Yolov7 기반의 공간 인식 시스템

A Yolov7-based Spatial Recognition System

Haichuan Chen
Dept. of AI Convergence Network, Ajou University
Suwon, Republic of Korea
c123156919@ajou.ac.kr

Gaoyang Shan
Dept. of Software and Computer Engineering, Ajou University
Suwon, Republic of Korea
shanyang166@ajou.ac.kr

Byeong-hee Roh
Dept. of AI Convergence Network, Ajou University
Suwon, Republic of Korea
bhroh@ajou.ac.kr

Abstract

Abstract—Computer vision has rapidly evolved into a critical field that has garnered significant attention due to its applications i recognition, human body analysis, automatic driving, indoor positioning, and other domains. The accuracy and speed of object det have become a primary focus in computer vision research. Among the notable architectures, YOLO stands out, as it delivers rema speed that is 300 times faster than Fast-RCNN while maintaining comparable accuracy. In this paper, we proposed the topic of recognition using the YOLO architecture. Specifically, we propose a solution that utilizes indoor video footage to identify objects in extract their spatial information, and store them in a database for matching and identifying spaces. We also introduce a new fingerprint method that leverages monocular vision and YOLO algorithm to assist users in determining their location and space. Our study pr valuable insights and directions for future spatial recognition research.

Keywords: Computer vision, YOLO , spatial recognition , Fast-RNN, Artificial Intelligence

1. Introduction

The Spatial Recognition System is a technological innovation that helps users orient themselves in various spaces. In today's era, with the continuous construction of urban buildings and subway stations, spatial recognition is becoming more and more important and convenient for citizens. There are various positioning methods available, which can be roughly divided into two types. The first is sensor-based positioning, which involves the use of various sensors, such as infrared positioning, ultrasonic positioning, Bluetooth positioning[1], RFID positioning, Wi-Fi positioning to determine the user's current location. These methods have been found to be highly accurate,but their high cost and specialized equipment requirements make their use in shopping centers and civil buildings impractical.The second is vision-based positioning[2], which has lower equipment requirements, lower cost, simpler installation process, and is more suitable for fire scenes where civil buildings and field equipment may be damaged.

In this paper, an algorithm for intelligent fingerprint entry is proposed. By reducing the noise of the model output results and extracting completes spatial information, the data volume of the digital map is reduced, and the judgment speed and accuracy are higher.

This paper is structured as follows. In the part Yolov7-based approach, a method of entering fingerprints based on yolov7 is proposed and a Knn-based template matching is provided. In order to verify the feasibility of the algorithm in this paper, some tests are set up in the experiment setup part to verify the performance of the method, and the final experiment result part fully reflects the performance of the algorithm proposed in this paper.

2. Related works

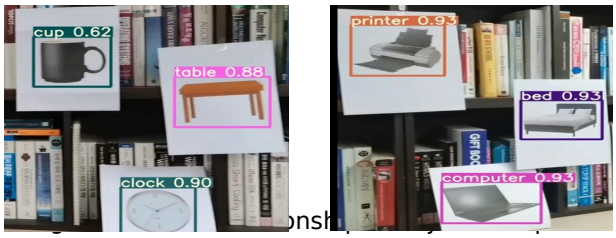
Jing, Li et al.[3] utilized SLAM to generate 3D models of buildings for user localization. In Bin, Yu et al.[5], a visual SLAM system based on ORB-SLAM2 is proposed, which reduces the impact of dynamic objects on pose estimation accuracy. L. Guanqi et al.[4] recorded the spatial relationship of objects in the photo, and recorded the position information of the current location.

3. Yolov7-based approach

This chapter is divided into object detection, noise removal, data sorting, building digital maps, and spatial recognition. It briefly introduces what algorithm is used for target detection and how to collect data and fine-tune it. In the collected data, it needs to be sorted first, what noise is included in the sorted data, and how to remove the noise. In the last part, how to use the processed data to convert it into a digital map, and what algorithm to use to match the space and judge the space where the user is located.

3.1. Object detection

In the first step of space recognition, we first need to recognize objects. In object recognition, we use the yolov7 algorithm. At this stage, we need to collect pictures of objects in the laboratory and label the collected data sets. We inject the marked database into the yolov7 network, and finally get a fine-tuned yolov7 network.



3.2. Noise removal and data sorting

When using yolov7, the output results are out of order, so we need to sort the output results from left to right and from top to bottom. When the coordinates of the objects are all the same, we sort by the first letter of the name. When the first letter

is the same, we sort by the second letter, and so on.

After obtaining these data, we need to process these data to facilitate our next step of denoising. At this point, we need to classify the currently owned data, so that the data of each object is arranged according to the object category and time order, that is, the label of the object, and the coordinates of the object are classified according to the order of the frame.

The fine-tuned yolov7 algorithm can detect objects in space more accurately, but the yolov7 target recognition algorithm still has invalid detection. There are two main types of invalid detection: one is recognition errors of different types, and the other is recognition errors of the same category. For different types of recognition errors, we will directly delete them. For the same type of recognition errors, we will set a threshold. When the number of frames does not reach this threshold, we will judge it as an invalid recognition.

3.3 Build a digital map

After we have collected the data, we need to convert the collected data into the spatial relationship of items in the real world. After we have collected the sorted data, we need to find out the number of frames where each two objects appear in the database at the same time, and then calculate the average coordinates of each object in the number of frames that appear at the same time. Then use the first object as a benchmark to standardize the following objects, then use the second object to standardize the third object, and so on to standardize all objects. As shown in Figure 3, we can get the fingerprint of an area. In each area, we collect fingerprints in the same way. A space is composed of multiple areas as shown in Figure 4. Therefore, a digital map is composed of multiple fingerprints, and we can Get a digital map of this space.

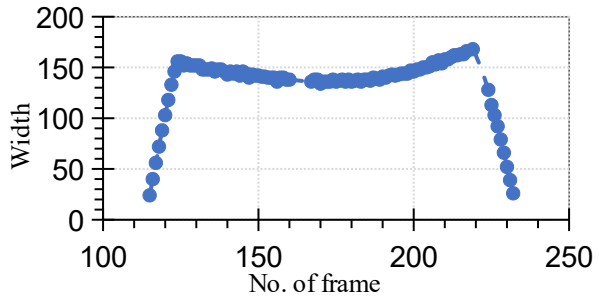


Fig.2. Width dimension of the object from appearing to disappearing (Object fully appears in frames 112-233)

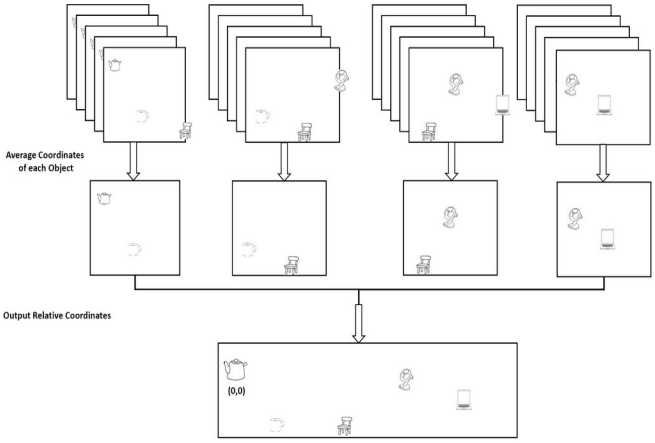


Fig.3. Make Fingerprints(Find the number of frames where each object appears completely and find its average coordinates)

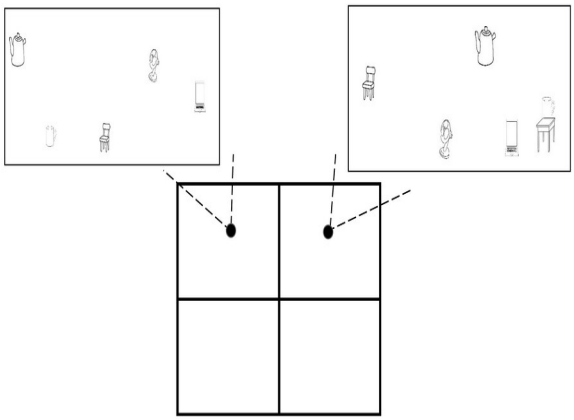


Fig.4. Construction of Digital Map

3.4 Spatial recognition

After the digital map is made in the offline stage, the user uploads the video taken to the server in the online stage, and uses the scheme we proposed to make the video taken by the user into a fingerprint. The digital map of each space in the digital map is divided into several areas in each space. We template-matched the fingerprint with each part of each space in the digital map, and used the Knn algorithm to find the area with the highest similarity.

4. Experiment setup

In order to verify the algorithm, we conducted an experiment in the Paldal Hall of Ajou University. For the convenience of the experiment, we printed different objects and took 577 pictures with a smartphone camera to provide model training. Models trained with these datasets by our algorithm can detect 8 different objects in videos and pictures. Table 1 shows the objects and their corresponding labels.

In addition, we simulated two spaces, divided each space into 4 areas, placed 3-4 objects in each area, collected 8 videos in 360° shooting in each area, and took 25 pictures for testing, each pictures including 2-4 objects.

5. Experiment result

Table 1. Model performance comparison

Algorithm	Prob.of success
Proposed	84%
Former[4]	66.6%

We used the same amount of data, labels, and parameters to fine-tune the two models. There is a big gap in the success rate of the two models. From the above table, we can see that the recognition success rate of yolov7 is 39.4% higher than that of Faster-Rcnn. Yolov7 is only 3% on the recognition omission, while Faster-Rcnn is as high as 53.3%.

Table 2. Solution types and rates

Model	Prob.of success	Prob.of lost
Yolov7	86%	3%
Faster-RCNN	46.6%	53.3%

From the table1, we used a yolov7 for the two algorithms, and did not use the Faster-Rcnn in [4]. From the table, we can find that in the same recognition framework, the success rate of our proposed method is 17.4% higher than that of [4].

6. Conclusions

In this paper, we propose an algorithm based on the object recognition space, which is designed to take into account the noise effect of the entered fingerprint. Experimental results show that compared with other methods, this method improves the correctness and efficiency of fingerprint entry.

Acknowledgement

“This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2023-2018-0-01431) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)”

References

[1] G. Shan, B.-h. Park, S.-h. Nam, B. Kim, B.-h. Roh, and Y.-B. Ko, “A 3-dimensional triangulation scheme to improve the accuracy of indoor localization for iot services,” in 2015 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM) , 2015, pp. 359–363

[2] D.-h.Yoo,G. Shan, and B.-h. Roh, “A vision-based indoor positioning systems utilizing computer aided design drawing,” in Proceedings of the 28th Annual International Conference on Mobile Computing And Networking , 2022, pp. 880–882.

[3] J.Li,C.Wang,X.Kang,and Q.Zhao,“Camera localization for augmented reality and indoor positioning: a vision-based 3d feature database approach,” International journal of digital earth, 2019.

[4] L.Guanqi,X.Haowei,and L. Chenning,“An improved indoor navigation method based on monocular vision measuring and region based convolutional network,” in 2019 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC).IEEE, 2019, pp. 1–6

[5]余,斌, and 王晨捷, “基于空洞全的 slam 方法,”信息控制, vol. 51, no. 3, pp. 330–338, 20