Image retrieval using BIM and features from pretrained VGG network for indoor localization

Inhae Ha, Hongjo Kim, Somin Park, Hyoungkwan Kim

School of Civil and Environmental Engineering, Yonsei University, South Korea

ARTICLE INFO

Keywords:
Image-based indoor localization
Cross-domain image retrieval
BIM
Feature extraction
Global descriptor

ABSTRACT

Various devices that are used indoors require information regarding the user’s position and orientation. This information enables the devices to offer the user customized and more relevant information. This study presents a new image-based indoor localization method using building information modeling (BIM) and convolutional neural networks (CNNs). This method constructs a dataset with rendered BIM images and searches the dataset for images most similar to indoor photographs, thereby estimating the indoor position and orientation of the photograph. A pretrained CNN (the VGG network) is used for image feature extraction for the similarity evaluation of two different types of images (BIM rendered and real images). Experiments were performed in real buildings to verify the method, and the matching accuracy is 91.61% for a total of 143 images. The results also confirm that pooling layer 4 in the VGG network is best suited for feature selection.

1. Introduction

Indoor localization is the process of understanding the location and orientation of an object in an indoor environment. Indoor localization methods are mainly used for augmenting visual information [1], indoor navigation [2], and tracking the indoor location of a person [3]. In the architectural, engineering, and construction (AEC) industries, indoor localization technology is used for various purposes, including facility maintenance using augmented reality [4], evacuation route guidance in emergency situations [5], and the tracking of workers or equipment [6–9].

In indoor environments, it is difficult to receive global navigation satellite system (GNSS) signals, which are mainly used for outdoor localization. Therefore, indoor localization methods using various sensors are being studied. Radio-based indoor localization is a method of estimating the position by transmitting and receiving signals. Radio-based indoor localization methods include radio-frequency identification (RFID) [7,10–13], Bluetooth [8,14–16], wireless local area networks (WLANs) [6,17–20], and ultra-wideband (UWB) technology [21–24]. However, these methods require preliminary tasks such as attaching a sensor in advance to the object whose indoor position is to be determined or making a map of the signal strength by location. In addition, radio-based methods are not very suitable for indicating the direction of an object because the position is determined by the signal [25].

Vision-based indoor localization has attracted attention as a cost-effective method [26]. Image-based indoor localization can utilize an image dataset for the prebuilt indoor environment and indoor photographs taken with a camera at the user’s location. As the dataset has the prior information (position and orientation) necessary for localization, a person’s indoor position is determined by extracting the image most similar to the photograph taken indoors from the dataset. The types of images in the dataset are monocular images [2,25,27–31] and omnidirectional images [32–34]. The use of an omnidirectional image is more advantageous for estimating the trajectory and orientation than the use of a monocular image, but it requires more cameras than a monocular image, which increases the cost and makes photographing more difficult.

In order to determine the indoor location information, a map of the indoor environment is required. The map is usually obtained from a three-dimensional (3D) model, which requires a number of indoor images or point clouds obtained from equipment such as laser scanners. Indoor images can be used for the retrieval of similar images in indoor localization and are also used in 3D modeling through structure from motion (SfM) algorithms [28,30,31,35]. These methods require a separate process to take photographs or point clouds to build a model including every location where indoor localization is desired.

Another way to use images for indoor localization is to construct a dataset by rendering a view in the form of a two-dimensional (2D) image from the 3D model of a building and to compare it with a photograph taken in a real indoor environment [27,33,36]. In Refs. [27,33,36], linear features are extracted from images to identify the
locations related to the features. The lines of an image representing an indoor scene are related to the structural characteristics of each location; therefore, they can be useful for determining the location. Although these studies have significantly advanced the body of knowledge in image-based indoor localization, there remains room for improvement. Higher levels of performance may be possible with additional information such as the color or texture of an image. Simpler or lighter equipment settings may provide the user with a more cost-effective solution for indoor localization.

Building information modeling (BIM) is increasingly being used in a wide variety of areas in the construction industry owing to its effectiveness [37–39]. Once a BIM model is developed, it is easy to obtain a desired map from any position; therefore, there have been attempts to utilize BIM in applications that use localization [7,8,24,29,40]. For localization, BIM is mostly used with sensors such as RFID [7], Bluetooth [8], radio-frequency beacons [40], and UWB [24]. Application-wise, BIM is used to provide geometric information [8,40], robot path extraction [24], and the integrated management and visualization of the information provided by the sensors [7]. The study presented in Ref. [29] proposes a method for recognizing the corner of a tile in an indoor floor photograph and identifying the indoor location using the coordinate information of the tile in BIM. On the basis of the utility of BIM, this paper presents an image-based indoor localization method that retrieves rendered BIM images most similar to the photographs taken from the user’s location.

As shown in Fig. 1, the visual characteristics of an image rendered from the 3D model differ from the photograph obtained in the actual indoor environment. It is therefore difficult to perform image retrieval by conducting a cross-domain comparison between images. Previous studies that aim to estimate the indoor location by image retrieval use the photographs acquired in the indoor environment for both the query image and the images in the dataset; scale-invariant feature transform (SIFT) [25,30,34] or speeded-up robust features (SURF) [32] have been used for feature extraction in the same domain for comparison purposes. Fig. 2 shows that SIFT works well for images of the same domain but can have problems if the image types are different.

In this study, a convolutional neural network (CNN) is used to extract features for cross-domain image retrieval. CNNs are deep learning networks suitable for image processing. Unlike existing handcrafted feature extraction methods, CNNs have exhibited excellent performance in various fields of computer vision, such as classification, segmentation [41], and image-based localization [42,43], using the ability to learn visual features by itself. CNNs trained for image classification have proven to be a powerful tool for extracting the generic features from an image [44]. In fact, the intermediate output of each layer can be used as a feature map in the image retrieval process [45,46]. Thus, the proposed method utilizes a CNN trained for classification. The CNN is pretrained by ImageNet [47], a large labeled image dataset for feature extraction, and the extracted features are used for image retrieval.

2. Methodology

Fig. 3 shows the proposed method of image-based indoor localization. The core idea is that images are processed through the CNN to produce their multiscale features. The features from the rendered BIM images and the indoor photograph are vectorized so that they can be compared with each other. Once the image most similar to the user’s photograph is retrieved from the dataset, the user’s indoor position is estimated.

CNNs consist of multiple layers of neurons that learn to extract the meaningful features of an image to perform a desired task. The types of layers that make up a CNN are typically divided into convolutional,
Fig. 2. Examples of feature matching with SIFT for two pairs of images: (a) BIM images in the same domain and (b) a BIM image and an indoor photograph at the same position.

Fig. 3. Proposed image-based indoor localization method.
pooling, and fully connected (FC) layers. The network receives the 2D raw image as the initial input, and each layer receives the output from the previous layer as its input. Each convolutional layer has a predetermined filter size; the depth of the output is equal to the number of filters for the layer. When performing the convolution operations, each filter slides over the input by the amount of stride. Because image data tend to have a higher correlation in a local area rather than in a global area, the convolutional computation can extract the prominent characteristics of the receptive field, which is the portion of the input to which the filter is applied [48]. Therefore, the output from a convolutional layer can be thought of as a feature extracted from each position of the input image and is also called a feature map [49].

Once the features of each part of the image are found in the convolutional layer, the pooling layer merges the semantically similar features, and the performance of the network becomes invariant to image distortion and movement [48]. The pooling layer also has the effect of reducing the dimensions of the feature map; the area of the input image represented by the unit area of the feature map becomes larger than before. When the resultant feature map is input to the next convolutional layer, a new feature map is extracted. Although the new feature map has a lower dimension, it represents the features of the entire input image. Therefore, as the operation moves downstream in the network, more global features are extracted.

The proposed method uses a feature map that passes through each layer of a pretrained CNN as a feature for cross-domain image retrieval. A CNN trained for object classification with the ImageNet dataset (a large image dataset) [47] is known to perform well for the image retrieval of other datasets [44]. Although the CNN has been trained for classification based on a different dataset, the pretrained CNN now has filters that can extract the key features from any image; it can extract meaningful features even when applied to other datasets. Therefore, a pretrained CNN, which shows excellent performance in extracting generic features, was used to match the indoor photographs with the corresponding BIM images.

One CNN type, the VGG network [50], was used in the proposed method; the VGG16 (Fig. 4) and VGG19 networks, which exhibited the highest performance in Ref. [50], were used. In both networks, multiple convolutional layers and max pooling are repeated five times, and there are three FC layers at the end of the network. The result of an FC layer is not used as a feature in the proposed method because the features of the FC layer do not have positional information. The five feature maps after each instance of max pooling are used as the features for image retrieval.

The dataset for image retrieval consists of BIM images rendered similar to a pedestrian’s viewpoint at the designated indoor locations. The BIM model can produce a pedestrian’s perspective in the form of a 2D image inside the building. Ten view angles were predetermined for each indoor location to represent the diverse views of a pedestrian. However, when a certain view angle produced an image with no distinctive features, the view was not adopted as part of the dataset. For example, if an image of a plain wall was shown from the view angle, the view image was dropped from the dataset. In addition, the pedestrian’s view in the BIM model was rendered in accordance with the view angle range of the camera to be operated by the user.

As previously mentioned, the BIM images and indoor photographs are processed by the VGG network to be transformed into features. It was assumed that the images of the dataset were transformed into features before indoor localization occurred. That is, when the user of a camera wishes to know his/her location, only the real photograph needs to be processed by the VGG network for feature extraction. Because the VGG network receives images with a resolution of \(224 \times 224\) as an input, the BIM images and photographs were resized to \(224 \times 224\) pixels. Then, image retrieval was performed by evaluating the similarity between the feature map of the indoor photograph, which is the query image, and the feature map of the BIM dataset images. The similarity was evaluated through the cosine distance with vectorized feature maps. The cosine distance is the cosine value of the angle between two vectors, which can be used to evaluate the similarity between vectors and has the advantage of being naturally normalized by the norm [51].

3. Experiments

Experiments were conducted in two indoor environments with different structures. The first experimental site was an indoor corridor with offices on both sides. In the experiment, the most suitable feature extraction layer of the VGG network was determined, and the performance of the proposed model was measured using four datasets with different numbers of test images. The second experiment was carried out in a building hall, which is larger and wider than the corridor. This experiment verified the applicability of the method in indoor spaces. BIM models for the experiments were produced on the basis of the design documents of the buildings and have a level of development (LOD) of 200–300 according to the American Institute of Architects (AIA) guide.

3.1. Experiment in the corridor

The first experiment was conducted on one floor (the fifth floor) of the north wing of the first engineering building of Yonsei University, Seoul, Korea. The total area of the fifth floor is approximately 517m². Fig. 5 shows the experimental site in two formats: a real photograph and BIM model. The locations for rendering the view for dataset construction were arbitrarily selected from the experimental site (north wing) and are shown in Fig. 6. The views at each point were taken using various orientations of the user when moving in the building. In addition, the indoor photographs, which are the query images for image retrieval, were taken indoors at the same positions and orientations as the BIM model renderings.

According to the degree of difficulty of the dataset, the dataset was classified into four levels: Levels 1-1 and 1-2 and Levels 2-1 and 2-2. Level 1-1 has a total of 54 images rendered at 9 locations; a relatively smaller amount of overlap exists in the dataset of Level 1-1. However, Level 1-2 has a total of 86 images with the addition of 32 images; the additional images have higher levels of view overlap. Likewise, Level 2-2 has 58 more images compared to Level 2-1; the dataset of Level 2-2 has higher levels of view overlap than Level 2-1 (see Table 1). The difference between Level 1 and Level 2 is the number of locations for

![Fig. 4. Architecture of the VGG16 network.](image-url)
image acquisition, as shown in Fig. 6; 14 locations were needed for Level 2, as compared with 9 locations for Level 1. Fig. 7 shows the different degrees of view overlap.

The features of the indoor photographs and dataset images are obtained from the outputs of the pooling layers as each image or photograph passes through the pretrained VGG network. The VGG network has five pooling layers; therefore, a total of five features are extracted from one image or photograph. In order to retrieve the image most similar to the query image in the dataset, the similarity between images is evaluated as the cosine distance between features. The similarity evaluation is performed among the features obtained from the pooling layer of the same order. That is, a feature of the query image extracted from the nth pooling layer is compared with the features of the dataset images extracted from the nth pooling layer. When a BIM image extracted at the same position and orientation as the query image is retrieved, the proposed method has correctly performed indoor location estimation.

Table 2 presents the results of image retrieval based on the features obtained from the different pooling layers for the four datasets. When the feature maps obtained from pooling layer 4 were used, the most accurate image matching results were obtained for all datasets. Moreover, the results when using features obtained from the fourth pooling layer were not significantly affected by the difficulty of the dataset, and all four datasets showed a high accuracy of more than 90%. Although the experimental results of VGG16 and VGG19 networks showed slight differences for each dataset, the overall results show similar trends and values. With the outputs obtained from pooling layer 4, the times to extract features of an indoor photograph and retrieve a single picture from the Level 2-2 dataset were 0.0411 and 0.0455 s on average for the VGG16 and VGG19 networks, respectively.

3.2. Experiment in the hall

Since the method was verified in the first experiment, the applicability of the method to other types of indoor spaces was confirmed in the second experiment. The method of building a BIM image dataset, taking a query image, and evaluating the similarity between images for image retrieval was the same as that of the first experiment. The second experiment was conducted in the hall in the first basement level of the third engineering building of Yonsei University, Seoul, Korea. The total area of the hall is approximately 330 m². As shown in Fig. 8, 14 points were designated as the locations to render the views, and images were acquired at about eight orientations per location. A total of 102 BIM images were prepared in the dataset for this experiment, and the same number of indoor photographs was taken at the designated positions. On the basis of the results of the first experiment, the features of an image extracted from pooling layer 4, which had best performance for image retrieval, were used in the second experiment. For the 102 hall images, the matching accuracy was 76.47% for both the VGG16 and VGG19 networks, which was lower than the performance for the corridor.
4. Discussion

The proposed method of vision-based indoor localization was verified on the basis of its high performance for image retrieval. Each image in the dataset was stored along with the extracted indoor location and orientation information. Accordingly, when the indoor photograph was correctly matched with the BIM image obtained at the same point and orientation, the indoor location information of the user who has acquired the photograph can be confirmed. If a BIM model is created and used from the design stage, the method does not require a separate 3D model for indoor localization. In addition, unlike existing vision-based localization methods, it is not accompanied by the work needed to construct a dataset by photographing various places in advance.

This study shows which layer of the VGG network is most suitable for matching the indoor image and the rendered BIM image; when extracting features from pooling layer 4 and performing image retrieval, better results were obtained than those from other layers. Fig. 9 illustrates how the BIM images and photographs are converted to feature maps by different pooling layers of the VGG16 network. For example, Fig. 9a shows a photograph and BIM image at location 5 of Fig. 6 and the five feature maps extracted from pooling layers 1–5. As the layer becomes deeper, the VGG network is trained to extract features at increasingly larger scales of the image. That is, the network starts extracting the local descriptors and gradually extracts the global descriptors. In the feature maps visualized in Fig. 9a, b, and c, the results from the later pooling layers (pooling layers 3, 4, and 5) show the structural characteristics rather than the detailed characteristics in the image. Evaluating the results from the front layers (pooling layers 1 and 2), the color arrangements show detailed features such as the grid pattern of the tiles, the nameplates, and illumination.

In the proposed method, the overall structure of the images rather than detailed features proved most effective in the comparison because image retrieval was performed on the basis of two different types of images. In fact, the experimental results confirm that the features from pooling layer 4, which essentially extracts global descriptors, is the most suitable for cross-domain indoor image matching. As can be seen in Fig. 9, the feature maps of the two matching images are most similar in color arrangement when obtained from pooling layer 4. The feature map obtained from pooling layer 5 extracts the features of a larger area than that obtained from pooling layer 4. However, as the accuracy of image retrieval based on pooling layer 5 is rather low, it is assumed that too much approximation of image information has a negative effect.

The proposed method exhibited robust performance, irrespective of the number of images or the different degrees of view overlap. As in the cases of Level 1-2 and Level 2-2, images of similar views extracted with slight angle differences were discriminated well. These results confirmed that the proposed method should be able to achieve superior performance even in places larger than the place where the experiment was conducted in this study.

The time performance of the proposed method is largely determined by two factors. The first is the time that it takes to obtain the output of a

<table>
<thead>
<tr>
<th>Dataset (number of images)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pooling layer 1</td>
</tr>
<tr>
<td></td>
<td>VGG 16</td>
</tr>
<tr>
<td>Level 1-1 (54)</td>
<td>64.81</td>
</tr>
<tr>
<td>Level 1-2 (86)</td>
<td>54.65</td>
</tr>
<tr>
<td>Level 2-1 (85)</td>
<td>55.29</td>
</tr>
<tr>
<td>Level 2-2 (143)</td>
<td>45.45</td>
</tr>
</tbody>
</table>
Fig. 8. Floor plan of the first basement level and the points of data acquisition.

Fig. 9. Paired input images and the visualization of the feature maps: (a) location #5, (b) location #11, and (c) location #14 in Fig. 6.
specific layer by passing an indoor photograph through the pretrained CNN. The second is the time required to evaluate the similarity between the features of the photograph and the features of the BIM image dataset prepared in advance. The first factor is a fixed time, but the second factor increases in proportion to the number of images in the dataset. This means that as the number of dataset images increases, the speed of the method decreases. However, in the experiment, the average time required to evaluate the similarity between a single photograph and a single BIM image was only 0.000143 s. This means that the proposed method is sufficiently fast to handle a large-scale dataset.

The second experiment, which was conducted to evaluate whether the proposed method is applicable in various indoor places, showed lower results compared to the results in the corridor. Excluding the area of the offices in the north wing, the area of the hall was substantially larger than the area of the corridor where the first experiment was conducted. The hall structure was also monotonous and symmetrical owing to its architectural characteristics; compared with the corridor, the structural features of the hall were not clear, and the area represented by the same-size image was widened. Therefore, even if the images were acquired at different positions, the structural differences in the views tended to be relatively small. As a result, the proposed method achieved an accuracy of 76.47% for image retrieval, which is lower than the accuracy for the corridor (91.61%).

Fig. 10 shows examples where the method failed to retrieve the BIM image that matched the query image. Fig. 10a shows a case where the degree of similarity between the indoor photograph and the matching BIM image is not high owing to a lack of completeness in the BIM model; if the indoor image is visually compared with the correct answer, which is the query image, the door on the left wall of the indoor photograph does not exist in the BIM image. As only visual information is used for localization, a high-quality BIM model similar to the actual indoor environment is essential. Fig. 10b presents a challenging example in which an incorrect BIM image is selected as the image with the highest degree of similarity to the query image. However, when looking at the query image and incorrect image, there is a high degree of linear similarity in the overall structure, although the details of the interior configuration are different. This example demonstrates a limitation of the proposed method; one photograph may not be sufficient to accurately identify the location of interest. Multiple photographs may be required to provide more robust and reliable performance for indoor localization.

5. Conclusions

This paper proposes a new vision-based indoor localization method for identifying the location and orientation of mobile device users in an indoor environment. In the proposed method, a dataset is constructed with the images rendered from a BIM model at each point similar to the user’s field of view. An image that was most similar to the query indoor photograph was retrieved from the dataset, and the indoor position was estimated on the basis of the location and orientation at which the BIM image was rendered. The features of images extracted from a pretrained VGG network were used to evaluate the similarity between cross-domain images. Experiments have confirmed that the features extracted from a deep neural network (the VGG network) are suitable for indoor localization.

The specific contributions of this study to the body of knowledge are as follows. First, the proposed method can be more efficient in terms of time and cost than existing image-based indoor localization methods because a BIM model from the design stage is used to construct the dataset for localization. Moreover, only monocular images are used for location confirmation, and other sensing information is not required. Second, the proposed method uses the VGG network, a type of CNN, for feature extraction and demonstrates which pooling layer is most suitable for feature extraction. The fourth pooling layer extracts the best global descriptor from images for image retrieval. The effectiveness of this method has been verified by a high matching accuracy of 91.61% when an indoor photograph is supplied as a query image in a total of 143 paired indoor photographs and BIM images. Finally, this method utilizes a pretrained VGG network trained with other image datasets that are less relevant to this application area. In other words, the network is equipped with the ability to extract the generic features of images regardless of image type. This strength implies that the proposed method would show robust performance in other applications of indoor localization without requiring a separate training process.

In order for the proposed method to be widely applied in practice, there are some technical problems that should be addressed. First, if there is a large object such as a picture frame or poster that did not exist in the BIM map, it may negatively affect the accuracy of the method for estimating the position with visual similarity. Second, because images representing similar indoor structures may necessarily exist, there may be a limit to finding a user’s position using only one query image in the entire dataset. These problems naturally lead to future research topics. Future studies such as performing localization through multiple images observed at the same point or performing current localization using existing localization information can greatly improve the proposed method. The improved method is expected to advance the traditional maintenance processes of various facilities requiring knowledge of the user’s indoor location.
Acknowledgments

This work was supported by a grant from the National Research Foundation of Korea funded by the Korean government (MSIP) (No. 2018R1A2B208600).

References


