시각적 랜드마크 지도기반 공간인식을 위한 YOLOv9 성능 평가

여요쉰, 첸하이첸, 산가오양, 임정현, 노병희*

*아주대학교, 상하이하당지능과기유한공사

{yxyao, shanyang166, wjdguszoqt, *bhroh} @ ajou.ac.kr, 123156919@qq.com

Performance Evaluation of YOLOv9 for Visual Landmark Map-Based Spatial Recognition

Youxun Yao, Haichuan Chen, Gaoyang Shan, Junghyun Lim, Byeong-hee Roh*

*Ajou univ., Shanghai Hetang Intelligent Technology

Abstract

This paper evaluates the performance of YOLOv9 for visual landmark map-based spatial recognition. We compare YOLOv9's accuracy, inference time, and memory usage against YOLOv7. Our results show that YOLOv9 provides significant improvements in accuracy and efficiency, making it a viable option for real-time indoor spatial recognition.

Ⅰ. Introduction

Visual landmark map-based spatial recognition is crucial for indoor localization and navigation. Object detection models like YOLOv7 have been used extensively for this purpose. This paper explores the performance of the latest YOLOv9 model in this context, aiming to provide insights into its efficacy and potential benefits.[1]

YOLOv7, the previous state-of-the-art model in the YOLO series, has demonstrated impressive capabilities in various applications, including visual landmark map-based spatial recognition. However, the recent development of YOLOv9 introduces several enhancements that promise to further improve performance. YOLOv9 incorporates the Programmable Gradient Information (PGI) mechanism and the Generalized Efficient Layer Aggregation Network (GELAN) architecture, which address issues related to information bottlenecks and parameter utilization in deep neural networks.

The primary objective of this paper is to evaluate the performance of YOLOv9 in the context of visual landmark map-based spatial recognition. We compare YOLOv9's accuracy, inference time, and memory usage against YOLOv7. Our custom dataset comprises annotated images of indoor environments, captured using a monocular camera on a mobile phone. We aim to provide insights into the efficacy of YOLOv9 and its potential benefits for real-time indoor localization applications.

Ⅱ. Method

We used a custom dataset comprising annotated images of indoor environments with various objects. The custom dataset is collected using a monocular camera on a mobile phone with a pixel number of 48 million, recorded with 1920*1080 resolution in 60 frames per second. Then the videos are split into images every 10 frames. After that, every single image is manually labeled by drawing the boxes and naming the labels. We selected 11 different labels for this experiment, since too many labels can slow down the training process, leading to delayed convergence; while too few labels can also affect the model's ability to localize the objects.

The location where the dataset was collected is a hallway of a building, and the objects contain doors, door plates, trash cans, etc. The selection of the objects are considered adequate to be a visual landmark, and is common in everyday practice.

YOLOv9 was trained on the custom dataset using the following hyperparameters: epochs: 600; batch-size: 8; imgsz: 640. The training was conducted on an NVIDIA RTX3080 for 600 epochs. The YOLOv9 model was integrated into the existing spatial recognition system. The original code was implemented for YOLOv7, and to achieve better performance, we modified the code to handle preprocessing, inference, and detection processing using YOLOv9.

The performance evaluation is essentially a comparison between the proposed method implemented in YOLOv9 and YOLOv7, which is the state-of-the-art solution for visual landmark map-based spatial recognition using a monocular camera. Not only are we comparing the spatial recognition accuracy, but we are also taking into account the data sizes required to maintain visual landmark maps in subspaces. To achieve this, we conducted numerous groups of controlled experiments.

Ⅲ. Results and Discussion

The result comparison between the proposed method implemented in YOLOv7 and YOLOv9 is analyzed in two dimensions. First, we compare the raw performance of the object detection capability of the two models.

First, we compare the raw performance of the object detection capability of the two models.

In terms of accuracy, YOLOv9 achieved an accuracy of 92.5% compared to YOLOv7's 88.3%, as shown in Table 1. This improvement is primarily due to the advanced PGI mechanism and GELAN architecture.

When it comes to inference time, YOLOv9's average inference time was 27 milliseconds per frame, demonstrating a 4% improvement over YOLOv7, with the inference time of 28 milliseconds. This reduction in inference time enhances the real-time performance of the system, making it more suitable for applications requiring fast processing.

As for the system resource consumption, particularly memory usage, YOLOv9's memory usage was observed to be 320 MB, which is 10% lower than YOLOv7, which consumes about 355 MB of memory. This decrease in memory usage allows for deployment on devices with limited resources, broadening the potential applications of the model.

Model	Overall Accuracy	Small Objects	Large Objects
YOLOv7	88.3%	85.0%	91.6%
YOLOv9	92.5%	90.2%	94.3%

Table 1. Comparison of Raw Performance of YOLOv7 and YOLOv9

The enhancements in YOLOv9 provide a more robust and efficient solution for real-time indoor spatial recognition. The improvements in accuracy and speed, combined with lower memory usage, make it an excellent choice for applications such as autonomous navigation and augmented reality.

Fig 1. Snapshot of the Labeling of Dataset for Detection

Ⅳ. Conclusion

YOLOv9 demonstrates significant improvements in accuracy, inference time, and memory usage for visual landmark map-based spatial recognition. These findings suggest that YOLOv9 is well-suited for real-time indoor localization applications. The enhanced detection capabilities, particularly for small objects, and the reduced computational load make YOLOv9 a valuable tool for various indoor navigation and localization tasks.

Future work will focus on further optimizing the model, exploring its performance in diverse environments, and integrating additional sensors for improved robustness. We also plan to investigate the application of YOLOv9 in other domains such as augmented reality and robotics to fully leverage its capabilities.

ACKNOWLEDGMENT

REFERENCES

[1] Chen, Haichuan, et al. "Visual Landmark Map-Based Spatial Recognition Using a Monocular Camera." 2024 IEEE International Conference on Consumer Electronics (ICCE). IEEE, 2024.

[2] Wang, Chien-Yao, I-Hau Yeh, and Hong-Yuan Mark Liao. "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information." arXiv preprint arXiv:2402.13616 (2024).

[3] Yuming Chen, Xinbin Yuan, Ruiqi Wu, Jiabao Wang, Qibin Hou, and Ming-Ming Cheng. YOLO-MS: rethinking multiscale representation learning for real-time object detection. arXiv preprint arXiv:2308.05480, 2023.

[4] Lin Huang, Weisheng Li, Linlin Shen, Haojie Fu, Xue Xiao, and Suihan Xiao. YOLOCS: Object detection based on dense channel compression for feature spatial solidification. arXiv preprint arXiv:2305.04170, 2023.

[5] Diwan, Tausif, G. Anirudh, and Jitendra V. Tembhurne. "Object detection using YOLO: Challenges, architectural successors, datasets and applications." multimedia Tools and Applications 82.6 (2023): 9243-9275.

[6] Park, Young-Kook, et al. "Traffic landmark matching framework for HD-map update: Dataset training case study." Electronics 11.6 (2022): 863.

[7] Li, Guanqing, Zhiyong Song, and Qiang Fu. "A new method of image detection for small datasets under the framework of YOLO network." 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, 2018.