Journal of Korea Robotics Society (2022) 17(3):334-338
https://doi.org/10.7746/jkros.2022.17.3.334 ISSN: 1975-6291 / eISSN: 2287-3961

동적 도시 환경에서 의미론적 시각적 장소 인식

Semantic Visual Place Recognition in Dynamic Urban Environment

사바 아르샤드[1] · 김 곤 우[†]
Saba Arshad[1], Gon-Woo Kim[†]

Abstract: In visual simultaneous localization and mapping (vSLAM), the correct recognition of a place benefits in relocalization and improved map accuracy. However, its performance is significantly affected by the environmental conditions such as variation in light, viewpoints, seasons, and presence of dynamic objects. This research addresses the problem of feature occlusion caused by interference of dynamic objects leading to the poor performance of visual place recognition algorithm. To overcome the aforementioned problem, this research analyzes the role of scene semantics in correct detection of a place in challenging environments and presents a semantics aided visual place recognition method. Semantics being invariant to viewpoint changes and dynamic environment can improve the overall performance of the place matching method. The proposed method is evaluated on the two benchmark datasets with dynamic environment and seasonal changes. Experimental results show the improved performance of the visual place recognition method for vSLAM.

Keywords: Semantics, Simultaneous Localization and Mapping, Visual Place Recognition

1. Introduction

SLAM enables the mobile robots to move autonomously in an unknown environment by generating map along the trajectory and correctly identifying the already visited places. This recognition of places is performed by the loop closure detection system, one of the components of SLAM, which performs visual place recognition to support robot’s relocalization in the environment map and reduction in map drift which may occur during robot motion.

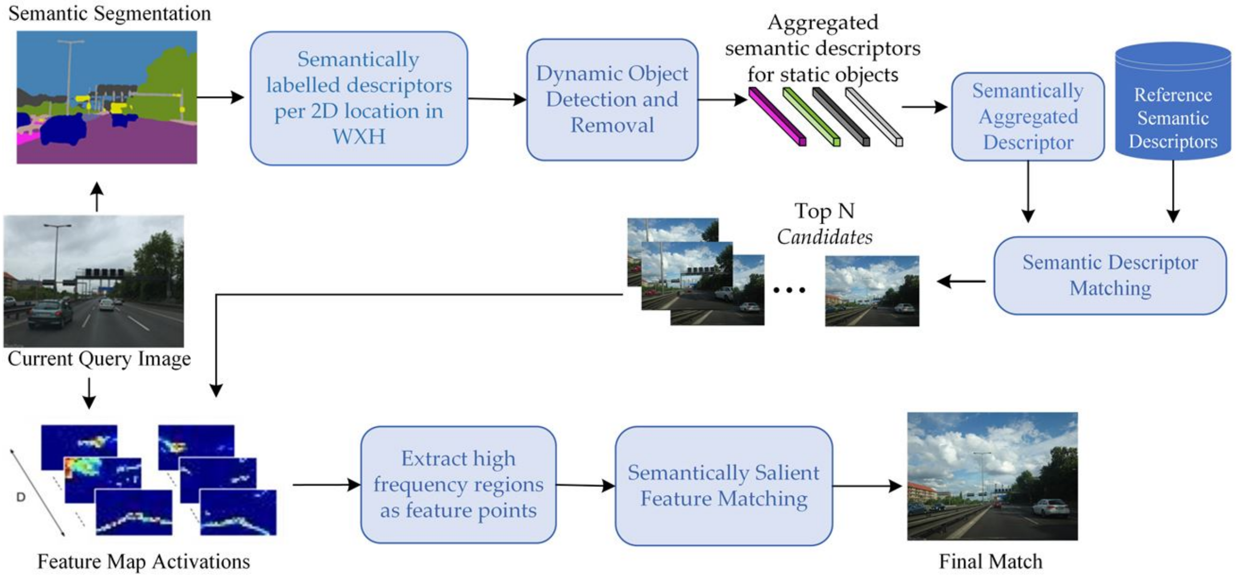
The traditional handcrafted features-based visual place recognition methods have shown good performance in large-scale localization, however they lack resilience to severe appearance changes[1]. Also, image matching using these features is often a bottleneck for mobile robot operation in real-time scenarios. While, the appearance-based approaches such as sequence searching[2], appearance prediction[3], shadow removal[4] and illumination invariance[5] are robust to conditional variations yet their performance suffers from viewpoint dependency and velocity sensitivity. Recently, with the advancements in deep learning, numerous deep neural networks have been trained and used for image representation and place matching and detection tasks[6,7]. The features extracted from convolutional networks are deeper and more abstract, hence being comparatively more robust to environmental conditions[8,9]. A very little attention has been paid to semantics-aware loop closure detection. The use of semantics for place recognition have shown improvement in overall performance[10-12].

This paper specifically addresses the problem of visual place recognition in the presence of dynamic objects in an urban

Received : Mar. 10. 2022; Revised : Mar. 22. 2022; Accepted : Mar. 23. 2022
※ This research was supported in part by the MSIT (Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2022-2020-0-01462) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation), and in part by the Technology Innovation Program (or Industrial Strategic Technology Development Program-ATC+) (20009546, Development of service robot core technology that can provide advanced service in real life) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea)

1. PhD Candidate, Control and Robot Engineering, Chungbuk National University, Cheongju, Korea (sabarshad1000@gmail.com)
† Professor, Corresponding author: Department of Intelligent Systems and Robotics, Chungbuk National University, Cheongju, Korea (gwkim@cbnu.ac.kr)

Copyrightⓒ KROS


Semantic Segmentation
Semantically labelled descriptors per 2D location in WXH
Dynamic Object Detection and Removal
Aggregated semantic descriptors for static objects
Semantically Aggregated Descriptor
Reference Semantic Descriptors
Top N Candidates
Semantic Descriptor Matching
Current Query Image
D
Feature Map Activations
Extract high frequency regions as feature points
Semantically Salient Feature Matching
Final Match

[Fig. 1] Semantic visual place recognition pipeline

environment. In[13] moving objects have been considered as occlusions, however, the dynamic objects may not necessarily be always moving. Mostly, the dynamic objects might be temporarily static such as parked vehicle, people sitting somewhere and pedestrians standing on traffic signals etc. These dynamic objects that are in temporarily static state may not be present at the same location when robot revisits the same place later leading to the false correspondences because a temporarily static mobile object might appear at different locations. Also, the detection of moving objects from one scene to other overburdens the processor as it stores visual information of multiple scenes in memory.

Recent works applying semantic labels extracted through semantic segmentation networks have shown improved performance in drastic appearance changes[14,15]. Garg et al.[14] proposed a local semantic tensor descriptor and fused pixel-wise semantic labels with deep features building for robust place matching under extreme illumination and viewpoint changes i.e., 180-degree. However, it is typical developed for place recognition in opposing viewpoints while focusing only three semantic class labels i.e., building, road and vegetation. Due to the selected semantic labels, the performance degrades in the environment with semantically similar scenes. A subsequent study[15] proposed a two nearest neighbour local sensor tensor based on TNNVLAD. Thus, instead of relying on a single nearest neighbour match, the algorithm improves the overall accuracy by comparing with the second nearest neighbour descriptor. Similarly,[16] used the LoST in combination with semantic edge features extracted from semantic segmentation mask to achieve loop closure detection. Some researchers have used pixelwise semantic labels for building the topological connectivity graph of semantic objects in a scene to extract the spatial information[17,18]. This topological graph is fused with handcrafted features-based BoW model to achieve robustness in drastic appearance changes.

2. Semantic Visual Place Recognition

[Fig. 1] illustrates the working pipeline of hierarchical visual place recognition method. The proposed method uses a hierarchical approach where top matching candidates are selected based on the global semantic similarity between image pairs and then CNN feature based visual-semantic verification is performed to find the best match.

Initially the semantic information is extracted through pixelwise semantics segmentation using RefineNet[19] trained on cityscapes dataset. The semantic segmentation produces the semantic label scores of 1/4th of the input image size in width and height while depth corresponds to the number of semantic classes. Each pixel is associated with the semantic label based on the probability score. Aggregating the semantic label scores for each static objects available in an image while removing the dynamic object semantics, a semantic descriptor is computed as done in[14]. In a dynamic environment, the feature occlusion occurs due to the moving objects such as person, car, leading to the reduces robustness. Including those semantic objects for place matching causes low accuracy as they will increase non-overlap contents between the image pairs. Thus, we keep only static object to construct the semantic image descriptor vector V. V is obtained through concatenation of L2-normalized semantic descriptors for the class labels given as V = {$V_{road} + V_{building} + V_{vegetation} + V_{pole} + V_{traffic\_light}$}. The semantic descriptor of the reference dataset is extracted as an offline step. The semantic descriptor of each query image is compared with the reference database through cosine similarity δ as given in Eq. (1). Higher the cosine similarity value, higher will be the semantic similarity between the image pairs.

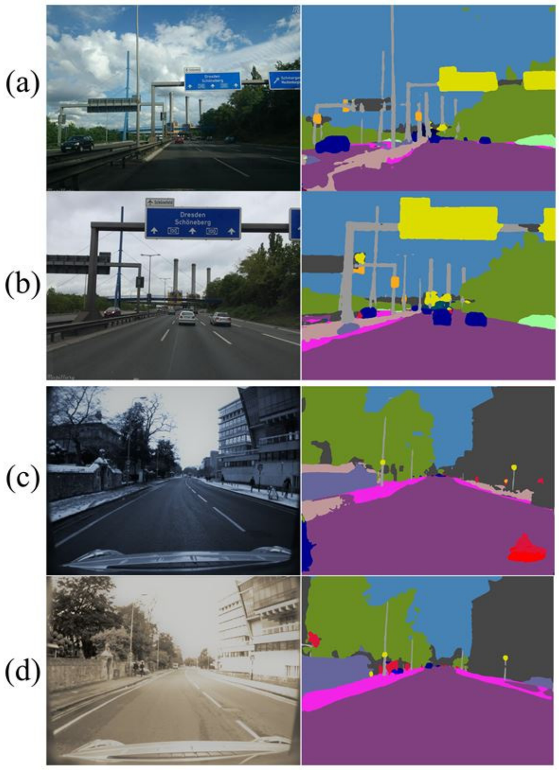
$$\cos(m,n) = 1 - \frac{m \cdot n}{|m| * |n|} \tag{1}$$

After the global scene similarity computation, top 10 reference images with the highest similarity score are selected as candidate images for further processing to obtain the final match.

In the second phase, the feature map activations of the query image are obtained as keypoints. Each keypoint is associated with the semantic labels. Similar to[14] the activations of the query image are compared with those of candidate images. Instead of matching the feature points of all the semantic class labels, the proposed method excludes the feature points from the dynamic objects while only matching the features from the static objects. This greatly enhances the place matching performance while reducing the number of feature correspondences resulting in lower computational cost.

3. Experiments and Results

This section presents the qualitative and quantitative results. The evaluation is performed on the publicly available benchmark datasets captured in different environments with challenging environmental conditions i.e., Berlin A100 and Oxford Robotcar summer vs. winter. Both of the datasets comprise of images captured in dyna mic urban environment with many confusing objects i.e., pedestrians, vehicles etc. These dynamic objects cause dynamic interference making them challenging datasets. Furthermore, the Berlin A100 dataset exhibits strong viewpoint changes while Oxford RobotCar dataset extreme seasonal changes and illumination changes with slight viewpoint variation. The semantic segmentation results obtained from the RefineNet are shown in [Fig. 2]. It can be observed that segmentation network is able to generate fine pixelwise segmentation results. Images of [Fig. 2(a)-(b)] presents the same place in Berlin A100 while with large appearance changes. Similarly, [Fig. 2(c)-(d)] present images from the same place taken in winter and summer seasons. Though the appearance change causes the feature-based method to fail in place recognition however the semantics remains the same. Thus, utilizing useful semantics can significantly improve the overall performance of visual place recognition method even in such challenging environments.



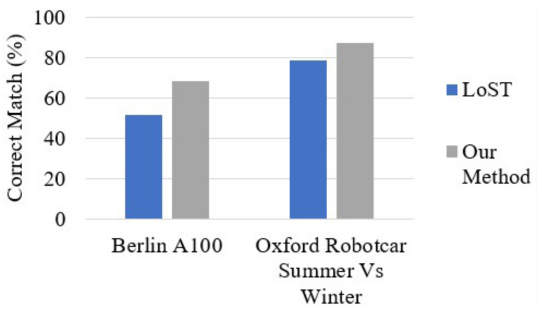
[Fig. 2] Sample images for pixelwise semantic segmentation using RefineNet for benchmark datasets i.e., (a) Berlin A100 Traverse 1; (b) Berlin A100 Traverse 2; (c) Oxford RobotCar Winter Season and (d) Oxford RobotCar Summer Season

The semantic class labels generated by the RefineNet for static objects includes road, building, vegetation, pole, traffic_ light, traffic_sign, terrain, wall and fence. Here, we analyse the significance of each class labels for correct detection of a place. Using the semantic labels, the global semantic descriptors are generated $V_1 \cdots V_6$, as given below.

· V_1 = {road, building, vegetation}
· V_2 = {road, building, vegetation, pole}
· V_3 = {road, building, vegetation, pole, traffic_light}
· V_4 = {road, building, vegetation, pole, traffic_light, traffic_sign}

[Table 1] F1-Score obtained using semantic descriptors on Berlin A100 and Oxford RobotCar dataset

Global Semantic Descriptor Vectors	Berlin A100	Oxford Robotcar Summer Vs. Winter
V_1	0.68	0.88
V_2	0.71	0.89
V_3	0.77	0.94
V_4	0.75	0.89
V_5	0.69	0.92
V_6	0.71	0.91



[Fig. 3] The performance comparison of the percentage of correctly matched places by the baseline method and our algorithm for the benchmark datasets

- V_5 = {road, building, vegetation, pole, traffic_light, traffic_sign, terrain}
- V_6 = {road, building, vegetation, pole, traffic_light, traffic_sign, terrain, wall, fence}

The F1-score obtained using each of the global semantic vector is given in [Table 1]. It can be observed that semantics included in V_3 are significant for scene understanding in an urban environment. The correct place detection performance of the proposed method is compared with the baseline method[14], as shown in [Fig. 3]. Utilizing useful semantics for the specific environment helps improve the place recognition performance of the proposed method while outperforming the[14].

4. Conclusion

This research proposes a hierarchical visual place recognition for visual SLAM based systems. The proposed method analyzes the scene semantics to attain the scene understanding and local and global level to improve the overall performance of the semantic visual place recognition method. The semantics being invariant to the dynamic environment and viewpoint changes helps to improve the overall performance of the system in challenging environmental conditions.

References

[1] M. Cummins and P. Newman, “Appearance-only SLAM at large scale with FAB-MAP 2.0,” *The International Journal of Robotics Research*, vol. 30, no. 9, pp. 1100-1123, Aug., 2011, DOI: 10.1177/0278364910385483.

[2] E. Pepperell, P. I. Corke, and M. J. Milford, “All-environment visual place recognition with SMART,” *2014 IEEE International Conference on Robotics and Automation (ICRA)*, Hong Kong, China, 2014, DOI: 10.1109/ICRA.2014.6907067.

[3] P. Neubert, N. Sunderhauf, and P. Protzel, “Appearance change prediction for long-term navigation across seasons,” *2013 European Conference on Mobile Robots*, Barcelona, Spain, 2013, DOI: 10.1109/ECMR.2013.6698842.

[4] P. Corke, R. Paul, W. Churchill, and P. Newman, “Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation,” *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Tokyo, Japan, 2013, DOI: 10.1109/IROS.2013.6696648.

[5] C. McManus, W. Churchill, W. Maddern, A. D. Stewart, and P. Newman, “Shady dealings: Robust, long-term visual localisation using illumination invariance,” *2014 IEEE International Conference on Robotics and Automation (ICRA)*, Hong Kong, China, 2014, DOI: 10.1109/ICRA.2014.6906961.

[6] Z. Chen, O. Lam, A. Jacobson, and M. Milford, “Convolutional Neural Network-based Place Recognition,” *16th Australasian Conference on Robotics and Automation 2014*, 2014, [Online], https://eprints.qut.edu.au/79662.

[7] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, “NetVLAD: CNN Architecture for Weakly Supervised Place Recognition,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 40, no. 6, 2018, DOI: 10.1109/TPAMI.2017.2711011.

[8] T.-H. Wang, H.-J. Huang, J.-T. Lin, C.-W. Hu, K.-H. Zeng, and M. Sun, “Omnidirectional CNN for Visual Place Recognition and Navigation,” *2018 IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, QLD, Australia, 2018, DOI: 10.1109/ICRA.2018.8463173.

[9] B. Chen, D. Yuan, C. Liu, and Q. Wu, “Loop Closure Detection Based on Multi-Scale Deep Feature Fusion,” *Applied Sciences*, vol. 9, no. 6, Mar., 2019, DOI: 10.3390/app9061120.

[10] A. Mousavian and J. Kosecka, “Semantic Image Based Geolocation Given a Map,” *arXiv:1609.00278*, Sep., 2016, Accessed: Mar. 25, 2021, [Online], http://arxiv.org/abs/1609.00278.

[11] Y. Hou, H. Zhang, S. Zhou, and H. Zou, “Use of Roadway Scene Semantic Information and Geometry-Preserving Landmark Pairs to Improve Visual Place Recognition in Changing Environments,” *IEEE Acces*s, vol. 5, pp. 7702-7713, 2017, DOI: 10.1109/ACCESS.2017.2698524.

[12] T. Naseer, G. L. Oliveira, T. Brox, and W. Burgard, "Semantics-aware visual localization under challenging perceptual conditions," *2017 IEEE International Conference on Robotics and Automation (ICRA)*, Singapore, 2017, DOI: 10.1109/ICRA.2017.7989305.

[13] H. Xu, H. Zhang, E. Yao, and H. Song, "A Loop Closure Detection Algorithm in Dynamic Scene," *International Conference on Computer, Communication and Network Technology (CCNT 2018)*, 2018, DOI: 10.12783/dtcse/ccnt2018/24714.

[14] S. Garg, N. Sunderhauf, M. Milford, and N. Suenderhauf, "LoST? Appearance-Invariant Place Recognition for Opposite Viewpoints using Visual Semantics," *arXiv :1804.05526*, 2018, Accessed: Nov. 23, 2018, [Online], https://github.com/oravus/lostX.

[15] P. Wu, J. Wang, C. Wang, L. Zhang, and Y. Wang, "A novel fusing semantic-and appearance-based descriptors for visual loop closure detection," *Optik*, vol. 243, Oct., 2021, DOI: 10.1016/J.IJLEO.2021.167230.

[16] B. Chen, X. Song, H. Shen, and T. Lu, "Hierarchical Visual Place Recognition Based on Semantic-Aggregation," *Applied Sciences*, vol. 11, no. 20, Oct., 2021, DOI: 10.3390/APP11209540.

[17] G. Singh, M. Wu, S.-K. Lam, and D. Van Minh, "Hierarchical Loop Closure Detection for Long-term Visual SLAM with Semantic-Geometric Descriptors," *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, Indianapolis, IN, USA, 2021, DOI: 10.1109/ITSC48978.2021.9564866.

[18] Z. Yuan, K. Xu, X. Zhou, B. Deng, and Y. Ma, "SVG-Loop: Semantic-Visual-Geometric Information-Based Loop Closure Detection," *Remote Sensing*, vol. 13, no. 17, Sep., 2021, DOI: 10.3390/RS13173520.

[19] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, DOI: 10.1109/CVPR.2017.549.

Saba Arshad

2015 Department of Computer Science, Pir Mehr Ali Shah Arid Agriculture University, Pakistan(Bachelor)

2017 Department of Computer Science, COMSATS University Islamabad, Pakistan(Master)

2018~Present Department of Control and Robot Engineering, Chungbuk National University, Korea(Ph.D.)

Interests: SLAM, Robot Vision, Loop Closure Detection, Map Data Representation

Gon-Woo Kim

2008 Assistant Professor, Electronics and Control Engineering, Wonkwang University

2012 Assistant Professor, School of Electronics Engineering, Chungbuk National University

2014 Associate Professor, School of Electronics Engineering, Chungbuk National University

2021~Present Professor, Department of Intelligent Systems and Robotics, Chunbuk National University

Interests: Navigation, Localization, SLAM