



Moving objects extraction for building environmental maps describing human walking activity using a 3D LiDAR

Takumi Nishio & Mihoko Niitsuma

To cite this article: Takumi Nishio & Mihoko Niitsuma (2022) Moving objects extraction for building environmental maps describing human walking activity using a 3D LiDAR, *Advanced Robotics*, 36:23, 1291-1304, DOI: [10.1080/01691864.2022.2126732](https://doi.org/10.1080/01691864.2022.2126732)

To link to this article: <https://doi.org/10.1080/01691864.2022.2126732>



© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



[View supplementary material](#)



Published online: 01 Oct 2022.



[Submit your article to this journal](#)



Article views: 2048



[View related articles](#)



[View Crossmark data](#)



Citing articles: 4 [View citing articles](#)

Moving objects extraction for building environmental maps describing human walking activity using a 3D LiDAR

Takumi Nishio^a and Mihoko Niitsuma^b

^aKawasaki Heavy Industries, Ltd., Hyogo, Japan; ^bDepartment of Precision Mechanics, Chuo University, Tokyo, Japan

ABSTRACT

This study proposes an indoor movable static object extraction approach for building environmental maps that describe indoor human activities. In particular, it is essential to distinguishing people from non-human movable objects. To address this issue, we classify static objects into three categories based on their features. Methods for extracting movable static objects using each feature are presented. We conducted experiments and confirmed that the movable static objects were successfully detected and redefined as backgrounds, and only people were determined as dynamic objects.

ARTICLE HISTORY

Received 31 March 2022
Revised 28 June 2022
Accepted 19 August 2022

KEYWORDS

Service robot; environmental map; indoor human activity observation; 3D LiDAR

1. Introduction

Sensing technology, represented by high-performance cameras and LiDAR, is developing rapidly and has been introduced into various fields, such as automated driving technology for automobiles. In recent years, research has been actively conducted to understand the space by acquiring the movement of people in the area, analyzing the behavior of users in the environment, and acquiring the features of their usage.

In such research, there are two phases: a dynamic object extraction phase in which dynamic objects are extracted from sensor data and an information acquisition phase in which information such as motion information is acquired for the extracted dynamic objects. Most of the information obtained from sensors is about objects in the environment other than people, such as walls and desks (hereafter referred to as 'environmental objects'). Information on dynamic objects such as people (hereafter referred to as 'human objects') is often acquired in smaller amounts. Some parts are missing owing to occlusions from environmental objects and other effects. Therefore, a dynamic object extraction phase is required to distinguish human objects from environmental objects while considering these effects. Then, in the information acquisition phase, various types of helpful information are acquired for the extracted human objects, such as movement paths and time spent. It is used for spatial understanding and behavior analysis.

Skeleton tracking [1] and template matching [2] are widely used methods for human extraction using webcams and depth cameras, which have good extraction accuracy. However, their measurement range is narrow, and multiple devices and locations must be considered when conducting global observations. In addition, the relative positions of human and environmental objects are also crucial in dynamic object extraction for spatial understanding. It is thus desirable to be able to acquire the positions of the environmental objects. Therefore, LiDAR sensors, which obtain the distance to an object by irradiating a laser beam, is often used in large-scale spatial understanding research.

The simplest method of dynamic object extraction using LiDAR is the background subtraction method [3], which is based on the difference between point clouds. However, in indoor environments, such as office spaces or stores, some people may remain seated or standing and stay in one place for a while. There is a high possibility that they cannot be extracted as dynamic objects by the difference. Therefore, in recent years, many 3D-LiDAR point clouds have been used for object identification and dynamic object extraction using supervised machine learning. These methods can be divided into those that process a point cloud as an image [4–9] and those that classify it directly [10–17]. However, because both focus on classifying the point cloud at a given instant, the processing cost is very high, making it challenging to apply them to the information acquisition

CONTACT Mihoko Niitsuma  niitsuma@mech.chuo-u.ac.jp Department of Precision Mechanics, Chuo University,  1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8551, Japan

 Supplemental data for this article can be accessed here. <https://doi.org/10.1080/01691864.2022.2126732>

© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

phase, in which information is acquired while extracting dynamic objects sequentially.

In contrast to these methods, object identification from 3D point clouds without applying machine learning is often based on features between point clouds, such as the surface shape of an object or the density of points in a voxel [18,19]. Lv et al. [20] voxelized point clouds at intersections and extracted dynamic objects based on their density changes. Although these approaches, including [21], are less accurate than machine learning methods, they reduce processing costs and enables the extraction of dynamic objects in a continuous time series. However, these methods are unsuitable for indoor observations because they cannot extract moving objects in an environment where people do not move. Preparing teaching data for diverse environments, such as indoor office environments, is also impractical.

Thus, conventional dynamic object extraction methods can only be applied to environments where accurate teacher data are available or where there are no stationary people and cannot be used in the indoor environment envisioned in this study. In addition, there are many movable static objects indoors, such as chairs, that change their position with human intervention, and there is no method to detect their movement and distinguish them from dynamic objects.

Therefore, the objective of this study was to develop a dynamic object extraction method that can acquire human activity indoors, even with movable static objects. The proposed dynamic object extraction method and the information acquisition phase based on this process can be executed continuously using long time-series data. They should be completed within 100 ms, the sensor period of 3D LiDAR. Therefore, this study utilizes an environmental map building system [22] that represents an environment with a collection of 10 cm

square cells. This system reduces the number of point clouds to be handled by projecting a point cloud onto the appropriate cell and using the center point of the cloud to reduce the processing costs. In addition, the initial background based on the newly proposed initial background extraction method is registered on the map in this study. The basic strategy is dynamic object extraction based on the difference between the initial and registered backgrounds, thereby realizing real-time processing.

The extracted dynamic objects include two possibilities: people and movable static objects. The most common objects among movable static objects present indoors are chairs. It is very likely that a person will move a chair after a series of actions such as sitting, standing up, and leaving. Next are objects such as partitions and file cabinets with wheels. Most movable static objects have a flat surface. Other objects, such as houseplants, trash cans, and unattended chairs, have different shapes and conditions for movement depending on their environment. Therefore, in this study, movable static objects are classified into three major categories, as shown in Figure 1, and an algorithm for extracting movable static objects is developed using each feature. On this basis, the movable static objects are successively and locally redefined as backgrounds, and only people are dynamic objects, which are used to build an environmental map describing indoor human activities.

The observation of human activities in daily environments, such as offices and stores, is essential for robotic systems to provide appropriate assistance. The proposed approaches play a significant role in the environmental map building system for the iSpace-based future convenience store presented by HARChuo for the customer interaction task at the World Robot Summit, Future Convenience Store Challenge 2020 [24].

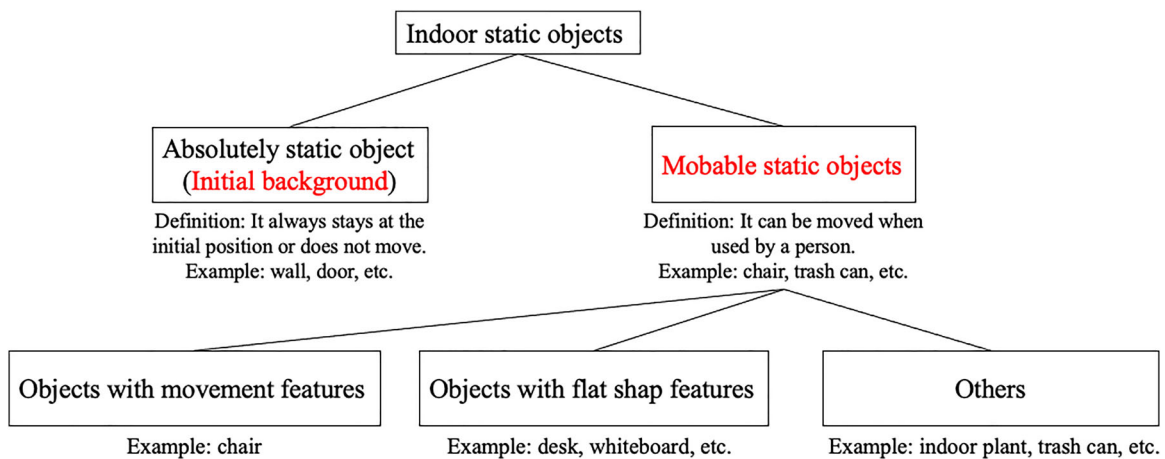


Figure 1. Classification of indoor static objects.

We believe that the proposed approach will be helpful not only for the offered future convenience store services but also for broader purposes that require building environmental maps expressing indoor human activities. The methods for describing human walking activity in an environmental map and the results of the map were presented in [22,23].

2. Initial background extraction using angles between adjacent points

The indoor activities of the people to be observed include the positions and movements of dynamic objects and their spatial connectivity with other objects such as counters and shelves. Therefore, it is necessary to extract dynamic objects in a continuous time series and acquire the positions of stationary objects that affect the behavior of active objects. We focused on the background subtraction method to enable real-time acquisition of these data.

Dynamic object extraction by background subtraction is commonly used in point clouds. In simple differencing, moving objects in a time series are removed as dynamic objects, whereas other objects are drawn as static objects. Dynamic objects do not always move in indoor environments, for example, sitting people. Owing to the presence of stationary objects, they may be removed as static objects. Therefore, when using differencing in an indoor environment, it is necessary to develop an algorithm to extract a person who is not moving as a dynamic object.

In this study, to extract a person who is not moving as a dynamic object, the standard deviation of the angle of each point in the point cloud cluster and the result of the background subtraction are used as features. The cost is defined based on the feature values. According to the cost, it distinguishes between stationary people and static objects, such as walls and shelves. The static objects identified by the proposed algorithm are registered in the environmental map as elements that create space without being removed to acquire indoor activities. The spatial relationships between a dynamic object (a person) and static objects can be obtained.

In this study, after the initial background is determined, dynamic object extraction is performed by subtracting the initial background from the point cloud at each time point. Thus, the accuracy of the initial background extraction is critical. Therefore, dynamic object extraction and activity information are not performed until the initial background is determined and only the initial background is acquired. This strategy aims to reduce the possibility of erroneous information accumulation in the environmental map and achieve an accurate initial background extraction.

2.1. Initial background extraction method using the angular variation of each point in the cluster

As described above, when extracting the initial background object from all points obtained using a 3D-LiDAR sensor, it is difficult to process it within 100 ms, the sensor period. However, it is possible to achieve a high accuracy. Therefore, the strategy used in this study is to reduce the number of point clouds to be processed to extract stationary objects that serve as the initial background from the point cloud without compromising the real-time performance. In other words, after the point clouds are clustered, all clusters with point clouds that are higher than the sensor position are defined as static objects as shown in Figure 2.

In studies using 3D - LiDAR for spatial understanding, it is common to place the sensors above the height of a person's head. As it is assumed that there is no point cloud representing a person above the sensor location, this process does not affect the accuracy of person extraction. However, it is practical to reduce the number of point clouds to be processed.

After the above process, the remaining point cloud cluster i in the space has two possibilities: a static object identified as background or a dynamic object, a person. Because people may remain seated indoors, a method of object identification other than using the motion of the point cloud is necessary. In this study, to address this, the point cloud clusters are first divided into virtual 10 cm square voxels. Let $\mathbf{P}_{\text{center}} = (P_{\text{center},x}, P_{\text{center},y})$ be the center point of all points in a voxel (p, q, r) , and define the xy-plane angle θ_n and its mean value $\bar{\theta}$ with each point $\mathbf{P}_n = (P_{n,x}, P_{n,y})$ in the voxel using the maximum number of points N and their numbers $n = 1, \dots, N$ in the following equation. Here, $\mathbf{P}_{\text{center}}$ and $\mathbf{P}_n$ are projected onto the xy-plane.

$$\theta_n = \tan^{-1} \left| \frac{P_{n,y} - P_{\text{center},y}}{P_{n,x} - P_{\text{center},x}} \right|, \quad \bar{\theta} = \frac{\sum_{k=1}^N \theta_k}{N} \quad (1)$$

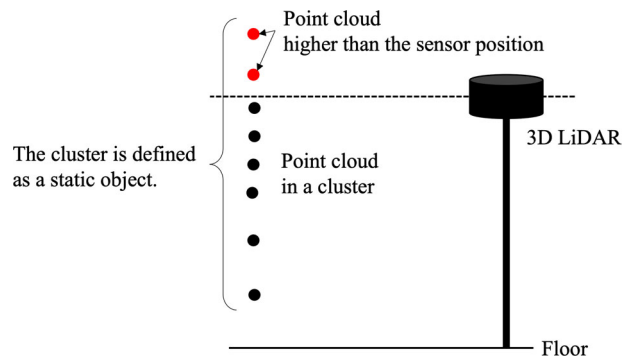


Figure 2. Example of the static object detection based on the height of the point cloud.

Using the angle θ_n and its mean value $\bar{\theta}$ determined in the above equation, their angular standard deviation $\sigma_\theta(p, q, r)$ is calculated. Using $\sigma_\theta(p, q, r)$, we define the background weight $\omega_{BG}(p, q, r)$ for each voxel as follows:

$$\omega_{BG}(p, q, r) = \begin{cases} 1.0, & \text{if } 0 < \sigma_\theta(p, q, r) \leq \alpha_1 \text{ or } \beta_1 < \sigma_\theta(p, q, r), \\ 0.0, & \text{else if } \alpha_1 < \sigma_\theta(p, q, r) \leq \beta_1. \end{cases} \quad (2)$$

In the above equation, α_1 and β_1 are design parameters. The background weights $\omega_{BG}(p, q, r)$ are affected by sensor errors that cause variations in the acquired point set. Through various analyses, we empirically found that the standard deviation for dynamic objects such as people tends to converge between 0.1 and 0.5. Based on this knowledge, we set $\alpha_1 = 0.10$, $\beta_1 = 0.50$ in this study. The background weights $\omega_{BG}(p, q, r)$ are obtained for all voxels in a point cluster, and the average of $\omega_{BG}(p, q, r)$ of point cluster i is incremented to the background cost $C_{cluster}(i)$ of the point cluster at each cycle.

The above process of obtaining the background cost $C_{cluster}(i)$ is continuously performed for an arbitrary t_{th} after starting the 3D LiDAR sensor. t_{th} s have passed, if $C_{cluster}(i) > C_{th}$ is satisfied, then point cluster i is defined as a static object, which constitutes the initial background. Here, the threshold time $t_{th} = 120$, and threshold of the background cost $C_{th} = 960$ are used.

2.2. Evaluation of initial background extraction

This section describes the experiments and results to verify the accuracy of the initial background extraction method using the standard deviation of the angles in the clusters described above. In a space with static objects such as walls and motionless people, we verified whether the proposed method can extract only stationary objects as initial backgrounds by performing initial background extraction based on the proposed method. This experiment used a portion of the measured data from a meeting room, as shown in Figure 3 (a). The data include walls and desks as static objects and people sitting in chairs as dynamic objects in the meeting room. In this experiment, the first 120 s were used for the initial background extraction, and the dynamic objects were extracted from the point cloud at 120.1 s. To verify the accuracy of the method described in the previous section, we manually labeled the point cloud at 120.1 s, as shown in Figure 3 (b), and compared the extraction results with it. After labeling, we confirmed that the points in the area circled in red in Figure 3 (b) indicated a person, and all the other points were backgrounds.

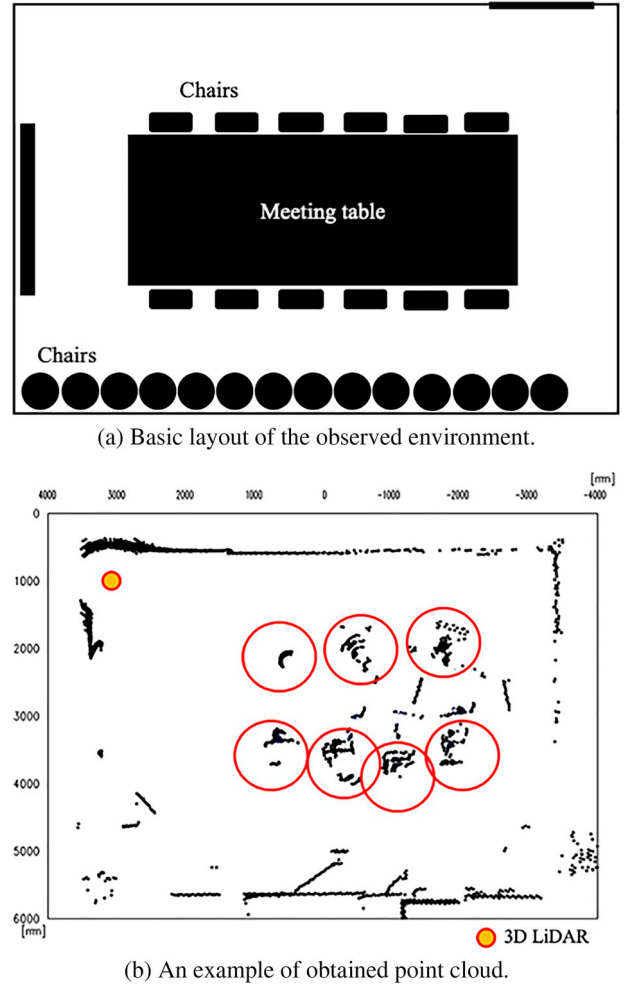


Figure 3. Experiment environment of foreground extraction experiment. (a) Basic layout of the observed environment. (b) An example of obtained point cloud.

2.3. Results and discussion of initial background extraction experiment

The results of dynamic object extraction by defining an initial background using the proposed method based on 120 s of data and then differencing the initial background from the 120.1s point cloud data are shown in Figure 4. In this figure, each cell in black represents a cell extracted as a static object and is defined as the initial background. The numbers on the map indicate the ID assigned to each object. The red circles in Figure 4 (a) and blue circles corresponding to the ID number in Figure 4 (b) represent the center point of the object extracted as a dynamic object.

Figure 4 shows that objects such as the wall surrounding the environment and the chair to the right of the object with ID No. 5 do not appear circled as dynamic objects, indicating that these objects can be extracted as the initial background. In addition, there were seven people in this environment. Each person was extracted as a dynamic object and assigned an ID, confirming

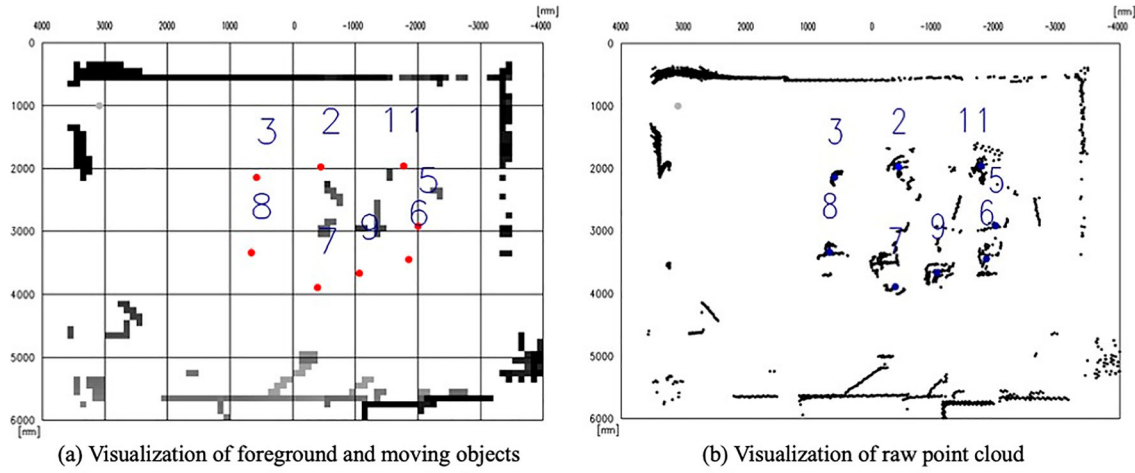


Figure 4. Extraction results of foreground and moving objects.

that people were not mistakenly recognized as the initial background. Here, the object in No. 5 is a false positive because the ID is assigned to a nonhuman object. During the 120 s, there were many occlusions by people in Nos. 2 or 3, so the point cloud of the object in No. 5 was insufficient. Therefore, the cost was insufficient to define No. 5 as background. In addition, IDs are assigned to possible persons even during the 120 s of the background definition. Therefore, IDs 1, 4, and 10 were assigned once during the 120 s background definition. However, they were redefined as background using the proposed background extraction method and the IDs were eliminated.

The above results indicate that objects acquired as point clouds during initial background extraction can be distinguished from people and other objects. Only static objects were defined as the initial background. Meanwhile, objects for which an appropriate point cloud could not be obtained owing to occlusion or sensor noise and static objects that were moved after the initial background was determined were extracted as dynamic objects. Therefore, a method that can successively redefine them as background is needed.

3. Method for extracting movable static objects using features of point cloud data

In an indoor environment, the object most likely to change its position is an object that moves with a person, such as a chair. A chair is expected to be moved around the position where a person is during a series of actions, such as sitting, standing up, and leaving. In this study, we focus on a chair as a movable static object that moves simultaneously with a person and investigate the solutions. Because the series of movements to use a chair is invariant regardless of the environment, a person's sitting position is detected, and dynamic objects in

the local area around the person who has transitioned from the sitting to the standing state are extracted as movable static objects. Subsequently, local background updating is performed to register the movable static objects as background.

3.1. Sitting position detection

To realize an algorithm for detecting the human sitting position, we focus on the height $h(t, i)$ of the dynamic object i at time t and the maximum height $h_{\max}(i)$ up to that time. To reduce the computational cost, the dynamic object in this study is represented by a set of cells in which a group of three-dimensional points is projected onto a 10 cm square cell. Each cell stores the largest z-coordinate of the points in that cell, and $h(t, i)$ is the largest value among the multiple cells representing dynamic object i . The maximum height $h_{\max}(i)$ is the largest height of the acquired dynamic object i up to time t .

In this study, the sitting decision is based on a comparison of a person's sitting height $h_{\text{sit}}(i)$ and $h(t, i)$ based on $h_{\max}(i)$. $h_{\text{sit}}(i)$ is determined based on the average sitting height of Japanese people according to the Ministry of Health, Labor and Welfare [25], which is 53% to 55% of the person's height, and the ideal ergonomic sitting height [26], which is 10 mm lower than 25% of the person's height. Based on the above findings, $h_{\text{sit}}(i)$ is defined by (3).

$$h_{\text{sit}}(i) = 0.55h_{\max}(i) + (0.25h_{\max}(i) - 10) \quad (3)$$

If the height $h(t, i)$ of dynamic object i at time t satisfies $h(t, i) < h_{\text{sit}}(i)$, then the dynamic object i is in a seated position. Otherwise, the user is assumed to be in the standing position. The sitting height $h_{\text{sit}}(i)$ is redefined each time $h_{\max}(i)$, is updated.

3.2. Extraction of movable static objects based on sitting position detection

When a dynamic object i is detected as seated, the height $h(t, i)$ and history of the seated detection are stored in all cells within a radius of 50 cm from its center. The average value $\bar{h}(p, q)$ of the heights accumulated in any cell (p, q) is then computed.

When a person detected in sitting position shifts to a standing position and leaves the scene, a chair or a new person sitting in a chair is expected to be present around the cell (p, q) as a dynamic object k . If the dynamic object k is a chair, its height $h(t, k)$ is likely to be smaller than $h(p, q)$. Therefore, considering the sensor noise and other aspects, if the height $h(t, k)$ of the dynamic object k satisfies $h(t, k) + \alpha < h(p, q)$, the dynamic object k is a candidate for a movable static object, and the time $t_{\text{chair}}(k)$ when the inequality is satisfied is recorded. Considering the time from when dynamic object i is determined to be standing to when it leaves its place, the dynamic object k is defined as a movable static object if and only if $t_{\text{chair}}(k) > \beta$ is satisfied within an arbitrary time t_{th} . In this study, $\alpha = 100\text{mm}$, $\beta = 0.1$, and $t_{\text{th}} = 120\text{s}$ were set empirically.

All the cells occupied by objects defined as movable static objects are immediately redefined locally as background. By performing the above process, a chair can be extracted as a movable static object after a person sits on it, and the background can be locally updated while maintaining the real-time performance.

3.3. Movable static object extraction based on shape features

In addition to chairs, display board stands, cart wagons and whiteboards are movable static objects in indoor environments such as offices or stores. Such objects have the common feature that their surfaces are often linear in shape. When such an object moves, it casts a shadow on the initial background as shown in Figure 5. This feature is used to detect the candidates for movable static objects. Therefore, in this study, the shape cost is obtained based on the surface shape of the extracted object, and the background cost is defined based on the shape cost to extract movable static objects and perform local background updating. The background cost value is used to distinguish people from movable static objects. By applying local background updating only to the appropriate regions where the extracted movable static objects are present, human detection and background updating can be performed without compromising real-time performance.

3.3.1. Obtaining shape scores for static objects

When a movable static object moves, its occlusion creates a shadow in the initial background. If the newly created shadow exists in the same area for more than 10 s, the dynamic object i between the shadow and the 3D LiDAR sensor is defined as a candidate moveable static object, and the linearity μ of its surface shape is obtained.

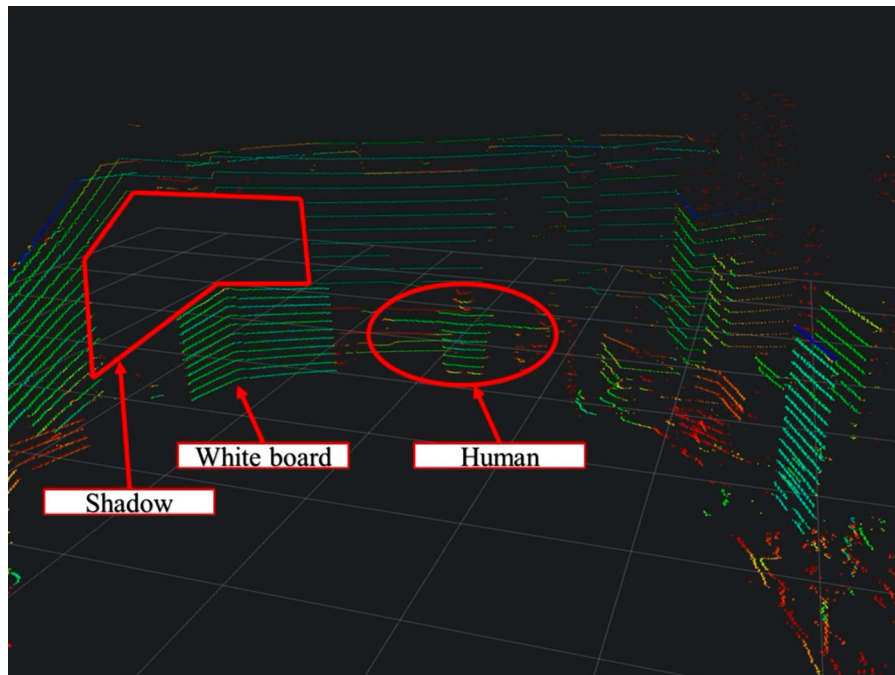


Figure 5. 3D point cloud view of objects and shadows obtained by a 3D LiDAR sensor.

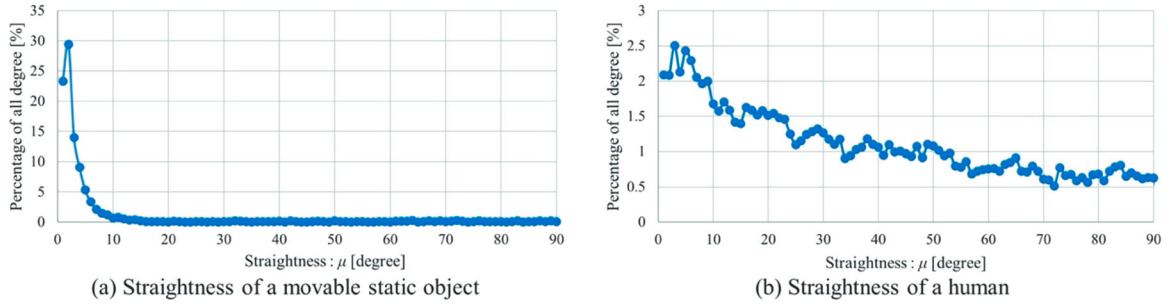


Figure 6. Straightness distribution of human and movable static object.

To obtain the linearity μ , we obtain the center point P_{cell} of the point group contained in the cell (p, q) constituting the dynamic object i and calculate the Euclidean distance d_n from the center point of the nearest neighbor cell. n is the number of the cell that forms dynamic object i . This process is applied to all the cells occupied by the dynamic object to obtain the mean value $\bar{d}$ of the distance d_n .

Dynamic object i is classified into two clusters, one with a large distance and the other with a small distance, based on $\bar{d}$, and the angle θ_m between the center points of the cells in the clusters is calculated. m is the number of cells in the cluster and M is the maximum number of cells. Based on the angle θ_m , the linearity μ of the cluster is defined by the following equation to obtain the linearity in each of the two clusters.

$$\mu = \frac{\sum_{k=3}^M |\theta_k - \theta_{k-1}|}{M - 2} \quad (4)$$

For each cluster, we obtain the linearity defined by Equation (4). The closer the value of linearity is to 0, the straighter the surface. Based on this definition, linearity was obtained from a movable static object and a static person. In Figure 5, 1200 linearity distributions are shown. If the surface is flat, the distribution is as shown in Figure 6 (a). If it is not flat, the distribution is as shown in Figure 6 (b). This distribution indicates that the degree of linearity of a movable static object tends to be low at high rates. In contrast, the variability of linearity tends to be large in the case of humans. Based on this knowledge, in this study, we define the shape score ϵ for object identification based on linearity μ as follows:

$$\epsilon = \alpha_2(90.0 - \mu)^2 + \beta_2 \quad (5)$$

The above equation is a simple quadratic function that determines that the linearity distribution can be approximated as inversely proportional. More specifically, this equation is designed to express the degree of straightness of an object by considering the angle between the two clusters of the object. This means that the closer μ is to 90° , the less straight the object is.

When determining the score based on the linearity of Equation (5), 90 is used to weight the score according to the degree of approximation from 90° . α_2 and β_2 in this equation are design parameters; in this study, $\alpha_2 = -0.0006$ and $\beta_2 = 5.0$ were determined empirically. The shape scores of the two clusters obtained using Equation (5) are defined as ϵ_{max} and ϵ_{min} , depending on their magnitude.

3.3.2. Movable static object identification and background update by shape score

Sensor errors cause variations in the acquired point clouds. Therefore, the shape feature can be expressed by setting a range for the defined shape score. For example, ϵ_{max} and ϵ_{min} of the movable static object with straight surfaces targeted here are $0 < \epsilon_{\text{max}} < 2.0$ and $0 < \epsilon_{\text{min}} < 1.5$, while objects that do not contain linear shapes, such as people, are represented in the range of $3.0 < \epsilon_{\text{max}} < 5.0$ and $2.0 < \epsilon_{\text{min}} < 4.0$. Based on these findings, we define the background cost $C_{\text{BG}}(t, i)$ of dynamic object i at time t according to its shape score based on the following conditions:

$$C_{\text{BG}}(t, i) = \begin{cases} C_{\text{BG}}(t-1, i) + 2.0, & \text{if } 0 < \epsilon_{\text{min}} \leq 1.0, \\ C_{\text{BG}}(t-1, i) + 1.0, & \text{else if } 1.0 < \epsilon_{\text{min}} \leq 2.0, \\ C_{\text{BG}}(t-1, i) - 2.0, & \text{else if } 4.0 < \epsilon_{\text{max}}. \end{cases} \quad (6)$$

Furthermore, if the object velocity is 0.1 m/s for more than 1 s, the object is likely to be a person; therefore, the background cost $C_{\text{BG}}(t, i)$ is subtracted by 1.0. Following the above definition, the background cost $C_{\text{BG}}(t, i)$ of the dynamic object i is summed up; if $C_{\text{BG}}(t, i)$ satisfies $C_{\text{BG}}(t, i) > \gamma$ within any time t_{th} s, the object is defined as a movable static object and is registered as a background. In this study, $t_{\text{th}} = 120$ s and the threshold of the background cost $\gamma = 960$ are empirically assumed. This process distinguishes people from movable static objects with linear shapes and updates the background sequentially in real time.

3.4. Method for extracting movable static objects based on their active movement

Houseplants, trash cans, and unattended chairs are difficult to extract using the methods proposed in the previous sections because their shapes and conditions of movement vary depending on the environment. The most remarkable difference between these objects and people is the presence or absence of active movement. Therefore, to extract such objects as movable static objects, we use the position distribution of the highest point of the object and the change in the number of voxels occupied by the object at an arbitrary time t_{th} s. This method is applied only to objects detected by the process described in Subsection 3.3 and extracted as dynamic objects.

3.4.1. Object classification and background update based on the distribution of the highest points of movable static objects

In situations in which a person is sitting or working stationary, active motion is most noticeable in the human head. However, movable static objects do not move actively by nature, and the position distribution of the point cloud does not change except under the influence of sensor noise. Therefore, to use these differences as features, the method stores the positions (p, q) of cells with height $h(t, i)$ of dynamic object i at a certain time t . If there are multiple cells, they are saved, and the process continues for t_{th} s. After t_{th} s, the recording time for each position of all stored cells is obtained. If there are one or more cells that have been stored continuously for t_{th} s, dynamic object i is determined to be a movable static object. All cells occupied by the movable static objects considered to be stationary are defined as background and registered on the environmental map. This series of processes realizes local background updating only for the location of the movable static objects.

3.4.2. Object classification and background update based on changes in the number of voxels occupied by movable static objects

The method described in the previous section may cause variations in the distribution of the highest point positions of movable static objects and dynamic objects, such as people, depending on the positional relationship between the object and the LiDAR sensor. In such cases, it is not possible to extract the movable static object, and it is necessary to use other means to address the problem. Therefore, we propose a method that uses the changes in the number of voxels occupied by an object as a second feature to express whether an object is actively moving.

As shown in Figure 4, there is always a gap in the vertical axis direction in the point cloud obtained from the 3D LiDAR sensor used in this study. The gap size is determined by the angle between the scanning laser and distance to the object. Suppose that only voxels in the region where a point exists are defined as voxels occupied by the object. The number of voxels is smaller than the number of voxels occupied by the object. In indoor environments, such as offices and eating spaces, where there are many short objects, only the point cloud of the upper body can be acquired owing to the occlusion caused by tables and chairs.

To deal with these effects, we define the number of voxels $V_{cell,i}(t, n)$ occupied by a cell (p, q) constituting a dynamic object i at time t using the maximum $h_{max,i}(p, q)$ of the point cloud contained in the cell and the edge length l_{th} of the voxel. n is the number of cells constituting dynamic object i .

$$V_{cell,i}(t, n) = \frac{h_{max,i}(p, q)}{l_{th}} \quad (7)$$

Voxels of size l_{th} per side in the space between the highest point and ground stored in a cell (p, q) , with or without a point cloud, are generated. Equation (7) defines the sum as the number of voxels in the cell. The number of voxels $V(t, i)$ occupied by the dynamic object i at time t is obtained from Equation (8).

$$V(t, i) = \sum_{k=1}^n V_{cell,i}(t, k) \quad (8)$$

By obtaining the number of voxels $V(t, i)$ using Equation (8), it is possible to represent the size of an object in voxels even when some point clouds are missing, such as the lower body of a person shaded by a table when seated. In this study, $l_{th} = 100$ mm is empirically obtained. The number of voxels $V(t, i)$ is continuously acquired for t_{th} s, and the most frequent value of the number of voxels $V_{mode}(i)$ during that period is determined. The difference between $V_{mode}(i)$ and $V(t, i)$ is defined as the voxel change $V_{diff}(t, i)$. The standard deviation $\sigma_{diff}(i)$ of the voxel change for t_{th} seconds is calculated.

The standard deviation $\sigma_{diff}(i)$ is expected to differ between a movable static object and a person who moves actively. Therefore, we analyzed the data of movable static objects and seated workers acquired in various environments. The movable static objects analyzed here are only those that the previously described methods cannot extract. Through various data analyses, we found that the standard deviation $\sigma_{diff}(i) < 5.0$ is satisfied in the case of a movable static object and $6.0 \leq \sigma_{diff}(i)$ in the case of a person. Based on this finding, we define a movable static object when $\sigma_{diff}(i) < 5.0$ and a dynamic object

when $6.0 \leq \sigma_{\text{diff}}(i)$. We also confirmed that the standard deviation of movable static objects might be distributed with low frequency in the range of $5.0 \leq \sigma_{\text{diff}}(i) < 6.0$. Because the objective of this study is to obtain indoor activities of people, the most important thing to avoid is the incorrect extraction of people as movable static objects and the loss of their information to be stored on the environmental map. Therefore, if a standard deviation in the range $5.0 \leq \sigma_{\text{diff}}(i) < 6.0$ is obtained, the above process is performed for another $t_{\text{th}}/2.0$ s, and the target object is determined to be a movable static object only if the obtained standard deviation satisfies $\sigma_{\text{diff}}(i) < 5.0$.

A local background update is applied to the extracted movable static objects based on the above methods. All the cells of the object are registered as a background on the environmental map. This process redefines movable static objects as background and allows only information on dynamic objects, such as people, to be stored in an environmental map.

4. Evaluation experiment of movable static object extraction

The aim of this experiment is to verify the accuracy of the proposed series of movable static object extraction methods. The accuracy of extracting movable static objects is evaluated based on three factors: the time required to detect objects, the extraction results of each process, and the number of incorrect detections in the result.

4.1. Experimental environment setup and evaluation criteria

In this experiment, six types of movable static objects (chair A, Chair B, whiteboard, desk, camera tripod, and houseplants) and two dynamic objects (a seated person and a standing person) were selected as targets objects. Chairs A and B are chairs sat on and left unseated by a person, respectively. Figure 7 shows the experimental environment. The 3D LiDAR sensor (vlp-16) was placed 1.8 m above the ground. All the target objects were placed within 8.0 m of the sensor. A total of 500 s of point cloud data were acquired in one trial. We obtained the point cloud data for four trials by changing the orientation and distance of each object to the sensor. Each object was measured four times. In two of the four trials, the objects were set in the environment before the initial background was defined, and moved to other locations after the initial background was determined. The objects were placed in the environment after the initial background was defined for the remaining two times.

Table 1. Proposed methods to extract movable moving objects.

Abbreviation	Methods
ST	Extraction method based on sitting determination
SS	Extraction by using shape score
PD	Extraction based on position distribution of maximum height
VC	Extraction based on voxel changes

A table was placed as the static object in the middle of the observation target area. Two of the four times were taken on the right side of the table (sensor side), and the other two on the left side (wall side). The locations of the objects were changed to account for the presence of occlusion by the table. When the objects were installed on the right side (sensor side), they were placed immediately next to the table, and when they were on the left side (wall side), they were placed in a position where the occlusion effect was likely to be significant.

The method used for extraction consisted of a series of the proposed methods, as listed in Table 1. ST indicates extraction based on sitting position judgment; SS, extraction using shape scores; PD, extraction based on the position distribution of the highest point; and VC, extraction based on voxel change. The time required for object identification using each method, its accuracy, and the final extraction results were used to evaluate the accuracy of the proposed series of movable static object extraction methods. The time at which the moving object was detected was set to 0. The time required to determine whether the object is dynamic is defined as the shape determination time. In addition, we evaluated the representation on the environmental map and its differences depending on whether the local background is updated based on the extraction of movable static objects.

4.2. Results and discussion of evaluation experiments

Table 2 shows the results of the dynamic and movable static object extraction. The ‘Result’ column in Table 2 shows the number of false positives for each object. Specifically, it is the number of times a movable static object is detected as a person, or a person is detected as a static object, that is, background. The evaluation process consists of three steps: evaluation by ST, evaluation by SS, and evaluation by VC or PD. No further processing was performed if the object was redefined as background in any evaluations. The proposed method always assigns IDs to objects other than the initial background regardless of whether they are people or movable static objects. Therefore, the moment the ID is assigned is set as the 0th s. The time until the object is determined to be either a

movable static object or a person is measured and defined as the shape determination time.

The shape determination time of chair A is the time measured from the moment when a seated person is determined to have stood up to the moment when the object is concluded to be a movable static object. In the case of other objects, when a person moves them, they are extracted as moving objects, and then the processes to remove movable static objects are started. The shape determination time indicates the time taken to conclude that they are movable static objects. The extraction results of each process and the number of false detections are

Table 2. Extraction result based on the proposed methods.

		False extraction					Shape determination time [s]
Object		ST	SS	VC	PD	Result	
Movable static objects	Chair A	0 / 4				0 / 4	17.8 (ST)
	White board		0 / 4			0 / 4	53.4 (SS)
	Desk		0 / 4			0 / 4	58.6 (SS)
	Camera stand		0 / 4			0 / 4	96.2 (SS)
	Chair B		3 / 4	0 / 3	2 / 3	0 / 4	55.4 (SS)
	Indoor plant		4 / 4	4 / 4	0 / 4	0 / 4	177.9 (VC, PD)
	Standing		0 / 4	0 / 4	0 / 4	0 / 4	191.0 (PD)
Moving objects (Human)	Sitting	0 / 4	0 / 4	0 / 4	0 / 4	0 / 4	

listed in Table 2. The experimental results showed that all movable static objects were redefined as background by one of the evaluations, and people were recognized as persons.

Chair A was extracted as a movable static object in a short time without false detection by detecting the state transition of a person in a sitting/standing position and performing extraction based on this transition. These results indicate that the extraction of movable static objects based on the sit-stand decision requires a shorter time than the other methods. The whiteboard, desk, and tripod were extracted as movable static objects using SS, without false positives. The tripod has fewer straight surfaces than the whiteboard and desk, and the difference in shape is reflected in the shape determination time. Meanwhile, chair B could only be extracted once as a moveable static object based on its shape score because the obtained point cloud of the chair might include arm-rests and seat surfaces. The above results indicate that the proposed method can be used to extract movable static objects with linear surface shapes.

The houseplant was extracted as a movable static object using PD with no false positives, and chair B was detected as a movable static object once using PD and twice using VC. When processing with PD and VC, more time is required to determine the shape than that needed with other methods because they use data for 120 s after the object is extracted as a dynamic object using SS. Because the object is considered dynamic during this period, it is necessary to shorten this time in the

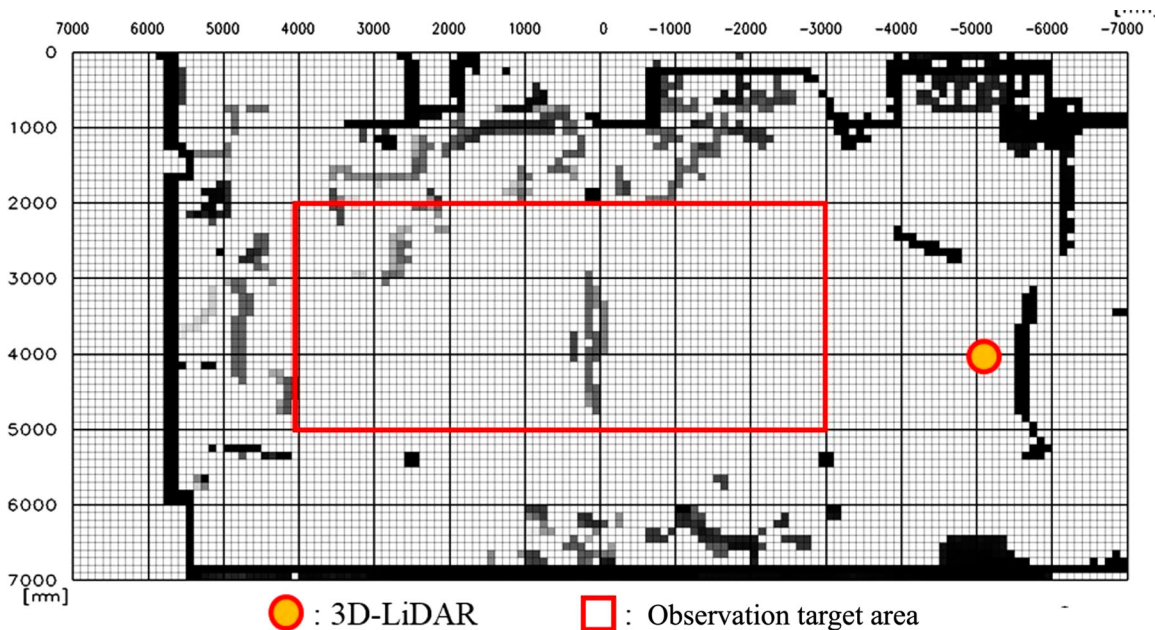


Figure 7. Designed experiment environment.

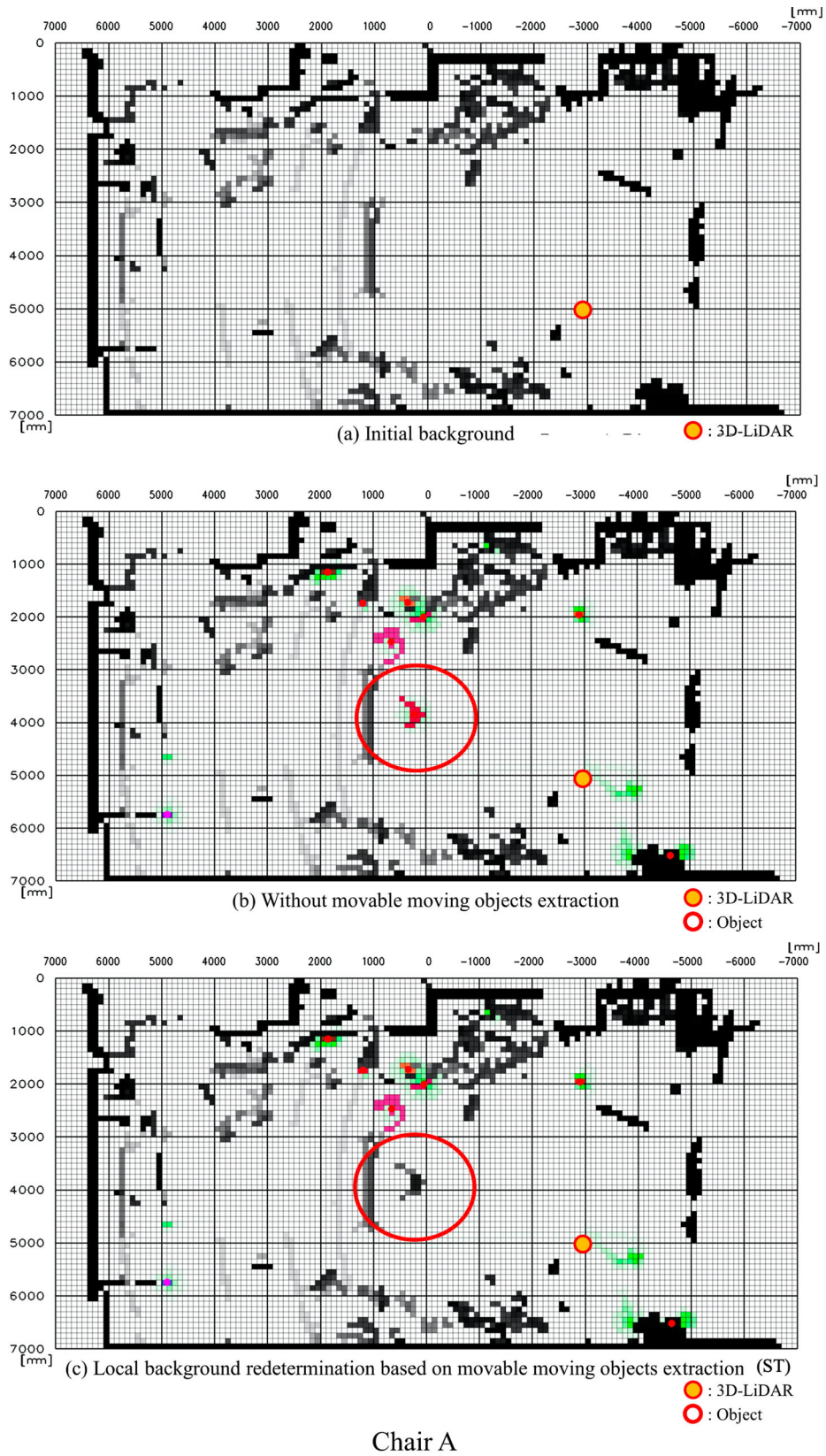


Figure 8. Chair A at $t = 17.9$ s.

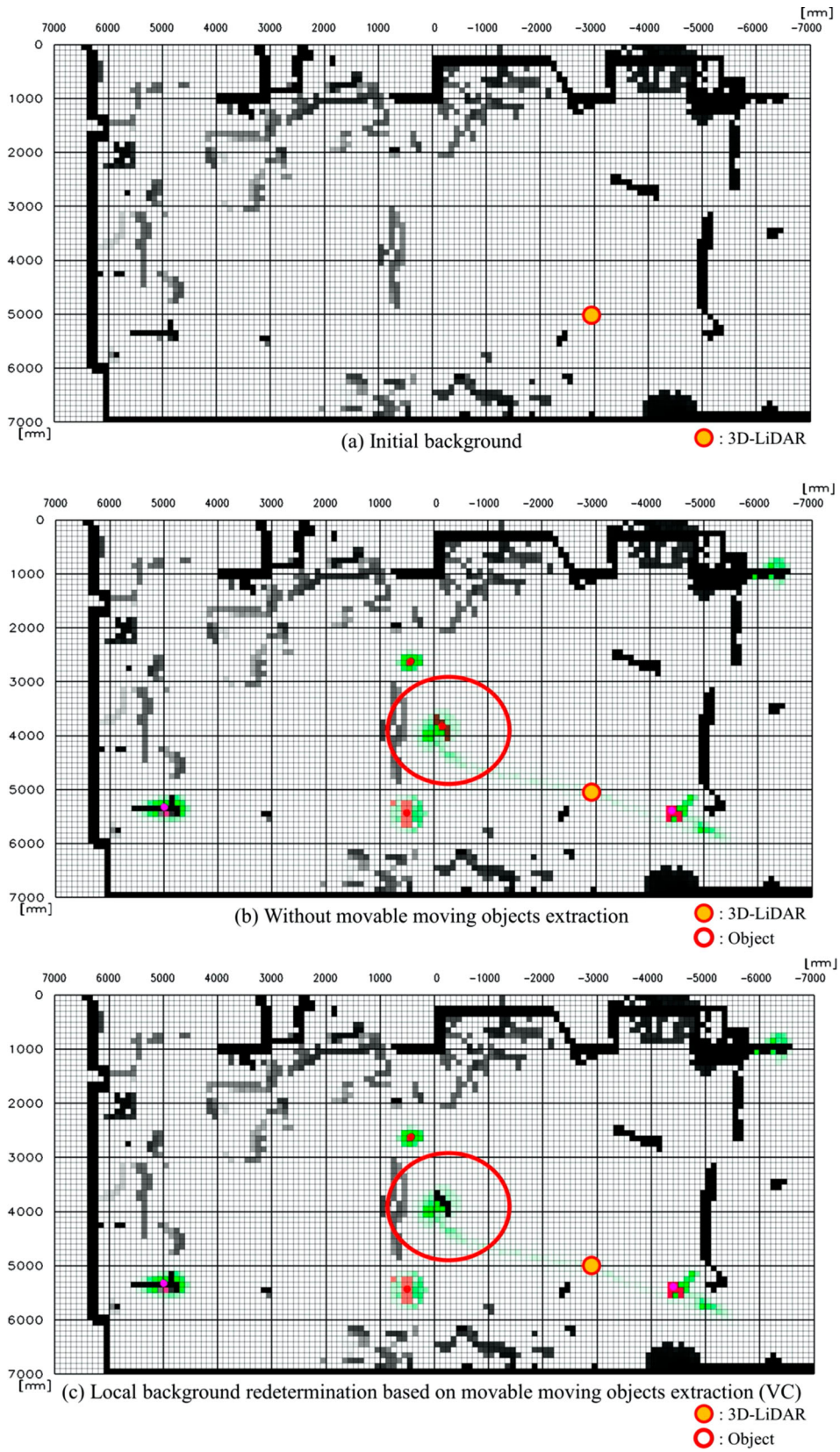


Figure 9. Chair B at $t = 178.0$ s.

future. However, these objects were eventually detected as movable static objects through the PD and VC processes.

In addition, people were detected as dynamic objects in all the trials, regardless of their state. Because the method defines them as dynamic objects during shape determination and acquires information on their position and movement, no data are missing during shape determination. These results show that the proposed method can successfully detect objects as movable static objects, although it takes time to determine their shapes.

Figures 8 and 9 show examples of changes in an environmental map owing to local background updating based on the extraction of movable static objects. The prepared objects were placed in the experiment after determining the initial background. Figures 8 and 9 (a) show the acquired initial background. Each figure is represented by the point cloud information obtained 0.1 s after the shape determination time indicated in Table 2. In Figures 8 and 9 (b) and (c), the object in the area circled in red is the movable static object of interest; the randomly colored cells in (b) indicate the presence of a point representing the movable static object. In all the figures, black or its gradient color represents a cell defined as background, and the closer the cell maximum height is to 1.8 m, which is the height of the sensor, the more intense the color.

In Figures 8 and 9 (b), no movable static object extraction is performed; therefore, each object is extracted as a dynamic object. In Figures 8 and 9 (c), the same area is colored in black. This result indicates that the proposed method removes the target objects as movable static objects and then locally redefines all cells occupied by these objects as background.

In this study, each variable was uniquely defined. These values were determined empirically based on the detection accuracy and processing time by verifying the results when other values were used. The proposed method was used to detect people and build environmental maps of various commercial buildings. We confirmed that the proposed method is effective in multiple environments, without the need to adjust the design parameters. However, the design parameters can be optimized as more data on various environments become available.

5. Conclusion

This study proposed an indoor movable static object extraction approach and a local background updating method for building environmental maps describing indoor human activities. From the evaluation experiments, we confirmed that the proposed movable static object extraction method based on different features

of the objects was able to extract only movable static objects and the local background update was conducted appropriately. Consequently, only people were identified as dynamic objects. In the future, to further improve the accuracy, we will develop an efficient detection method for objects to which the proposed method is applied and reduce the time required to determine their shape.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Takumi Nishio received his B.E. and M.E. degrees from the Department of Precision Mechanics, Chuo University, in 2018 and 2020, respectively. He has worked with the Robot Division of Kawasaki Heavy Industries, Ltd., Hyogo, Japan, since 2020. His research interests include observation of human walking activity using LiDAR sensors and environment map building.

Mihoko Niitsuma received her Ph.D. in electrical engineering from the University of Tokyo, Japan, in 2007. She is a professor in the Department of Precision Mechanics, Faculty of Science and Engineering, Chuo University. Her research interests include human-robot interaction/collaboration and human assistive robotics based on the intelligent space (iSpace) approach.

References

- [1] Shuai L, Li C, Guo X, et al. Motion capture with ellipsoidal skeleton using multiple depth cameras. *IEEE Trans. Visual Comput. Graphics*. 2017;23(2):1085–1098. <https://doi.org/10.1109/TVCG.2016.2520926>.
- [2] Hasanuzzaman M, Ampornaramveth V, Zhang T, et al. Real-time vision-based gesture recognition for human robot interaction. *IEEE International Conference on Robotics and Biomimetics*; 2004. p. 413–418.
- [3] Jung E, Lee J, Yi B, et al. Development of a laser-range-finder-based human tracking and control algorithm for a marathoner service robot. *IEEE/ASME Trans. Mechatron*. 2014;19(6):1963.
- [4] Xiao J, Quan L. Multiple view semantic segmentation for street view images. In *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE; 2009. p. 686–693.
- [5] Lim E, Suter D. Multi-scale conditional random fields for over-segmented irregular 3D point-clouds classification. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*; 2008. p. 1–7.
- [6] Lim E, Suter D. 3D terrestrial LIDAR classifications with super-voxels and multi-scale conditional random fields. *Computer-Aided Design*. 2009;41(10):701–710.
- [7] Wang Z, Zhang L, Fang T, et al. A multiscale and hierarchical feature extraction method for terrestrial laser scanning point cloud classification. *IEEE Trans Geosci Remote Sens*. 2015;88(3-4):4839–4854.
- [8] Lodha S, Fitzpatrick D, Helmbold D. Aerial lidar data classification using AdaBoost. *Sixth International Conference*

- on 3-D Digital Imaging and Modeling (3DIM 2007); 2017. p. 435–442.
- [9] Golovinskiy A, Kim V, Funkhouser T. Shape-based recognition of 3D point clouds in urban environments. 2009 IEEE 12th International Conference on Computer Vision; 2009. p. 2154–2161.
 - [10] Lodha S, Kreps E, Helmbold D, et al. Aerial LiDAR data classification using support vector machines (SVM). Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06); 2006. p. 567–574.
 - [11] Chehata N, Guo L, Mallet C. Airborne lidar feature selection for urban classification using random forests. *Int Arch Photogramm Remote Sens Spatial Inf Sci.* 2009;38(3):207–212.
 - [12] Niemeyer J, Rottensteiner F, Soergel U. Conditional random fields for lidar point cloud classification in complex urban areas. *ISPRS Ann Photogramm Remote Sens Spatial Inf Sci.* 2012;1–3:263–268.
 - [13] Horn B. Extended Gaussian images. *Proc IEEE.* 1984; 72(12):1671–1686.
 - [14] Hebert M, Ikeuchi K, Delingette H. A spherical representation for recognition of free-form surfaces. *IEEE Trans Pattern Anal Mach Intell.* 1995;17(7):681–690.
 - [15] Weinmann M, Jutzi B, Mallet C. Feature relevance assessment for the semantic interpretation of 3D point cloud data. *ISPRS Ann Photogramm Remote Sens Spatial Inf Sci.* 2013;II-5/W2:313–318.
 - [16] Johnson A, Hebert M. Using spin images for efficient object recognition in cluttered 3D scenes. *IEEE Trans Pattern Anal Mach Intell.* 1999;21(5):433–449.
 - [17] Caceres J, Slatton K. Improved classification of building infrastructure from airborne lidar data using spin images and fusion with ground-based lidar. *Proc Urban Remote Sens Joint Event.* 2007; 1–7. <https://doi.org/10.1109/URS.2007.371862>.
 - [18] Fang L, Yang B, Chen C, et al. Extraction 3D road boundaries from mobile laser scanning point clouds. 2015 2nd IEEE Int. Conf. on Spatial Data Mining and Geographical Knowledge Services (ICSDM); 2015. p. 102–106.
 - [19] Zhang Z, Zheng J, Wang X, et al. Background filtering and vehicle detection with roadside lidar based on point association. 2018 37th Chinese Control Conference (CCC); 2018. p.7938–7943.
 - [20] Lv B, Xu H, Wu J, et al. Raster-based background filtering for roadside LiDAR data. *IEEE Access.* 2019;7: 76779–76788.
 - [21] Endres F, Plagemann C, Stachniss C, et al. Unsupervised discovery of object classes from range data using latent Dirichlet allocation. *Proc Robot Sci Syst.* 2009. <http://www.roboticsproceedings.org/rss05/p15.pdf>
 - [22] Nishio T, Niitsuma M. Environmental map building to describe walking dynamics for determination of spatial feature of walking activity. The 28th IEEE International Symposium on Industrial Electronics; 2019. p. 2515–2520.
 - [23] Uenoyama N, Niitsuma M. Temporal segmentation of environmental map based on changing rate of activity areas and improvement of tracking method using LRF. 2015 IEEE/SICE Int. Symp. on System Integration; 2012. p. 57–62.
 - [24] Shiga J, Maruyama G, Kato S, et al. Intelligent space-based future convenience store for customer interaction service with new experience, safety, and flexibility. 2022 IEEE/SICE International Symposium on System Integration; 2022. p. 1064–1071.
 - [25] e-stat. [cited 2021 Jan 26]. Available from: <https://www.e-stat.go.jp/>
 - [26] modern-deco. [cited 2021 Jan 26]. <https://www.modern-deco.jp/blog/1092>