoustic feedback to users according to potential hazards and the feasible path that it has identified. The main algorithms of the proposed framework are described in detail in the following subsections.

A. DEPTH PRE-PROCESSING

The raw depth data output by the RGB-D camera may have “holes” and noise, which negatively affects system performance. Therefore, the pre-processing on depth is necessary to get better performance of the point cloud. This module comprises a decimation filter, a spatial filter, and a temporal filter to reduce data amount, fill holes, and improve noise while preserving features.

Since the amount of data plays a crucial role in application performance, downsampling could reduce the density of points in very dense areas, but has little effect on sparse areas. Non-zero downsampling is performed by taking the average of neighbor pixels while ignoring zeros. The bilateral filter is used to smooth depth noise while preserving edges by calculating the one-dimensional exponential moving average (EMA). The recursive equation of EMA is represented as:

\[
S_t = \begin{cases} 
Y_t & t = 1 \\
\alpha Y_t + (1 - \alpha)S_{t-1} & t > 1 \text{ and } S_{t-1} - S_t < \delta \\
Y_t & t > 1 \text{ and } S_{t-1} - S_t > \delta 
\end{cases}
\]
where the coefficient $\alpha$ represents the degree of the weighting factor. $Y_t$ is the instantaneous value, $\delta$ is the threshold, and $S_t$ is the value of the EMA at time $t$. In a stereoscopic system, holes exist when either data being unavailable or not having met a confidence metric (e.g., occlusion, lack of texture, etc.). In this situation, the camera sets the depth of these holes to zero instead of providing the wrong value. As shown in Fig. 3 (b), the holes are displayed as black dots. The method of filling holes is a combination of spatial and temporal filters. The holes are filled first with valid adjacent pixels within the specified radius in the same frame. Besides, if holes still occur, they are filled with the last valid (if they exist) value by looking at several frames back in history.

**TABLE 2. Parameters of the depth pre-processing module.**

| Filters   | Parameters                        |
|-----------|-----------------------------------|
| Decimation| sub-sample = 5                    |
| Spatial   | EMA: $\alpha = 0.25$, $\delta = 20$ |
| Temporal  | number of adjacent pixels for hole filling = 2 |

Based on the recommendation from [40], the RealSense D435i camera is streaming depth data at 840 by 480 resolution and with the infrared sensor on to have the best quality of the depth image. In the depth pre-processing module, the parameters of filters are listed in Table 2. The raw depth image is shown in Fig. 3 (b), which has noises and holes. After processing by decimation filter, spatial filter, and temporal filter, the quality of the depth images is improved with few or no holes, while preserving features, shown in Fig. 3 (c).

**B. GROUND REGION PROPOSAL**

Since the point cloud has the feature of 3D Euclidean measurements in front of the camera, it is essential to calculate the relative position between the camera and the object to understand the absolute position of objects. We provide real-time orientation and altitude trackers to help calculate the relative position. In our implementation, we only consider the pitch and roll angles. Then, the scene is reconstructed based on the relative position. Besides, the potential ground area is suggested based on the height of the plane in the reconstructed 3D scene.

1) **ORIENTATION AND HEIGHT TRACKER**

Orientation and height estimation within the world coordinate space play a principal role in proposing the region of interest. As shown in Fig. 4, the camera is centered at the right-handed Cartesian coordinate system, where the positive directions of the X, Y, and Z axes are defined as the left, up, and facing directions of the camera. We use $A_{X,OUT}$ as the acceleration along the X-axis, subsequently $A_{Y,OUT}$, and $A_{Z,OUT}$ are accelerations in Y and Z axes. The initial orientation of the camera is calculated by decomposing gravity $g$ from three-axis accelerations, which is represented as:

$$\text{Pitch} = \tan^{-1}(\frac{\sqrt{A_{Z,OUT}^2 + A_{Y,OUT}^2}}{A_{X,OUT}})$$

$$\text{Roll} = \tan^{-1}(\frac{\sqrt{A_{Y,OUT}^2 + A_{Z,OUT}^2}}{A_{X,OUT}})$$

In theory, the real-time orientation can be gauged through integrating the output of the gyroscope. However, the estimation results suffer from the integration of drift over time. To help fix the time drift of the gyroscope, we use the complementary filter to allow short-duration signals from the gyroscope to pass through while blocking steady signals. The noise and horizontal acceleration dependency are also alleviated by the complementary filter, which is formulated as:

$$\text{Angle}(t) = \alpha\text{Angle}(t) + \int \text{gyro}(t) dt + (1 - \alpha)\text{acc}(t)$$

where $\alpha$ is the filtering weight, $\text{gyro}(t)$ is the output of the gyroscope at time $t$, and $\text{acc}(t)$ is the output of the accelerometer at time $t$.

The only knowledge about the camera pose is the approximate position of the camera on the chest because the subject’s height and movement are continually changing. The initial camera height estimation is equivalent to calculating the distance from the camera position to the ground. The RANSAC program is executed in the initial system setup to search the ground. During the system setup, the camera is facing down statically, and an additional pass filter on the y-axis can be used to limit the search of the ground plane. Repeat the process until the floor is detected. Since the user should stand on the floor, this process will not last long.
Once the ground plane is determined, it can be represented as (5), where A, B, C, and D are the coefficients of the ground plane. From the point-plane distance, the height of the camera \( H_{\text{camera}} \) can be computed as (6).

\[
Ax + By + Cz + D = 0
\]

\[
H_{\text{camera}} = \frac{D}{\sqrt{A^2 + B^2 + C^2}}
\]

2) 3D RECONSTRUCTION

As long as the orientation and the height of the camera are estimated, the 3D scene can be reconstructed. The rotation angles of interest are pitch angle \( \theta \) and roll angle \( \alpha \) corresponding to the x-axis and the z-axis. The rigid transformation from the camera coordinate system to the world coordinate system is realized by a combination of a rotation matrix \( R \) and a transition vector \( t \). The rotation matrix is used to make the normal vector \( n \) parallel to \((0, 1, 0)\), and the transition matrix is used to move a point to the origin. The affine transformation matrix \( T \) in the homogeneous coordinate is represented as:

\[
T = \begin{bmatrix} R & t \\ 0^T & 1 \end{bmatrix} = \begin{bmatrix} R'_x & R'_z \\ 0^T & 1 \end{bmatrix} = \begin{bmatrix} \cos \theta & -\sin \theta & \sin \theta \sin \alpha & 0 \\ \sin \theta & \cos \theta \cos \alpha & -\cos \theta \sin \alpha & 0 \\ 0 & \sin \alpha & \cos \alpha & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}
\]

where:

\[
R'_x = \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}
\]

\[
R'_z = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \alpha & -\sin \alpha \\ 0 & \sin \alpha & \cos \alpha \end{bmatrix}
\]

\[
t = \begin{bmatrix} -H_{\text{camera}} \\ 0 \end{bmatrix}
\]

After the original point cloud is reconstructed based on the real-time orientation and height trackers, the absolute position of objects in the scene is obtained. Since the ground has the following characteristics: parallel to the XZ plane and the intersection point on the y-axis is 0, a band-pass filter is used to propose a ground ROI along the y-axis, for example, from \(-0.1\) meters to 0.1 meters. The advantage of searching the ground in the ROI is that it dramatically reduces the 3D search space for the RANSAC algorithm and saves the calculation time.

C. GROUND SEGMENTATION

Ground segmentation is essential to understand the surrounding environment because the segmented ground can be severed as the three purposes:

- Wayfinding for users and use the segmented ground area as the free movable area;
- The parameters of the ground plane can be used to calibrate the orientation and height tracker;
- Applying object detection over the segmented ground.

Algorithm 1 Ground Segmentation on Point Cloud Using RANSAC Algorithm

\textbf{Input:}

\( c \): input point cloud
\( n \): normals of the input point cloud
\( \text{threshold} \): distance threshold to the model
\( \text{axis} \): surface normal for plane estimation
\( \theta_{th} \): the maximum allowed difference between the model normal and the given axis
\( M \): max iterations

\textbf{Output:}

\( \text{inliers} \): the point list of the segmented plane
\( \text{bestPlane} \): the coefficients of the segmented plane

\textbf{procedure} \text{RANSAC}(c, n, \text{threshold}, \text{axis}, \theta_{th}, M)

\textbf{initialization:}

\( i \leftarrow 0, \text{bestSupport} \leftarrow 0; \)
\( \text{bestPlane} \leftarrow [0, 0, 0, 0], \text{bestStd} \leftarrow \infty; \)

\textbf{while} \( i < M \) \textbf{do}

\( j = \text{sample} \; 3 \; \text{points} \; \text{randomly} \; \text{among} \; c; \)
\( \text{plane} = \text{points2plane}(j); \)
\( \text{diff} = \text{angle}(\text{plane}, \text{axis}); \)

\textbf{if} \( \text{diff} < \theta_{th} \) \textbf{then}

\( \text{dis} = \text{dis2plane}((\text{plane}, c) \)
\( \text{inliers} = \text{find}(\text{abs}(\text{dis}) <= \text{threshold}) \)
\( \text{st} = \text{standard deviation}(\text{inliers}) \)

\textbf{if} \( \text{length}(\text{inliers}) > \text{bestSupport} \) \text{or} \( \text{(length}(\text{inliers}) = \text{bestSupport and} \)
\( \text{st} < \text{bestStd}) \) \textbf{then}

\( \text{bestSupport} = \text{length}(\text{inliers}); \)
\( \text{bestPlane} = \text{pl}; \)
\( \text{bestStd} = \text{st}; \)

\textbf{else}

\( i = i + 1; \)

\textbf{return} \text{bestPlane}

1) GROUND SEGMENTATION

To segment the ground, we adopt the RANSAC algorithm by adding additional constraints to the ROI out of the region proposal module. Algorithm 1 specifies the pseudo-code of the RANSAC algorithm for ground segmentation. The segmentation starts from randomly choosing a set of 3D points to estimate the parameters A, B, C, and D in (5). The function points2plane calculates the distance from a point to the
estimated plane. The remaining points are evaluated through (9) to count the number of inliers (T is the distance threshold). After certain iterations, the ground plane is determined, which has the most inlier points. The number of iterations is represented by k, as shown in (10).

\[
\frac{|Ax + By + Cz + D|}{\sqrt{A^2 + B^2 + C^2}} < T \quad (9)
\]

\[
k = \frac{\log(1 - p)}{\log(1 - \text{w}^n)} \quad (10)
\]

where \(1 - \text{w}^n\) is the probability that at least one of the n points is failing the computation of outliers, \(1 - p\) is the probability that is independently selecting a set of n points are not all inliers. The maximum number of iterations can be set to stop the process.

In the sampling stage, some sampling errors are eliminated by setting the tilt angle threshold (as in (11)) of the ground plane. The perpendicular plane model is adopted to restrict the wall-like vertical plane by using the angle function to check the angle between the normal direction of the detected plane and a given axis. Even though the proposed ROI restricts the vertical height range to eliminate the error-prone planes (e.g., table surfaces), the distance of the detected plane from the origin is checked to make detection results more reliable.

\[
\theta = \arccos \frac{|B|}{\sqrt{A^2 + B^2 + C^2}} \quad (11)
\]

2) FEEDBACK CALIBRATION

As mentioned in section III-B1, the real-time height estimation is performed after the coefficients of the ground plane were determined. The height is calibrated during movement by updating the transform matrix using the intersection D of the plane. At the same time, we also update the transform matrix to ensure the normal vector \(n\) of the ground is parallel to \((0,1,0)\). Therefore, the long-term drift of the gyroscope is further reduced. The closed-loop feedback mechanism makes the system robust when used for long periods under walking conditions.

D. OBJECT DETECTION

Two approaches for object detection were built in the system. One used a CNN-based object recognition method; the other used simpler and faster imminent danger detection.

1) CNN-BASED OBJECT RECOGNITION

Based on the research on the trade-off between speed and accuracy for modern convolutional object detectors [41], we choose the combined detector, SSD_MobileNet_V2, as the object recognizer running on the mobile device. By using depthwise separable convolutions to reduce both numbers of parameters and computational cost, the MobileNetV2 [42] achieves a better balance between lightweight and performance than other networks [37], [39], [43], [44]. The detector uses MobileNetV2 as the feature extractor from images, then generates anchors using SSD [44]. The detector is trained on the Common Objects in Context (COCO) dataset\(^1\) from Microsoft, which includes 91 classes (e.g., person, chair, table), to help visually impaired people have a general perception of surroundings. However, the “2D” object detector cannot provide the distance information for blind navigation.

Therefore, the result of the CNN object detector is fused with the depth map to get the distance information. The distance of the detected object is computed by averaging the value in the corresponding depth map area. Thus, rich information about surrounding objects is provided, such as the category and the distance. Because the depth is not considered in advance, a painted object on the ground might still be detected as an obstacle, which will lead to visually impaired people making wrong decisions. (see Fig. 5). In order to eliminate this false positive detection result, the ground area identified from the point cloud is mapped to the original image according to the transformation matrix, and the corresponding ground area is masked from the image. The procedure is shown in Fig. 6.

2) IMMINENT DANGER DETECTION

The CNN-based object detector is not able to recognize some objects, and they may pose a danger to the user. Our system features an imminent danger (e.g., hanging objects) detection that can detect any potentially dangerous obstacles and output their 3D bounding box. The imminent danger detection is performed using Euclidean clustering on the 3D point cloud. Rather than utilizing whole point cloud as input, the points belong to the ground area are removed. Besides, a passthrough filter is performed to focus on specific areas. For example, the area within 1 meter of the user is considered, because obstacles in this range may pose a danger to the user. In addition, when the distance from the user is less than 0.5 meters, the detection area is further reduced in the

\(^1\) http://cocolodataset.org/
horizontal field of view (FoV) (Fig. 7). This strategy allows the algorithm to focus on the user’s straight direction to reduce false alarms, such as the door frame might be detected as an obstacle when crossing a door.

The k-dimensional (three-dimensional in our example) tree is used for range search and nearest neighbor search. Euclidean cluster extraction divides the input point cloud into a cluster list, which represents the detected obstacles. As described in Algorithm 2, the output of a set of 3D bonding boxes illustrate the position and size of the identified obstacles.

**E. ACOUSTIC FEEDBACK**

As described in related work [20], [45], the human reaction of audible emergency warnings has been investigated. We had a series of people who are actually blind. They were the ones who suggested feedback abilities. This call occurred and that many of the goals were shaped on the potential end-user feedback. The designed sound emission policy includes different levels are presented in Table. 3.

![Algorithm 2 Hazard Detection Based on Euclidean Clustering Approach](image)

The low emergency sound warns that a collision may occur between 0.3 and 1 meters. The beep sound with low tone is spatialized in the 3D audio environment to guide visually impaired people. From the 3D reconstruction module, the user’s position and orientation in the world coordinate were acquired. The output from speakers is determined by the relative position of the detected object to the user. One example is illustrated in Fig 9. The user is oriented along the positive Z-axis, and his head is pointing towards up. If the position of the detected object on the X-axis has a positive value, the sound emitted from the location of the object will be heard from the left speaker. The volume of the sound emitted is controlled by attenuation (a multiplicative factor). The higher the attenuation, the less sound the user hears. If a warning occurs, the user should slow down the walking speed to continue.

If the detected obstacle is within 0.3 meters or no available path is identified, the system will continue to issue an alarm to notify the user. In this case, the user should stop and then turn left / right to search for the best walking direction.
until the warning stops. In addition to providing warnings, voice prompts are used for users to improve their situational awareness. For example, if the user enters an unfamiliar environment, the user can activate the system to convert the result of the CNN object detector into speech to describe the scene to the user.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

Experiments were conducted to evaluate the performance of the main modules under real indoor scenarios. Due to limitations caused by the ongoing COVID-19 pandemic, the test was performed only in the home of one author. Once the pandemic is under control, we expect to conduct more tests in more scenarios.

A. EXPERIMENTS ON GROUND SEGMENTATION

1) ORIENTATION TRACKING

Sensitivity, zero-offset, and bias are the intrinsic parameters of IMU, and they attempt to convert the raw sensor data into real measurements. These intrinsic parameters were calibrated to improve the accuracy of the angle estimation. First, IMU data in 6 different static positions (upright outward, upward and downward, etc.) were collected. We calibrated these parameters to ensure the accelerometer’s output is consistent with the gravity decomposition on three axes. Then, the quantitative analyses of the real-time orientation tracker were performed.

We evaluated the orientation tracker in a pre-set direction and compared the estimation results with the ground truth in three cases: using only the gyroscope, using complementary filters on the gyroscope and accelerometer, and applying feedback calibration to the complementary filters. The comparisons of pitch angle estimation and roll angle estimation are shown in Fig. 10. The gyroscope based estimations could not be used in practice because of the time drift. The complementary filters illustrate good results; however, mechanical devices inevitably have biases and noises. Our feedback calibration expresses excellent results, because similar to the Kalman filter’s mechanism, another discrete resource is invoked to correct bias.

2) GROUND SEGMENTATION

As depicted in Fig. 11, three situations in the test scene were tested to evaluate the performance of ground segmentation. The first case is on the left, with no obstacles on the floor. The ground had been successfully segmented, but several spots were not included. We adjusted parameters using loose constraints of the distance and angle. The missing spots can be segmented, but the whole ground area invaded to the object surface adjacent to the ground. Since the floor near the door in the test scene is not flat, and the in distance area does not affect the whole system performance. We set the angle threshold to 5 degrees and the distance threshold to 0.1 meters to get balance results. The parameters of the RANSAC algorithm are listed in Table. 4. These parameters are selected according
to the performance of the segmentation, that is, including as many true-positive points as possible and few false-positive points. Then, scenes with obstacles in the distance and nearby were evaluated. We can see that the proposed algorithm has robust segmentation results in these cases.

### B. EXPERIMENTS ON OBJECT DETECTION

As shown in Table 5, the default configuration of SSD_MobileNet_V2 was utilized. The depth multiplier was applied to scale the number of channels in each layer. The minimum depth parameter ensures that all layers have that many channels after applying the depth multiplier. To provide regularization and reduce generalization errors, we used L2 regularizer and batch normalization. Finally, the RMSprop optimizer with momentum was applied to classification loss and localization loss for training.

During reference, the confidence threshold was set to 0.5 to filter out low-confidence results. The point clouds were downsampled to 160 by 120 for imminent danger detection to reduce the computation cost. The distance threshold for point cloud clustering was set to 0.2 meters to separate point clusters.

Three scenarios were tested in an indoor environment to verify our object detection method. As shown in Fig. 12, we first tested the ability to detect large and small everyday objects. In cases 1 and 2, imminent danger detection from point clouds and CNN-based object detection from images can successfully identify the bottle and the chair. For exceptional cases, a medium-sized unusual object was tested. The CNN model could not identify the object because there was no category with a confidence score higher than 0.5. However, our imminent danger detection method could still detect the missing object. Regardless of the size and type of object, the imminent danger detection is a reliable method. It can identify objects that are missing from machine learning-based methods and ensure that the object detection can be performed reliably by blind people.

### C. EXPERIMENTS ON ACOUSTIC FEEDBACK

Acoustic feedback is the main method of interaction between the prototype system and the end-user. We used the text-to-speech to provide informative human voice feedback and a 3D audio engine to emit spatial beeps. The designed scenario of Fig. 12 case 2 (where the way was blocked by a chair) was used to evaluate the sound emission policy and trigger conditions. In this test, we assumed that a blindfolded person was trying to reach the destination through this blocked path.

---

**FIGURE 11.** Results of the ground segmentation. (a) Open space without obstacles; (b) Space with obstacles in the distance; (c) Space with an obstacle nearby. Note: The blue region represents the segmented ground.

**TABLE 4.** Parameters of RANSAC algorithm for ground segmentation.

| Model Type | NORMAL PLANE |
|------------|--------------|
| Max iterations | 200 |
| Input normal | True |
| Distance threshold | 0.1m |
| Angle threshold | 3° |
| Max distance from origin | 0.1m |

**TABLE 5.** Hyperparameters of SSD_MobileNet_V2 for object detection on images.

| Architecture hyperparameters: | depth multiplier | 1 |
|-------------------------------|-----------------|---|
| minimum depth                 | 16              |
| L2 regularizer weight         | 0.000004        |
| batch normalization            | 0.997           |
| input size                    | 300 * 300       |
| batch size                    | 24              |
| RMSProp Optimizer             | exponential decay-rate: 0.004 |
|                              | decay steps: 800720 |
|                              | decay factor: 0.95 |
|                              | momentum optimizer-value: 0.9 |
|                              | decay: 0.9 |
|                              | epsilon: 1      |

**FIGURE 12.** Results of the object detection. Case 1: Small object (a bottle) detection; Case 2: Large object (a chair) detection; Case 3: Unrecognized object by the machine learning model. Note: (a)s are RGB images; (b)s are results of the imminent danger detection from point clouds, the direction is left and right reversed relative to RGB images; (c)s are results of the CNN-based object detection from images. The green regions represent the segmented ground from point clouds; The blue regions represent the segmented ground in images.
During the system setup process, the system prompted “The system is setting” to notify the user to remain static. After initialization, the system informed the user of the estimated height of the camera above the ground. If the user judged that the height was incorrect, the setting process could be restarted. When the chair was detected within 1 meter, the 3D position of the estimated bounding box centroid of the chair and the height of the camera were used as parameters for invoking the 3D audio engine. Since the volume and sound channel of the beep sound depend on the relative position of the chair and the camera, the user could obtain the spatial information of the chair. When the subject continued to advance to the chair within 0.3 meters, the system beeped with a high tone to warn of an impending collision. In addition, from the camera’s point of view, since the chair occupied most of the scene and the ground could not be detected, “Path not found” was issued. At this point, the user should immediately stop and change the direction to search for available paths. However, the user could activate the CNN-based object detection, and the text-to-speech engine converted the results of the CNN object detector into speech (“chair 0.3 meters ahead”). Therefore, the subject could simply move the chair and continue to move. This can prevent the visually impaired from detouring and increase their environmental awareness. Hence, our dedicated audio feedback can help visually impaired people improve travel efficiency and bring them a better travel experience.

### D. COMPUTATIONAL COST

The hardware includes an Intel RealSense D435i camera, a Raspberry Pi model b with 4Gb memory, and an ANKER PowerCore 10000 mobile power supply. Point Cloud Library (PCL) version 1.9.0 is used to process point clouds, and Open Multi-Processing (OpenMP) is applied to achieve multi-threaded parallelism. The machine learning model for object detection is optimized using quantization to reduce the reference time. In the sound feedback module, the Flite text-to-speech engine and the Open AL audio engine are adopted to provide human speech and 3D audio feedback.

All algorithms run on the local Raspberry Pi. As depicted in Table 6, the average computational time among 100 epochs for the three cases in Fig. 12 were calculated. We show that the time spent on ground segmentation and imminent danger detection has a positive correlation with the size of the object. Compared with case 1, case 2 requires more than 26% of the time to segment the ground. Because the smaller the percentage of the ground area in the regional proposal, the more time it takes to segment it from random sampling. Besides, due to the need to cluster more points for larger objects, case 2 takes 2.03 milliseconds, 3.44 times longer than case 1. Since CNN-based object detection is only activated when a visually impaired user wishes to understand the surrounding environment, the total time of all algorithms is about 358 milliseconds on average. Therefore, the proposed system can provide real-time assistance to the visually impaired in typical walking situations.

### E. DISCUSSIONS

Through quantitative and qualitative evaluation, our wearable system can be a useful tool to guide the visually impaired to avoid obstacles in the indoor environment using low-power embedded devices. Different from works [9], [28] that segment ground in the whole search space, or search the area below the estimated ground height [27], we propose a novel architecture that can segment the ground on the proposed region based on real-time orientation and altitude tracker. This method reliably proposes the ground ROI and greatly reduces the 3D search space of the RANSAC algorithm. The performance improvement is highly dependent on the concrete scenes because of the irregularity of the point cloud. We compared the ground segmentation time in the whole space and the proposed region of the three typical indoor scenes in Fig. 11. In scenario (a), the average time of region-based segmentation is shortened by an average of 73.73%, in scenario (b) by 50.85%, and in scenario (c) by 57.78%. The results are depicted in Fig. 13, which demonstrates the efficiency of our method on ground segmentation using a low power device.

The object detection from the point cloud is a guaranteed method that can reliably detect any potential hazards. Still, it is not efficient to identify the category of objects through modeling. Machine-learning-based methods can provide the category information about detected objects, but they may miss objects, although they may still pose a danger. Another advantage of our approach is that object detection is performed from point clouds and images. Such unrecognized objects can be detected by the imminent danger detector to help users avoid collisions. Simultaneously, machine learning methods can be used to improve the situational awareness of blind people. Since machine learning is a data-driven model that predicts objects based on the trained data, we can retrain the machine learning model by capturing images of the target environment to improve accuracy.

Instead of sensing the reflected sound from the parametric speaker [9] to help users understand the angle of the object, we use 3D audio feedback to provide not only the direction

### TABLE 6. Average running time among 100 epochs. Note: The total and average time are calculated with CNN-based object detection excluded.

| Stage                        | Case 1     | Case 2     | Case 3     |
|------------------------------|------------|------------|------------|
| Depth preprocessing          | 114.34 ms  | 118.32 ms  | 120.70 ms  |
| 3D scene reconstruction      | 61.57 ms   | 61.74 ms   | 60.28 ms   |
| Normal estimation            | 37.88 ms   | 39.92 ms   | 39.87 ms   |
| Ground segmentation          | 122.91 ms  | 155.11 ms  | 148.57 ms  |
| Imminent danger detection    | 0.59 ms    | 2.03 ms    | 1.06 ms    |
| CNN-based object detection   | 6.89 ms    | 7.03 ms    | 7.02 ms    |
| Total                        | 337.29 ms  | 365.12 ms  | 370.48 ms  |
| Average                     | 357.68 ms  | 369.66 ms  | 373.78 ms  |

1. [https://www.intelrealsense.com/depth-camera-d435/](https://www.intelrealsense.com/depth-camera-d435/)
2. [http://www.raspberrypi.org/products/raspberry-pi-4-model-b/](http://www.raspberrypi.org/products/raspberry-pi-4-model-b/)
3. [http://www.anker.com/products/variant/powercore-10000-pd/A1235011](http://www.anker.com/products/variant/powercore-10000-pd/A1235011)
4. [http://www.festvox.org/flite/](http://www.festvox.org/flite/)
5. [https://www.openal.org/](https://www.openal.org/)
but also the distance information of the object. Besides, if the object’s surface is not flat, the parametric speaker may not provide the correct guidance. Our system can help build a more accurate understanding of the surrounding environment by offering objects with a sense of distance and direction to avoid obstacles.

The system runs on the Raspberry Pi at about 3 Hz in real-time. By overclocking the CPU frequency to a maximum of 2 GHz, the operating speed is slightly increased. By further downsampling the point cloud to a smaller size, the processing time can be reduced, which is a compromise between detection accuracy and speed. We also tested the system on a desktop computer with an Intel 8700K CPU and 16 Gb RAM, and the results showed an approximately 10-fold increase in update speed. Although the hardware prototype is less powerful, it consumes less power and can still provide timely feedback in terms of typical walking situations of visually impaired people.

Our method still has limitations though. The neural network is only trained for common objects. For special objects, it cannot provide the correct semantic category information for the end-user, for example, case 3 in Fig. 12. Since we do not search the space under the ground in the point cloud, the imminent danger detector is almost impossible to detect down-stairs. The acoustic feedback may distract blind people because the sound is the main source of blind people’s perception of the environment.

V. CONCLUSION AND FUTURE WORK
This article proposes a wearable device for indoor imminent danger detection that can assist people with severe visual impairments to move around independently. The system features include wayfinding, imminent danger detection, object recognition, and 3D acoustic feedback. The user’s surroundings can be reconstructed accurately by tracking the orientation and altitude of the camera. Thus, we proposed a region-based ground segmentation, which reduces the computational cost and increases robustness. The imminent danger detector provides a guaranteed method for identifying obstacles that are unrecognizable by the machine learning method but still might pose a danger. By utilizing an optimized 3D audio feedback mechanism based on the detected potential threats, our system can help users prevent imminent hazards and improve their situational awareness.

We conducted experimental evaluations in common indoor environments. Experimental results prove that our effective technology can be a useful tool to help blind people reduce travel injuries. Due to the COVID-19, the university restricted access to labs. It was difficult to invite blind people to campus for system evaluation. In the future, more evaluations involving additional scenarios and settings will be conducted, and a tactile wrap will be added to the wrist to improve the feedback mechanism. In addition, we plan to apply deep neural network models directly on the 3D point cloud to estimate object-oriented 3D bounding boxes and semantic categories to improve performance further.

REFERENCES
[1] World Health Organization. Blindness and Vision Impairment. Accessed: Apr. 29, 2020. [Online]. Available: https://www.who.int/news-room/factsheets/detail/blindness-and-visual-impairment
[2] S. Bhatlawande, M. Mahadevappa, J. Mukherjee, M. Biswas, D. Das, and S. Gupta, “Design, development, and clinical evaluation of the electronic mobility cane for vision rehabilitation,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 22, no. 6, pp. 1148–1159, Nov. 2014.
[3] H. Zhang and C. Ye, “An indoor wayfinding system based on geometric features aided graph SLAM for the visually impaired,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 25, no. 9, pp. 1592–1604, Sep. 2017.
[4] S. Shoval, J. Borenstein, and Y. Koren, “The NavBelt—A computerized travel aid for the blind based on mobile robotics technology,” IEEE Trans. Biomed. Eng., vol. 45, no. 11, pp. 1376–1386, Nov. 1998.
[5] F. Shaikh, M. A. Meghani, V. Kuvar, and S. Pappu, “Wearable navigation and assistive system for visually impaired,” in Proc. 2nd Int. Conf. Trends Electron. Informat. (ICOEI), May 2018, pp. 747–751.
[6] M. Brock and P. D. Kristensson, “Supporting blind navigation using depth sensing and sonification,” in Proc. ACM Conf. Pervas. Ubiquitous Comput. Adjunct Publication UbiComp Adjunct, New York, NY, USA: Association for Computing Machinery, 2013, pp. 255–258, doi: 10.1145/2494091.2494173.
[7] H.-C. Huang, C.-T. Hsieh, and C.-H. Yeh, “An indoor obstacle detection system using depth information and region growth,” Sensors, vol. 15, no. 10, pp. 27116–27141, Oct. 2015.
[8] V. Kulyukin, C. Gharpure, J. Nicholson, and G. Osborne, “Robot-assisted wayfinding for the visually impaired in structured indoor environments,” Auto. Robots, vol. 21, no. 1, pp. 29–41, Aug. 2006, doi: 10.1007/s10514-006-7223-8.
[9] H. Takizawa, Y. Kuramochi, and M. Aoyagi, “Kinect cane system : Recognition aid of available seats for the visually impaired,” in Proc. IEEE 1st Global Conf. Life Sci. Technol. (LifeTech), Mar. 2019, pp. 189–193.
H.-C. Wang, R. K. Katzschmann, S. Teng, B. Araki, L. Giarre, and D. Rus, “Expanding the detection of stairs and doorways,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Oct. 2016, pp. 473–479.

A. Pérez-Yus, G. López-Nicolás, and J. J. Guerrero, “Detection and modelling of staircases using a wearable depth sensor,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2014, pp. 449–463.

T. Westfechtel, K. Ohno, B. Mertsching, D. Nickchen, S. Kojima, and S. Takeda, “Robust and lightweight stairway detection and localization for mobile robots,” in Proc. IEEIEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Oct. 2016, pp. 3926–3934.

R. Hulik, M. Spanel, P. Smrz, and Z. Materna, “Continuous plane detection in point-cloud data based on 3D Hough transform,” J. Vis. Commun. Image Represent., vol. 25, no. 1, pp. 86–97, Jan. 2014, doi: 10.1016/j.jvcir.2013.04.001.

R. A. Zeinelddin and N. A. El-Fishawy, “Fast and accurate ground plane detection for the visually impaired from 3D organized point clouds,” in Proc. SAI Comp. Conf. (SAI), Jul. 2016, pp. 373–379.

J. Strom, A. Richardson, and E. Olson, “Graph-based segmentation for colored 3D laser point clouds,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst., Oct. 2010, pp. 2131–2136.

D. Holz, S. Holzer, R. B. Rusu, and S. Behnke, “Real-time plane segmentation using RGB-D cameras,” in Robot Soccer World Cup. Cham, Switzerland: Springer, 2011, pp. 306–317.

J. Redmon, S. Divvula, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 779–788.

A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “MobileNets: Efficient convolutional neural networks for mobile vision applications,” 2017, arXiv:1704.04861. [Online]. Available: http://arxiv.org/abs/1704.04861.

J. Podeeom and R. Huang, “YOLO-LITE: A real-time object detection algorithm optimized for non-GPU computers,” 2018, arXiv:1811.05588. [Online]. Available: http://arxiv.org/abs/1811.05588.

X. Zhang, X. Zhou, M. Lin, and J. Sun, “ShuffleNet: An extremely efficient convolutional neural network for mobile devices,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 6848–6856.

Intel Corp. (2018). Depth Post-Processing for Intel RealSense d400 Depth Cameras. Accessed: Apr. 29, 2020. [Online]. Available: https://www.intel.com/content/dam/support/us/en/documents/emerging-technologies/intel-realsense-technology/IntelRealSense-DepthPostProcess.pdf.

J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy, “Speed/Accuracy trade-offs for modern convolutional object detectors,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 7310–7311.

M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 4510–4520.

S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 91–99.

W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “SSD: Single shot multibox detector,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 21–37.

J. Edworthy, S. Loxley, and I. Dennis, “Improving auditory warning design: Relationship between warning sound parameters and perceived urgency,” Hum. Factors, J. Hum. Factors Ergonom. Soc., vol. 33, no. 2, pp. 205–231, Apr. 1991, doi: 10.1177/001872089103300206.

ZHONGEN LI received the B.S. degree in information and computing science from Anhui Jianzhu University, Hefei, China, in 2013, and the M.S. degree in integrated circuit engineering from the University of Science and Technology of China, Hefei, in 2016. He is currently pursuing the Ph.D. degree in computer engineering with Ohio University, Athens, OH, USA. His current research interests include object detection in 2D and 3D, and natural language processing.
FANGHAO SONG (Graduate Student Member, IEEE) received the B.S. degree in heat energy dynamics engineering (thermal engineering) from Shenyang Aerospace University, Shenyang, China, in 2014, and the M.S. degree in mechanical engineering from the University of Southern California, CA, USA, in 2016. He is currently pursuing the Ph.D. degree in computer science with Ohio University, Athens, OH, USA. His current research interests include health applications and indoor navigation software development.

BRIAN C. CLARK received the B.S. degree in biology from Western Carolina University, Cullowhee, NC, USA, in 1998, and the M.S. degree in exercise physiology and the Ph.D. degree in neuromuscular physiology from Syracuse University, Syracuse, NY, USA, in 2001 and 2005, respectively.

He is currently the Clybourne Endowed Research Chair and a Professor of Physiology and Neuroscience with the Department of Biomedical Sciences, Ohio University, Athens, OH, USA, where he is also a Principal Investigator and the Executive Director of the Ohio Musculoskeletal and Neurological Institute. The overall goal of his research is to develop effective and implementable interventions that increase muscle function and mobility in patients with orthopedic and neurologic disabilities. He has published more than 135 peer reviewed articles in clinical, basic science, and engineering journals.

Dr. Clark’s awards and honors include the All-University Doctoral Prize from Syracuse University, in 2006, serving on dozens of national and international grant review panels, including being a Stand Member of the NIH’s Motor Function and Speech Rehabilitation Study Section (2014–2020), and serving on numerous expert panels and round tables for federally- and industry sponsored projects.

DUSTIN R. GROOMS received the bachelor’s degree in athletic training from Northern Kentucky University, Highland Heights, KY, USA, in 2008, and the M.S. degree in kinesiology from the University of Virginia, Charlottesville, VA, USA, in 2009, and the Ph.D. degree in health and rehabilitation sciences from Ohio State University, Columbus, OH, USA, in 2015.

Prior to pursuing doctoral studies, he was a Rehabilitation Clinician and an Instructor with the College of Mount Saint Joseph, Cincinnati, OH, USA. He is currently an Associate Professor with the Division of Athletic Training, School of Applied Health Sciences and Wellness, Ohio University. His current research interests include how the brain and movement mechanics change after injury and the design of novel therapies.

CHANG LIU (Senior Member, IEEE) received the Ph.D. degree in information and computer science from the University of California at Irvine, CA, USA, in 2002.

He is currently a Professor of Computer Science with Ohio University. His research interests include software engineering and its applications in various domains, including learning and medical applications.