Place Recognition Enhancement for Autonomous Mobile Robot: Box Bounding vs. Pixel Segmentation

Yu-Jen Li and Chih-Hung G. Li*, *IEEE Member*

***Abstract*—Visual detection has been widely used for place recognition and localization of autonomous mobile robots (AMRs). The visual factors will prompt AMRs to produce corresponding calculation results. For AMRs working in a human-centric environment, it is inevitable to encounter pedestrians, and the presence of pedestrians may confuse the visual detector and reduce the accuracy of localization. In this article, we proposed a method of image conditioning by eliminating the pedestrian pixels and restoring the scene with the inpainting technique. We studied the quality of the conditioned images produced by box bounding and pixel segmentation. We conducted a series of experiments in corridor environments and compared two pedestrian detection methods – YoLo-v4 and Mask R-CNN. Their performances affecting the success of place recognition were compared. The results showed that Mask R-CNN appeared to lead to a better quality of restored images and also result in a higher accuracy of place recognition.**

I. Introduction

Autonomous mobile robots (AMRs) have been widely used in various industries, such as logistics, manufacturing, medical care, and agriculture. They learn about the surrounding environmental variables through sensors and calculate mobile strategies. Modern AMRs in various industries utilize multiple ways for sensing the environment, including magnetic tape, QR code, RFID, laser scanning [1], and computer vision [2-4].

For AMRs working in human-centric environments, pedestrians appearing in front of the robot are inevitable and may cause errors in the AMR's environmental recognition [5]. Previously, Li et al. proposed a topological localization method (TLM) based on surrounding RGB inputs and Convolutional Neural Network (CNN) [4]; it was found that the accuracy and recall rate can reach 90% when driving normally. However, when pedestrians entered the robot's view, the accuracy of place recognition may drop significantly, indicating a potential threat to place recognition raised by the appearance of pedestrians or other dynamic objects. For this reason, we proposed a method to improve the accuracy of place recognition when there is pedestrian interference (see Fig. 1). Using object detection to detect pedestrians and removing the pedestrian pixels, the images taken by the robot are repaired and the original scenes without pedestrians are restored. By integrating the series of image processing actions, views of the environment are restored automatically and the results are inputted to TLM for robot localization.

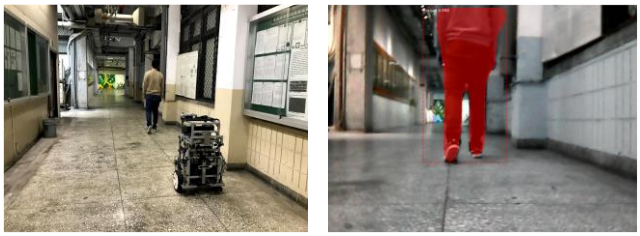

Fig. 1. Example of AMR conducting visual navigation with the view occluded by nearby pedestrians.

II. Related Works

A. Object Detection

To identify pedestrians in the robot's view, visual object detection methods are adopted. Recently, many state-of-the-art object detection methods are CNN-based image processors. Based on the strategies adopted by these methods, they can be divided into two main categories: one-stage and two-stage methods. The one-stage methods use a single neural network to detect the location and identification of the item, with the weight of the reason, including YoLo-v4 [6], SSD [7], and Retinanet [8]. The two-stage methods consider the detection accuracy, including R-CNN [9], FASTER R-CNN [10], and Mask R-CNN [11].

B. Image Inpainting

The emergence of Generative Adversarial Networks (GAN) has provided new ways of data synthesis and applications in data augmentation. Frid-Adar et al. [12] utilized GAN for data enhancement. Li and Huang [13] used pix2pix GAN to impose shadow effect on object images. Yu et al. [14] proposed a free-form image inpainting algorithm called SN-PatchGAN, which can perform object deletion and free editing. In their example, free-form patching was performed on the image to remove unwanted pixels. The gated CNN they proposed is characterized by learning each spatial information of all layers and helping users to quickly repair pictures through a given mask. In this article, we proposed a method to improve the accuracy of AMR place recognition due to the occlusion of pedestrians. In the proposed framework, SN-PatchGAN is used to repair the image where the pedestrian pixels have been deleted. The pedestrian pixels were first detected by the object detectors such as YoLo-v4 and Mask R-CNN. As the performance of the object detector affects the quality of the

This work was supported by the Ministry of Science and Technology of the Republic of China, Taiwan under Contract No.: MOST109-2637-E-027-008-.

Yu-Jen Li is with the Graduate Institute of Mechatronic Engineering, National Taipei University of Technology, Taipei, Taiwan, R.O.C. (phone: +886-2-2771-2171; fax: +886-2-2776-4889; e-mail: nidavidce@gmail.com).

Chih-Hung G. Li* is with the Graduate Institute of Manufacturing Technology, National Taipei University of Technology, Taipei, Taiwan, R.O.C. (phone: +886-2-2771-2171; fax: +886-2-2776-4889; e-mail: cL4e@mail.ntut.edu.tw).

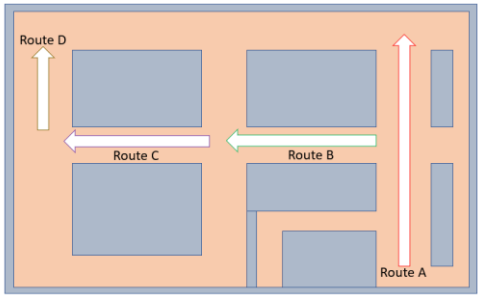

Fig. 2. Illustration of the four routes in the field test.

processed images, it also affects the quality of the repaired images and the subsequent place recognition. In this article, we focused on comparing the performances of YoLo-v4 and Mask R-CNN and how the results affect the success rate of place recognition.

III. Proposed Method

TLM developed in [4] was used for testing the performance of the proposed place recognition enhancement method. Four corridor paths denoted as A, B, C, and D as shown in Fig. 2. were defined as the experimental site. Each path contains 21, 21, 17, and 7 nodal locations, respectively; each node is separated by 2 m. In TLM, Each path was trained to be an independent model; the training images do not contain any pedestrians. To test the influence of pedestrians on TLM, at each nodal location, 81 photos containing pedestrians in the view were acquired. These images were directly inputted to TLM as benchmark samples. For the proposed method, the images were first inputted to a pedestrian detector for identification of the existence and location of the pedestrians on the image. Then, the pedestrian pixels were reset to zero. Finally, the images were repaired by SN-PatchGAN before sending to TLM for place recognition.

A. Pedestrian Detection

The first step of the proposed method needs to accurately detect the locations of pedestrians in the image. We adopted two different object detection modules for the task, YoLo-v4 and Mask R-CNN. YoLo-v4 outputs rectangular bounding boxes to frame pedestrians, whereas Mask R-CNN performs segmentation on the pixel level and provides masks over the shape of the pedestrian. It is an extension of the Faster R-CNN architecture, combining Faster R-CNN with Fully Convolutional Networks (FCN) so that Mask R-CNN can output object detection and semantic segmentation simultaneously.

B. Image Processing

The image processing flow is shown in Fig. 3. After pedestrian detection, the pixel ranges of the pedestrian in the image are obtained. Those pixels are reset to RGB white (255, 255, 255) for subsequent image inpainting. To match the TLM format, The 640×480 input image is cropped at the center by 400×400 and then resized to 32×32. Finally, the processed image is sent to SN-PatchGAN for image repair; a repaired image without pedestrians will be obtained.

C. Place Recognition

The images repaired by SN-PatchGAN are sent to TLM for place recognition, as shown in Fig. 3. TLM is a CNN-based image classifier that can recognize the nodal locations along a predefined route. The image for recognition at each node is composed of photos taken by 4 RGB cameras, placed to monitor the front, back, left, and right scenes. At the detection phase, the CNN classifier receives the image of the four views in real-time and directly outputs a prediction on the nodal location. If a route is defined with 21 nodes, the output of TLM provides 21 classes.

IV. Results and Discussion

The impacts of pedestrian pixels on TLM accuracy are shown in Table I. It can be seen that Routes A, B, and D have relatively small impacts due to the appearance of pedestrians. In those cases, TLM results showed an accuracy higher than 95% and reaching 100%. In contrast, Route C appeared more affected by pedestrians; the accuracy is only 90.3%. In the subsequent experiments, we processed the same pedestrian images following the procedures depicted in Fig. 3 before sending them to TLM for recognition. The results are also summarized in Table I. One may find that for all four routes tested, the proposed method enhanced the TLM accuracy. The most profound influence exists in Route C, where an improvement of 8.2% was generated with Mask R-CNN as the pedestrian detector. The original 90.3% accuracy of Route C is improved to be 98.5%.

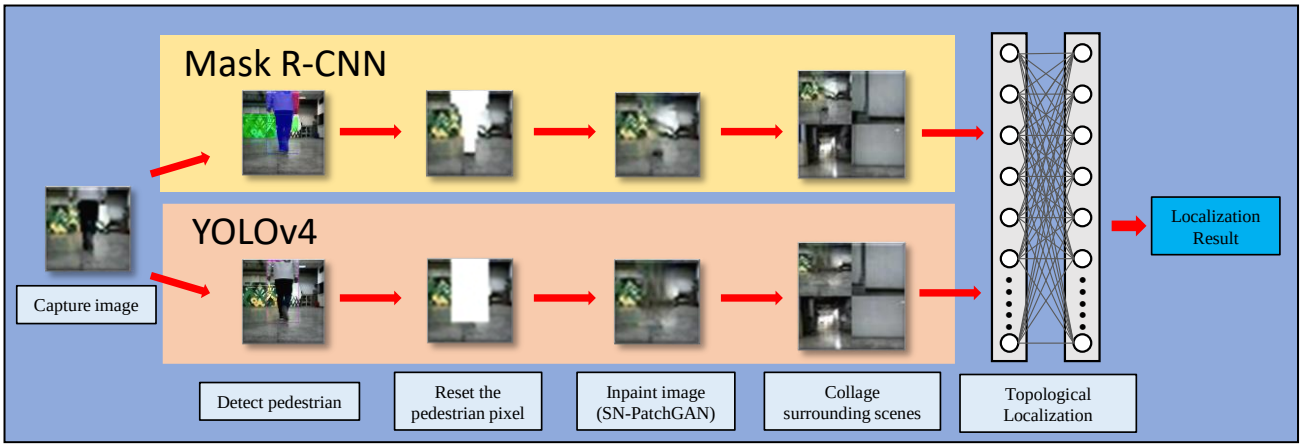

Fig. 3. Procedures of the proposed pedestrian removal and subsequent topological localization.

TABLE I. COMPARISON OF THE PEDESTRIAN DETECTION RESULT

Type/ Correct percentage	Route A	Route B	Route C	Route D
Original	99.5%	95.2%	90.3%	100%
YoLov4	99.8%	96.2%	94.8%	100%
Mask R-CNN	100%	99.1%	98.5%	100%

Fig. 4. Comparison of the pedestrian detection results. *Left*: Mask R-CNN, *Right*: YoLo-v4 .

A. *YoLo-v4 Box Bounding*

YoLo-v4 shows the result of object detection in the form of a bounding box defined by (x, y, w, h), where (x, y) represents the normalized center coordinates and (w, h) denotes the normalized width and height. As the shape of the bounding box is always rectangular and outlines the maximum area of the pedestrian region, when the pixels inside the bounding box are reset to RGB white, inevitably, there will be background pixels other than the pedestrians reset during the process. (see Fig. 4). As a result, SN-PatchGAN retains less information on the background surrounding the pedestrians and is less likely to perform perfectly on image inpainting. Nevertheless, successful image inpainting was witnessed and so was the improvement in TLM accuracy demonstrated in Table I.

B. *Mask R-CNN Pixel Segmentation*

Mask R-CNN presented the results of pedestrian detection in the form of instance segmentation, and the appearance of pedestrians is masked in pixels instead of rectangular bounding boxes. Thus, one may more accurately process the pedestrian region on the image without falsely altering the adjacent background as shown in Fig. 4. Unsurprisingly, the results of image inpainting by SN-PatchGAN were better than the YoLo-v4 cases as indicated in Table I. We summarized the performance results of Mask R-CNN in Table II. It can be seen that 187 images originally failed by TLM were correctly predicted after the proposed procedure; on the contrary, only one image originally predicted by TLM was failed by the proposed procedure. Overall, the accuracy of the four paths was between 98%~100%. Compared with YoLo-v4, the accuracy has increased by nearly 4%.

V. CONCLUSION

In an attempt to remove pedestrian pixels for improvement on place recognition of AMR localization, YoLo-v4 and Mask R-CNN were adopted for identification of pedestrian regions and the subsequent image inpainting by SN-PatchGAN and robot localization by TLM. It was found that Mask R-CNN provides instance segmentation on the contour of the pedestrian and less affects the adjacent pixels of the background. As a result, the performance of SN-PatchGAN on restoring the environmental scene was more successful and led to higher accuracy of place recognition by TLM. On the contrary, YoLo-v4 generates rectangular bounding boxes which include more background pixels other than the pedestrian. Resetting of those background pixels resulted in more difficulty of correct image inpainting and adversely affected the accuracy of TLM. Our test results showed that the accuracy of TLM is significantly improved, especially on Route C from 90.3% to 98.5%. Specifically, among 134 originally miss-detected images by TLM, 114 were corrected by the proposed procedure using Mask R-CNN for pedestrian segmentation and SN-PatchGAN for image inpainting.

ACKNOWLEDGMENT

This work was supported by the Ministry of Science and Technology of the Republic of China, Taiwan under Contract No.: MOST109-2637-E-027-008-.

REFERENCES

[1] J. J. Leonard and H. G. Durrat-Whyte, "Mobile robot localization by tracking geometric beacons," *IEEE Trans. Robot. Auto*., vol. 7, no. 3, pp. 376–382, 1991.

TABLE II. RESULT SUMMARY OF TLM USING MASK R-CNN

Route	Testing images	Original Correct predictions	Original Correct percentage	Corrected by procedures	Failed by procedures	Correct predictions after procedures	Correct percentage after procedures
A	1701	1693	99.5%	8	0	1701	100%
B	1701	1620	95.2%	65	0	1685	99.1%
C	1377	1243	90.3%	114	1	1356	98.5%
D	567	567	100%	0	0	567	100%
Sum	5346	5123	95.8%	187	1	5309	99.3%

[2] R. Mur-Artal, J. M. M. Montiel and J. D. Tardós, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, 2015.
[3] N. Piasco, D. Sidibé, C. Demonceaux, and V. Gouet-Brunet, "A survey on visual-based localization: on the benefit of heterogeneous data," *Pattern Recognition*, vol. 74, pp. 90–109, 2018.
[4] C. G. Li, Y.-F. Hong, P.-K. Hsu, and T. Maneewarn, "Real-time topological localization using structured-view ConvNet with expectation rules and training renewal," *Robot. Auto. Syst.*, vol. 131, https://doi.org/10.1016/j.robot.2020.103578, 2020.
[5] T. Fan, X. Cheng, J. Pan, P. Long, W. Lin, R. Yang, and D. Manocha, "Getting robots unfrozen and unlost in dense pedestrian crowds," *IEEE Robot. Auto. Letters*, vol. 4, no. 2, pp. 1178–1185, 2019.
[6] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YoLo-v4: optimal speed and accuracy of object detection," arXiv:2004.10934.
[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. "SSD: Single shot multibox detector," arXiv:1512.02325.
[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. "Focal loss for dense object detection," arXiv:1708.02002.
[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," arXiv:1311.2524.
[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 91–99, 2015.
[11] K. He, G. Gkioxari, P. Dollár, R. Girshick, "Mask R-CNN," arXiv:1703.06870.
[12] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using GAN for improved liver lesion classification," in *Proc. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, Washington, DC, 2018, pp. 289–293.
[13] C. G. Li and Y.-H. Huang, "Deep-trained illumination-robust precision positioning for real-time manipulation of embedded objects," *Int. J. Adv. Manuf. Technol.*, vol. 111, pp. 2259–2276, 2020.
[14] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," arXiv:1806.03589.