






# Interactive Indoor Localization Based on Image Retrieval and Question Response

Xinyun Li<sup>1</sup><sup>a</sup>, Ryosuke Furuta<sup>2</sup><sup>b</sup>, Go Irie<sup>1</sup><sup>c</sup>, Yota Yamamoto<sup>1</sup><sup>d</sup> and Yukinobu Taniguchi<sup>1</sup><sup>e</sup>

<sup>1</sup>Department of Information and Computer Technology, Tokyo University of Science, Tokyo, Japan

<sup>2</sup>Institute of Industrial Science, The University of Tokyo, Tokyo, Japan

**Keywords:** Indoor Localization, Image Recognition, Similarity Image Search, Scene Text Information.

**Abstract:** Due to the increasing complexity of indoor facilities such as shopping malls and train stations, there is a need for a new technology that can find the current location of the user of a smartphone or other device, as such facilities prevent the reception of GPS signals. Although many methods have been proposed for location estimation based on image search, accuracy is unreliable as there are many similar architectural indoors, and there are few features that are unique enough to offer unequivocal localization. Some methods increase the accuracy of location estimation by increasing the number of query images, but this increases the user's burden of image capture. In this paper, we propose a method for accurately estimating the current indoor location based on question-response interaction from the user, without imposing greater image capture loads. Specifically, the proposal (i) generates questions using object detection and scene text detection, (ii) sequences the questions by minimizing conditional entropy, and (iii) filters candidate locations to find the current location based on the user's response.

## 1 INTRODUCTION

The number of pedestrians getting lost has been increasing because of the increasing complexity of indoor facilities such as shopping malls. Accordingly, there is need for a new technology for smartphones or other devices that can easily determine the user's current location. The indoor location market is projected to grow from USD 8.8 billion in 2022 to USD 24.0 billion by 2027, at a Compound Annual Growth Rate (CAGR) of 22.4% during the forecast period (MarketSandMarkets, 2022).


To achieve this goal, many localization methods based on image recognition, such as (Torii et al., 2015), have been proposed and studied in the area of computer vision for a long time. 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 architectural features, such as a cluster of restaurants and clothing stores as shown in Figure 1, and there are few unambiguous cues available for localization. Therefore, a single query image is usually not enough to achieve high accuracy.


To solve the indoor specific problems, among the many solutions, (Chiou et al., 2020; Li et al., 2021) have been proposed to improve accuracy by increasing the number of capture directions used as input. However, taking multiple query images is time-consuming, and increases the issue of legal and ethical considerations, such as the difficulty of taking pictures when a person is in front of the camera.

To reduce the number of query images while keeping the accuracy, we propose a method for accurately estimating the current indoor location by asking the user to answer generated questions. The main contribution of this paper are as follows: (i) Our proposal yields highly accurate interactive localization by short question-response sequences with the user. The question(s) are generated by object detection and scene text detection. (ii) To reduce the number of question-responses, we propose a question selection

<sup>a</sup> <https://orcid.org/0000-0002-8920-942X>

<sup>b</sup> <https://orcid.org/0000-0003-1441-889X>

<sup>c</sup> <https://orcid.org/0000-0002-4309-4700>

<sup>d</sup> <https://orcid.org/0000-0002-1679-5050>


<sup>e</sup> <https://orcid.org/0000-0003-3290-1041>



Figure 1: (a) and (b) are similar, but they are actually about 10 m apart.

method based on conditional entropy. This reduces the number of needed query to 1, and the user need to respond with only “yes” or “no” to an average of 2.75 questions to get location information. (iii) Experiments on two shopping mall datasets show that the proposed method can achieve better accuracy than the non-interactive method compared (Chiou et al., 2020; Li et al., 2021), which captures images in multiple directions while keeping the number of images taken to a single shot.

## 2 RELATED WORKS

### 2.1 Localization Based on Image Search

Many indoor localization methods using images have been proposed. Examples (Gao et al., 2016; Dong et al., 2019) include using images captured by a smartphone in combination with sensors to identify the location. Although these methods are able to reduce the cost of creating reference images, in complex indoor environments the signal is subject to strong interference which degrades accuracy.

As a method that uses only images, and so dispenses with sensors, (Taira et al., 2018) use dense features such as DenseSIFT (Liu et al., 2008) and estimate 6DoF camera poses to achieve highly accurate localization on a large scale. (Li and He, 2021) proposed a robust localization method for dynamic environments that uses video images and text information. The method proposed by (Wang et al., 2015) uses text detection, shop facade segmentation, and map information. In addition, (Radenović et al., 2018) proposed Generalized Mean Pooling (GeM Pooling), which generalizes the pooling layer calculation for similar image search when extracting image features using convolutional neural networks (CNN); they demonstrated high accuracy.

In this paper, we use image features from GeM Pooling for a similar image search and improve the accuracy by adding interaction by question and response.

### 2.2 Localization Using Multi-View Images

(Liu et al., 2017) proposed a method for estimating the current location from images and geomagnetic signals by processing multi-view images in Multi-view Graph (MVG). They conducted extensive experiments on three types of buildings and showed that an accuracy of 1 m was successfully achieved even when noise and outliers occupied 30% of the data. However, the method does not account for differences between views, and thus cannot capture robust local representations.

To solve this problem, (Chiou et al., 2020) proposed Graph Location Networks (GLN), a new architecture based on Graph Convolutional Networks (GCN). This method extracts features from multi-view images using ResNet152 trained on ImageNet and uses GCN, whose nodes represent the locations of image capture points, to connect location information and image features; it offers robust estimation of correct location. Furthermore, they use a zero-shot learning approach to reduce the labor cost of taking reference images allowing the system to be deployed in large indoor environments.

Considering the problem that there are many similar architectural features indoors, (Li et al., 2021) proposed a method that uses multi-view images with four shooting directions (front, behind, left, and right) as the query and introduced the term of multi-view image distance to effectively evaluate the dissimilarity between query and reference images. Although we obtained higher accuracy than (Chiou et al., 2020), it did not solve the problem that users have to take many images, which is not only time-consuming but also burdensome in terms of legal and ethical considerations.

To solve this problem, this paper focuses on estimating the current location by using the user’s responses to questions.

## 3 PROPOSED METHOD

To improve the accuracy of current location estimation, this paper proposes a method that combines image search using a single query image and location filtering by question responses (Figure 2). By asking the user about the presence or absence of store signs

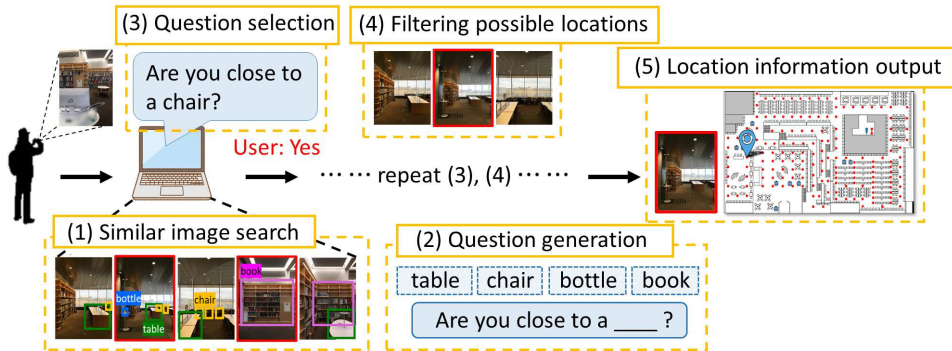


Figure 2: Flowchart of localization based on question response.

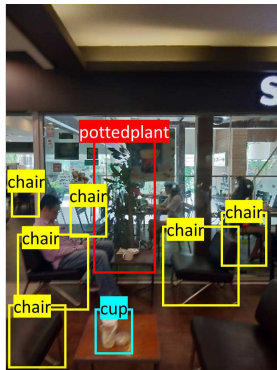


Figure 3: The result of object detection, which is used to generate the questions.

and objects in similar architectural features, the current location can be identified more precisely. The flow of the proposed method is described below.

### 3.1 Similar Image Search

From a single query image taken by the user, locations linked to the top  $K$  similar images from the database are selected as possible locations using similarity-based image search. In addition, the database includes reference images  $\mathbf{R}_k = \{R_{ka}\}_{a=1,2,3,4}$ , which are taken in 4 directions at every location in the facility.  $k = 1, 2, \dots, N_R$ , and  $N_R$  are the number of reference images.

### 3.2 Question Generation

Our method generates questions  $Q_i (i = 1, 2, \dots)$  from all reference images  $\mathbf{R}_k$  in the following way.

**Object Label.** As shown in Figure 3, we use an object detector to find objects (desk, chair, etc.) and generate questions such as “Are you close to a (object label)?” Objects that change in a short span of time (e.g., person, bag) and objects that do not exist



(a) The result of scene text detection.



(b) The image to be presented to the dr at the same time as the generated question.

Figure 4: The scene text detection is to detect the position of scene text from reference images and generate the most effective questions.

indoors (e.g., dog) are not included in the search. In the case of Figure 3, the questions “Are you close to chairs?”, “Are you close to a potted plant?” and “Are you close to a cup” are generated.

**Scene Text Information.** As shown in Figure 4, the position of scene text is detected from the reference image (Figure 4(a)) and a text image (Figure 4(b)) is presented to the user. At the same time the question “Can you see this signboard?” is generated.

### 3.3 Question Selection

To get the correct answer with fewer questions, we define conditional entropy  $H_{Q_i}$ , a measure of the amount of information included in question  $Q_i$ , and find the question that has the minimum conditional entropy:

$$H_{Q_i} = - \sum_{k=1}^K \sum_{j \in \{\text{Yes, No}\}} P(B_{ij}) P(A_k | B_{ij}) \log_2 P(A_k | B_{ij}), \quad (1)$$

where  $A_k$  is the event that possible location  $k$  is correct,  $P(B_{ij})$  is the posterior probability when user responds to question  $Q_i$  with  $j \in \{\text{Yes, No}\}$ . Its proba-

bility is defined as:

$$P(B_{ij}) = \begin{cases} \frac{1}{K} \sum_k score_{ik}, & j = \text{Yes} \\ \frac{1}{K} \sum_k (1 - score_{ik}), & j = \text{No} \end{cases} \quad (2)$$

$score_{ik} \in [0, 1]$  is defined as the confidence score if object class  $i$  is detected in the reference image  $R_k$ , and 0 if not detected. Also,  $score_{ik} = 1$  if scene text is detected, and  $score_{ik} = 0$  otherwise.

$P(A_k | B_{ij})$  is the posterior probability when the user is at possible location  $k$  and responds  $j \in \{\text{Yes}, \text{No}\}$  to question  $Q_i$ . It is defined as:

$$P(A_k | B_{ij}) = \begin{cases} \frac{score_{ik} \cdot e^{S_k}}{\sum_{l=1}^K score_{il} \cdot e^{S_l}}, & j = \text{Yes} \\ \frac{(1 - score_{ik}) \cdot e^{S_k}}{\sum_{l=1}^K (1 - score_{il}) \cdot e^{S_l}}, & j = \text{No} \end{cases} \quad (3)$$

where  $S_k$  is the similarity between reference image  $R_k$  and the query image.

### 3.4 Filtering Possible Locations

Question  $Q_i^*$  that has minimum conditional entropy

$$Q_i^* = \arg \min_i H_{Q_i} \quad (4)$$

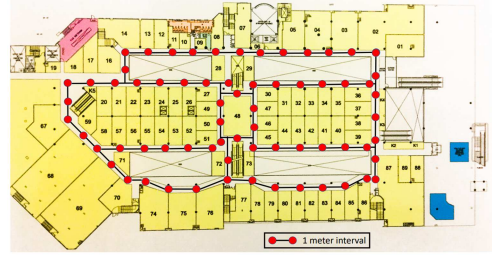
will be asked to the user, and we remove the possible locations that do not match the user's response. The process stops when the number of possible locations is reduced to one, or when there are no more questions to ask the user. If all the possible locations have been removed, the location with maximum image similarity (top 1) is output.

## 4 EXPERIMENTS

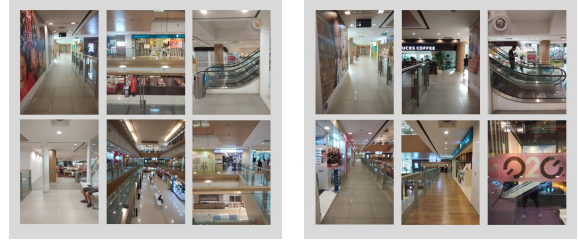
To evaluate the effectiveness of the proposed method, the following methods were compared.

- Similar image search: Feature vectors are extracted using GeM Pooling (Radenović et al., 2018). To measure the similarity between images, cosine similarity is used. See Section 4.2 for details.
- Multi-view image distance (Li et al., 2021): Multi-view images with 4 different shooting directions (front, behind, left, and right) are used as the query. The multi-view image distance is the summation of the Euclidean distances of the query and reference image pairs, which are created without duplication.
- Ours

Moreover, we conducted an ablation study to evaluate the



(a) An illustration of reference image locations. The map is for illustrative purposes only (Chiou et al., 2020).



(b) Reference images.

(c) Query images.

Figure 5: West Coast Plaza (WCP) Dataset (Chiou et al., 2020).

- relationship between the number of initial possible locations and accuracy.
- effect of each question generation module.
- effectiveness of conditional entropy in question selection.

## 4.1 Dataset

### (a) West Coast Plaza Dataset

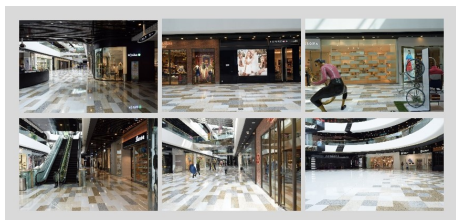
WCP Dataset (Chiou et al., 2020) is a public dataset of images taken at a shopping mall in Singapore (floor area: 15,000m<sup>2</sup>). Examples of the reference images and query images are shown in Figure 5(b) and Figure 5(c), respectively. We have reference images of 316 locations  $\times$  4 directions (1,264 images in total) taken at about 1 m intervals, as shown in Figure 5(a), with a Vivo Y79 and query images of 78 locations  $\times$  4 directions (312 images in total) were taken at random locations with a Vivo Y79.

### (b) Mall Dataset

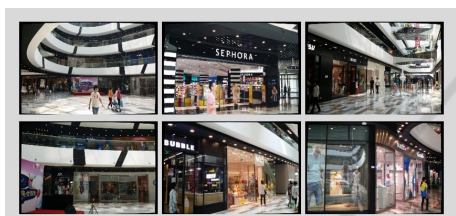
Mall Dataset (Sun et al., 2017) is image data taken at a shopping mall (Figure 6(a)) in Jiangsu, China. Examples of reference images and query images taken are shown in Figure 6(c) and Figure 6(d). The reference images are 682 images taken with a Nikon D5300 at about 1 m intervals (Figure 6(b)). The query images are images taken by six different smartphone cameras.



(a) Captured point cloud in birds-eye view. (Sun et al., 2017)  
 (b) Close-up of the camera poses for capturing database images. (Sun et al., 2017)



(c) Reference images.



(d) Query images.

Figure 6: Mall Dataset (Sun et al., 2017).

However, in this experiment, only a subset of images (80 images) taken with an iPhone4s were tested.

## 4.2 Experimental Setting

**Similar Image Search.** For feature extraction, we use GeM Pooling (Radenović et al., 2018) As the backbone network, we used ResNet152, which was trained on google-landmarks-2018 (Noh et al., 2017), and includes whitening. We determine similar images by cosine similarity, and the top 5 retrieved images are considered as possible images.

**Object Detection.** Cascade Mask-RCNN (He et al., 2017) with Swin Transformer (Liu et al., 2021) backbone was used for object detection. Of the 36 types of labels detected, we kept 33 object labels that have been present in indoor facilities for a long time (desks, chairs, etc.) and removed 3 other labels (people, dogs, bags).

**Scene Text Detection.** EAST (Zhou et al., 2017) was used. Furthermore, we perform scene text recognition (Bautista and Atienza, 2022) on the detected bounding box. Images that were successfully recognized



(a) The image that could be presented to the user. Output: STEVENCHAN  
 (b) The image to be deleted. Output: -

Figure 7: The result of scene text recognition.

(Figure 7(a)) are presented to the user, while images that failed (Figure 7(b)) are deleted.

**Question Response.** The 4 direction query images have the correct object label and scene text information manually assigned. Since the location and time of capture of the query image and the reference image do not match, the final estimation result does not always match the correct answer, even if the user’s answer is accurate. In addition, the question “Are you close to \*\*?” was answered with “yes” regardless of the distance to the objects or signboards.

## 4.3 Evaluation Metrics

We tested the effectiveness of our proposed method using two evaluation metrics and the average number of questions as described below.

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

$$\text{Accuracy} = \frac{\sum_{q=1}^{N_{\text{query}}} C(I_q)}{N_{\text{query}}}, \quad (5)$$

where,  $C(I_q)$  is set to 1 if the distance from the query images to detected location is within 1 m, and 0 otherwise.  $N_{\text{query}}$  is the number of query locations,  $I_q$  is the  $q$ -th query image.

**Cumulative Distribution Function of Localization Error at distance  $x$  (CDF@ $x$ ).** The percentage of query images where the distances between the estimated location and the ground truth location are within  $x$  m is reported as the second evaluation metric. The percentage of correct answers is calculated in the same way as in equation (5). However,  $C(I_q)$  was set to 1 if the distance from the query images to detected location was within  $x$  m, and 0 otherwise.

**Average number of Questions.** To evaluate the efficiency of the proposed approach, we determined

Table 1: Comparison of proposed and conventional methods.

(a) WCP Dataset			
Method	Direction(s)	Avg. # questions	Accuracy[%]
image search	1	-	73.1
Ours	1	2.75	<b>86.2</b>
GLN (Chiou et al., 2020)	4	-	79.9
Multi-view distance (Li et al., 2021)	4	-	84.0

(b) Mall Dataset		
Method	Avg. # questions	Accuracy[%]
image search	-	27.5
Ours	2.73	<b>51.3</b>

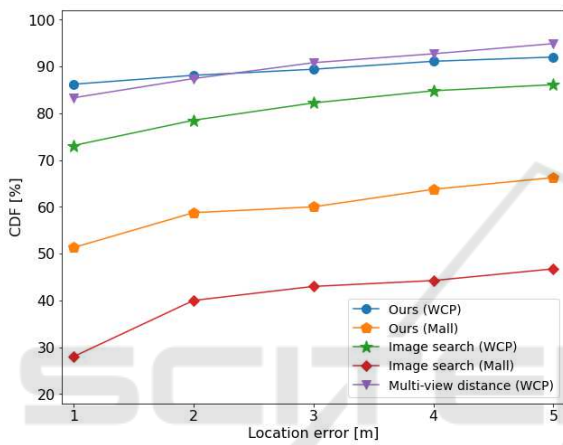


Figure 8: The CDF curves of the localization error of the previous and our approaches.

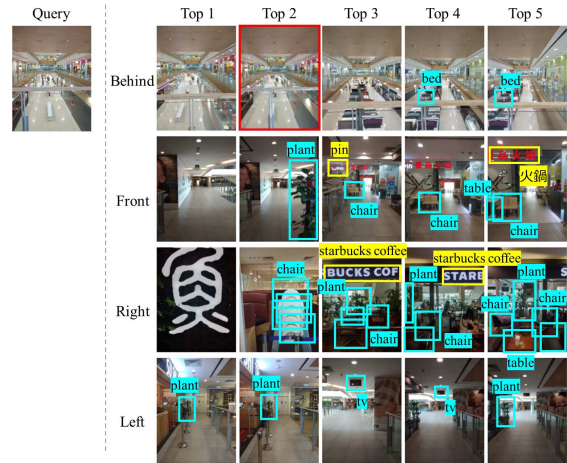
the number of questions posed to the user before there was only one possible location or no more questions were possible.

## 5 RESULTS AND DISCUSSIONS

### 5.1 Accuracy and Average Number of Questions

#### (a) West Coast Plaza Dataset

The results are shown in Table 1(a). The proposed method improved accuracy by 13.1 points compared to the conventional method (similar image retrieval with 1 shooting direction). In other words, the proposed method achieves the same level of accuracy as the conventional method (Li et al., 2021) with 4 directions while requiring the user to answer an average of 2.75 questions. This is 6.3 points better than GLN and 2.2 points better than the multi-view image distance approach, both of demand 4 query images from



(a) The results of possible images, object detection and scene text detection. The correct image is in red. The results of scene text detection are in cyan, and the results of object detection are in yellow.



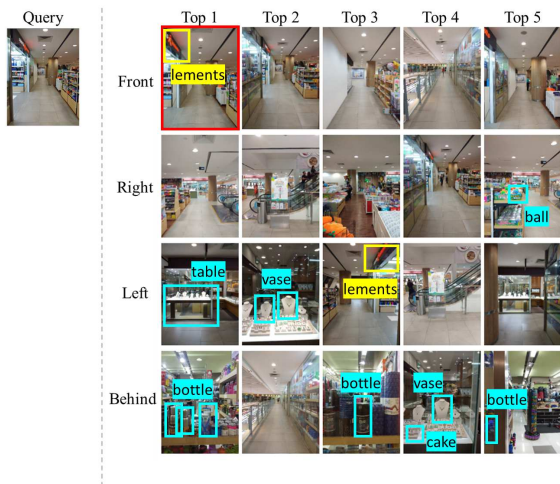
(b) Question and response with the user. The correct location is in red.

Figure 9: Localization success by proposal.

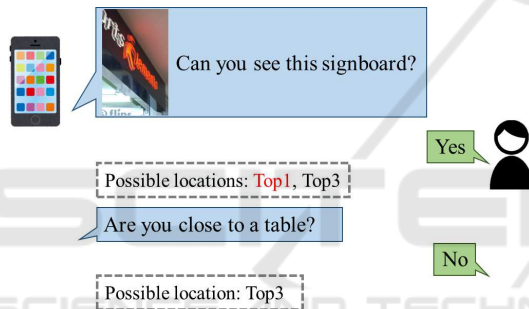
different directions. Therefore, it can be said that proposal offers greater accuracy with far less user effort.

The CDF@x result, shown in Figure 8, confirm that the proposed method achieved higher accuracy than image search even when the location error between the estimated location and the ground truth was increased to 5 m. This fact proves the effectiveness of the proposed method. However, the proposed method achieved lower accuracy than multi-view distance, when the location error was more than 3 m. The reason may be that the generated questions do not take into account the relationship between views.

As shown in Figure 9(a), although this query image failed to get the correct answer at the top position using similarity image search, the correct answer



(a) The results of possible images, object detection and scene text detection. The correct image is in red. The results of scene text detection are in cyan, and the results of object detection are in yellow.



(b) Question and response with the user. The correct location is in red.

Figure 10: Failures of the proposal.

could be attained by generating questions like Figure 9(b) and Q&A with the user.

However, as shown in the failure example (Figure 10(a)), when object detection fails or the correct image is missing, images that were actually correct were deleted during Q&A with the user (Figure 10(b)).

### (b) Mall Dataset

The results are shown in Table 1(b). Due to the change of object by time zone difference and the capture of many images at an oblique direction, this dataset achieved only 27.5% correct answers in similar image retrieval. We improved the accuracy by 23.8 points by posing an average of 2.73 questions to the user.

Table 2: Accuracy vs. number of initial proposed locations. (a) WCP Dataset

initial images	Avg. # questions	Accuracy[%]
3	1.91	84.3
5	2.75	86.2
7	3.35	86.8
10	3.96	87.1

(b) Mall Dataset

initial images	Avg. # questions	Accuracy[%]
3	2.11	50.0
5	2.73	51.3
7	3.23	52.5
10	3.49	56.3

Table 3: Impact of question generation modules.

(a) WCP Dataset

Method	Avg. # questions	Accuracy[%]
object only	3.26	76.9
scene text only	2.31	84.0
scene text + object	2.75	86.2

(b) Mall Dataset

Method	Avg. # questions	Accuracy[%]
object only	2.43	45.0
scene text only	2.08	38.8
scene text + object	2.73	51.3

## 5.2 Accuracy vs. Number of Initial Possible Locations

Table 2 lists the relationship between the number of possible images and accuracy. For both datasets, the accuracy improves with the number of initial images. Thus, it is possible to obtain the correct image by responding to the questions. However, as this increases the burden on the user, it is important to decide on the optimum number of possible images.

## 5.3 Impact of Question Generation Modules

As shown in Table 3, if all object labels are correctly detected, the result is more accurate than the result of similar image search. However, in locations with a lot of restaurants and clothing stores, judgments based on objects alone are not enough. Similarly, the detection of signboard information by scene text direction achieve high accuracy, but still lower than that of multi-view image distance (84.0%). We can see that the combination of object label and scene text information provides the highest accuracy.

Table 4: Comparisons of the results of selecting questions randomly vs. according to conditional entropy

(a) WCP Dataset		
Method	Avg. # questions	Accuracy[%]
random	3.01	81.3
entropy	2.75	86.2
(b) Mall Dataset		
Method	Avg. # questions	Accuracy[%]
random	3.12	41.3
entropy	2.73	51.3

## 5.4 Impact of Conditional Entropy

To show the effect of conditional entropy, we compared the results achieved with randomly selected questions. As shown in Table 4, WCP Dataset and Mall Dataset improved accuracy by 4.9 points and 10.0 points when using conditional entropy rather than randomly selected questions. In selecting questions, we let the questions of scene text information, which is comparatively easy to locate, be more likely to be selected, while questions that have low confidence object labels are less likely to be selected. Therefore, localization can be more accurate. In addition, the average number of questions was also reduced by about 0.4.

## 6 CONCLUSIONS

In this paper, in order to reduce the burden on the user and at the same time achieve highly accurate indoor localization, we proposed a method that generates questions from reference images and filters the possible locations based on responses from the user to questions used by the method. The results of experiments on two datasets showed that even in the case of extremely low accuracy in similar image retrieval, an average of 2.75 responses, without increasing the number of captured query images needed, resulted in higher accuracy than the conventional method.

As a future challenge, methods such as fine-tuning using indoor datasets will be considered to improve the accuracy of object detection. Furthermore, to generate questions that users are comfortable responding to, and questions that consider the difference between views, it is worth checking Visual Question Generation (VQG) as it can be adapted for localization. Last but not least, the problem of the fall in accuracy due to changes in stores or objects because of timezone differences should be resolved.

## REFERENCES

- Bautista, D. and Atienza, R. (2022). Scene Text Recognition with Permuted Autoregressive Sequence Models. In *ECCV*, Cham. Springer International Publishing.
- Chiou, M. J. et al. (2020). Zero-Shot Multi-View Indoor Localization via Graph Location Networks. In *ACM Multimedia*, pages 3431–3440.
- Dong, J. et al. (2019). ViNav: A Vision-Based Indoor Navigation System for Smartphones. *IEEE Trans Mob Comput*, 18(6):1461–1475.
- Gao, R. et al. (2016). Sextant: Towards Ubiquitous Indoor Localization Service by Photo-Taking of the Environment. *IEEE Trans Mob Comput*, 15(2):460–474.
- He, K. et al. (2017). Mask R-CNN. In *ICCV*, pages 2961–2969.
- Li, S. and He, W. (2021). VideoLoc: Video-based Indoor Localization with Text Information. In *INFOCOM*, pages 1–10.
- Li, X. et al. (2021). Accurate Indoor Localization Using Multi-View Image Distance. *IEVC*.
- Liu, C. et al. (2008). SIFT Flow: Dense Correspondence Across Different Scenes. In *ECCV*, pages 28–42. Springer.
- Liu, Z. et al. (2017). Multiview and Multimodal Pervasive Indoor Localization. In *ACM Multimedia*, pages 109–117.
- Liu, Z. et al. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *arXiv preprint arXiv:2103.14030*.
- MarketsandMarkets (2022). *Indoor Location Market by Component (Hardware, Solutions, and Services), Technology (BLE, UWB, Wi-Fi, RFID), Application (Emergency Response Management, Remote Monitoring), Organization Size, Vertical and Region - Global Forecast to 2027*. MarketsandMarkets.
- Noh, H. et al. (2017). Large-Scale Image Retrieval with Attentive Deep Local Features. In *ICCV*, pages 3456–3465.
- Radenović, F. et al. (2018). Fine-Tuning CNN Image Retrieval with No Human Annotation. *TPAMI*, 41(7):1655–1668.
- Sun, X. et al. (2017). A Dataset for Benchmarking Image-Based Localization. In *CVPR*, pages 5641–5649.
- Taira, H. et al. (2018). InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. In *CVPR*, pages 7199–7209.
- Torii, A. et al. (2015). 24/7 Place Recognition by View Synthesis. In *CVPR*, pages 1808–1817.
- Wang, S. et al. (2015). Lost Shopping! Monocular Localization in Large Indoor Spaces. In *ICCV*, pages 2695–2703.
- Zhou, X. et al. (2017). EAST: an Efficient and Accurate Scene Text Detector. In *CVPR*, pages 5551–5560.