Robust Visual Place Recognition Method for Robot Facing Drastic Illumination Changes

Zhen Li¹, Lin Zhou² and Zeqin Lin³*

¹School of Cosmetics and Artistic Designing, Guangdong Food and Drug Vocational College, GuangDong, 510520, China.
²Department of Mechatronics Engineering, Foshan University, GuangDong, 528000, China.
³School of Mechatronic Engineering and Automation, Foshan University, GuangDong, 528000, China
*Corresponding author’s e-mail: martinkeel@foxmail.com

Abstract. The robustness of visual place recognition determines the accuracy of the SLAM to construct the environmental map. However, when the robot moves in the outdoor environment for a long time, it must face the challenge of drastic illumination changes (time shift, season or rain and fog weather factors), which leads to the robot’s ability to identify places is greatly restricted. This paper proposes a method for visual place recognition that is more robust to severe illumination changes. First, a generative adversarial network is introduced into visual SLAM to enhance the quality of candidate keyframes, and consistency of geographic locations corresponding to images before and after quality enhancement is evaluated, and then the image descriptor with illumination invariance is established for the robot’s new observation image and keyframes of the map. Finally, the performance of the method in this paper is tested on the public dataset. The experimental results show that this method is conducive to improving the quality of environmental map nodes, and enables the robot to show a highly robust ability for visual place recognition in the face of severe illumination changes, which provides a powerful tool for robot to build an accurate environmental map.

1. Introduction
Visual simultaneous location and mapping (VSALM) is the key technology for mobile robots to implement autonomous navigation. It mainly consists of four parts: estimation of front-end, optimization of back-end, loop detection and construction of a map [1]. Visual SLAM uses the inter-frame estimation of the visual odometry and the existing map to achieve the robot's own location. Based on the results of location, it achieves an incremental map construction. However, the motion estimation errors of front-end based on adjacent frames will continue to accumulate and gradually increase the trajectory drift, which will lead to the construction of a wrong environment map. In addition, the robot may repeatedly visit the place that it has been to. Therefore, loop detection is required to perform online place recognition to determine whether a closed-loop situation occurs currently. The final estimation of front-end and results of loop detection will be used as input information for optimization of back-end. In the end, the output result is an environment map consistent with the real scenario. In the entire visual SLAM process, it can be seen that visual place recognition has a crucial impact on building an accurate environment map.
When a mobile robot moves in an outdoor environment, it will inevitably undergo the influence of changes in ambient illumination caused by time shifts, seasonal and weather conditions, which has an impact on the robot's visual place recognition. As shown in Fig. 1, when the illumination at the same place changes dramatically in different times, the scene shows great differences, which easily causes false positive and false negative problems. The purpose of this paper is to propose a highly robust method of visual place recognition to enhance the adaptability of visual SLAM to complex illumination changes.

![Illumination changes at the same place](image)

Therefore, this paper proposes a method for visual place recognition with strong robustness in the face of severe illumination changes and presents three contributions as follows:

- A generative adversarial network is introduced into visual SLAM for the first time to enhance the image quality of candidate keyframe of severe illumination changes, which lays the foundation for generating a high-quality environment map.
- Proposing a consistency evaluation method of the corresponding geographic location before and after the enhancement of the candidate keyframe’s quality, which provides a guarantee for the authenticity expression of the map nodes.
- An invariant image descriptor is constructed to ensure high robustness of image matching and visual place recognition.

The rest of this paper is organized as follows. Section II analyzes the current research status at home and abroad; Section III introduces the methods of this paper; Section IV explains the experimental scheme, results and conclusions of this paper; Section V summarizes the paper and discuss the future work.

2. Related Work

The existing representative visual SLAM systems (FAB-MAP [2], ORB-SLAM [3], SeqSLAM [4]) mainly generate image descriptors based on traditional hand-crafted features, and then calculate the similarity of image descriptors to accomplish the task of place recognition. Among them, the FAB-MAP system uses the SIFT feature to construct a bag-of-words model for online place recognition. It has achieved good results in a relatively stable outdoor environment. However, traditional hand-crafted features represented by SIFT is too weak to illumination changes, which lead to the FAB-MAP system can’t adapt to the changing environment of illumination. In addition, the ORB-SLAM system used the bag-of-words model constructed by the FAST feature extractor and BRIEF feature descriptor also has the same shortcomings as FAB-MAP when place recognition in changing illumination conditions. Compared to the first two systems that use single image descriptor for place recognition, the SeqSLAM system uses a longer image sequence for visual place recognition and achieves better robustness, but the ability to adapt to severe illumination changes still exists major limitations.
In recent years, with the development of deep learning method and theory of computer vision, researchers have tried to use deep convolutional neural networks for autonomous learning of image features [5], and applied these features to visual SLAM localization and place recognition or loop detection. The robustness of SLAM is further improved compared to the traditional hand-crafted features. Chen, et al. [6] proposed a CNN-based visual place recognition algorithm, which is better than the method based on features of hand-crafted. Gao, et al. [7] proposed an unsupervised learning method to extract image features and used the similarity matrix between image sequences to detect closed loop, it has achieved good results in public datasets; Hou, et al. [8] used the PlaceCNN model to extract image features and had certain robustness under varying illumination conditions; Bai, et al. [9] proposed a method of SeqCNNSLAM, which combines the deep learning model AlexNet and SeqSLAM methods to simultaneously solve the robustness of view and illumination in the scene, but this method requires online adjustment of relevant parameters when place changes and the process is tedious.

Existing visual SLAM systems have done a lot of research that directly affect the accuracy of visual place recognition, such as feature extraction and image descriptor’s generation. However, the robustness of the method to improve the place recognition from the perspective of enhancing the quality of candidate keyframe is seldom seen. At present, generative adversarial network(GAN) [10] is used to generate image. Isola, et al. [11] proposed a solution to the generation of original images to target images, which named pix2pix, requiring paired training samples in the training set; Zhu, et al. [12] proposed a CycleGAN model, which does not need to establish one-to-one mapping between training data, and had achieved good results. Anoosheh, et al. [13] proposed an unsupervised learning method-TodayGAN, which can restore night images to day images and apply them to locate a place. Although this method verifies the model's ability to recover the scene, it does not evaluate the consistency of the geographic location of the image before and after the scene is restored, and uses the VLAD [14] algorithm to encode the SIFT features. The generated image descriptor has obvious limitations in the robustness of image matching under changing environment of illumination.

3. Proposed Approach

This paper proposes a visual place recognition method with high robustness against severe illumination changes. As shown in Fig. 2, the method’s design and implementation are specifically divided into four sub-modules: first, the generative adversarial network to enhance the candidate keyframe; Second, global image descriptors for assessment of geographical locations’ consistency corresponding to the image before and after quality enhancement; third, construction of illumination invariant image descriptors; fourth, image matching and place recognition. A detailed description of each sub-module will be completed in Section III.A and Section III.B.
3.1 Enhancement of Illumination Quality of Candidate Keyframes and Evaluation of Geographical Authenticity

The quality of the environment map constructed by the robot depends on the quality of the candidate keyframes. Therefore, the newly observed image sequence of the robot will be further enhanced after being filtered by the candidate keyframe method. Then the quality of subsequent maps will be guaranteed and the image matching process that plays a decisive role in place recognition is also performed between high-quality keyframes, thereby reducing the degree of interference of illumination changes on place recognition. In addition, it is a critical step whether the geographical location described by the candidate keyframe image with enhanced illumination quality is the same as the original real geographic location. The enhancement about candidate keyframe and the assessment of geographical location authenticity are described as follows:

3.1.1 The method uses the TodayGAN to enhance the illumination quality of candidate keyframes.

In the process of visual place recognition, it is inevitable to encounter scenes with strong illumination changes. This paper uses the TodayGAN model to enhance the illumination quality of candidate keyframes. As shown in Fig. 3, the generator network G performs quality enhancement on the current candidate keyframe image x, and generates an enhanced image G(x). In order to ensure that the enhanced image and the real candidate keyframe image are consistent in content, the generator network F performs an inverse process on the generated enhanced image G(x) to obtain an image F(G(x)), and ensures the image F(G(x)) is basically the same as the image x. At the same time, the discriminator network D discriminates the enhanced image G(x) and the map keyframe image y. The discriminator network are mainly composed of three parts: a discrimination network $D_t$ about image texture, a discrimination network $D_c$ about image color, and a discrimination network $D_g$ about image gradient. Three parts work together to ensure that the illumination quality of candidate keyframe images are effectively enhanced.
3.1.2 Evaluation of the authenticity of the geographic location of the candidate keyframe image expression after the enhancement of the illumination quality

If the candidate keyframe images after enhancing illumination quality can still express the true geographical location, then the images before and after the quality enhancement corresponding to this location should maintain a very high similarity. The multiple similarity curves calculated by candidate keyframe image sequences before and after enhancement should show highly consistent waveforms and peak positions. Based on this fact, the method uses the global image descriptor GIST [15] to calculate the similarity matrix of the candidate keyframe image sequence before and after the enhancement of the illumination quality. Next, the method analyzes each row of the similarity matrix to evaluate whether the candidate keyframe image can maintain the authenticity of the original geographical location after being enhanced by illumination quality. This is verified in Section IV.B of the experiment.

3.2 Illumination Invariant Image Descriptor and Place Recognition Method

The method in this paper uses the local feature package with illumination invariance to generate robust image descriptors, and then performs image matching by calculating the similarity of image descriptors to achieve the purpose of place recognition. This method extracts the DELF[16] feature with high invariance to illumination. The learning process of this feature is mainly connected to the attention mechanism module at the output of the conv4-x layer of ResNet50, so that the keypoints of the image are provided with robustness to changes in the external environment. The image descriptor is shown in Equation 1:

\[ D(i) = \{f(1), f(2), ..., f(i)\} \]  

Among them: \( D(i) \) represents the image descriptor, and \( f(i) \) represents the vector of the \( i^{th} \) feature point in the image.

Image descriptor the similarity between image descriptors is equivalent to image matching and place recognition, as shown in Equation 2:

\[ S(A, B) = \frac{M(A,B)}{N(A)+N(B)-M(A,B)} \]

Among them: \( M(A, B) \) represents the number of matching, \( N(A) \) represents the number of feature points of image \( A \), \( N(B) \) represents the number of feature points of image \( B \), the value range of similarity \( S(A,B) \) is \([0,1]\).

4. Experimental Evaluation

4.1 Datasets

In order to verify the robustness of the illumination invariance of the method in this paper, the Oxford RobotCar Dataset [17] was used as the test image for visual place recognition. The dataset contains
image sequences collected by a camera for a long time along a predetermined route in the urban area of
Oxford, England. The dataset is more complex, including the dramatic changes in illumination caused
by changes throughout the four seasons, rain and snow, etc. Scenarios are more in line with the
requirement of this experiment to verify the robustness of the illumination changes. In this paper, 100
day and night images are selected from the entire test dataset to form two subsets. The image sequences
in both subsets describe the same observation path, and the frame-by-frame images in the two sequences
correspond to each other.

4.2 Evaluation of the Consistency of the Geographic Locations Described before and after the
Enhancement of the Illumination Quality of Candidate Keyframes

Whether the corresponding geographic locations of the candidate keyframes are consistent before and
after the enhancement of the illumination quality can be achieved by using human eye observation
method and image descriptor-based evaluation method. As shown in Fig. 4, the experiment performed
a batch of enhancements on the image quality of the night image data to obtain a corresponding high-
brightness image sequence. The research team used a random detection method to compare the images
content before and after the quality enhancement and the consistency of geographical location with
human eyes. There is no distortion in the description of the geographical location of the candidate
keyframes. The comparison results preliminarily verify the validity of the method, and then the
experiment continues to use the image descriptor-based evaluation method to further verify the validity
of the method.

In fact, if the image descriptors before and after the enhancement of the illumination quality of the
candidate keyframes can still maintain a high degree of similarity, it will indicate that the geographic
locations described by successive images are consistent. Based on this logic, this experiment uses GIST
as the global descriptor of the image, and calculates the similarity matrix of the two image sequences
before and after the quality enhancement. The horizontal axis represents the night image sequence, and
the vertical axis represents the night image sequence with enhanced quality. As shown in Fig. 5, the
similarity matrix has a very continuous and significantly higher brightness than the diagonals of the
outer area, which indicates that the geographic location described by the night image after illumination
enhancement still maintains a high consistency with the previous one.

Fig. 4. Before and after enhancement of illumination quality of candidate keyframe images
Fig. 5. Image similarity matrix before and after candidate keyframe’s enhancement

(a) The images matching similarity curves of 6,16,26,36,46

(b) The images matching similarity curves of 56,66,76,86,96

Fig. 6. Similarity matrix quantitative sampling analysis: geographical authenticity evaluation of candidate keyframes with enhanced illumination quality
4.3 Verification of the Method’s Robustness

In the previous Section, it was verified that the candidate keyframes would not have a distorted description of the geographical position after the enhancement of the illumination quality in this paper, which provides a prerequisite guarantee for the further visual place recognition. This Section continues to verify the robustness of the proposed method in visual place recognition, and performs necessary experiments. First, verifying the robustness improvement with enhanced candidate keyframe quality on visual place recognition. Second, verifying that the proposed method performs better than other representative methods in robust place recognition.

In the experiment, the illumination-invariant image descriptor proposed in Section III.B was used to quantify and express the night image sequence A, the enhanced night image sequence A', and the day image sequence B. Then, the original feature package based on the image matching method calculates A and B, A' and B, and obtains an image similarity matrix, that is, judgment matrices M1 (as shown in Fig. 7) and M2 (as shown in Fig. 8) for visual place recognition. The horizontal axis is the day image sequence, and the vertical axis is the enhanced night image sequence.

It can be concluded from Fig. 7 and Fig. 8 that after the illumination quality enhancement of the candidate keyframes, a higher judgment matrix can be obtained when used for visual place recognition calculation. Compared with the intermittent diagonal of matrix M1, the diagonal of matrix M2 is more continuous and has higher brightness, which shows the method with a generative adversarial network introduced into visual SLAM for candidate keyframe's enhancement has a positive effect on improving the robustness of visual place recognition.

In order to verify the robustness of the method in visual place recognition, the experiment compared the method in this paper with the representative method. First, the candidate keyframes are subjected to the calculation process described in Section III.A to enhance the illumination quality, and then the image descriptors are constructed based on the existing typical hand-crafted features SIFT and SURF, and the existing representative deep learning feature SuperPoint [18]. According to the generation process of the discrimination matrix M2, similar discriminant matrices for visual place recognition are sequentially calculated.

It can be seen from Fig. 9 that the identification degree of the place recognition shown by the matrix DELF_GAN and the matrix SUPERPOINT_GAN calculated based on the learning feature package is much better than the previous two matrices SFI T_GAN and SURF_GAN based on the hand-crafted feature package. When the illumination changes drastically, the place recognition’s ability shown by Fig.
9(a) and Fig. 9(b) is almost ineffective. Moreover, the matrix DELF_GAN has continuous highlighted diagonal area than the matrix SUPERPOINT_GAN, which means that the proposed method in this paper performs better in robust visual place recognition.

![Matrix Comparison](image_url)

Fig. 9. Comparative analysis of similarity matrix generated by our method and other methods.

The experiment has quantitatively analyzed and compared the row data of the above four matrices. As shown in Fig. 10, it shows that the method of this paper has obvious advantages over other methods in visual place recognition at a certain location.
Fig. 10. Sampling for comparative analysis of the similarity matrix for visual place recognition: take the location number 49 as an example.

5. Conclusion
In view of the limitations of the robustness of the method for visual place recognition of mobile robots under severe illumination changes, this paper proposes a robust place recognition method based on the enhancement of candidate keyframes' illumination quality. First, the candidate keyframes are enhanced with illumination quality and authenticity assessment of geographical location, followed by similarity calculations using the proposed invariant image descriptors as the basis for place recognition. Finally, on the dataset of complex environment with drastic changes of illumination, it is verified that the effectiveness of the image after illumination quality enhancement to the real geographical location expression and the significance of the enhancement of illumination quality to the robustness of place recognition. At last, confirming the robustness of the method in this paper is better than other typical methods. This paper provides a valuable solution for the robust place recognition of mobile robot visual navigation systems in the face of drastic illumination changes.

The future work will focus on how to obtain a new generative adversarial network model with better performance based on NAS technology to enhance the keyframe’s quality of the map constructed by the robot. In addition, semantic information will also be integrated into the map to enhance the ability of human-computer interaction.

Acknowledgements
Science and Technology Projects in Guangzhou(202102080591);Characteristic Innovation Project of Colleges and Universities in Guangdong Province(2020KTSCX254);Guangdong Medical Research Fund Project(A2020550);Natural Science Foundation of Guangdong Food and Drug Vocational College(2019ZR04);The Natural Science Foundation of Guangdong under Grant(2020A1515110255);

References
[1] Lowry, S. , Sunderhauf, N. , Newman, P. , et al. (2016) Visual place recognition: a survey. IEEE Transactions on Robotics, 32(1), 1-19.
[2] Cummins, M. , & Newman, P. . (2008). Fab-map: probabilistic localization and mapping in the space of appearance. Int.j.robot.res, 27(6), 647-665.
[3] Mur-Artal, R. , Montiel, J. M. M. , & Tardos, J. D. . (2015) Orb-slam: a versatile and accurate monocular slam system. IEEE Transactions on Robotics, 31(5), 1147-1163.
[4] Milford, M. J., & Wyeth, G. F. (2012). SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In: Robotics and Automation (ICRA). IEEE International Conference on. IEEE North Rhine-Westphalia. pp. 1643-1649.

[5] Zheng, L., Yang, Y., & Tian, Q. (2017). Sift meets cnn: a decade survey of instance retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(5): 1224-1244.

[6] Chen, Z., Lam, O., Jacobson, A., & Milford, M. (2014). Convolutional neural network-based place recognition. Computer Science.

[7] Xiang, G., & Tao, Z. (2017). Unsupervised learning to detect loops using deep neural networks for visual slam system. Autonomous Robots, 41:1-18.

[8] Hou, Y., Zhang, H., & Zhou, S. (2015). Convolutional neural network-based image representation for visual loop closure detection. In: 2015 IEEE international conference on information and automation. IEEE. LiJiang. pp. 2238-2245.

[9] BAI, Dongdong, WANG, Chaoqun, ZHANG, & Bo, et al. (2018). Cnn feature boosted seqslam for real-time loop closure detection. Chinese Journal of Electronics, v.27(03), 48-59.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. (2014) Generative adversarial nets. In: Advances in neural information processing systems. Montreal. pp. 2672-2680.

[11] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2016). Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision & Pattern Recognition. Las Vegas. pp. 1125-1134.

[12] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. Venice. pp. 2223-2232.

[13] Anoosheh, A., Sattler, T., Timofte, R., Pollefeys, M., & Gool, L. V. (2019). Night-to-Day Image Translation for Retrieval-based Localization. In: 2019 International Conference on Robotics and Automation (ICRA). Montreal. pp. 5958-5964.

[14] Jegou, H., Douze, M., Schmid, C., & Perez, P. (2010). Aggregating local descriptors into a compact image representation. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE. San Francisco. pp. 3304-3311.

[15] Oliva, A. (2006). Building the gist of a scene: the role of global image features in recognition. Progress in Brain Research, 155(2), 23-36.

[16] Noh, H., Araujo, A., Sim, J., Weyand, T., & Han, B. (2017). Large-Scale Image Retrieval with Attentive Deep Local Features. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE. Venice. pp. 3456-3465.

[17] Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., & Stenborg, E., et al. (2018). Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. Salt Lake City. pp. 8601-8610.

[18] Detone, D., Malisiewicz, T., & Rabinovich, A. (2018). Superpoint: self-supervised interest point detection and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City. pp. 224-236.