Navigation requires self-localization, which GPS typically performs outdoors. However, GPS frequently introduces significant errors in environments where radio reflection occurs, such as in urban areas, which makes precise self-localization occasionally impossible. Meanwhile, a human can use street view images to understand their surroundings and their current location. To implement this, we must solve image matching between the current scene and the images in a street view database. However, because of the wide variations in the field angle, time, and season between images, standard pattern matching by feature is difficult. DeepMatching can precisely match images with various lighting and field angles. Nevertheless, DeepMatching tends to misinterpret street images because it may find unnecessary feature points in the road and sky. This study proposes several methods to reduce misjudgment: (1) gaining image similarity with features such as buildings by excluding the road and sky, and (2) splitting the panoramic image into four directions and matching in each. This study shows the results of each method and summarizes their performance using various images and resolutions.

Keywords: autonomous robot, view-based navigation, self-localization, image matching, human cognition

1. Introduction

1.1. Background

Service robots that carry loads and provide guidance have recently been the subject of extensive research. Navigation is important in these tasks to reach the destination automatically.

Self-localization is one of the most important factors in navigation. GPS is frequently used to identify the locations during outdoor navigation. However, satellite visibility and radio wave conditions, such as indoors and in tunnels, can both affect GPS and result in errors in tens of meters. Particularly in environments where radio waves are easily reflected by objects, such as buildings or trees in urban areas, GPS is prone to significant errors. This can cause problems when a less-than-ten meter accuracy is required, such as when locating a building or entrance.

However, regarding location identification by humans, if they have the street view image in their memory, they can identify their current location by comparing the surroundings with the street view image in their memory.

In this study, we focus on the cognitive function in humans. When we use a robot to implement this location identification method, we first allow the robot to memorize the street view image of the desired destination in its memory. Next, the robot is allowed to conduct self-localization by matching the images obtained from the camera while running with the memorized images. Then, the robot is allowed to decide the direction of movement based on the matching result. The robot can avoid spending time in advance setting up the scenery of the route to the destination using a road scenery image database, such as Google’s Street View (Fig. 1). In this study, the term ‘Street View’ refers to the image database associated with the location on a map, which includes the Google Street View service.

However, performing image matching between the image of the current surrounding and the street view image using general image features is difficult.

It is caused by (1) the change in viewpoint, (2) the differences in the angle of view and image types (panoramic or 360° image), and (3) differences in shooting time and season.
However, DeepMatching can handle the displacement of objects between images [1]. For this reason, we believe that it can handle the change in viewpoint between camera-generated images and street view images. Additionally, even when images have different view angles and hitting lights, DeepMatching can still match them with high accuracy.

However, when used to match sidewalk images with street view images, it often results in incorrect location determination because of problems such as finding feature points on part of sky and road that have no correspondence with their locations. Additionally, when misjudgment occurs, the coordinates of matching points are sometimes far apart from one another. These will often happen when two images taken at different locations are matched.

1.2. Purpose of this Paper

This study aims to reduce misjudgments associated with the use of DeepMatching. There are several causes of misjudgment, but we believe that one of the main causes is the discovery of feature points in the roads and sky areas. Additionally, the similarity of matching distant points is another cause of misjudgment.

Considering these two reasons, we propose two main methods (concepts).

The first method (concept) is to remove part of the road and sky from a street view image and a scenery image. This reduces unnecessary feature point findings in the roads and sky areas.

By removing these rapidly changing parts from the comparison, the DeepMatching method will concentrate on features specific to the location, such as buildings and trees, which will result in a stable location matching.

There are three specific methods for removing parts of the road and sky: 1. filling part of the road and sky, 2. cutting off everything else except the road and sky, and 3. removing what is left of the road and sky.

The second method (concept) involves dividing panoramic images into four parts according to the directions, and summarizes the similarities of each part.

We believe this will restrict the search area and reduce the number of distant matching points.

Additionally, because a road image significantly differs from the directions, it will reduce mismatch in similar characteristics at different directions.

In this study, we verify the validity of the proposed method by comparing the similarity between the processed panoramic street view image and peripheral images. Additionally, we investigate appropriate parameters through experiments using various formats and resolutions, such as 360° and panoramic images as the images for comparison with the street view images.

2. Related Work

The topic of human navigation is discussed in “Differences in spatial knowledge acquired from maps and navigation” [2]. In their study, they distinguished between several aspects of what people learn from maps and navigation. Our study closely resembles the process of learning from navigation described in the paper because we use actual landscape images as reference images. Our study aims to develop view-based navigation that is similar to how humans can estimate their current location based on memorized images and information obtained from their vision.

One of the examples of view-based navigation is “View-based navigation using an omniview sequence in a corridor environment” [3]. In their study, omniview sequence was proposed as a visual model of the path. The images of the omniview sequence are memorized by running an indoor route in advance. Then, self-localization and controlling direction are conducted by appropriately using global and local template matching.

“Outdoor autonomous navigation using SURF features” [4] is one example of research on view-based localization using SURF features. In their study, view-based localization is performed by image matching using SURF between the memorized images and the camera-acquired images. Each memorized image is used to conduct for the first time in localization. Subsequently, a search is conducted from the next three memorized images.

There is also an example of drone navigation based on the brain mechanism of a rat [5]. In their paper, scan-line intensity (SI)-grid-based motion statistics (GMS) algorithm is proposed. They used SI to quickly match, and they used GMS to check the matching points for accuracy. The algorithm outputs the latitude and longitude coordinates based on the template.

There have been some previous studies, such as “View-based localization in outdoor environments based on support vector learning” [6], that address the issue of taking seasonal and weather changes into account during vision-based navigation. In their study, machine-learning based object recognition is used for self-localization to deal with the change in season or weather conditions. By identifying objects rather than the entire scenery, it can exclude easily changeable features such as clouds, leaves, and roads.

Additionally, a study titled “SIFT, SURF and seasons: Long-term outdoor localization using local features” [7], is relevant to our research because it takes a seasonal change into account. It uses various local features, such as SIFT, SURF, U-SURF, and SURF-128. However, in the study, the authors concluded that localization from panoramic images with seasonal changes is challenging.

A study related to ours titled “View-sequence based indoor/outdoor navigation robust to illumination changes” [8] also considered illumination change. The study considered illumination changes in terms of the image-based navigation, and proposes the use of the accumulated block matching (ABM) method, which is robust to illumination change. Additionally, global and local image features were used correctly for robust localization in illumination changes in the study titled “A hybrid approach for vision-based outdoor robot localization using...”
global and local image features” [9]. They pointed out that the local image features should be used when the illumination changes significantly.

However, in these studies, a robot was required to memorize images with its camera by running a route to the destination beforehand. This is costly and ignores viewpoint changes, which are required for using external databases, such as street view. However, our method identifies sky and road areas in the external reference database and input images and excludes them so that only stable features will be used.

One of the examples that considered viewpoint changes is “SeqSLAM++: View-based robot localization and navigation” [10]. In their study, view-based navigation was conducted by comparing several reference images with input images.

However, for the reference images to consider viewpoint change, the robot must run in a zigzag pattern. Additionally, the study excludes viewpoint changes such as the change between sidewalks and roadways. In our method, we use DeepMatching method to incorporate viewpoint changes without preparing many samples.

3. Methods

In this study, we use DeepMatching as an image matching technique.

DeepMatching uses SIFT feature values. The similarity is calculated for each patch, which can move around. It searches within a certain range and moves to the location with the maximum similarity. Therefore, it can deal with the movement of objects between images. This led us to consider that DeepMatching can deal with a viewpoint change between the street view image and sidewalk image at the same location. For example, the position of the same object would differ between images if it appeared in both the street view and sidewalk images because there is a change in viewpoint. DeepMatching can handle viewpoint changes and find corresponding points for the matching objects whose locations vary between images, as in this example.

Our goal is to implement functions such as self-localization and movement direction determination using the evaluation function of DeepMatching for image similarity. Particularly, a robot-taken camera image and candidate street view images, both of which have been preprocessed using our proposed method, will be compared. Then, the location of the robot will be determined based on which street view image has the highest similarity value.

However, as a preliminary investigation, this study describes the results of the matching performance verification using the proposed method as follows:

1. Removal of sky and road areas.
2. Division of panoramic image into 1/6 in the horizontal direction, and assembling them according to the left, right, back, and forward directions.

The purpose of the first method (concept) is to reduce the number of feature points in areas such as roads and skies. There are three specific methods related to this: filling the road and sky areas with white, excluding the road and sky areas, and removing the remaining road and sky areas after excluding them. These will be discussed in further detail as follows. The first involves filling the road and sky areas with white. The roads and sky areas are manually filled with white after being manually identified by humans. Fig. 2 shows the concept of this method.

However, it was difficult to perfectly fill the whole road and sky areas in white. Therefore, some areas have been left unchanged. Owing to the difficulty of such manual filling, we developed a way to consider the method for automatically removing the road and sky areas. The method removes these two areas according to the predetermined region. Then, these cut-off areas are connected with each other. Fig. 3 shows the image that we used as a reference to decide the range.

We selected this image because the building is evenly sized on both sides when it is photographed. Therefore, we believe that the features of the building can be retained even if the cutting off was applied to other images. Fig. 4 shows the concept of this method. However, this method was insufficient to cut off all roads and sky areas; therefore, we developed a program that automatically removes roads and sky areas after cutting off. By drawing a yellowish rectangle, the area surrounding the road and sky is removed. This is because we wanted to reduce feature point findings to the road and sky areas by using an artificial color that would not exist in many of the outdoor scenery. Fig. 5 shows a reference image for drawing the rectangle. Fig. 5 shows the concept of this method.

The purpose of the second method (concept) is to nar-
row the search area and reduce the number of matching points that are located far away from each other. This concept can be implemented using a method that divides a panoramic image. This method divides the image into six equal parts and then groups them into four directions: left, right, forward, and behind. Fig. 6 shows a concept of this method.

Additionally, Fig. 7 illustrates the relationship between the methods used to prevent misjudgment. We will investigate appropriate values for the following parameters of DeepMatching: “downscale” that describes the degree of shrinking the input image, and “ngh_rad” that specifies the search range.

4. Experiments

4.1. About the Correct/wrong Judgment of the Matching

We describe the criteria for determining whether two images were taken at the same or different locations. DeepMatching can obtain the similarity of each matching point. In this study, the sum of similarities was used as the score for image matching. The correct or incorrect conclusion was made based on the score of image matching, which is the sum of the similarities. If the score of two images taken at the same location is higher than the score of two images at different locations, this is considered the correct judgment.

4.2. Proposed Method

We propose the following four methods:

1. Fill the road and sky areas with white.
2. Cut off all other areas except the road and sky areas.
3. Remove the remaining road and sky areas after cutting-off.
4. Divide the panoramic image into the direction of movement, left, right, and behind, and match between the images in each direction.

4.3. Experiments 1 (Sky and Road Area Removal)

First, Figs. 8 and 9 provide illustrations of matching using a general image format (not a panoramic or 360°).
The top two images in Fig. 8 are taken at almost the same places and viewing direction. The left image was taken by the iPhone 7 plus camera at a sidewalk, and the right image was screen-captured from a Google’s Street View image. Although these two images differ from their view angles and the formats, it is observed that the buildings in the back, center, and left are the same. These are easily identified by their distinctive textures.

The outcomes of DeepMatching between the above two images are displayed in the bottom two images. A group of points in the same color indicate the approximate corresponding points.

The top two images in Fig. 9 were taken at different locations. The left image in Fig. 8, was taken by the camera, and the right image was obtained from the street view. Furthermore, the bottom two images show the DeepMatching results.

The score representing the degree of similarity is shown in Fig. 8 as 893.07, and the number of matching points is 245. Fig. 9 shows a score of 648.72 and 196 matching points. In these examples, the same location has a higher similarity score than a different location. Thus, the DeepMatching’s location judgment is expected to be correct for images that have explicit features, such as buildings. As shown in Fig. 8, most of the matching points are on buildings.

Next, a comparison of panoramic images without the proposed method is shown in Fig. 10 (at the same location and view direction) and Fig. 11 (different locations).

The left image was taken using a panoramic camera (Omicam), and the right image is from a street view image capture (as a panoramic picture). Note that the camera has artificially filled the lower part of the images with black areas and lines to indicate its specifications. Additionally, the panoramic pictures for the Google’s Street View were created from the pictures obtained using the APIs (Application Programming Interface).

Figure 10 shows a matching score of 295.31, and 84 matching points. Fig. 11 shows a score of 340.40 and 105 matching points. In this example, the score for the same location is lower than the score for different locations, which shows misjudgment. The panoramic image appears small with features such as buildings compared to the original images.
Fig. 12. Matching of the same location using panoramic images when the road and sky areas are filled with white (score: 392.23, number of matching points: 84).

Fig. 13. Matching of different locations using panoramic images when the road and sky areas are filled with white (score: 286.67, number of matching points: 62).

Fig. 14. Matching of the same location using a cut-off method (score: 325.48, number of matching points: 106).

Fig. 15. Matching of different locations using a cut-off method (score: 329.62, number of matching points: 110).

Figures 12 and 13 show examples of the same images with the road and sky areas filled with white pixels. Figure 12 shows a score of 392.23 and 84 matching points. Fig. 13 shows a score of 286.67 and 62 matching points. In this example, the same location’s score is higher than the score of a different location. Thus, the location judgment is correct. The filled area contains most of the matching points.

Figures 14 and 15 are based on the same image, but the area corresponding to the road and sky are cut away. This is done by cutting off pixels from 0 to 65 and 150 to 250 in the vertical direction, and horizontally from 220 to 280 pixels for the 500×250 pixel image. Figure 14 shows a score of 325.48 and 106 matching points. Fig. 15 shows a score of 329.62 and 110 matching points. In this example, the same location’s score is lower than the score of a different location. This is a misjudged case. The sky area in the left image contains numerous mismatching points.

Then, another example of the same method under snow conditions is shown in Figs. 16 and 17. The left image was taken by the camera at the same location as the right street view image. However, it was captured during winter, when snow mostly obscured the roads. Figure 16 shows a score of 292.02 and 77 matching points. Fig. 17 shows a score of 188.27 and 54 matching points. In this example, the same location’s score is higher than the score of a different location. The judgment of the location is correct.

Figure 18 shows a score of 414.22 for the same location and 129 matching points. Additionally, Fig. 19 shows a score of 386.41 for the different locations and 124 matching points. Therefore, the results show a correct location judgment.

To verify the effect of seasonal change, we compared two images of the same location (Fig. 20) and two images of different locations (Fig. 21), both with and without snow. Figure 20 shows a score of 504.31 for the same location and 129 matching points. Fig. 21 shows a score of 358.79 for different locations and 111 matching points. In this example, the same location’s score is higher than the score of a different location. Therefore, the location judgment is correct given the change in season.
Fig. 16. Matching of the same location using a cut-off method under snow conditions (score: 292.02, number of matching points: 77).

Fig. 17. Matching of different locations using a cut-off method under snow conditions (score: 188.27, number of matching points: 54).

Fig. 18. Matching of the same location using a method that removes the remaining road and sky area after cutting-off (score: 414.22, number of matching points: 129).

Fig. 19. Matching of different locations using a method that removes the remaining road and sky area after cutting-off (score: 386.41, number of matching points: 124).

Fig. 20. Matching of the same location by removing the remaining road and sky area after cutting-off under snow conditions (score: 504.31, number of matching points: 146).

Fig. 21. Matching of different locations by removing the remaining road and sky area after cutting-off under snow conditions (score: 358.79, number of matching points: 111).
4.4. Experiment 2 (Panoramic Image Matching by Directionally Divided Sections)

The next idea of matching camera and street view images is to compare directionally divided sections of panoramic images. The goal is to prevent mismatch between distant image regions that appear similar to one another but do not belong to the same object.

We divide the panoramic images into the direction of movement, left, right, and behind, and match the divided images in the same split direction. Figs. 22–25 show the results of matching the same location using this method.

Additionally, the matching of the different locations is shown in Fig. 26–29. Figures 22–25 show the results of matching for the same location. Even though the weather makes the images from the panoramic camera and street view images look different, DeepMatching can accurately match the same buildings and trees. Figs. 22–25 show a total score of 345.70 and 126 matching points.

In contrast, the total score and matching points are 320.64 and 114 for Figs. 26–29, respectively, which represent the matching for different locations. In this example, the total score of the same location is higher than the total score of a different location, which means correct judgment.

4.5. Summary of the Probability of Misjudgment by Proposed Methods

In this study, the probability of misjudgment is defined as follows:

\[
\text{Probability of misjudgment} = \frac{\text{Number of misjudgment}}{\text{Number of comparison of the same location}} \times 100
\]

The probability of misjudgment are summarized in Table 1.

5. Discussion

Figures 10–13 show that misjudgments are eliminated when the image is filled with white pixels. However, there are also some examples of severe misjudgment. According to Figs. 14 and 15, it is obvious that the images in Fig. 15 are in a different location. However, even these
apparent examples were often misjudged because of the unintended similarities of the textures.

The use of the number of unmatched points or the Gist features to determine the overall image similarity and the use of DeepMatching when the similarity is above a certain value are two concepts that can be used to filter out the obvious dissimilar places.

The probability of misjudgment indicates that general image formats are preferable to panoramic image formats. However, whenever general image formats are used, it is necessary to determine whether the image is facing the same direction as the reference image.

As shown in Table 1, the panoramic image division method performed well. This is because the image division prevents the mismatch of points of distant objects, which are often observed when matching panoramic images, as shown in Fig. 10.

### 6. Conclusion

This study sought to clarify the navigation method by matching the camera images of the mobile robot with street view images memorized beforehand. It is based

### Table 1. Probability of misjudgment.

<table>
<thead>
<tr>
<th>Image type</th>
<th>Misjudged images</th>
<th>Total number of comparison</th>
<th>Probability of misjudgment</th>
</tr>
</thead>
<tbody>
<tr>
<td>General image format</td>
<td>2</td>
<td>38</td>
<td>5.26%</td>
</tr>
<tr>
<td>Panoramic image</td>
<td>8</td>
<td>19</td>
<td>42.11%</td>
</tr>
<tr>
<td>Removing road and sky area</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Filling the road and sky areas with white pixel</td>
<td>7</td>
<td>19</td>
<td>36.84%</td>
</tr>
<tr>
<td>2. Cutting-off all other areas except the road and sky areas</td>
<td>9</td>
<td>48</td>
<td>18.75%</td>
</tr>
<tr>
<td>3. Removing the remaining road and sky areas</td>
<td>14</td>
<td>48</td>
<td>29.17%</td>
</tr>
<tr>
<td>Panoramic image division</td>
<td>4</td>
<td>14</td>
<td>28.57%</td>
</tr>
</tbody>
</table>
on how humans use street view images to remember their current location and control the moving direction in urban areas.

In this study, we investigated a better image matching method by modifying the input street images in various ways.

We believe that the matching method using general image features is difficult to handle the changes in scenery, such as weather, time zone, and angle of view between camera-taken and street view images.

We used DeepMatching, a method for matching images that can deal with object movement, differences in lighting, and viewpoint angles.

However, DeepMatching tends to find feature points in parts of the road and sky, and a distance from each other. Therefore, we proposed methods to improve similarity by removing parts of the roads and sky and dividing the panoramic image into different directions.

Then, we verified the validity of the proposed methods.

We verified three methods for removing road and sky areas: painting the removed areas with white pixels, removing the fixed regions that correspond to road and sky areas, and removing the remaining road and sky areas from the image after applying the second method. While misjudgment could not be avoided, the second method had the lowest probability of misjudgment.

The method that divided the panoramic image and applied one to one image matching with the corresponding directions resulted in a low probability of misjudgment.

In the proposed method, the task of filling the road and sky areas with white pixels was conducted manually at this point. We will investigate automated methods to complete this task. Additionally, we will continue to investigate the actual implementation of this method for navigating autonomous mobile robots.

References:


