Digital Life

Journal of Digital Life 2025,5,S3

Special Issue: Measurement, Control, and Analysis of Motion Using ICT and AI

Article

Research on Indoor Self-Location Estimation Technique Using Similar Image Retrieval Considering Environmental Changes

Masaya Nakahara [1], Yoshinori Tsukada [2], Yoshimasa Umehara [3] and Shota Yamashita [4]

[1] Faculty of Information Science and Arts, Osaka Electro-Communication University, 1130-70 Kiyotaki, Shijonawate-shi, Osaka 575-0063, Japan
[2] Faculty of Engineering, Reitaku University, 2-2-1 Hikarigaoka, Kashiwa-shi, Chiba 277-8686, Japan
[3] Faculty of Business Administration, Setsunan University, 17-8 Ikedanakamachi, Neyagawa-shi, Osaka 572-8508, Japan
[4] Graduate School of Information Science and Arts, Osaka Electro-Communication University, 1130-70 Kiyotaki, Shijonawate-shi, Osaka 575-0063, Japan

Abstract
In Japan, the shortage of human resources due to the declining birthrate and aging population is becoming a social problem. Particularly in the security industry, the irregular working hours and associated risks are making it increasingly challenging to secure workers. This has led to a rise in use of security systems that utilize security cameras and drones. However, in factories and other buildings with a lot of equipment and intricate structures, there is the problem of blind spots caused by occlusion. This situation necessitates the use of automated drone patrols, and a problem arises when self-position estimation fails in areas where acquiring feature points is difficult, such as corridors. To solve these problems, in a previous study, we devised a technique for position estimation using a method that can calculate similarity based on changes in the distribution of color information across the entire image. In this study, we propose a method that can cope with environmental changes caused by object movement while combining feature point-based methods.

Keywords: *automated patrol, drone, position estimation, image search*

1. Introduction

In Japan, the shortage of human resources has become a social problem due to the declining birthrate and aging population. This has had a serious effect on the security industry because of the irregular hours of work, the danger involved in responding to suspicious persons, and the large number of personnel required to patrol large facilities (Ministry of Health, Labour and Welfare, 2024). These factors necessitate the development of security systems that utilize security cameras and drones as a solution to the shortage of human resources. However, security systems that utilize security cameras face several challenges. For example, when monitoring areas with many pieces of equipment and intricate locations, such as factories, there are concerns that the number of cameras installed will increase and blind spots will occur due to the effects of occlusion. Therefore, it is expected that drones and robots that can move autonomously can mount and move cameras to patrol and monitor these areas, thereby reducing the number of personnel required for security.

For example, Skydio 2+ is a drone that can fly autonomously using camera images. It uses Visual SLAM to estimate its own position with high accuracy even in non-GNSS space, based on the images from multiple cameras installed on the drone. This system enables safe navigation in narrow, complex structures with steel or concrete frames under bridges and in wide-area shooting. As methods for estimating self-position using camera images in fields other than drones, “A method for estimating self-position by feature point matching” (Okamoto et al., 2012; Yamazaki et al.,

Received: 21 November 2024, Revised: 20 January 2025, Accepted: 15 February 2025, Published: 26 March 2025
* Correspondence: nakahara@oecu.jp
Publisher’s Note: JOURNAL OF DIGITAL LIFE. stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

2019) and “A method for feature point matching on a 3D map generated from RGBD camera shots” (Matsumoto et al., 2024; Yang et al., 2020) have been proposed. The former calculates features from a set of previously captured images and compares them with features obtained from the captured images to estimate the location of the captured images. In “A method for feature point matching on a 3D map generated from RGBD camera shots,” a 3D map is created using an RGBD camera to obtain the positional relationships of characteristic structures and objects in a building. Then, based on the created 3D map, the positions where the features match the input image are searched for using deep learning and other methods.

All existing methods estimate self-location under the assumption of many common features in the video, regardless of the time of year. Therefore, when targeting narrow indoor areas with many plain walls, such as those patrolled by security drones, there is “the problem of failing to estimate self-position due to the small number of features” and “the problem of failing to estimate self-position when the feature object itself is moving” in places, such as factories. For example, in the case of Skydio 2+, when flying over a complex structure or a wide area, one of the cameras will always reflect a feature object or landmark location, enabling highly accurate self-position estimation. However, in the case of an indoor hallway between plain walls, the distance from the camera to the plain walls on both sides of the subject is short, and it may be difficult for each camera to always have sufficient features for self-position estimation during flight. Conversely, in “A method for estimating self-position by feature point matching” and “A method for feature point matching on a 3D map generated from RGBD camera shots,” position estimation is based on previously obtained features. Therefore, when applied to locations such as factories, where objects such as tools and instruments are easily moved, the number of commonly obtained features decreases, and a completely different position may be misestimated as the self-position.

In our previous study (Yamashita et al., 2024), we proposed a method using not only a feature point-based approach but also dHash (Figure 1), an algorithm that searches for overlapping images based on changes in the distribution of color information across the entire image. The dHash method calculates a hash value based on the distribution of color changes in the entire image using the luminance gradient of each segmented image area in relation to adjacent areas. Using this algorithm, hash values similar to those in daylight can be calculated based on small differences in luminance even in environments with insufficient light. Therefore, it is highly possible to generate hash values that approximate the nearest pre-captured image even at nighttime, if the features within the shooting range are visible. This method can be used to capture color changes common to images with location information that have been previously captured and images used for self-location estimation, even with few obtained features. In addition, because dHash utilizes information from the entire image, it effectively suppresses the effects of changes in local features caused by object movement better than feature point-based methods. For example, in the case of a hallway, opening and closing doors may cause environmental changes. Therefore, the method can address existing methods problems, such as those of “failing to estimate self-position due to the small number of features” and of “failing to estimate self-position when the feature object itself is moving.” However, demonstration experiments showed that the estimation results are prone to errors on straight sections. However, while the drone used for automatic patrol does not need to change the direction of travel in the straight sections, it needs to change the direction of travel significantly in the straight sections near the curve points. Therefore, the number of images taken in advance must be denser when the drone is close to a curve, and more accurate position estimation is required than in existing studies. In the existing method, no feature change occurs between the images taken before and after the straight section except for the distance from the wall at the end of the curve point, and it is said that there is little difference in the similarity in the straight section near the curve point (Figure 2).

In this study, we propose a method that selects multiple candidate images similar to the input image using scale-invariant feature transform (SIFT) features and then estimates similar images using dHash among them. This method is expected to improve the estimation results for straight sections by considering the distribution of local features influenced by columns, windows, and other factors.

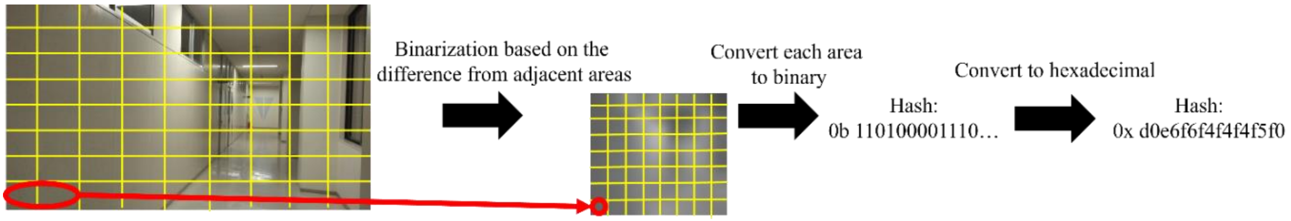

Binarization based on the difference from adjacent areas
Convert each area to binary
Hash:
0b 110100001110…
Convert to hexadecimal
Hash:
0x d0e6f6f4f4f4f5f0

Figure 1. Processing steps of dHash

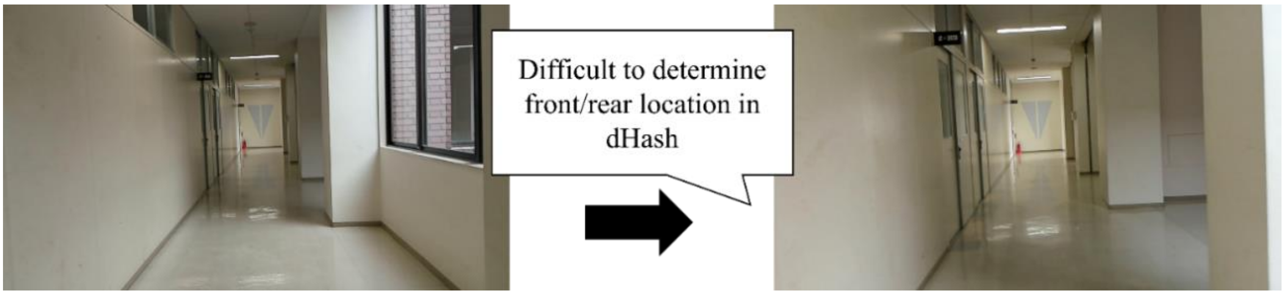


Figure 2. Examples of straight sections that are prone to estimation failures

2. Methods

2.1 Overview of Methodology

Based on the issues discussed in Section 1, we propose a self-positioning estimation technique that considers the similarity of images taken in straight sections, such as corridors, indoors, and in factories, where security drones target many plain wall surfaces. Figure 3 shows the process flow of the proposed method, which consists of the “Candidate Image Selection Function,” “Similar Image Retrieval Function,” and “Location Estimation Function.” The input data of the proposed method consists of “camera images for location estimation,” “images and location information on the patrol route taken in advance,” and “a map of the patrol route composed of point cloud data.” The output data is the “coordinates of the estimated location,” which can be displayed on a map. In this case, the images on the travel route in the input data are stored with the coordinate values on the map of the travel route corresponding to the shooting position of each image in advance (Figure 4). In this method, images on the patrol route are collected at regular intervals in straight sections, while images are collected at denser intervals in curve sections. This is because the estimated position acquired by this method is used to provide movement instructions to the drone, so it is necessary to collect images at high density in the curve section.

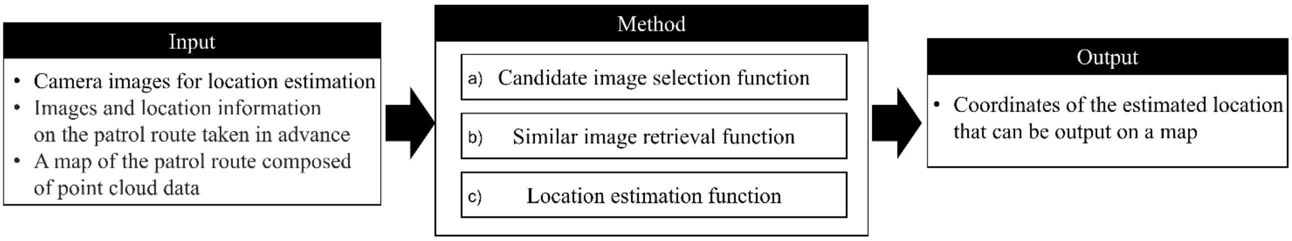

Input
• Camera images for location estimation
• Images and location information on the patrol route taken in advance
• A map of the patrol route composed of point cloud data
Method
a) Candidate image selection function
b) Similar image retrieval function
c) Location estimation function
Output
• Coordinates of the estimated location that can be output on a map

Figure 3. Flow of the proposed method

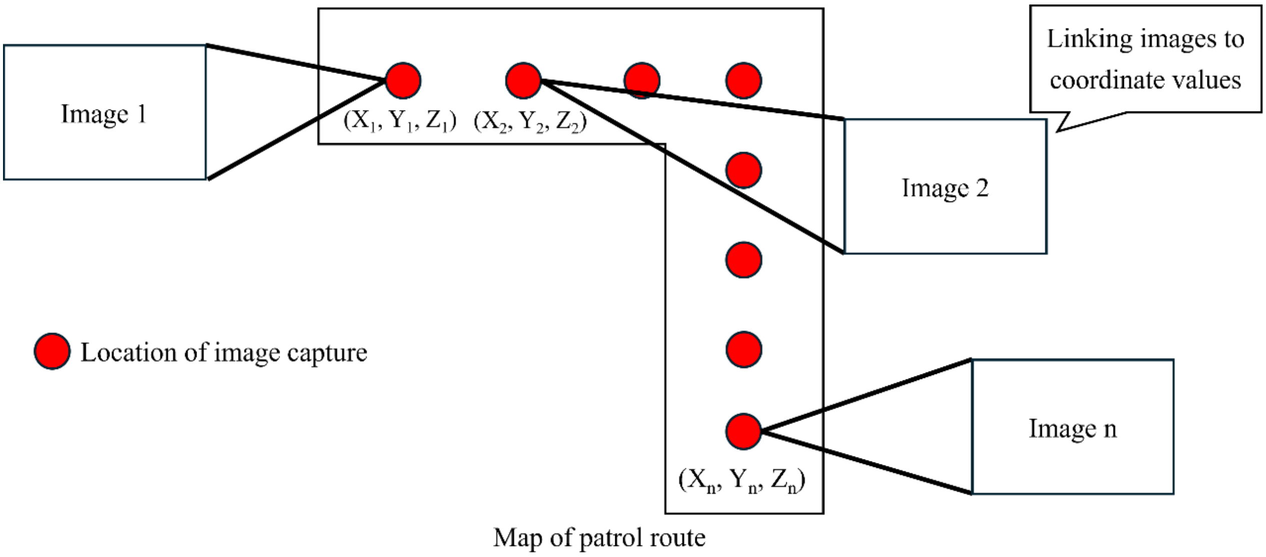


Figure 4. Image diagram of input data

2.2 Candidate Image Selection Function

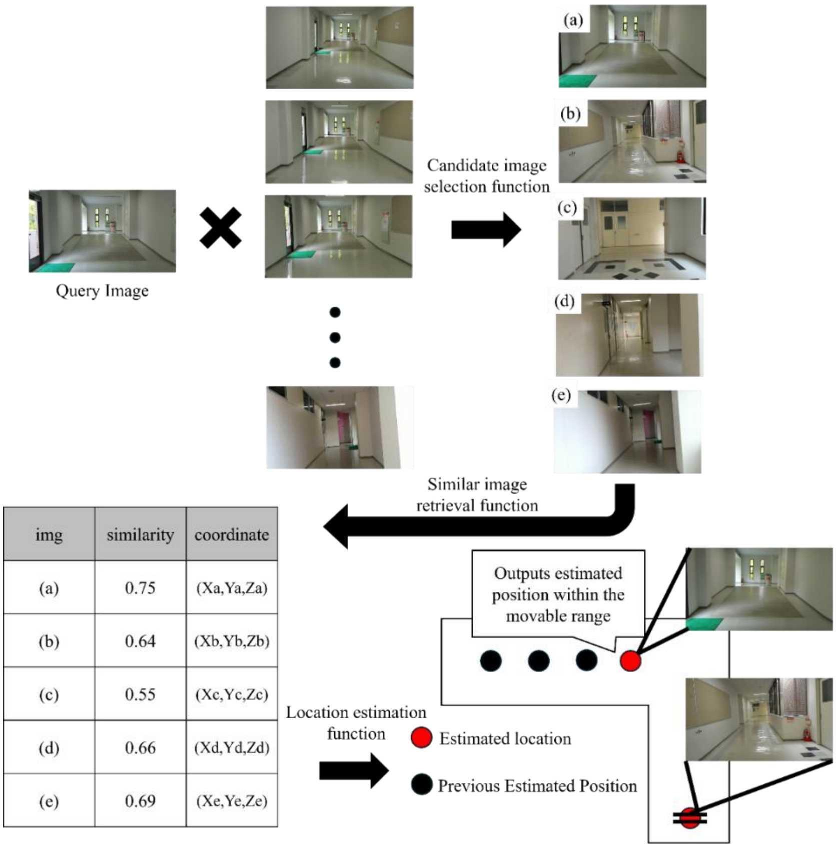
In the “Candidate Image Selection Function,” the distance to the feature point obtained by SIFT is used as the basis for calculating the similarity between each image and the target images to be searched for. First, local features are calculated using SIFT for images on the traversing path and images used for self-position estimation. However, if local features of the images on the traversing path have already been calculated, they are calculated only for the image used for self-position estimation. The Hamming distance to the corresponding feature is then calculated using Brute-Force Matcher, and the average of all Hamming distances calculated for each image is obtained. This selects a group of images with a certain number of matching features.

2.3 Similar Image Retrieval Function

The “Similar Image Retrieval Function” uses dHash to estimate and output images with a threshold level of similarity or higher with respect to the images selected in the “Candidate Image Selection Function.” First, a hash value is obtained from each image using dHash. Specifically, the grayscale image is divided into regions of a certain size, and the difference in luminance between adjacent regions is calculated. Then, based on the calculated results, the lightness and darkness of the left and right areas are expressed as a string of 01 and output as a hash value. Next, the hash values of the images on the traversing path are compared with the hash values from the images used for self-position estimation, and the Hamming distance is calculated from the XOR operation results. The calculated Hamming distance is normalized in the range of 0–1, and the value is used as the similarity. Furthermore, images on the traversing path with similarity above a threshold value are output as similar images.

2.4 Location Estimation Function

The “Location Estimation Function” estimates the most appropriate location on the map of the traversing route from the images with high similarity estimated by the “Similar Image Retrieval Function.” First, the system obtains the coordinates associated with the images on the traversing path that have a similarity greater than a threshold. Then, using the acquired coordinate values and the estimation result of the previous shooting position, the system outputs the coordinate values associated with the image with the highest similarity within the range where the drone can move from the previous position to the current shooting position(Figure 5). However, in the absence of a previous estimated position, the system outputs the position of the image with the highest similarity as the current shooting position.


Query Image
Candidate image selection function
(a)
(b)
(c)
(d)
(e)
Similar image retrieval function
Outputs estimated position within the movable range
Location estimation function
● Estimated location
● Previous Estimated Position

img	similarity	coordinate
(a)	0.75	(Xa,Ya,Za)
(b)	0.64	(Xb,Yb,Zb)
(c)	0.55	(Xc,Yc,Zc)
(d)	0.66	(Xd,Yd,Zd)
(e)	0.69	(Xe,Ye,Ze)

Figure 5. Diagram of each function

3. Results

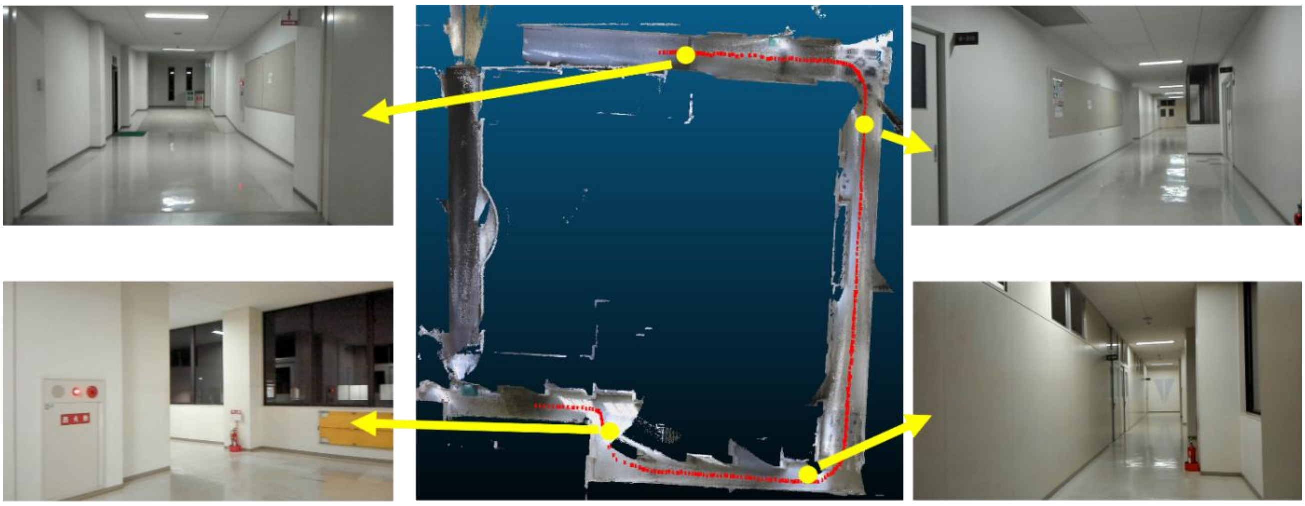
3.1 Verification Experiments

We applied the proposed method to an indoor corridor at night, assuming two types of situations: one in which the situation is the same as when the image was taken beforehand, and another where the situation has changed. Then, we verified the applicability of the proposed method to self-position estimation for automatic patrols by security drones. The experiment location was an L-shaped corridor on a university campus, which has many plain walls and straight sections that are difficult to estimate using the proposed method and previous research methods. For the input data, a 3D map consisting of point cloud data was constructed using a unit that can measure point cloud data by SLAM, as used in previous work (Kajitani et al., 2024). The measurement unit recorded images of the patrol route in advance and linked the coordinate values to the map (Figure 6).

In the present experiment, to compare with the previous study and to verify the effects of changes in local features, we also verified the case in which the objects that are features, such as door opening and closing and installation locations, which are likely to be features in self-position estimation, are varied in each case of the previous study (Yamashita et al., 2024) and the proposed method. In addition, each input image must match the shooting conditions of a small drone that can fly indoors. To simulate flight, we raised a hand-held web camera capable of capturing RGB images to the same height as the small drone's flight altitude. During the evaluation, we compared the self-positions estimated by each method with the actual shooting positions. We then compared the percentage of correctly estimated positions to the shooting positions at all locations, thereby confirming the usefulness of the proposed method. Furthermore, we verified the applicability of the proposed method from the viewpoint of applying it to automatic patrols by security drones.

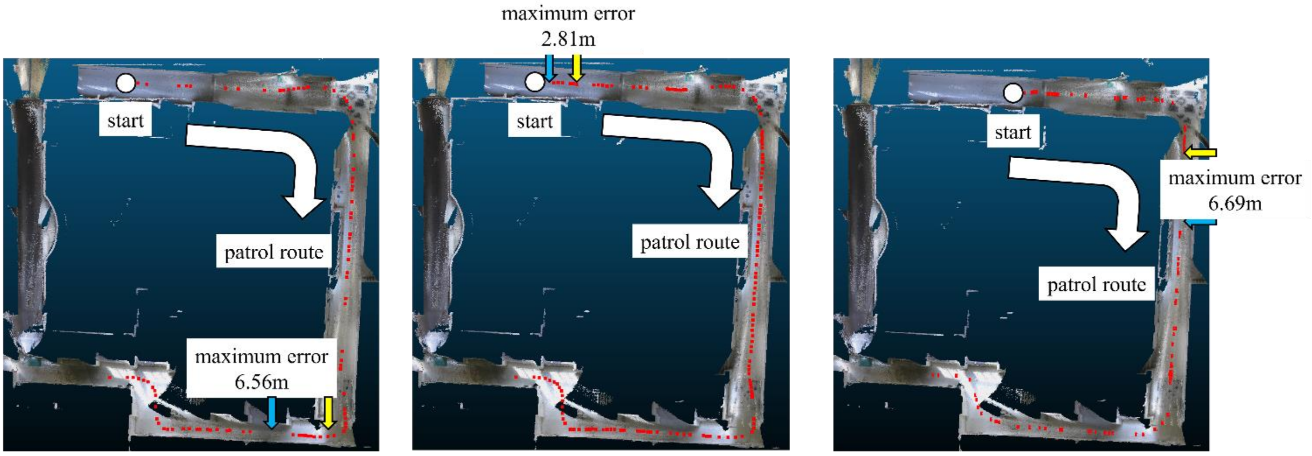
3.2 Experiments Results

Figure 7(a) shows the visualization results of position estimation using only dHash without changing local features; Figure 7(b) shows the visualization results of position estimation using the proposed method; and Figure 7(c) shows the visualization results of position estimation using the proposed method with changing local features. Table 1 shows the percentage of correct responses, the average position error, and the maximum error for similar images in each result. Notably, the location estimation process, which solely relied on dHash and required input images with local feature changes, experienced significant failure. Consequently, we didn't verify the visualization results or compute the error amount.



Red Point： Location of images on the tied patrol route

Figure 6. Map visualization of patrol routes



(a) dHash・No change in local features
(previous study)

(b) dHash+SIFT ・No change in local features
(proposed method)

(c) dHash+SIFT ・changes in local features
(proposed method)

Red Point : Estimated shooting position
Blue Arrow : Actual shooting position Yellow Arrow : Wrongly estimated location

Figure 7. Location Estimation Visualization Results

Table 1. Percentage of correct answers in each result

Local feature	Method	Number of input images	Number of correct images	Correct responses	Average position error	Maximum error
No change	dHash	216	153	70.8%	2.54m	6.56m
	dHash+SIFT (Our Method)	216	209	96.3%	2.14m	2.81m
Change	dHash	129	36	27.9%	error	error
	dHash+SIFT (Our Method)	129	106	82.8%	3.37m	6.69m

4. Discussion

As shown by visualization results of position estimation in Figure 7(b) and (c), the position of the drone on its patrol path is generally estimated from the webcam image, even when SIFT is combined with SIFT. The comparison of the percentage of correct responses in Table 1 confirms that the combination of SIFT produces more accurate location estimation results in both cases, with and without changes in local features. This suggests that the method improves two issues in the existing approach. Comparing the respective results with no change in local features, the maximum values of the percentage of correct answers and position errors confirm the improvement in the accuracy of position estimation. Specifically, the calculation of more than 90% of the correct answers validates the feasibility of location estimation across all sections. Therefore, we confirm that combining a method such as dHash, which generously captures overall features, and SIFT, which is a feature point-based method, is useful even with few features, such as in a corridor or in a straight section, which has been an issue in previous studies.

The combined dHash and SIFT method was generally successful in estimating the input images when local features were changing. However, when the results were compared with and without local features, the percentage of correct responses decreased and the location error worsened. In fact, when we checked the location where the maximum error occurred in Figure 7(c), we found several locations where local features changed (Figure 8). Thus, it is unlikely that feature point-based methods alone can further improve estimation accuracy in locations with many changes in local features. It is expected to achieve higher accuracy by comparing the results of similarity calculations between the feature point-based method and dHash and using the more reliable estimation result as a reference for location estimation. However, this experiment assumed that the actual patrolling guards might be in a dark place with no lighting, despite taking the input images with the lighting on. Therefore, we need to enhance the method's sensitivity to light intensity changes using gamma correction or deep learning to produce images that mimic a quasi-bright environment.

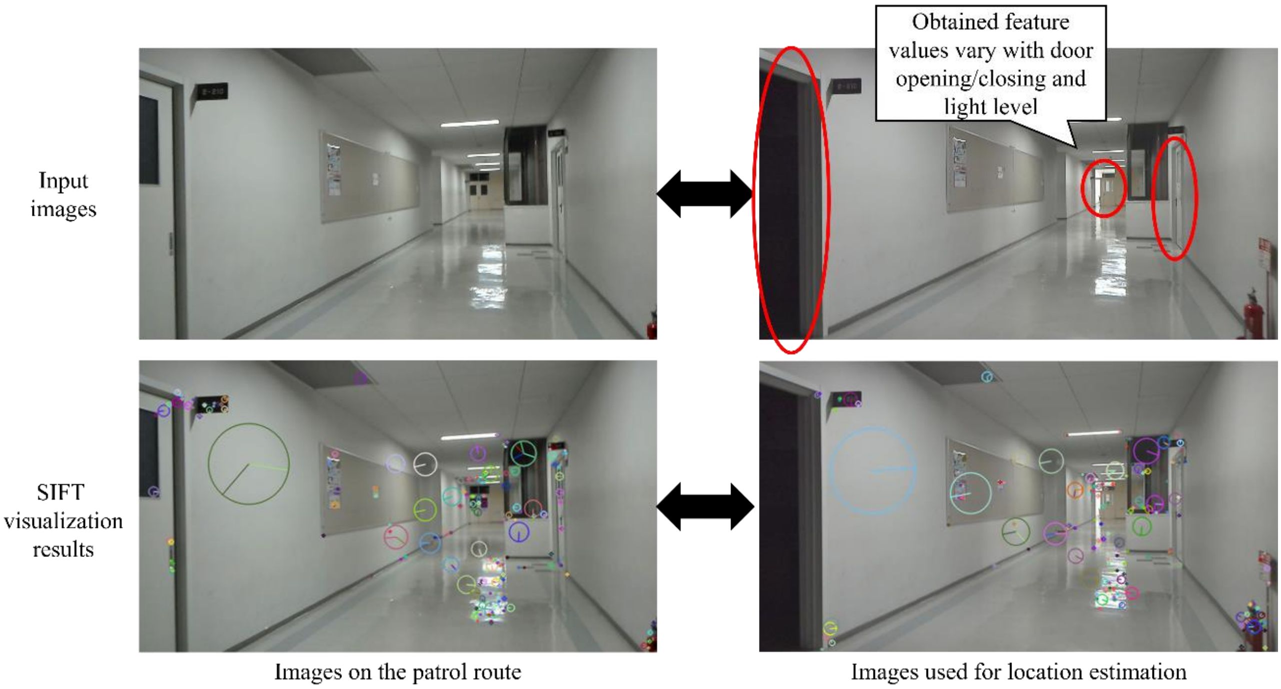

Obtained feature values vary with door opening/closing and light level
Input images
SIFT visualization results
Images on the patrol route
Images used for location estimation

Figure 8. Locations of maximum error in Figure 7(c) results

5. Conclusions

In this paper, we propose a method for estimating self-location by combining SIFT and dHash to search similar images for indoor areas with few features. Empirical experiments confirm that the overall accuracy can be improved, even for straight sections that previous studies found difficult to estimate. In the future, we aim to make this method robust to changes in light intensity and develop a self-position estimation technique that can be applied even at night.

Author Contributions

Conceptualization, M.N. and S.Y.; methodology, M.N., Y.T., Y.U. and S.Y.; software, S.Y.; validation, S.Y.; formal analysis, S.Y.; investigation, S.Y.; resources, M.N. and S.Y; data curation, S.Y.; writing—original draft preparation, S.Y.; writing—review and editing, M.N., Y.T. and Y.U; visualization, S.Y.; supervision, M.N., Y.T. and Y.U.; project administration, M.N.

Funding

This research received no external funding.

Informed Consent Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References:

Kajitani, M., Nakahara, M., Tsukada, Y., Umehara, Y., Nishida, Y., Shimizu, N. & Tanaka, S. (2024). Research on correction of point cloud data by handheld laser scanner with SLAM. *Proceedings of the 86th National Convention of IPSJ*, *86*(4), 805-806.

Matsumoto, Y., Nakano, G. & Ogura, K. (2024). Indoor visual localization using point and line correspondences in dense colored point cloud. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 3604-3613.

Ministry of Health, Labour and Welfare. Employment referrals for general workers. (2024). https://www.mhlw.go.jp/toukei/list/114-1.html

Okamoto, K., Kazama, H. & Kawamoto, K. (2012). A fuzzy codebook based image search method for visual localization in libraries. *28th Fuzzy System Symposium*, 28, 444-447.

Skydio. Skydio2+. (2024). https://www.skydio.com/skydio-2-plus-enterprise

Yamashita, S., Nakahara, M., Tsukada, Y. & Umehara, Y. (2024). Research on estimation technique of self-location indoors using similar image retrieval technique. *Proceedings of the symposium on civil engineering informatics*, 49, 205-208.

Yamazaki, K., Shishido, H., Kitahara, I. & Kameda, Y. (2019). Evaluation for harmonic location estimation system of image retrieval and SLAM. *International Workshop on Advanced Imaging Technologies 2020*

Yang, Y., Toth, C. & Brzezinska, D. (2020). A 3D map aided deep learning based indoor localization system for smart devices, *The International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences*., XLIII-B4-2020, 391–397.