ACCURATE INDOOR LOCALIZATION USING MULTI-VIEW IMAGE DISTANCE

Xinyun Li† Ryosuke Furuta†† Go Irie††† Yukinobu Taniguchi†
† Tokyo University of Science ††The University of Tokyo
†††NTT Communication Science Laboratories

ABSTRACT

Due to the increasing complexity of indoor facilities such as shopping malls and train stations, there is a need for a technology that can find the current location of a user using a smartphone or other devices, even in indoor areas where GPS signals cannot be received. Indoor localization methods based on image recognition have been proposed as solutions. While many localization methods have been proposed for outdoor use, indoor localization has difficultly in achieving high accuracy from just one image taken by the user (query image), because there are many similar objects (walls, desks, etc.) and there are only a few cues that can be used for localization. In this paper, we propose a novel indoor localization method that uses multi-view images. The basic idea is to improve the localization quality by retrieving the pre-captured image with location information (reference image) that best matches the multi-view query image taken from multiple directions around the user. To this end, we introduce a simple metric to evaluate the distance between multi-view images. Experiments on two image datasets of real indoor scenes demonstrate the effectiveness of the proposed method.

Keywords: indoor localization, multi-view image, image recognition, similarity image search

1. INTRODUCTION

Due to the increasing complexity of indoor facilities such as shopping malls and train stations, the number of pedestrians getting lost has been increasing. There is a need for technology for smartphones or other devices to easily localize the current location. Several localization methods have been developed that can be used indoors, where GPS signals cannot be received. For example, geomagnetism [1] takes advantage of the fact that the arrangement of steel in buildings is very unlikely to change in a short time. Pedestrian Dead Reckoning (PDR) [2] measures the position of a pedestrian relative to a reference point based on data obtained from several autonomous sensors, such as accelerometers and gyroscopes. However, the geomagnetic field is disturbed by the passage of large vehicles and other factors, which reduces accuracy. PDR requires a precise specification of the starting point, which may lead to a loss of accuracy due to errors in its setting. In addition, the installation and operation costs are high.

To solve these problems, many indoor localization methods based on image recognition, such as [3], have been proposed. This approach saves pre-captured images with location information (reference images) in a database, and the current location is estimated by comparing the user's image (query image) with the reference images in the database and identifying the closest match. This method does not require any special equipment and can be used in any location as long as a database of reference images is available. However, unlike outdoor locations, indoor locations have many similar objects such as walls and floors, and there are few cues available for localization.

In this paper, we propose an approach that use multi-view images with four shooting directions (front, back, left and right) as the query. The key is accurate evaluation of the distance between the multi-view images. Unfortunately, it is not sufficient to independently evaluate the distance between a query image and a reference image because different locations might be retrieved for each query image, even for the same location, which may confuse users. To address this issue, we introduce term of multi-view image distance to effectively evaluate the dissimilarity between query and reference images. Evaluation experiments using image datasets of two real scenes show that the method provides more accurate localization estimates than previous methods.

2. RELATED WORKS

2.1. Similarity Image Search Using Local Features

Philbin et al [4] used Scale-Invariant Feature Transform (SIFT) features [5] and Bag of Visual Words (BoVW) to reduce the processing time and improve the accuracy of similar image matching. However, this method has a problem in that the preprocessing of feature extraction and BoVW is expensive when the database is large. Moreover, because it only utilizes luminance information, it is difficult to distinguish the difference of color.

2.2. Similarity Image Search Using Global Features

Convolutional Neural Networks (CNN) are now used often in the field of similarity image search. Regional Maximum Activation of Convolutions (R-MAC) [6], which combines the feature vectors extracted from multiple regions in an image and Generalized Mean Pooling (GeM Pooling) [7], which generalizes the pooling layer calculation, have been proposed.

In this paper, we use the image features from GeM Pooling by Radenović et al. [7] for similar image retrieval.

2.3. Indoor Localization Using Graph Location Networks

Chiou et al [8] proposed a new deep Neural Network architecture for indoor localization, Graph Location Networks (GLN), based on Graph Convolutional Networks (GCN) and published the West Coast Plaza (WCP) dataset. In this method, features of multi-view images are extracted by ResNet152 trained on ImageNet. Their method uses GCN whose nodes represent the locations of image capture points to connect location information and image features, so it offers robust estimation of correct location. Furthermore, they use a zero-shot learning approach to reduce the labor costs of taking

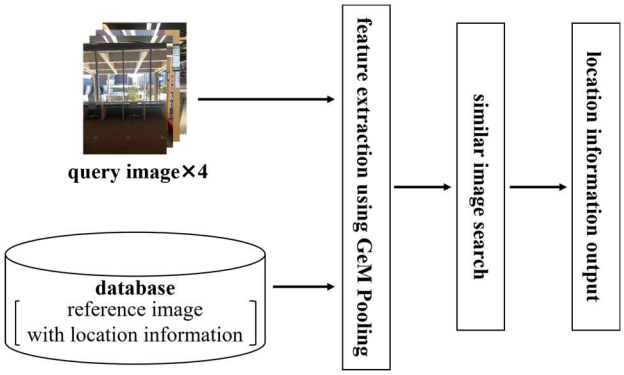


Fig. 1: Flowchart

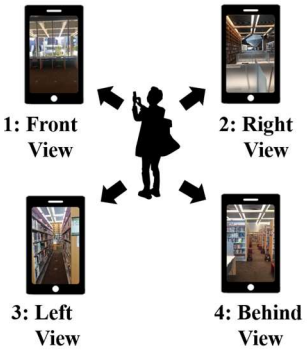


Fig. 2: The method of shooting query images in ours.

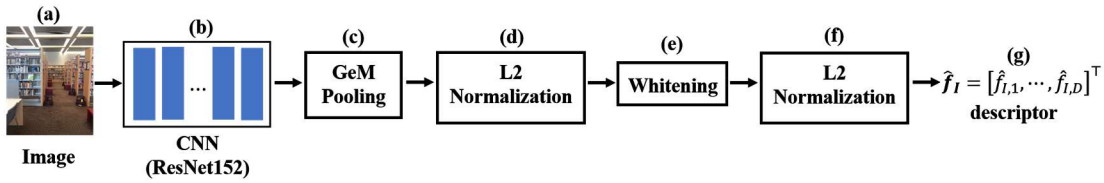


Fig. 3: Feature extraction using GeM Pooling.

reference images allowing the system to be deployed in large indoor environments.

While GLN also uses multi-view images as the query, our method predicts the location by using the proposed distance metric to calculate the dissimilarity between the multi-view query and reference images: GLN aggregates the features from the multi-view query and directly predicts the location with GCN. In this paper, we compare the accuracy of our method with GLN using the WCP dataset.

3. PROPOSED METHOD

The flowchart of the proposed method is shown in Fig. 1. To solve the problem that using just one query image makes it too difficult to achieve highly accurate estimates for indoor localization, we use multi-view images as the query image. Because it is infeasible to independently evaluate the distance between each of part of the multi-view query and the reference images, we propose an effective distance metric.

(I) Query images capture: In order to capture more information from around the current location, the shooting directions of the query images are defined as shown in Fig. 2. At one location, four images (front, back, left and right) are taken at 90° intervals while the camera is rotated horizontally.

(II) Database: At different locations in the facility $j \in \{1, \cdots, N_R\}$, reference images $\boldsymbol{R_j} = \{R_{ja}\}_{a=1,2,3,4}$ are captured in four directions in advance, just like the query image, and the reference location information is stored in the image database in the form of x-y coordinates.

(III) Feature extraction using GeM Pooling : In order to achieve higher accuracy in localization, we use GeM Pooling[7], which has achieved the highest level of accuracy among conventional methods. The procedure shown in Fig. 3 is used to extract 2048-dimensional feature vectors. The generalized mean of the feature map for each channel is defined as:

$$\boldsymbol{f_I} = [f_{I,1}, \cdots, f_{I,D}]^\top, \quad f_{I,d} = \left(\frac{1}{|\mathcal{X}_d|} \sum_{x \in \mathcal{X}_d} x^{p_d} \right)^{\frac{1}{p_d}}, \qquad (3.1)$$

where D is the number of dimensions, p_d is the pooling parameter and $\mathcal{X}_d$ is the feature map at d-th channel $d \in \{1, \cdots, D\}$.

(IV) Similar image search using multi-view image: Multi-view image distance $distance(\boldsymbol{Q}, \boldsymbol{R}_j)$ between query image set $\boldsymbol{Q} = \{Q_a\}_{a=1,2,3,4}$ and a reference image set $\boldsymbol{R_j}$ is defined as:

$$\text{distance}(\boldsymbol{Q}, \boldsymbol{R}_\text{j}) = \min_{\sigma \in S_4} \sum_{a=1}^{4} \text{dist}(Q_a, R_{j\sigma_a}), \qquad (3.2)$$

where a is the shooting direction shown in Fig. 2 and S_4 is a permutation of $\{1,2,3,4\}$, $\sigma = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$. The reason why we do not use the cyclic permutation of $\{1,2,3,4\}$ is because we assume that the query images and reference images are not limited to four per location, and the order in which they are captured is arbitrary. $\text{dist}(Q_a, R_{j\sigma_a})$ is the Euclidean distance between the feature vectors of query image Q_a and reference image $R_{j\sigma_a}$. We calculate the multi-view image distance between the query image set and all reference image sets in the database $\boldsymbol{R_j}(j = 1,2, \cdots, N_R)$ and determine the current location by finding the reference image set having the smallest distance.

$$j = \underset{j=1,2,\cdots,N_R}{\text{argmin}} \ \text{distance}(\boldsymbol{Q}, \boldsymbol{R_j}). \qquad (3.3)$$

4. EXPERIMENTS

In order to evaluate the effectiveness of our method, we compared its accuracy to the following methods:

(A) Local Feature: Point feature matching by SIFT features and Geometric Verification [4]. In this experiment, we do not generate the short list by BoVW.

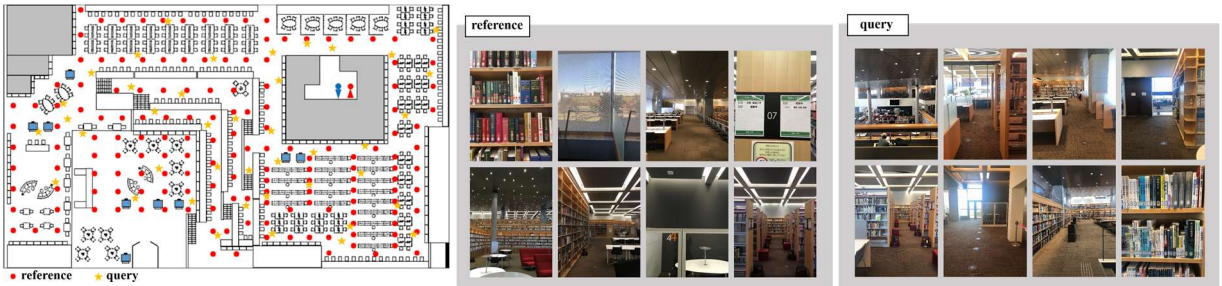
(B) GeM Pooling [7]: The feature vector is extracted by applying GeM Pooling to the feature map extracted from the query image. Cosine similarity is used for measuring the similarity between images.

(C) GLN [8]: GCN is used to estimate location with 4 query images.

(D) Our proposed method (number of directions $a = \{1, \cdots, 4\}$)

4.1. Experimental Conditions

As the backbone network, we used ResNet152, which was trained on google-landmarks-2018 [9], and includes whitening. The pooling parameter p_d in equation (3.1) was set to 3. We used two datasets as described below:



(a) The location of reference and query. (b) Examples of reference images. (c) Examples of query images.

Fig. 4: TUS Library Dataset

(a) An illustration of reference images' location. The map is for illustrative purposes only. [8] (b) Examples of reference images. (c) Examples of query images.

Fig. 5: West Coast Plaza (WCP) Dataset [8]

Table 1: Relationship between the number of directions and methods, and the accuracy of localization. The dataset of library was used. The higher accuracy of feature extraction method in blue. The highest accuracy results in **bold**.

Method	Distance	Direction(s)	Times[sec]	Accuracy [%]
SIFT	Cosine Similarity	1	132.23	47.62
GeM Pooling		1	0.76	61.31
Ours	Multi-View Distance	1	0.65	61.31
		2	1.93	71.43
		3	3.29	80.95
		4	3.95	**85.71**

Table 2: Accuracy comparison between GLN and ours. The highest accuracy results in **bold**. The dataset of WCP was used.

Method	Direction(s)	Accuracy [%]
GLN [8]	4	79.88
Ours	1	73.08
	2	78.21
	3	80.76
	4	**84.02**

4.1.1. TUS Library Dataset

TUS Library Dataset is our proprietary dataset: it is a set of images taken at the Tokyo University of Science (TUS) Katsushika Campus Library (floor area: 3,358 m^2). Examples of reference images and query images are shown in Fig. 4(b) and Fig. 4(c), respectively. As shown in Fig. 4(a), we captured reference images at 159 locations × 4 directions (636 images in total) taken at about 1[m] intervals by an iPhoneSE. Query images of 42 locations × 4 directions (168 images in total) were taken at random locations with an iPhone8Plus. All the images had size of 480 × 640[px].

4.1.2. West Coast Plaza (WCP) dataset [8]

WCP Dataset [8] is a public dataset of images taken at a shopping mall in Singapore (floor area: 15,000m^2). Examples of the reference images and query images are shown in Fig. 5(b) and Fig. 5(c), respectively. We have reference images of 316 locations × 4 directions (1264 images in total) taken at about 1 [m] intervals with a Vivo Y79 and query images of 78 locations × 4 directions (312 images in total) were taken at random locations with a Vivo Y79.

4.2. Evaluation Metrics

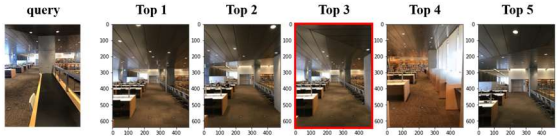
The percentage of query images where the distances between the estimated location and the ground truth location are within 1[m] is reported as One-Meter-Level Accuracy:

$$\text{Accuracy} = \frac{\sum_{i=1}^{N_Q} C(\boldsymbol{Q}_\text{i})}{N_Q}, \tag{4.1}$$

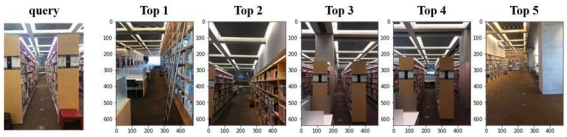
where, $C(\boldsymbol{Q}_\text{i})$ is set to 1 if the distance from the query images to detected location is within 1[m], and 0 otherwise, N_Q is the number of query locations, and $\boldsymbol{Q}_i$ is the set of query images at locations $i \in \{1, \cdots, \text{N}_\text{Q}\}$.

4.3. Results and Discussions on TUS Library Dataset

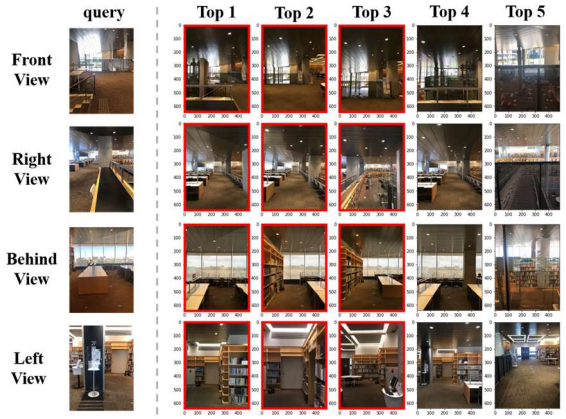
Table 1 shows the results of the experiments on the TUS Library Dataset. Comparing the results of SIFT features and GeM Pooling, the introduction of GeM Pooling increases the accuracy by 13.69 points, which shows the effectiveness of GeM Pooling. When the number of directions is increased from 1 to 4, the accuracy of the proposed method is increased by 24.41 points. The results show that as the number of directions increases, the accuracy improves. This demonstrates the effectiveness of using multi-view images and our distance measure.



(a) Localization by GeM Pooling and cosine similarity.

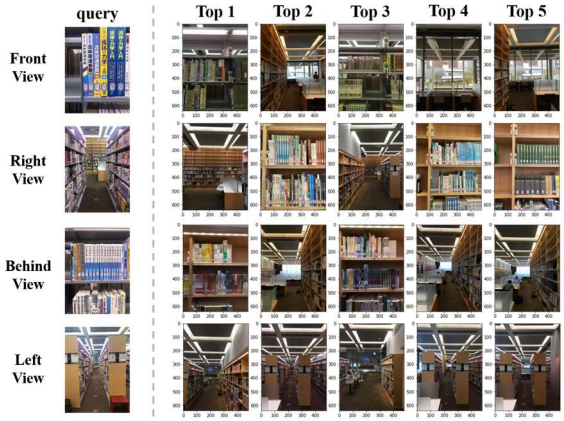


(b) Localization by GeM Pooling and cosine similarity.



(b) ours (success)

Fig. 8: Successful localization by proposal. The correct images, show in red, are those where the distances between the estimated location and the ground truth location are within 1[m].



(b) ours (fail)

Fig. 9: The failure of proposal.

To check not only the percentage of correct images in top1, but also their ranking, the results for the top 5 are shown in Fig.8 and Fig.9. As shown in Fig. 8, the method using GeM pooling with cosine similarity failed to retrieve the correct reference image (Fig. 8 (a)), while our method successfully retrieved the correct reference images (Fig. 8 (b)). Furthermore, the images ranked in the top 3 indicated the correct location. This is because by increasing the number of shooting directions, more information about the surrounding environment (such as black iron fences, windows, doors) is captured, providing more clues for localization.

The failure case (Fig. 9) is an example where the correct location could not be retrieved within the top 5 even after using multi-view images. This is because there were many similar images in all four directions, making it difficult to obtain effective features for localization.

4.4. Results and Discussions on WCP Dataset

The results of the experiments on the WCP dataset are shown in Table 2. Comparing GLN with the proposed method, the accuracy of the proposed method with four shooting directions was 4.14 points higher than that of GLN. It shows that our distance measure for multi-view image significantly contributes to accurate localization.

5. CONCLUSIONS

In this paper, we proposed a method using multi-view image distance to improve the accuracy of indoor localization based on image retrieval. Our proposed method uses the information of the surrounding area to solve the problem that a single query image cannot capture enough features. Experiments showed that the accuracy of the proposed method is improved by 24.41 points by using multi-view images. Furthermore, its accuracy was shown to be 4.14 points higher than that of GLN [8].

As a future challenge, considering the practicality of the system, it is necessary to reduce the number of query images taken while keeping the accuracy of localization, because requiring users to take four query images every time would reduce the convenience of the system. At the same time, it is also necessary to evaluate (i) performance in places where people often get lost, such as train stations, (ii) accuracy when query images include moving objects such as pedestrians, (iii) accuracy when the capture orientation of a query is different from that of the reference image, and (iv) to verify the accuracy when using single-shot panoramic images to reduce the cost of creating the reference images. Last but not least, the feature extractor should be trained on image data collected in a particular facility, so that it can learn the characteristics specific to the facility and to estimate the current position with higher accuracy.

REFERENCE

[1] Ho J. Jang, Jae M. Shin and Lynn Choi. Geomagnetic Field Based Indoor Localization Using Recurrent Neural Networks. In *GLOBECOM*, pp. 1-6, 2017.

[2] Beauregard Stephane, and Harald Haas. Pedestrian Dead Reckoning: A Basis for Personal Positioning. In *WPNC*, pp.27-35, 2006.

[3] Akihiko Torii, et al. 24/7 Place Recognition by View Synthesis. In *CVPR*, pp.1808-1817, 2015.

[4] James Philbin, et al. Object Retrieval with Large Vocabularies and Fast Spatial Matching. In *CVPR*, pp. 1–8, 2007.

[5] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, Vol. 60, No. 2, pp. 91–110, 2004.

[6] Jaeyoon Kim and Sung-Eui Yoon. Regional Attention Based Deep Feature for Image Retrieval. In *BMVC*, 2018.

[7] Filip Radenović, et al. Fine-Tuning CNN Image Retrieval with No Human Annotation. *TPAMI*, Vol. 41, No. 7, pp. 1655‑1668, 2019.

[8] Meng-Jiun Chiou, et. al. Zero-Shot Multi-View Indoor Localization via Graph Location Networks. In *ACM Multimedia*, pp. 3431-3440,2020.

[9] Google-Landmarks Dataset. https://www.kaggle.com/google/google-landmarks-dataset, May 2, 2018. Last access : 2020/12/20.