Demo: Image-based Indoor Localization using Object Detection and LSTM

Yuki Aoki
Keio Univerisity
t23011ya@sfc.keio.ac.jp

Naoki Kobayashi
Keio Univerisity
nk@keio.jp

Tadashi Okoshi
Keio Univerisity
slash@sfc.keio.ac.jp

Jin Nakazawa
Keio Univerisity
jin@sfc.keio.ac.jp

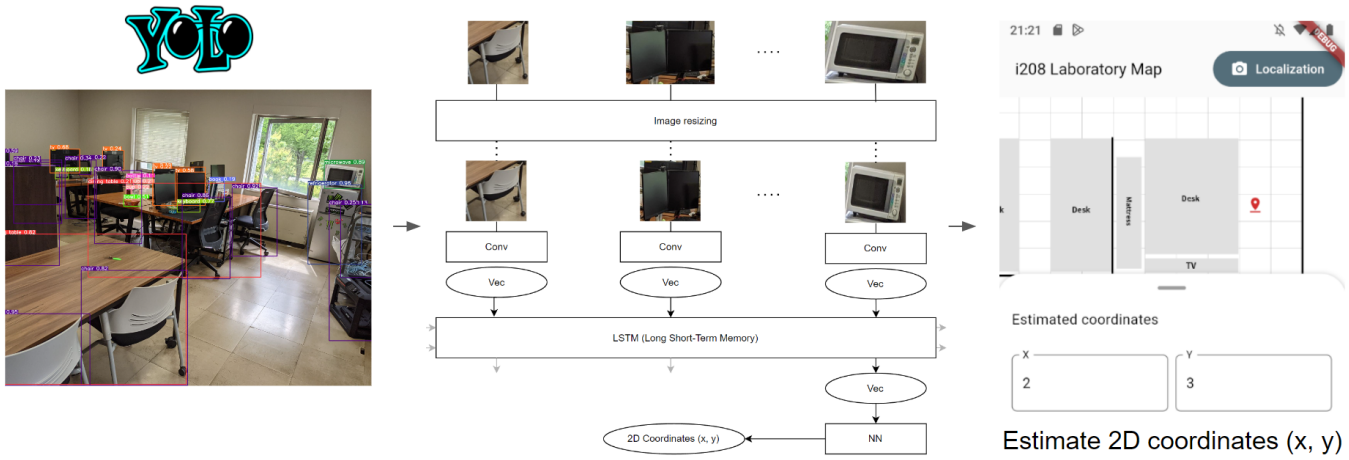

Figure 1: System Overview

ABSTRACT

In this work, we propose a novel model that focuses on object features by combining object detection with CNN and LSTM networks. In recent years, a multitude of deep learning-based methods for Visual Localization, have been extensively researched. However, conventional methods do not adequately account for object-level features. Therefore, it is difficult to use indoors where similar objects appear frequently. Our method applies CNN for feature extraction on detected objects cropped by YOLOv8, an object detection algorithm, and then integrates these features into a single feature vector using LSTM, enabling location estimation that takes into account multiple object features.

Experiments using the new indoor dataset of our laboratory room have revealed that our proposed method achieves a 19.0% higher accuracy compared to CNN models that input the whole image with the same number of layers. These results demonstrate the promising potential of exploring methods focused on object features for indoor localization. The codes are available at https://github.com/sakusaku3939/YoloLSTM.

KEYWORDS

Visual Localization, Image-based Localization, Indoor localization, Relocalization, Absolute Pose Regression

ACM Reference Format:
Yuki Aoki, Naoki Kobayashi, Tadashi Okoshi, and Jin Nakazawa. 2024. Demo: Image-based Indoor Localization using Object Detection and LSTM. In *The 22nd Annual International Conference on Mobile Systems, Applications and Services (MOBISYS '24), June 3–7, 2024, Minato-ku, Tokyo, Japan.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3643832.3661836

1 INTRODUCTION

In recent years, as smartphones have become more prevalent, navigation systems based on GPS have been widely used. However, GPS is particularly susceptible to interference from buildings and walls indoors, leading to inaccuracies in positioning.

In known spaces with existing 3D spatial maps or datasets, there are two main approaches: a two-step method that extracts image feature points and solves the Perspective-n-Point (PnP) problem, and direct methods using deep learning. The method for solving the PnP problem first extracts feature points using SIFT or other methods and then performs 2D-3D matching between the query image and the images in the database. Then, they estimate the

MOBISYS '24, June 3–7, 2024, Minato-ku, Tokyo, Japan

ACM ISBN 979-8-4007-0581-6/24/06
https://doi.org/10.1145/3643832.3661836

position by solving the PnP problem through optimization with RANSAC. In 2017, differentiable RANSAC was proposed, leading to the introduction of direct methods like DSAC and DSAC++[2], which perform feature extraction to 2D-3D matching directly.

Since 2015, numerous methods employing deep learning for end-to-end position estimation have been researched. These began with PoseNet[4], based on GoogLeNet, and have expanded to include methods like PoseLSTM[5], which combines GoogLeNet with LSTM. From 2018, many methods using Attention and Transformer models have also been proposed.

On the other hand, end-to-end methods using deep learning have been criticized for lower accuracy and scalability compared to methods solving the PnP problem. However, end-to-end estimation methods have the advantage of being significantly faster. For instance, DSAC++[2] which uses PnP, takes over 100ms per image, while PoseNet[4] can infer in less than 10ms per image[1]. Additionally, CNN relies solely on RGB images and does not depend on depth maps or SfM reconstructions, making them less constrained by hardware. Given these factors, end-to-end methods using deep learning hold promise for applications on end devices, such as smartphones.

2 METHOD

It is presumed that related works have difficulty achieving high accuracy for scenes with many similar objects, such as desks and chairs, in indoor location estimation. Moreover, convolution from the whole image focuses only on the few features involved in position determination. This makes it unstable to the movement of objects with important features.

Therefore, this work proposes a novel approach of image-based location estimation that focuses on the features and quantity of objects in the images. Figure 1 shows the system overview. Specifically, the YOLOv8[3] object detection model is used to crop objects in the image. Next, a CNN is used to extract features from the all cropped objects. LSTM is then used to convert these object features into a single feature vector.

By inputting the features of individual objects into LSTM instead of the whole image, it is possible to capture the characteristics of multiple objects as a single, comprehensive feature. Moreover, performing convolution on each object allows for the anticipation of higher accuracy in scenes where the types and numbers of objects are subject to significant changes.

3 EVALUATION

To assess the effectiveness of the proposed indoor method, we evaluated the accuracy of location estimation using a self-made laboratory dataset measuring 5.85m x 5.4m, shown in Figure 2. The dataset consists of 6,500 training datasets, segmented each 100 photo frames from video data for 65 locations where people can move. We divided the room into grids with the 0.45m x 0.45m floor tiles and determined the location label of each image. For the test datasets, we used 325 photos that were not included in the training. The test datasets comprised 325 photos that were not included in the training. To compare the performance of the models, we measured the average error (m) of a model of CNN with the same number of layers applied to the whole image, and the proposed method. The results are presented in Table 1.

Using the laboratory dataset, the proposed method achieved an average error of 0.929m on the test datasets. The proposed method's model achieved a 19.0% lower error than the model that applies the same CNN to the whole image.

Table 1: Average error comparison with CNN with the same number of layers in laboratory dataset

	CNN	**Proposed model**
Average errors [m]	1.147m	**0.929m**

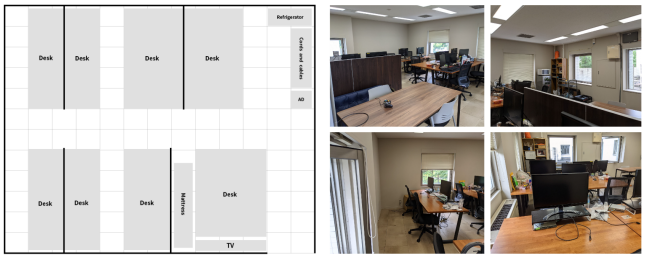

Figure 2: Map of laboratory dataset

4 DEMO

We will demonstrate real-time location estimation using the laboratory dataset on a mobile device. First, you will be asked to randomly select a photo of the our laboratory from an album on the application. Then, the server will return the results of the location estimation using the proposed method in real-time and display them as pin icons on the screen.

The front-end of the application is developed using Flutter, while the back-end is developed using Python and the Flask framework.

5 ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Number JP24K02935.

REFERENCES

[1] Hunter Blanton, Connor Greenwell, Scott Workman, and Nathan Jacobs. 2020. Extending Absolute Pose Regression to Multiple Scenes. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 170–178. https://doi.org/10.1109/CVPRW50498.2020.00027

[2] E. Brachmann and C. Rother. 2018. Learning Less is More - 6D Camera Localization via 3D Surface Regression. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 4654–4662. https://doi.org/10.1109/CVPR.2018.00489

[3] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. *Ultralytics YOLOv8*. https://github.com/ultralytics/ultralytics

[4] Alex Kendall, Matthew Grimes, and Roberto Cipolla. 2015. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 2938–2946. https://doi.org/10.1109/ICCV.2015.336

[5] F. Walch, C. Hazirbas, L. Leal-Taixé, T. Sattler, S. Hilsenbeck, and D. Cremers. 2017. Image-Based Localization Using LSTMs for Structured Feature Correlation. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 627–637. https://doi.org/10.1109/ICCV.2017.75