drones

MDPI

Article

Vision-Based Indoor Scene Recognition from Time-Series Aerial Images Obtained Using a MAV Mounted Monocular Camera

Hirokazu Madokoro *, Kazuhito Sato and Nobuhiro Shimoi

Faculty of Systems Science and Technology, Akita Prefectural University, Yurihonjo City, Akita 015–0055, Japan; ksato@akita-pu.ac.jp (K.S.); shimoi@akita-pu.ac.jp (N.S.)

* Correspondence: madokoro@akita-pu.ac.jp; Tel.: +81-184-27-2180

Received: 15 December 2018; Accepted: 20 February 2019; Published: 25 February 2019

Abstract: This paper presents a vision-based indoor scene recognition method from aerial time-series images obtained using a micro air vehicle (MAV). The proposed method comprises two procedures: a codebook feature description procedure, and a recognition procedure using category maps. For the former procedure, codebooks are created automatically as visual words using self-organizing maps (SOMs) after extracting part-based local features using a part-based descriptor from time-series scene images. For the latter procedure, category maps are created using counter propagation networks (CPNs) with the extraction of category boundaries using a unified distance matrix (U-Matrix). Using category maps, topologies of image features are mapped into a low-dimensional space based on competitive and neighborhood learning. We obtained aerial time-series image datasets of five sets for two flight routes: a round flight route and a zigzag flight route. The experimentally obtained results with leave-one-out cross-validation (LOOCV) revealed respective mean recognition accuracies for the round flight datasets (RFDs) and zigzag flight datasets (ZFDs) of 71.7% and 65.5% for 10 zones. The category maps addressed the complexity of scenes because of segmented categories. Although extraction results of category boundaries using U-Matrix were partially discontinuous, we obtained comprehensive category boundaries that segment scenes into several categories.

Keywords: category maps; counter propagation networks; leave-one-out cross-validation; micro air vehicles; self-organizing maps; unified distance matrix

1. Introduction

For environmental sensing used for autonomous locomotion robots, sensors that can obtain metric information are used not merely for laser range finders (LRFs) and stereo cameras, but also for depth and visual sensors, such as red, green, blue, and depth (RGB-D) cameras [1]. To estimate a self-position and to create an environmental map together from sensing signals with metric information, simultaneous localization and mapping (SLAM) [2] has been studied widely as an effective and useful approach for indoor environments without using a global positioning system (GPS) for localization. Actually, SLAM is used practically and widely for cleaning robots, pet robots, and guide robots, but ground locomotion robots in terms of unmanned ground vehicles (UGVs) using wheels or crawlers have the effect of block-sensing from static or dynamic objects of various types in our environments of everyday life. Therefore, the accuracy of the vertical-axis map tends to be lower than that of the horizontal axis.

Recently, unmanned aerial vehicles (UAVs), which can fly freely in three-dimensional (3D) environments, have become increasingly popular-not merely for use as a hobby, but also for industrial applications [3]. Assuming its main use for indoor flight with less wind influence, small airframe

UAVs have been designated as micro air vehicles (MAVs). Compared with UGVs, MAVs have excellent sensing capability for the vertical axis because of their advanced locomotion capability and freedom. Nevertheless, because of the limited payload, it is a challenging task for MAVs to equip LRFs or stereo cameras, which can directly obtain metric information. Therefore, an approach to construct a 3D map using a monocular camera with structure from motion (SfM) [4] has been attracting attention, especially in combination with MAVs [5].

For the use of SfM, map construction accuracy depends strongly on the similarity of camera movements, visual fields, and scene features [6]. Particularly, camera parameters and movements have the greatest effect among them. For example, a MAV obtains no metric information if flight patterns are less diverse. Moreover, 3D map creation using SfM is based on signal information that is similar to map creation using SLAM for position estimation [7]. Semantic scene recognition is set to another task for robot vision studies [8]. Improving autonomous flight accuracy for MAVs can be accomplished using a combination of 3D map creation with SfM and semantic scene recognition from appearance patterns on images with visual salience objects as visual landmarks [9]. Semantic scene recognition has been studied widely as a subject-not merely for the advancement of computer vision studies [10], but also for active robot vision studies, including MAV-based mobile vision studies.

As an approach based on appearance changes using machine learning algorithms, this paper presents a vision-based indoor scene recognition method, as depicted in Figure 1 for aerial time-series images with a feature of visualization using a category map based on supervised learning. We used original benchmark datasets obtained in an atrium at our university building. The benchmark datasets comprise 10 zones as ground truth (GT) on two flight routes: a round flight route and a zigzag flight route. We evaluated the effectiveness of our method to demonstrate the recognition results of similarity and relation in each zone for visualization as category maps and their boundaries.

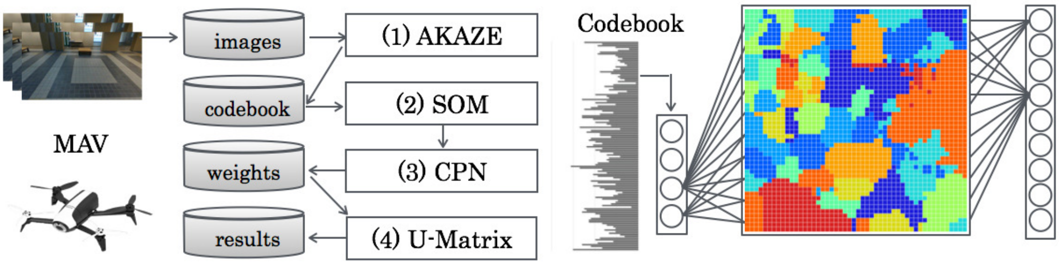


Figure 1. System structure of our proposed method.

The rest of the paper is structured as follows. In Section 2, related studies are presented, especially for monocular vision systems. Sections 3 and 4 present our proposed method based on machine learning and our original indoor datasets obtained using a MAV, respectively. Subsequently, Section 5 presents experimentally obtained results, including parameter optimization, confusion matrix analyses, and our discussion. Finally, Section 6 presents our conclusions and highlights our future work. Herein, we proposed this basic method in the proceedings [11]. For this paper, we present detailed results and a discussion in Section 5.

2. Related Studies

In computer vision studies, numerous methods to recognize semantic scene categories from large numbers of static or dynamic images have been proposed [12]. In general, their recognition targets were mainly in outdoor environments. Quattoni et al. [13] reported that recognition accuracy drops dramatically if outdoor scene recognition methods are applied to indoor scene recognition. Assuming applications for human symbiotic robots, they play actively not merely in outdoor environments, but also in indoor environments. Improving recognition accuracy is necessary for semantic scene recognition methods that are applicable in indoor environments, especially for MAVs. Moreover, the environment for MAV locomotion as aerial robots changes in real time according to human activities and the diverse lifestyles of various people. Therefore, it must be used for MAVs not merely for environmental recognition and understanding, but also for adaptation capability according to environmental changes.

Numerous generalized methods using machine-learning algorithms have been proposed [14] for environmental adaptation. Herein, machine learning is classified roughly into two types: supervised learning and unsupervised learning. Supervised learning requires training datasets annotated in advance with teaching signals such as GT. The burden related to teaching accompanies applying robots due to the obtaining of numerous training datasets in real time according to the tasks and actions. Moreover, the number of target recognition scene categories is set from teaching signals prepared in advance. In contrast, unsupervised learning classifies input datasets without teaching signals. After learning, semantic information is assigned from a part of datasets with GT for the use of a recognizer. For supervised learning methods, categories are extracted from various datasets obtained from environments. Compared with supervised learning methods, which require teaching signals in advance, the burden for teaching is reduced substantially, irrespective of the excessive number of learning data. Moreover, the number of categories was extracted automatically from training datasets according to the environments. Therefore, unsupervised learning has benefits for robotics applications, especially for robots collaborating with humans [15].

As explained herein, unsupervised learning can construct knowledge frameworks autonomously as an intelligent robot [16]. In actuality, neurons that selectively respond to human faces or cat faces were generated using deep learning (DL) frameworks [17], which are attracting attention as a representative of unsupervised learning [18]. We consider that unsupervised learning is useful for actualizing advanced communication and interaction between robots and humans.

The primary process for semantic scene recognition is to extract suitable scene features from image pixels in terms of brightness, color distribution, gradient, edge, energy, and entropy. As a global scene descriptor, Oliva et al. [19] proposed Gist, which has been used widely for perspective scene feature description. However, the recognition accuracy drops dramatically in indoor environments that include numerous small objects because Gist is used for describing comprehensive structures in terms of roads, mountains, and buildings in outdoor environments [13]. As an alternative approach, foreground and background features are used for context-based scene-recognition methods. Quattoni et al. [13] proposed a method to describe features of a spatial pyramid using scale-invalid feature transform (SIFT) [20] as foreground features and Gist [19] as background features. They evaluated their method using a benchmark dataset for 67 indoor categories. Although the maximum recognition accuracy in the scene category of inside a church was 63.2%, the recognition accuracies in the scene categories of a jewelry shop, laboratory, mall, and office were 0%.

Madokoro et al. [21] proposed a context-based scene recognition method using two-dimensional (2D) histograms as visual words (VWs) [22] for voting Gist and SIFT on a 2D matrix. They demonstrated the effectiveness on indoor semantic scene recognition and the superiority of recognition accuracy compared with existing methods evaluated using the KTH-IDOL2 benchmark dataset [23]. Moreover, they proposed a robust semantic scene recognition method for human effects using histogram of oriented gradient (HOG) [24] features [25]. However, both studies included evaluation experiments using a unmanned aerial vehicle (UGV)- based mobile robot. Although they evaluated occlusion and corruption among objects from the perspective of object recognition using an MAV [26], they presented no result obtained from applying their MAV-based platform and machine-learning-based methods for scene recognition. Anbarasu et al. [27] proposed a recognition method using support vector machines (SVMs) [28] after describing scene features using Gist and histogram of directional morphological gradients (HODMGs) for aerial images obtained using a MAV. They conducted evaluation experiments using their originally collected datasets. Although they obtained sufficient recognition accuracies, recognition targets were merely three scene categories: corridors, staircases, and rooms.

3. Proposed Method

Our proposed method based on part-based features and supervised learning, as depicted in Figure 1, comprises four steps:

(1) Feature extraction using AKAZE;
(2) Codebook generation using SOMs;
(3) Category map generation using CPNs;
(4) Category boundary extraction using U-Matrix.

First, part-based features were extracted from aerial time-series scene images using accelerated KAZE (AKAZE) descriptors [29]. Subsequently, VWs based on bags of keypoints [30] were generated using self-organizing maps (SOMs) [31] as codebooks with the integrated feature dimension, as depicted in Figure 2. For creating category maps, VWs were presented to counter propagation networks (CPNs) [32] as input features. For our method, scene relational features were visualized on category maps [33] because of topological mapping of spatial scene relations, as depicted in Figure 3. Finally, category boundaries were extracted using a unified distance matrix (U-Matrix) [34] from weights between the input layer and the mapping layer on CPNs for extracting categories according to scene features.

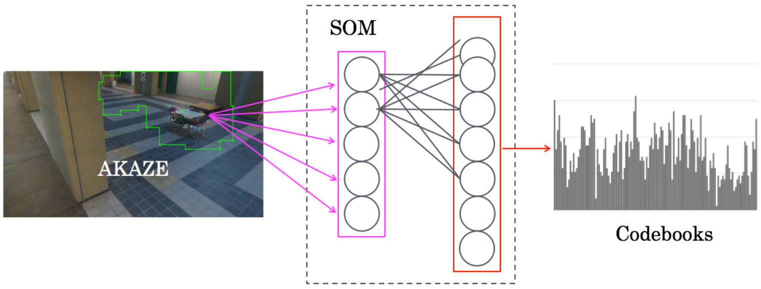


Figure 2. Procedure of generated visual words (VWs) using self-organizing maps (SOMs) as codebooks with the integrated feature dimension.

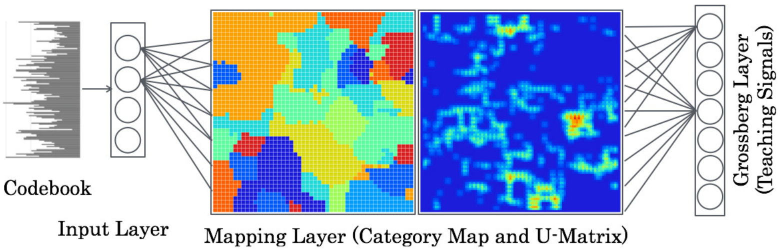


Figure 3. Procedure of visualized scene relational features on category maps as topological mapping of spatial scene relations.

Our method extracts scene categories for visualizing boundary depth after calculating the similarity of neighborhood categories mapped as category maps on CPNs. Detailed procedures of our method, which consists of three machine-learning algorithms with supervised and unsupervised modes, were presented in an earlier report [33] For this paper, we used the network as the supervised mode.

3.1. SOMs

As depicted in Figure 2, the network architecture of SOMs comprises two layers: an input layer and a mapping layer. Input signals are presented to the input layer. No teaching signals are presented to the mapping layer because of unsupervised learning.

The learning algorithm of SOMs is the following. $x_i(t)$ and $w_{i,j}(t)$ respectively denote input data and weights from an input layer unit i to a mapping layer unit j at time t. Herein, I, J respectively denote the total numbers of the input layer and the mapping layer. $w_{i,j}(t)$ were initialized randomly before learning. The unit for which the Euclidean distance between $x_i(t)$ and $w_{i,j}(t)$ is the smallest is sought as the winner unit of its index c as:

$$c = \underset{1 \le j \le J}{\operatorname{argmin}} \sqrt{\sum_{i=1}^{I} (x_i(t) - w_{i,j}(t))^2}. \tag{1}$$

As a local region for updating weights, the neighborhood region $N_c(t)$ is defined as the center of the winner unit c, as:

$$N_c(t) = \lfloor \mu \cdot J \cdot \left(1 - \frac{t}{O}\right) + 0.5 \rfloor. \tag{2}$$

Therein, $\mu (0 < \mu < 1.0)$ is the initial size of $N_c(t)$; O is the maximum iteration for training. Coefficient 0.5 is appended as the floor function $\lfloor \cdot \rfloor$ for rounding.

Subsequently, $w_{i,j}(t)$ of $N_c(t)$ was updated to close input feature patterns.

$$w_{i,j}(t+1) = w_{i,j}(t) + \alpha(t)(x_i(t) - w_{i,j}(t)). \tag{3}$$

Therein, $\alpha(t)$ is a learning coefficient that decreases along with the progress of learning. $\alpha(0)(0 < \alpha(0) < 1.0)$ is the initial value of $\alpha(t)$. $\alpha(t)$ is defined at time t as:

$$\alpha(t) = \alpha(0) \cdot \left(1 - \frac{t}{O}\right). \tag{4}$$

In the initial stage, the learning speed is higher when this rate is high. In the final stage, the learning converges while the range decreases.

For this module, the input features of I dimension are quantized into the J dimension, which is a similar dimension to the number of units on the mapping layer. The module output $y_j(t)$ is calculated as:

$$y_j(t) = \sqrt{\sum_{i=1}^{I} (x_i(t) \cdot w_{i,j}(t))^2}. \tag{5}$$

This module is connected to the labeling module at the training phase. For the testing phase, this module is switched to the mapping module. Moreover, this module is passed when input features are used without creating codebooks directly.

3.2. CPNs

As depicted in Figure 3, the network architecture of CPNs comprises three layers: an input layer, a mapping layer, and a Grossberg layer. The input layer and mapping layer resemble those of SOMs. Teaching signals are presented to the Grossberg layer.

The learning algorithm of CPNs is the following. Herein, for visualization characteristics of category maps, we set the mapping layer to a two-dimensional structure $X \times Y$ unit. For this paper, we set one dimension of the input and Grossberg layers, although they can take any structure. The numbers of units are I and K, respectively. $u_{i,j(x,y)}(t)$ are weights from an input layer unit i to a mapping layer unit $j(x,y)$ at time t. $v_{j(x,y),k}(t)$ are weights from a Grossberg layer unit k to a mapping layer unit $j(x,y)$ at time t. These weights are initialized randomly before learning. $x_i(t)$ are training data to present to the input layer unit i at time t. The unit for which the Euclidean distance between $x_i(t)$ and $u_{i,j(x,y)}(t)$ is the smallest is sought as the winner unit. c(x,y) is the index of the unit.

$$c(x,y) = \underset{(1,1)\leq j(x,y)\leq (X,Y)}{\text{argmin}} \sqrt{\sum_{i=1}^{I}(x_i(t) - u_{i,j(x,y)}(t))^2}, \tag{6}$$

The neighborhood region $N_{(c_x,c_y)}(t)$ around $c(x,y)$ is defined as:

$$N_{c(x,y)}(t) = \lfloor \mu \cdot (X,Y) \cdot \left(1 - \frac{t}{O}\right) + 0.5 \rfloor, \tag{7}$$

where $\mu (0 < \mu < 1.0)$ is the initial size of the neighborhood region, and O is the maximum iteration for training. $u^i_{n,m}(t)$ of $N_{c(x,y)}(t)$ were updated to close input feature patterns using Kohonen's learning algorithm as:

$$u_{i,j(x,y)}(t+1) = u_{i,j(x,y)}(t) + \alpha(t)(x_i(t) - u_{i,j(x,y)}(t)), \tag{8}$$

Subsequently, $v_{j(x,y),k}(t)$ of $N_{c(x,y)}(t)$ were updated to close teaching signal patterns using Grossberg's learning algorithm.

$$v_{j(x,y),k}(t+1) = v_{j(x,y),k}(t) + \beta(t)(T_k(t) - v_{j(x,y),k}(t)), \tag{9}$$

Herein, T_k are training signals obtained using ART-2. $\alpha(t)$ and $\beta(t)$ are learning coefficients that have decreasing values with the progress of learning. $\alpha(0)$ and $\beta(0)$ denote the initial values of $\alpha(t)$ and $\beta(t)$, respectively. The learning coefficients are given as:

$$\begin{bmatrix} \alpha(t) \\ \beta(t) \end{bmatrix} = \begin{bmatrix} \alpha(0) \\ \beta(0) \end{bmatrix} \cdot \left(1 - \frac{t}{O}\right). \tag{10}$$

In the initial stage, the learning is done rapidly when the efficiencies are high. In the final stage, the learning converges, although the efficiencies decrease. As the maximum number of $v_{j(x,y),k}(t)$ for the k-th Grossberg unit, category $L_k(t)$ is searched as:

$$L_k(t) = \underset{(1,1)\leq j(x,y)\leq (X,Y)}{\text{argmax}} v_{j(x,y),k}(t). \tag{11}$$

A category map is created after determining categories for all units. Test datasets are presented to the network that is created through learning. A mapping layer unit, which is the minimum of the Euclidean distance as the similarity of test data and feature patterns, is burst. Categories for these units are recognition results for CPNs.

4. Datasets

4.1. Experimental Environment

We obtained original aerial time-series image benchmark datasets in an atrium at our university building. Figure 4 depicts a photograph of the atrium from the ground floor to 74 m in the longitudinal direction. Its width and height are, respectively, 17 m and 18 m as a void space. The scene structure in the environment is holistically simple with generic objects of chairs, tables, bulletin boards, partitions, benches, trash boxes, desks, and posters. Occasionally, pedestrians were in the environment during image acquisition.

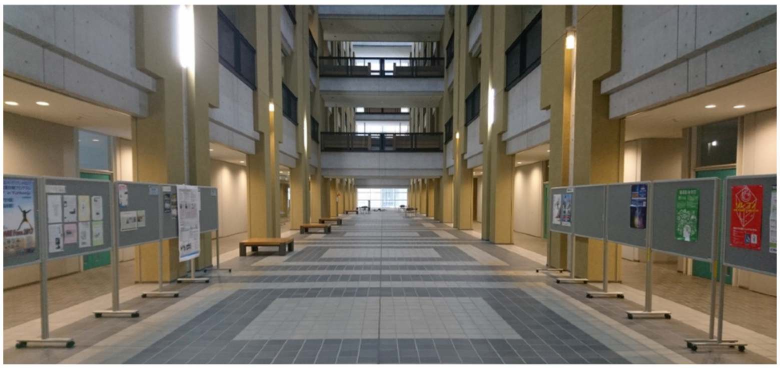


Figure 4. Overview of atrium as an experimental environment: 74 m long × 17 m wide × 18 m high.

4.2. MAV

For this experiment, we used Bebop2 by Parrot SA as depicted in Figure 5. The body size was 328 mm in length × 382 mm in width × 89 mm in height for the front side of the camera view range. The body weight was 500 g with an attached 2700 mAh LiPo battery. Although the payload was not disclosed by the manufacturer, a complementary metal-oxide semiconductor (CMOS) camera was mounted for the main vision sensor. The camera resolution was 4096 × 3072 pixels for still images and 920 × 1080 pixels for full high-definition (FHD) video images. For this experiment, we used the latter resolution for time-series images.

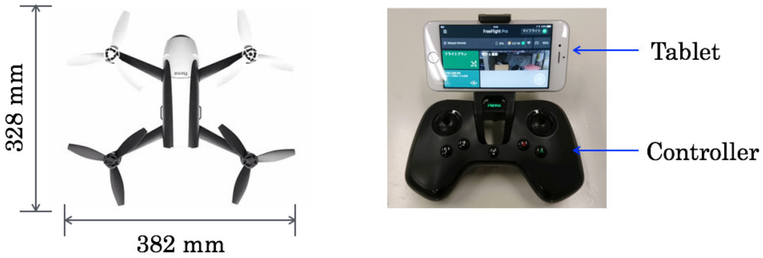


Figure 5. Our MAV and controller with tablet.

As flight performance, the maximum vertical and horizontal flight speeds were, respectively, 18 m/s and 6 m/s driven by four 1280 kV motors with six-inch rotors. For a measurement and positioning system, signal-receiving modules of GPS and global navigation satellite system (GLONASS) were embedded on the MAV. For indoor environments that are denied these positioning signals, precision operation is necessary for a steady and safe flight. Therefore, we operated the MAV manually using a dedicated controller and a tablet computer for avoiding a flyaway or a crash.

4.3. Obtained Datasets

In the atrium as our experimental environment, we set two flight routes: a round flight route and a zigzag flight route. We obtained flight scene images for five rounds on each flight route with an interval to charge the MAV battery. Figures 6 and 7 depict both flight routes with divided zones as recognition targets. We assigned zone colors using heat maps. According to the flight patterns and the similarity of appearance scenes, we divided flight routes into 10 zones labeled as Zones A–Z for both flight routes.

Figure 8 portrays sample images of round flight datasets (RFDs) on the upper row and zigzag-flight image datasets (ZFDs) on the lower row. The field of view (FOV) in RFDs includes numerous scene images for the longitudinal direction. For ZFDs, the FOV includes numerous scene images for the lateral direction. Table 1 presents the number of image frames in each dataset.

We converted the original frame rate of 30 fps to 1 fps by linear downsampling. After this conversion, the number of image frames in Figure 1 corresponds to the flight time in seconds. The number of RFD images is greater than that of ZFD images because of the differences in flight time and distance.

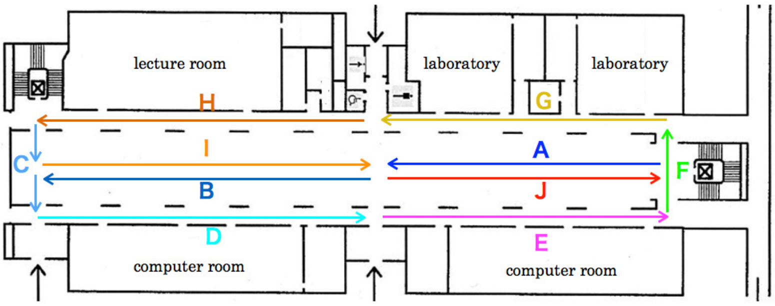


Figure 6. Round flight route and zones.

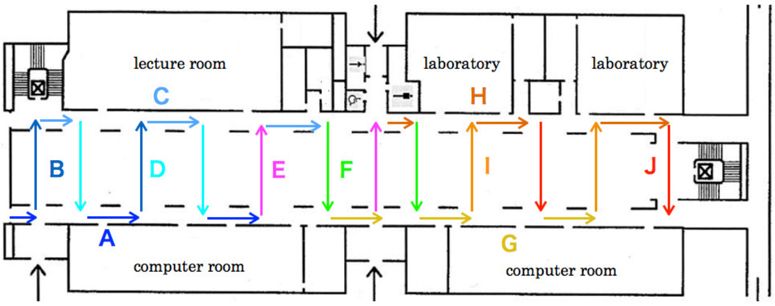


Figure 7. Zigzag flight routes and zones.

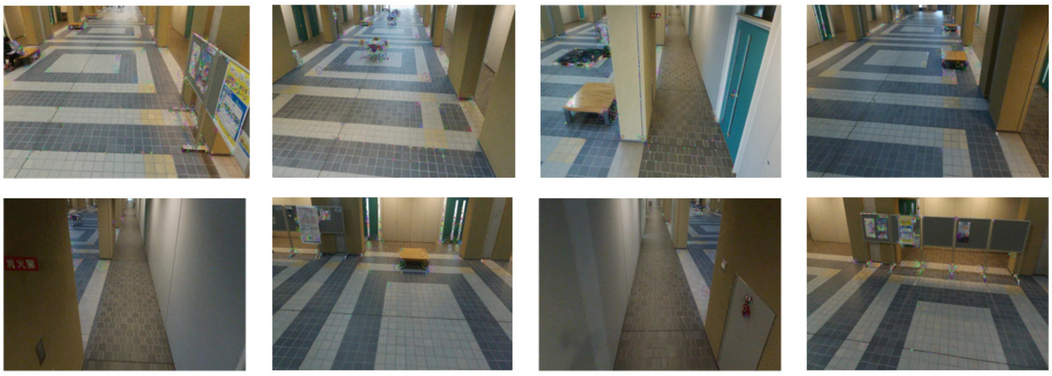


Figure 8. Sample images of round flight datasets (RFDs) (**upper**) and zigzag-flight image datasets (ZFDs) (**lower**).

Table 1. Number of images in respective datasets [images].

Dataset	1	2	3	4	5	Sum
Round flight	679	729	693	702	713	3516
Zigzag flight	595	565	554	559	537	2810

For the five datasets in each flight route, we evaluated our method with leave-one-out cross-validation (LOOCV) [35], which means the use of one dataset for validation and other four datasets for training [36]. Herein, we used just Dataset 1 for creating 3D environmental maps and parameter optimization for preliminary experiments. We used all five datasets as LOOCV for the scene recognition evaluation. Table 2 defines the combination of learning and validation datasets as validation patterns (VPs).

Table 2. Combination of learning and validation datasets as validation patterns (VPs).

VP	Learning Datasets	Validation Dataset
1	2, 3, 4, 5	1
2	3, 4, 5, 1	2
3	4, 5, 1, 2	3
4	5, 1, 2, 3	4
5	1, 2, 3, 4	5

4.4. Building Results of 3D Maps

We used PhotoScan by Agisoft Limited Liability Company for building 3D maps before the scene recognition experiment. Because of its simple user interface with automatic inner parameter estimation, SfM application software PhotoScan has been widely used, especially in remote sensing and civil engineering fields [37].

Figure 9 depicts the building results of 3D maps for both FDs. The upper and lower results respectively depict bird's-eye-view 3D maps and estimated camera positions with the respective maps. Although comprehensive 3D maps were generated automatically, numerous deficient pixels were locally apparent. The flight routes comprised straight flight and 90 degree rotations with a fixed altitude. However, the estimated camera positions included the movements of the horizontal and vertical directions. These experimental results demonstrate the limitation of SfM for indoor images obtained using a MAV without GPS signals. We considered that precise GPS signals are necessary for 3D map construction. Herein, indoor GPS technology using Wi-Fi is still quite costly for practical use. Therefore, this study examines the possibility of visual position estimation.

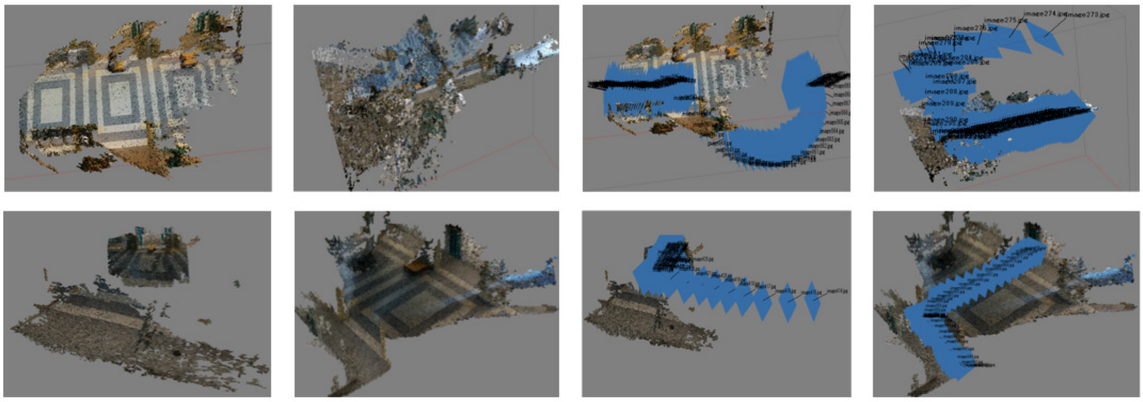


Figure 9. Building results of 3D maps for RFDs (**upper**) and ZFDs (**lower**).

5. Evaluation Experiment

5.1. Parameter Optimization

As a preliminary experiment, we optimized three major parameters that influence the recognition accuracy. For evaluation criteria, recognition accuracy recognition accuracy R [%] for a validation dataset is defined as:

$$R = \frac{C}{N} \times 100, \tag{12}$$

where C and N respectively represent the total numbers of validation images and correct recognition images that matched to zone labels such as GT.

We used RFDs for this optimization. The first optimization parameter was codebook dimensions of input features. Whereas the size of a category map and the number of learning iterations were set, respectively, to 50 × 50 units and 10,000 iterations, we changed the coefficient n of the codebook dimension 2^n from Steps 5 to 10 by 1. The optimization result, as denoted in Figure 10 revealed that the recognition accuracy increased according to increased n. The maximum accuracy was 71.7% in $n = 9$, which corresponds to 512 codebook dimensions.

The second optimization parameter is the number of category map units. Whereas the codebook size and the number of learning iterations were set, respectively, to 256 dimensions and 10,000 iterations, we changed the category map units from 10 × 10 units to 60 × 60 units in 10 × 10 unit intervals. The optimization result, as denoted in Figure 11 revealed that the recognition accuracy increased according to the category map size. The maximum accuracy was 71.7% in 50 × 50 category map units, which corresponds to the 512 codebook dimensions.

The third optimization parameter is the number of learning iterations. Whereas the codebook size and the size of a category map were set, respectively, to 256 dimensions and 50 × 50 units, we changed the learning iterations of five steps: 100, 1000, 5000, 10,000, 50,000, and 100,000. The optimization result, as denoted in Figure 12 revealed that the recognition accuracy increased according to the iterations up to 10,000. The maximum accuracy was 71.7% in 10,000, which corresponds to the 512 codebook dimensions and 50 × 50 category map units. The results of the above three preliminary experiments showed that we used 512 codebook dimensions, 50 × 50 category map units, and 10,000 learning iterations as optimal values.

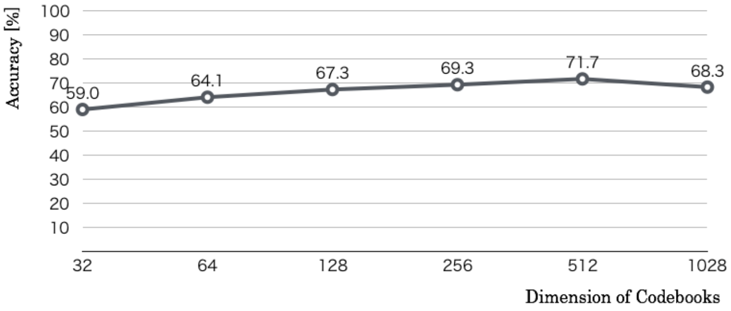


Figure 10. Relation between the codebook dimension and recognition accuracies.

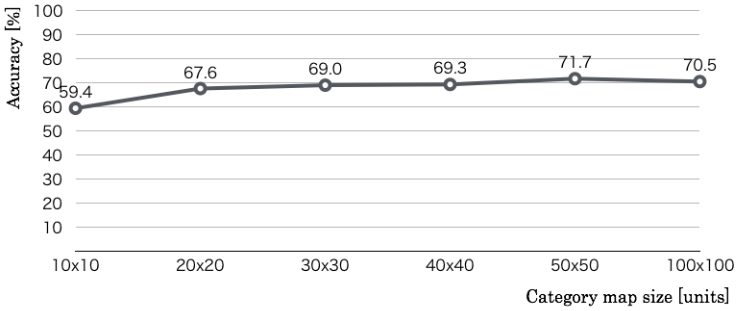


Figure 11. Relation between the size of category maps and recognition accuracies.

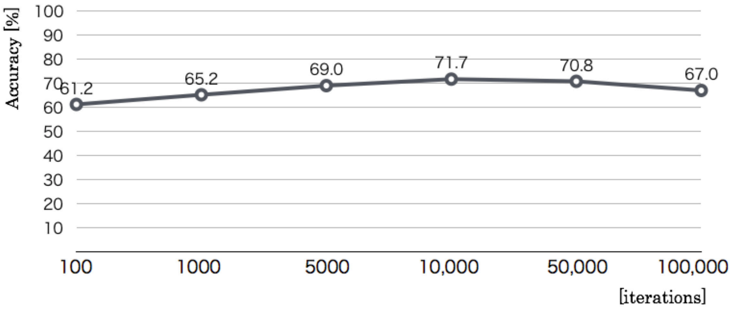
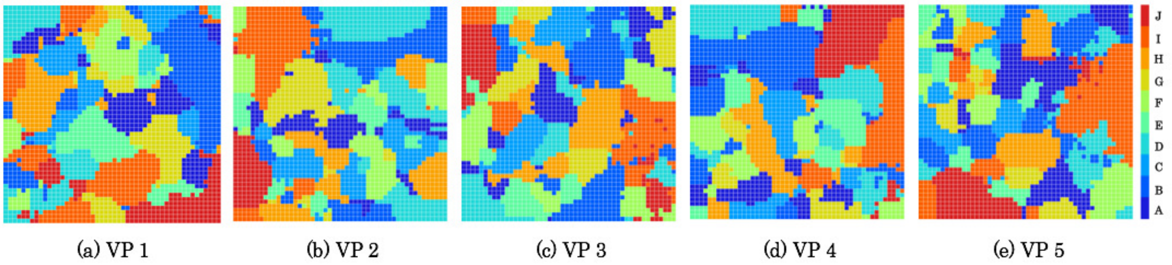


Figure 12. Relation between the size of learning iterations and recognition accuracies.

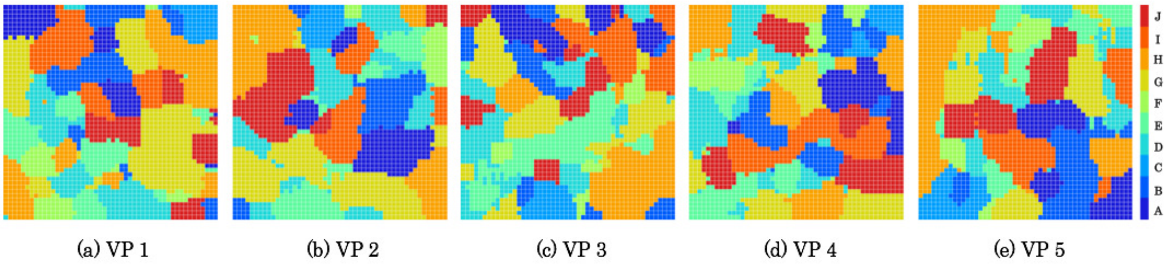
5.2. Created Category Maps

Figures 13 and 14 respectively depicts category maps created from RFDs and ZFDs. Categories are shown using heat maps according to color temperatures of ten steps. Zones A and J are allocated, respectively, to the lowest and the highest temperature colors. According to the progress of flight zones, as depicted in Figures 6 and 7, categories corresponding with zones were changed from low-temperature colored units to high-temperature colored units in heat maps. After learning, category maps are used for recognizers.



(a) VP 1 (b) VP 2 (c) VP 3 (d) VP 4 (e) VP 5

Figure 13. Category maps created from RFDs.



(a) VP 1 (b) VP 2 (c) VP 3 (d) VP 4 (e) VP 5

Figure 14. Category maps created from ZFDs.

Category maps were divided into several categories of respective zones with categorical representation reflected in scene complexity. The category maps of ZFDs denote higher scene complexity than those of RFDs because of segmented categories. Moreover, the category maps denote not merely high complexity of category boundaries, but also distributional characteristics to different positions with similar zones. Because category maps are created from training datasets, a single unit bursts from a test dataset based on the mechanism of winner-take-all (WTA) competition that obtains everything by a winner neuron. The label of a WTA unit is judged in terms of recognition results.

5.3. Recognition Results

Figure 15 depicts recognition accuracies for five LOOCV patterns. The mean recognition accuracy of RFDs was 71.7% for five datasets evaluated using LOOCV. The maximum and minimum recognition accuracies were, respectively, 81.1% for VP 4 and 51.0% for VP 2. The accuracy gap separating them was 27.8%. The recognition accuracy in Zone D of VP 4 is the lowest. Herein, we analyze false recognition tendencies from scene image features and flight routes. Numerous images in Zone D were falsely recognized to Zones E and I. Figure 6 shows that the flight route from Zone D to Zone E comprised a passage between pillars and walls. Feature changes in these zones were slight compared with those in other zones. Moreover, numerous similar scene features existed because green doors for the computer rooms are lined up continuously. The flight orientation in Zone I was the same as that in Zone D, which included salient objects in the atrium to the scene images. We consider that false recognition occurred from objects in terms of chairs and tables that exist in Zone I as landmarks.

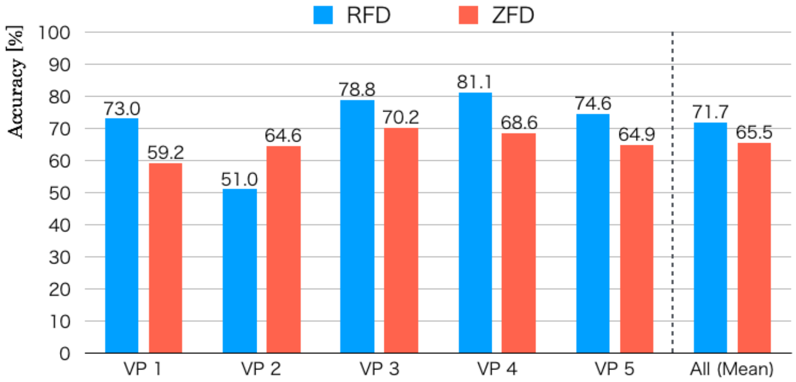


Figure 15. Recognition accuracies for all LOOCV patterns in respective VPs.

As the lowest recognition accuracy, images in Zones D, E, and G of VP 2 were falsely recognized as Zone J. Zones D, E, and G were a straight pathway between pillars and walls. False recognition images included a partial view to the center of the atrium that can be seen from pillars. In contrast, Zone J comprised a route, of which the MAV flew to the center of the atrium. For obtaining images of Dataset 2, the MAV flew to the pillar side compared with other datasets. Therefore, we infer that false recognition occurred from images between the pillars.

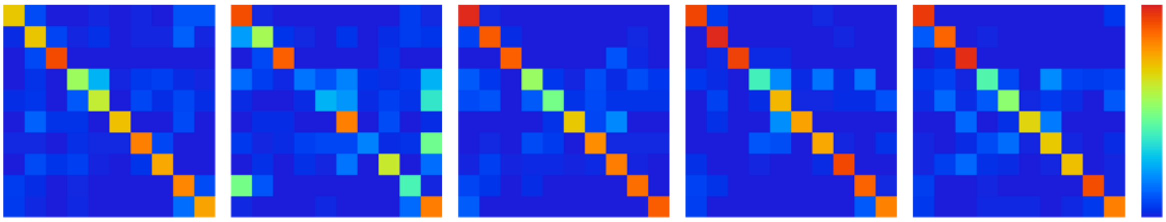
The mean recognition accuracy for five datasets of ZFDs was 65.5%, which was 6.2 percentage points lower than that of the RFDs. The scene complexity of the zigzag flight route was greater than that of the round flight route because it included numerous 90 degree rotations. The salient features in the ZFDs were fewer than in the RFDs because the MAV flew mainly in a lateral direction to both walls in the atrium. The maximum and minimum recognition accuracies were, respectively, 70.2% for VP 3 and 59.2% for VP 1. The gap separating them was 11.0%, which was lower than that of the RFDs.

False recognition was high for VP 3 in Zone H. As depicted in Figure 7, the flight orientation is similar in Zones H and G with different routes. Objects on the scene images in both zones included pillars and doors merely because the flight route existed between pillars and walls. We infer that false recognition occurred from numerous similar scene features and their slight changes. Numerous images in Zone H were falsely recognized as Zone A or C. We infer that false recognition occurred from scene feature similarity.

For VP 1, with the lowest recognition accuracy, numerous images between Zones H and G and between Zones B and F were falsely recognized. However, falsely recognized images in Zone G were slight, except those of Zone H. We consider that salient features of pedestrians, desks, and sign boards are included in Zone H. In Zones B and F, the MAV flew a straight route to both walls of the lateral side in the atrium in opposite directions. The scene images in Zone F that were falsely recognized as Zone F included white walls, of which numerous features were overlapped in the different zones. In contrast, false recognition images in Zone B were slightly allocated to other zones except for Zone F.

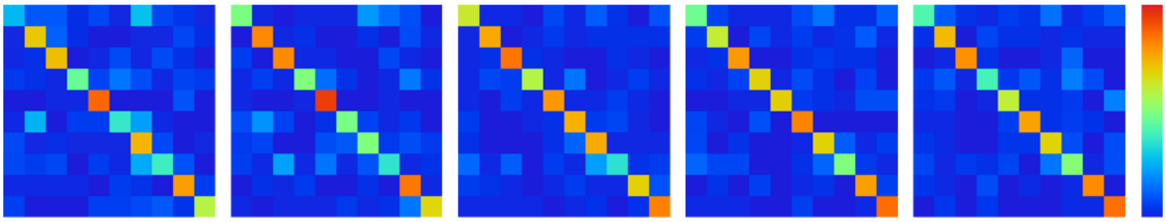
5.4. Confusion Matrix Analysis

Figures 16 and 17 respectively depict confusion matrixes as detailed results of recognition accuracies, as presented in Figure 15. The matrices comprise 10 rows × 10 columns that correspond to the number of segmented zones. The data existence rates on the matrixes are represented using heat maps. Correct numbers of images are represented on the diagonal matrix elements. The false recognition rate is high if the color temperature is high on the elements, except for the diagonal elements. For the base of the horizontal lines, false recognition images can be specified to the positions of vertical elements. A comprehensive tendency is that false recognition images are apparently numerous on the elements near the diagonal elements because scene features are changed sequentially according to the MAV flight.



(a) VP 1 (b) VP 2 (c) VP 3 (d) VP 4 (e) VP 5

Figure 16. Confusion matrixes for the round flight recognition results.



(a) VP 1 (b) VP 2 (c) VP 3 (d) VP 4 (e) VP 5

Figure 17. Confusion matrixes for the zigzag flight recognition results.

Table 3 presents a confusion matrix of VP 4, which has the highest recognition accuracy in the RFD. The bold numbers represent the maximum number of images in each zone. For this result, the maximum number of images is distributed on the diagonal elements. Particularly, false recognition images in Zone D were more numerous than in other zones. In all, 52 of 86 images in Zone D were falsely recognized to other zones, including 16 false recognition images to Zone E. In contrast, other zones were recognized correctly, especially in Zones B and C, which were recognized correctly except for the numbers of images for one image in Zone B and three images in Zone C.

Table 3. Confusion matrix of VP 4 on RFDs [images].

	A	B	C	D	E	F	G	H	I	J
A	**69**	4	0	0	0	0	1	0	0	0
B	0	**82**	0	0	0	0	0	1	0	0
C	0	1	**45**	1	1	0	0	0	0	0
D	4	2	0	**34**	16	2	13	2	13	0
E	0	5	0	2	**56**	1	1	2	2	7
F	0	1	0	0	6	**24**	0	0	2	0
G	0	0	0	8	7	0	**62**	1	0	0
H	3	0	0	1	0	0	2	**77**	0	0
I	5	3	0	0	0	0	0	0	**66**	1
J	4	0	0	0	0	0	0	0	7	**54**

Table 4 denotes a confusion matrix of VP 2, which has the lowest recognition accuracy. The bold numbers, which demonstrate the maximum numbers of images, are available on the diagonal elements. The numbers of false recognition images in Zones D, E, and G were, respectively, 89 of 105 images, 61 of 82 images, and 67 of 81 images. Particularly, 27, 30, and 38 images in these zones were commonly falsely recognized as Zone J. However, false recognition images between Zones C and F were merely six.

We obtained a similar tendency from detailed confusion matrixes in the ZFD. Therefore, we consider that the arbitrary categories selected were not sufficiently distinct as to be separable in image space.

Table 4. Confusion matrix of VP 2 on RFDs [images].

	A	B	C	D	E	F	G	H	I	J
A	**66**	17	0	0	1	0	1	0	4	1
B	16	**41**	3	1	0	3	0	2	4	3
C	0	4	**46**	0	0	0	0	2	0	0
D	14	6	2	16	10	18	4	3	5	**27**
E	2	0	0	2	21	17	2	5	3	**30**
F	0	1	1	0	0	**30**	0	3	0	1
G	4	1	2	5	6	6	14	2	3	**38**
H	0	3	0	3	1	13	0	**54**	0	12
I	**39**	8	0	0	0	0	0	0	33	1
J	1	0	0	0	1	0	0	0	8	**51**

5.5. Extracted Category Boundaries

Based on weights between the input layer and the mapping layer on CPNs, category boundaries were extracted with a U-Matrix that calculates the similarity of neighborhood units. Figures 18 and 19 respectively depict extraction results of category boundaries on category maps as depicted in Figures 13 and 14. The depth of category boundaries is depicted using heat maps. Category boundaries of much weight difference among units are depicted in a high-temperature near-red color. Herein, slight output signals shown as low-temperature near-blue color were reduced automatically using a threshold extracted using the Otsu Method [38].

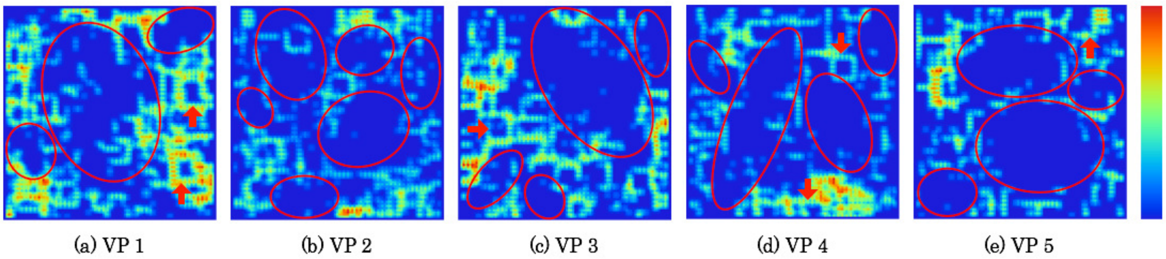


Figure 18. Extraction result of category boundaries on RFDs.

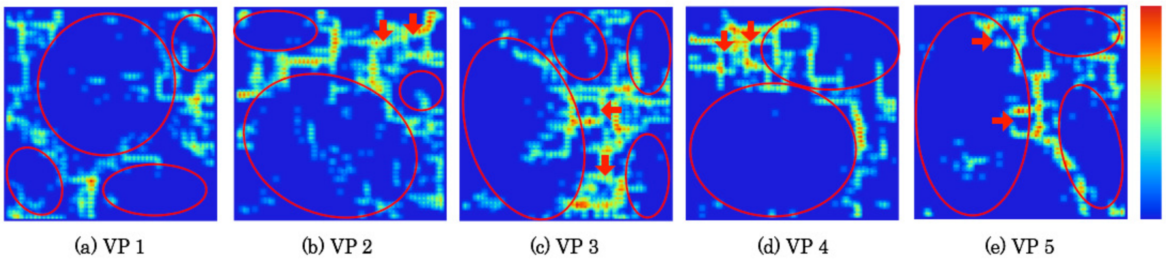


Figure 19. Extraction result of category boundaries on ZFDs.

The red ellipses on respective category maps denote comprehensive categories. The extraction results demonstrate that up to six categories were created with discontinuity boundaries and dependent units. The red-filled arrows denote local categories. As shown in category maps depicted in Figures 13 and 14, the first and second neighborhood regions up to 25 units correspond to the respective local categories. In the next section, we attempt to discuss how many actual categories there are in the image space.

5.6. Discussion

We attempted to analyze mapping characteristics of our method using superimposed scene images as a visual observation. As a mapping result, category boundaries were extracted using the U-Matrix. Up to eight typical scene images are shown on independent categories segmented by the category boundaries. Figure 20a depicts representative images in each category superimposed on category boundaries obtained from VP 4 of RFDs. The category map was segmented to two comprehensive categories, as shown in Figure 18d by red ellipses. We annotated numbers from the order of category sizes as Categories 1–4 in Figure 20a. Scene images related to the longitudinal direction were mapped to Categories 1 and 2. Numerous scene images that include objects are present in Category 1. Moreover, several scene images in which the MAV flew the lateral direction of the center of the atrium in Zone F were in Category 1.

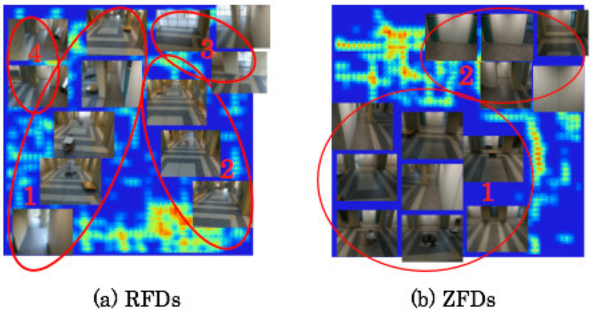


Figure 20. Representative images in each category superimposed on the category boundaries of VP 4 of RFDs and ZFDs.

In contrast, Category 2 had numerous scene images without objects. Although Category 2 included images with objects, these were not merely small objects because of their distant location, but also a part of the background, which is difficult to distinguish. We inferred that this feature difference divided images into two independent categories. In Category 3, some images encompassed outdoor environments through a transparent glass wall when the MAV flew in the direction of the wall in the north part of the atrium. An independent category was mapped on the category map because appearance features were widely different when outdoor scene features were included in the images obtained in indoor environments. Categories 1 and 4 were partially connected without a boundary. Therefore, Category 4 included images obtained in terminal areas in the atrium.

Figure 20b depicts representative images in each category superimposed on the category boundaries obtained from VP 4 of ZFDs. The category map was segmented to four comprehensive categories, as depicted in Figure 19d with red ellipses. We annotated numbers from the order of the category size as Categories 1–2 in Figure 20b. Scene images related to the longitudinal direction of the terminal flight routes in Zones A, C, G, and H were mapped to Category 1. Although Category 1 included flight images of the lateral direction in Zones B, D, E, F, I, and J, these were wide-view images after a 90 degree turn from the longitudinal direction. Category 2 included images near the walls in Zones B, D, F, I, and J of the flight in the lateral direction without images obtained from the aerial scenes of the longitudinal direction. We regard independent categories as created from Category 1 because the feature points were fewer according to the ratio of the walls in the view range. Herein, the effects of existing objects in these scene images were slight for this category map.

6. Conclusions

This paper presented a vision-based scene recognition method from indoor aerial time-series images using category maps that were mapped in topologies of features into a low-dimensional space based on competitive and neighborhood learning. The experimentally obtained results with LOOCV for datasets divided with 10 zones for both flight routes revealed that mean recognition accuracies for RFDs and ZFDs were 71.7% and 65.5%, respectively. The created category maps addressed the complexity of scenes because of the segmented categories. Although extraction results of category boundaries using U-Matrix were partially discontinuous, comprehensive category boundaries were obtained for scenes segmented into several categories.

As a subject of future work, we will integrate 3D maps created using SfM and scene recognition results obtained using our method. Moreover, we expect to fly a MAV using autopilot to obtain various datasets automatically in various environments to demonstrate the versatility and utility of our method, particularly localization and navigation. Furthermore, we will append a DL framework to CPNs, not merely for creating multilayer category maps according to respective layers, but also for learning contexts and time-series features to improve recognition accuracy.

Author Contributions: Conceptualization, H.M. and K.S.; methodology, H.M.; software, H.M.; validation, N.S.; formal analysis, H.M.; investigation, N.S.; resources, H.M.; data curation, K.S.; writing—original draft preparation, H.M.; writing—review and editing, H.M.; visualization, K.S.; supervision, H.M.; project administration, H.M.; funding acquisition, H.M.

Funding: This research was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 17K00384.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AKAZE	accelerated KAZE
CMOS	complementary metal-oxide semiconductor
CPN	counter propagation network
DL	deep learning
FOV	field of view
FHD	full high definition
GLONASS	global navigation satellite system
GPS	global positioning system
GT	ground truth
HODMG	histogram of directional morphological gradient
HOG	histogram of oriented gradient
LOOCV	leave-one-out cross-validation
LRF	laser range finder
MAV	micro air vehicle
RFD	round flight dataset
RGB-D	red, green, blue, and depth
SfM	structure from motion
SIFT	scale-invariant feature transform
SLAM	simultaneous localization and mapping
SOM	self-organizing map
SVM	support vector machine
UAV	unmanned aerial vehicle
UGV	unmanned ground vehicle
U-Matrix	unified distance matrix
VP	validation pattern
VW	visual word
WTA	winner-take-all
ZFD	zigzag flight dataset
2D	two-dimensional
3D	three-dimensional

References

1. Huang, A.S.; Bachrach, A.; Henry, P.; Krainin, M.; Maturana, D.; Fox, D.; Roy, N. Visual Odometry and Mapping for Autonomous Flight Using an RGB-D Camera. *Robot. Res.* **2017**, 235–252. [CrossRef]
2. Dissanayake, G.; Newman, P.; Clark, S.; Durrant-Whyte, H.F.; Csorba, M. An Experimental and Theoretical Investigation into Simultaneous Localisation and Map Building (SLAM). In *Experimental Robotics VI, Lecture Notes in Control and Information Sciences*; Springer: London, UK, 2000; pp. 265–274.
3. Hassanalian, M.; Abdelkefi, A. Classifications, Applications, and Design Challenges of Drones: A Review. *Prog. Aerosp. Sci.* **2017**, *91*, 99–131. [CrossRef]
4. Ullman, S. The Interpretation of Structure from Motion. *Proc. R. Soc. Lond.* **1979**, *203*, 405–426. [CrossRef] [PubMed]
5. Schonbergerab, J.L.; Fraundorfera, F.; Frahmb, J.M. Structure-from-Motion for MAV Image Sequence Analysis with Photogrammetric Applications. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2014**, *40*, 305–312. [CrossRef]
6. Clapuyt, F.; Vanacker, V.; Oost, K.V. Reproducibility of UAV-Based Earth Topography Reconstructions Based on Structure-from-Motion Algorithms. *Geomorphology* **2016**, *260*, 4–15. [CrossRef]
7. Davison, A.J.; Reid, I.D.; Molton, N.D.; Stasse, O. MonoSLAM :Real-Time Single Camera SLAM. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1052–1067. [CrossRef] [PubMed]
8. Wu, J.; Christensen, H.I.; Rehg, J.M. Visual Place Categorization: Problem, Dataset, and Algorithm. In Proceedings of the IEEE/RSJ International Conference Intelligent Robots and Systems, St. Louis, MO, USA, 10–15 October 2009; pp. 4763–4770. [CrossRef]
9. Wu, J.; Rehg, J.M. CENTRIST: A Visual Descriptor for Scene Categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *38*, 1489–1501. [CrossRef]
10. Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; Torralba, A. Places: A 10 million Image Database for Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1452–1464. [CrossRef] [PubMed]
11. Madokoro, H.; Ueda, S.; Sato, K. Semantic Indoor Scene Recognition of Time-Series Arial Images from a Micro Air Vehicle Mounted Monocular Camera. In Proceedings of the 16th International Conference on Control, Automation and Systems (ICCAS), PyeongChang, GangWon, Korea, 17–20 October 2018; pp. 265–270.
12. Siagian, C.; Itti, L. Rapid Biologically Inspired Scene Classification Using Features Shared with Visual Attention. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 300–312. [CrossRef] [PubMed]
13. Quattoni, A.; Torralba, A. Recognizing Indoor Scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 413–420. [CrossRef]
14. Torralba, A.; Murphy, K.P.; Freeman, W.T.; Rubin, M.A. Context-Based Vision System for Place and Object Recognition. In Proceedings of the IEEE International Conference Computer Vision, Nice, France, 13–16 October 2003; Volume 1, pp. 1023–1029. [CrossRef]
15. Koenig, S.; Simmons, R.G. Unsupervised Learning of Probabilistic Models for Robot Navigation. In Proceedings of the IEEE International Conference on Robotics and Automation, Minneapolis, MN, USA, 22–28 April 1996; Volume 3, pp. 2301–2308. [CrossRef]
16. Tsukada, M.; Utsumi, Y.; Madokoro, H.; Sato, K. Unsupervised Feature Selection and Category Classification for a Vision-Based Mobile Robot. *IEICE Trans. Inf. Syst.* **2011**, *E94-D-1*, 127–136. [CrossRef]
17. Hinton, G.E.; Salakhutdinov, R.R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504–507. [CrossRef] [PubMed]
18. Le, Q.V. Building high-level features using large scale unsupervised learning. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 8595–8598. [CrossRef]
19. Oliva, A.; Torralba, A. Building the Gist of a Scene: The Role of Global Image Features in Recognition. *Vis. Percept. Prog. Brain Res.* **2006**, *155*, 23–36. [CrossRef]
20. Lowe, D.G. Object Recognition from Local Scale-Invariant Features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; Volume 2, pp. 1150–1157. [CrossRef]

21. Madokoro, H.; Utsumi, Y.; Sato, K. Unsupervised Scene Classification Based on Context of Features for a Mobile Robot. In Proceedings of the 15th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, Kaiserslautern, Germany, 12–14 September 2011; Volume 1, pp. 446–455. [CrossRef]
22. Li, F.-F.; Perona, P. A Bayesian Hierarchical Model for Learning Natural Scene Categories. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; Volume 2, pp. 524–531. [CrossRef]
23. Luo, J.; Pronobis, A.; Caputo, B.; Jensfelt, P. Incremental learning for place recognition in dynamic environments. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, San Diego, CA, USA, 29 October–2 November 2007; pp. 721–728. [CrossRef]
24. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893. [CrossRef]
25. Ishikoori, Y.; Madokoro, H.; Sato, K. Semantic Position Recognition and Visual Landmark Detection with Invariant for Human Effect. In Proceedings of the IEEE/SICE International Symposium on System Integration (SII), Taipei, Taiwan, 11–14 December 2017; pp. 657–662. [CrossRef]
26. Kainuma, A.; Madokoro, H.; Sato, K.; Shimoi, N. Occlusion-Robust Segmentation for Multiple Objects using a Micro Air Vehicle. In Proceedings of the 16th International Conference on Control, Automation and Systems, Gyeongju, Korea, 16–19 October 2016; pp. 111–116. [CrossRef]
27. Anbarasu, B.; Anitha, G. Indoor Scene Recognition for Micro Aerial Vehicles Navigation using Enhanced-GIST Descriptors. *Def. Sci. J.* **2018**, *68*, 129–137. [CrossRef]
28. Vapnik, V.; Lerner, A. Pattern Recognition Using Generalized Portrait Method. *Autom. Remote Control* **1963**, *24*, 774–780.
29. Alcantarilla, P.F.; Nuevo, J.; Batoli, A. Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces. In Proceedings of the British Machine Vision Conference, Bristol, UK, 9–13 September 2013.
30. Csurka, G.; Dance, C.R.; Fan, L.; Willamowski, J.; Bray, C. Visual Categorization with Bags of Keypoints. In Proceedings of the International Workshop on Statistical Learning in Computer Vision, Prague, Czech Republic, 11–14 May 2004; pp. 1–22. [CrossRef]
31. Kohonen, T. *Self-Organizing Maps*; Springer Series in Information Sciences; Springer: Berlin/Heidelberg, Germany, 1995.
32. Hetch-Nielsen, R. Counterpropagation networks. *Appl. Opt.* **1987**, *26*, 4979–4983. [CrossRef] [PubMed]
33. Madokoro, H.; Shimoi, N.; Sato, K. Adaptive Category Mapping Networks for All-Mode Topological Feature Learning Used for Mobile Robot Vision. In Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication, Edinburgh, UK, 25–29 August 2014; pp. 678–683. [CrossRef]
34. Ultsch, A. Clustering with SOM: U*C. In Proceedings of the Workshop on Self-Organizing Maps, Paris, France, 5–8 September 2005; pp. 75–82. [CrossRef]
35. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In Proceedings of International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 20–25 August 1995; Volume 2, pp. 1137–1143. [CrossRef]
36. Arlot, S.; Celisse, A. A Survey of Cross-Validation Procedures for Model Selection. *Stat. Surv.* **2010**, *4*, 40–79. [CrossRef]
37. Verhoeven, G. Taking Computer Vision Aloft– Archaeological Three-dimensional Reconstructions from Aerial Photographs with PhotoScan. *Archaeol. Prospect.* **2011**, *18*, 67–73. [CrossRef]
38. Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans. Syst. Man Cybern.* **1979**, *9*, 62–66. [CrossRef]