Proceedings of the 10th World Congress on Mechanical, Chemical, and Material Engineering (MCM'24)
Barcelona, Spain - August 22-24, 2024
Paper No. ICMIE 139
DOI: 10.11159/icmie24.139

Hard Example Detection at Actual Environment for Semantic Segmentation Used in Visual Navigation

Yuriko Ueda[1], Marin Wada[1], Miho Adachi[1], Ryusuke Miyamoto[2]
[1]Dept. of Computer Science, Graduate School of Science and Technology
Meiji University, Kawasaki, Japan
ueda@cs.meiji.ac.jp; marin@cs.meiji.ac.jp; miho@cs.meiji.ac.jp;
[2]Dept. of Computer Science, School of Science and Technology
Meiji University, Kawasaki, Japan
miya@cs.meiji.ac.jp

Abstract **-** Visual navigation based on the results of semantic segmentation is a promising method as it does not require expensive sensors and detailed metric maps to actualize autonomous movement. However, navigation performance strongly depends on the accuracy of semantic segmentation for which high quality training datasets reflecting the target environment are indispensable. This paper presents a novel method to detect hard examples for semantic segmentation in the case of visual navigation. This method adopts interframe changes in the number of class labels and simple thresholding. Experimental results using training and testing data gathered during the course of the Tsukuba Challenge, a competition of autonomous moving robots held in Japan showed that the proposed method is feasible to detect hard examples for segmentation in visual navigation.

Keywords**:** hard example, semantic segmentation, visual navigation, actual environment

1. Introduction

Currently, three-dimensional (3D) LiDAR is the most popular device for external sensors used on autonomous moving robots. When 3D LiDAR is adopted on a robot, environmental maps composed of 3D point clouds are initially created to determine its location and to detect moving and static obstacles that should be avoided during autonomous movement [1]. This approach is quite popular but has several problems, including the difficulty of environmental recognition using only 3D point clouds, the necessity of map creation, and the high cost of 3D LiDAR. One method proposed to solve these problems is visual navigation based on semantic segmentation [2]–[5] categorized as mapless navigation [6]–[9].

The visual navigation method based on the results of our proposed semantic segmentation [2], [4], [5], [10], [11] enables road following by setting a target point toward which a robot moves on a region extracted as a traversable area. In addition, obstacle avoidance is performed simultaneously during the road following process; the extraction of traversable area is performed considering several kinds of obstacles. When a change of course at intersections is applied during road following, a robot can move autonomously in both indoor and outdoor environments [2]–[5].

The visual navigation method based on results of semantic segmentation seems feasible because it requires neither precise environmental maps nor expensive sensing devices. However, this method has a different problem that has significant influence on movement performance; it requires accurate segmentation results. Our previous research [12] has shown that segmentation accuracy strongly depends on the quality of the training dataset. In particular, we found that training samples collected in the target environment drastically improved classification accuracy [12]. Unfortunately, it is not feasible to construct a dataset including many samples collected at the target site because assigning pixel-wise class labels for a large image dataset requires significant human labor.

To solve this problem, this paper proposes a novel method to select hard samples that are difficult to classify accurately. This concept, called "hard example mining," was widely used for object detector training before the emergence of deep learning. For visual object detection, it is easy to extract hard examples from training samples because they are extracted using the locations of positive samples included in the training dataset. However, the same process cannot be applied to semantic segmentation in the target environment, where ground truth is not assigned to input images. Considering the characteristics of hard examples for semantic segmentation in the selection process, the proposed method extracts hard examples using a time-series analysis of inference results. To verify the proposed approach, experiments were conducted using a dataset used in our previous research [13] generated from 3D point clouds with color information.

2. Selection of Hard Examples from an Actual Image Sequence

This section proposes a novel method for selecting hard examples from an actual image sequence using changes in segmentation results along the time axis.

2.1. Idea of hard example mining to train a classifier for segmentation

Figure 1 shows the training workflow, where hard examples are added to training samples based on the inference results of a previously trained classifier. During hard example mining in visual object detection, newly selected samples can be extracted from training images using the locations of bounding boxes corresponding to positive samples [14]. However, it is impossible to obtain hard examples from training images in the case of semantic segmentation, as the number of training images is limited and does not have ground truth. Therefore, we must construct a method to select hard examples using actual input images without ground truth.

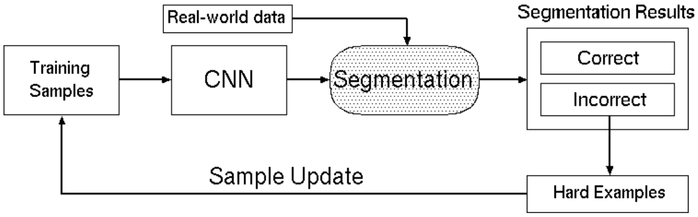

Fig. 1: Hard example mining to train a classifier for segmentation.

2.2. How to select hard examples

It is difficult to select hard examples for semantic segmentation in general applications. However, in the case of visual navigation, appropriate hard examples can be collected because of the characteristics of segmentation results; poor estimation results are obtained momentarily in subsequent input images. Figure 2 shows example inference results for subsequent images. As shown in the figure, segmentation results may change drastically even though the appearances of these images are quite similar.

To detect drastic changes in the segmentation results, the proposed method focuses on the change in pixel counts classified as a class. The changes in these values are shown in Fig. 3. Although this approach is not sophisticated, it can find scenes in which a currently trained classifier is insufficient for segmentation, even when simple thresholding is applied. Fortunately, we do not need to extract all hard examples for this purpose.

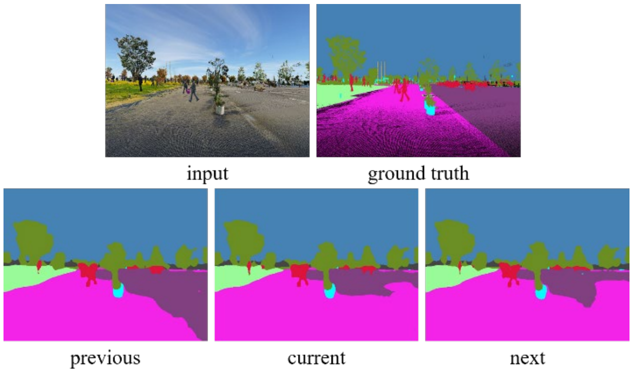

Fig 2: Inference images including large area of miss classification.

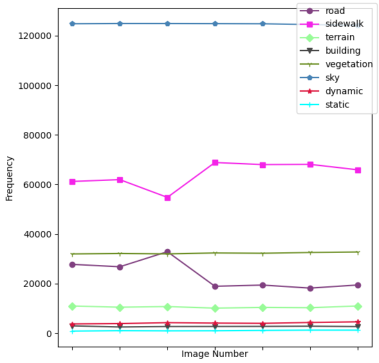

Fig 3: Changes of pixel numbers of class labels at miss classification.

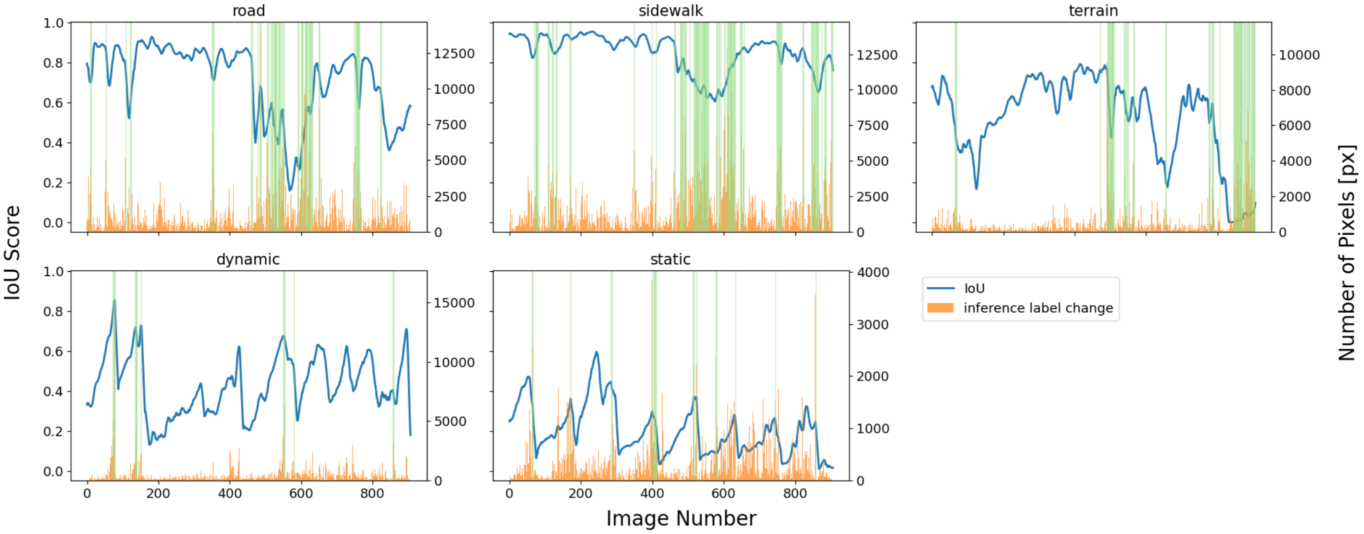

Fig. 4: IoU and interframe change of pixels of class labels for sequential images.

3. Evaluation

This section verifies the ability of the proposed method to extract hard examples from sequential images.

3.1. Test dataset

To perform quantitative evaluation, it is necessary to capture a series of images in a real-life environment. However, annotating these images manually is labor-intensive. Therefore, in this study we used a dataset generated from a 3D model. This 3D model consists of 3D point clouds with color information and corresponding class labels captured on the course of the Tsukuba Challenge, an autonomous mobile robot competition held in Japan. From this model, 907 images were generated for evaluation.

3.2. Model used for inference

For the experiment, PSPNet [15] with ResNext101 [16] were adopted as a baseline, which demonstrated good accuracy for test images captured in the real-world environment [12]. To train this model, actual images with class labels though test images were generated from a 3D model scanned at the same location. This procedure was used because this experiment required a moderate number of images that were misclassified.

3.3. Results

Figure 4 shows the experimental results, where blue and orange lines represent intersection over union (IoU) and interframe difference of pixels corresponding to the target class. In this study, we evaluated only the following classes (out of a total of eight classes): road, sidewalk, terrain, dynamic, and static. This is because a decrease in accuracy for these classes significantly impacts the navigation of the robot, making it necessary to prioritize their extraction. The valleys in the IoU indicate that the trained classifier did not estimate appropriately. In other words, the IoU valleys include hard examples for the current classifier. Focusing on the interframe difference of label images obtained at runtime, peaks emerge in frames where IoU decreases, indicating hard examples. As a result, the frame numbers corresponding to hard examples can be obtained by simple thresholding, except for the dynamic class. For this class, it is difficult to find hard examples using the proposed method. Modifying the proposed method to address the dynamic class is planned for future work.

4. Conclusion

This paper proposed a novel method to detect hard examples for semantic segmentation using actual images for use visual navigation. The proposed method finds hard examples using interframe differences of the number of the target which does not change drastically when sequential images are obtained from a camera mounted on a robot at a moderate framerate. Experimental results showed that peak interframe differences of class labels appeared when IoU decreased, indicating poor segmentation results. The results demonstrate that simple thresholding to obtain interframe difference of class labels is a more accurate method to detect hard examples.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP22K12082.

References

[1] M. Elhousni and X. Huang, "A Survey on 3D LiDAR Localization for Autonomous Vehicles," in Proc. IEEE Intelligent Vehicles Symposium, Oct 2020.

[2] R. Miyamoto, M. Adachi, H. Ishida, T. Watanabe, K. Matsutani, H. Komatsuzaki, S. Sakata, R. Yokota, , and S. Kobayashi, "Visual navigation based on semantic segmentation using only a monocular camera as an external sensor," J. Robot. Mechatron., vol. 32, no. 6, pp. 1137–1153, 2020.

[3] M. Adachi, K. Honda, J. Xue, H. Sudo, Y. Ueda, Y. Yuda, M. Wada, and R. Miyamoto, "Practical implementation of visual navigation based on semantic segmentation for human-centric environment," J. Robot. Mechatron., vol. 35, no. 6, pp. 1419–1434, 2023.

[4] R. Miyamoto, Y. Nakamura, M. Adachi, T. Nakajima, H. Ishida, K. Kojima, R. Aoki, T. Oki, and S. Kobayashi, "Vision-based road-following using results of semantic segmentation for autonomous navigation," in Proc. ICCE Berlin, 2019, pp. 194–199.

[5] M. Adachi, S. Shatari, and R. Miyamoto, "Visual navigation using a webcam based on semantic segmentation for indoor robots," in Proc. SITIS, 2019, pp. 15–21.

[6] R. Partsey, E. Wijmans, N. Yokoyama, O. Dobosevych, D. Batra, and O. Maksymets, "Is mapping necessary for realistic pointgoal navigation?" in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., jun 2022, pp. 17 211–17 220.

[7] H. Du, L. Li, Z. Huang, and X. Yu, "Object-goal visual navigation via effective exploration of relations among historical navigation states," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., jun 2023, pp. 2563–2573.

[8] A. Amini, G. Rosman, S. Karaman, and D. Rus, "Variational end-to-end navigation and localization," in Proc. IEEE Int. Conf. on Robotics and Automation, 2019, pp. 8958–8964.

[9] T. Ort, L. Paull, and D. Rus, "Autonomous vehicle navigation in rural environments without detailed prior maps," in Proc. IEEE Int. Conf. on Robotics and Automation, 2018, pp. 2040–2047.

[10] M. Adachi, K. Honda, and R. Miyamoto, "Turning at intersections using virtual lidar signals obtained from a segmentation result," Journal of Robotics and Mechatronics, vol. 35, no. 2, pp. 347–361, 2023.

[11] M. Adachi and R. Miyamoto, "Model-based estimation of road direction in urban scenes using virtual lidar signals," in Proc. IEEE Int. Conf. Systems, Man, and Cybernetics, 2020, pp. 4498–4503.

[12] R. Miyamoto, M. Adachi, Y. Nakamura, T. Nakajima, H. Ishida, and S. Kobayashi, "Accuracy improvement of semantic segmentation using appropriate datasets for robot navigation," in Proc. CoDIT, 2019, pp. 1610–1615.

[13] M. Wada, M. Adachi, Y. Ueda, and R. Miyamoto, "Dataset for semantic segmentation generated from 3D scanned data for visual navigation," in Proc. CoDIT, 2023, pp. 1711–1716.

[14] R. Miyamoto and T. Oki, "Soccer player detection with only color features selected using informed haar-like features," in Advanced Concepts for Intelligent Vision Systems, 2016, pp. 1751–1760.

[15] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 2881–2890.

[16] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in ´ Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5987–5995.