Image-Based Place Recognition Using Semantic Segmentation and Inpainting to Remove Dynamic Objects

Linrunjia Liu(✉), Cindy Cappelle, and Yassine Ruichek

CIAD, Univ. Bourgogne Franche-Comté, UTBM, Belfort, France
{linrunjia.liu,cindy.cappelle, yassine.ruichek}@utbm.fr

Abstract. Place recognition is an important step in intelligent driving, allowing the vehicle recognize where it is to plan its route. Obtaining distinguishable features can ensure the success of image-based place recognition. However, generating robust features across drastically appearance changing images is still a challenging problem. Deep features are frequently chosen instead of local features in the tasks of place recognition following the development of convolutional neural networks. But even the deep features generated by powerful neural models can cause unsatisfactory recognition results. This is perhaps due to a lack of information selecting process. The technology of semantic segmentation allows recognizing and classifying image information. Semantic segmentation followed by image inpainting provide a possibility of detecting, deleting and reconstructing annoying information.

This paper proves that dynamic information present in images such as vehicles and pedestrians damages the performance of place recognition and proposes a feature extraction system that includes a step to decrease the presence of dynamic information of an image. This system is composed of two stages: 1) dynamic objects detection and removing, 2) image inpainting to reconstruct the background of removed regions. Objects detection and removing consists of deleting unstable objects recognized by semantic segmentation method from images. Image inpainting and reconstructing deals with generating inpaint-images by repairing missing regions through image inpainting method. The robustness of the proposed approach is evaluated by comparing to the non-selecting deep feature based place recognition approaches over three datasets.

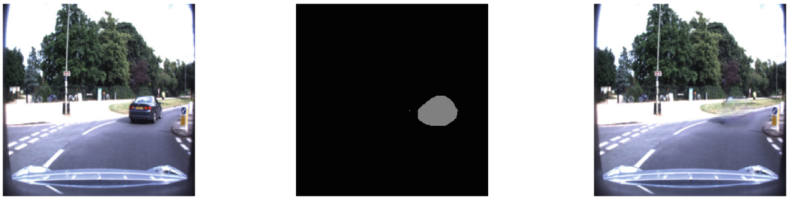
Keywords: Place recognition · Image inpainting · Semantic segmentation

1 Introduction

Place recognition for vehicle localization aims to select the image in a long-term robotcar dataset, which represents the same place as a query image. Extracting

Supported by CSC.

A. El Moataz et al. (Eds.): ICISP 2020, LNCS 12119, pp. 262–270, 2020.
https://doi.org/10.1007/978-3-030-51935-3_28



(a) Original-image (b) Mask (c) Inpaint-image

Fig. 1. (a) Original-image. (b) Mask generating by FCN based semantic segmentation. Regions depicted in grey are dynamic objects recognized by FCN and are seen as missing regions by EdgeConnect system based inpainting. (c) Inpaint-image getting from EdgeConnect system.

robust features to represent an image is then a critical step of place recognition. Since feature generation technique of Convolutional Neural Networks (CNNs) has been gradually applied into the tasks of intelligent vehicle [18], place recognition has also progressed significantly. Nevertheless, due to the severe appearance changes in long-term robot navigation, there are still many challenges and problems to achieve efficient image based place recognition. In addition to dramatic changes in weather conditions, illumination and viewpoint, dynamic objects, such as vehicles and pedestrians, also contribute to the appearance changes and make it difficult to distinguish the same place at different time.

Since learning invariant features of an image can improve the performance of place recognition, what about deleting all the dynamic objects in an image and extract features by the remaining stable background? The development of semantic segmentation and image inpainting approaches makes it possible.

The main goal of this paper is to provide a new method that addresses the challenges in traffic variations of image-based place recognition. Inpaint-images are generated from original images from which dynamic objects–objects classified as cars, buses, trucks, and pedestrians–are removed and the stable background is completed by a two-stages process (Fig. 1): 1) objects detection and removing and 2) image inpainting and reconstructing. Firstly, we use FCN (Fully Convolutional Networks) [8] to detect the dynamic objects in the original images and label them into masks. Then, the parts labeled as masks, which are then seen as missing regions, are filled by fine details using EdgeConnect [11] based inpainting approach. The outputs of Edge-Connect are called inpaint-images in this paper. A CNN network is then used to extract global features from these inpaint-images. Image retrieval process finally performed using Euclidean distance as a metric.

In this paper, the source of images for experiments are selected from three publicly available datasets: St.lucia, CMU, and Oxford RobotCar [10]. Three sequences of different routes are chosen in each dataset and their inpaint-images are generated independently. Then, four different CNNs [6,7,15,17] are used to learn the features of both inpaint-images and original-images (sequences without dynamic detection and inpainting). In each test, comparing by using CNN features of original-images in the process of image retrieval, a noticeable improvement is obtained on the task of place recognition performed using the inpaint-images CNN features.

2 Related Work

Generating distinguishable features to represent images with severe appearance variation determines the results of place recognition in changing environments.

Traditional methods design hand-craft local features that are illumination-, viewpoint-, and scale-invariant, such as speeded-up robust features (SURF) [1] and scale-invariant feature transforms (SIFT) [9].

Because of the generalization of features learned by representative Convolutional Neural Networks (CNNs), a recent trend is to exploit CNN features into place recognition. Papers [2, 14] use holistic features of images. [2] concludes that different layers of CNN features have different effect in changing environments. In the work of [14], different layers of CNN features are connected together to represent an image. This kind of holistic features contains more information but also increases computation complexity.

Though holistic CNN features outperform classical features, they sometimes fail to recognize correctly the same places when facing the realistic street scenes with significant appearance changes. Increasing the features complexity to learn as much as possible different appearance representations was the goal of many place recognition approaches. [4] observes the environmental changes over the course of day and night and combines the different features together. [12] learns systematic appearance changes so that to predict the changes under different environmental conditions. These works require lots of training data and didn't show generality abilities in the different conditions as with training data.

Some approaches choose to make the features invariable such as using landmarks features. [16] applies Edge Boxes to extract landmark proposals and then learns the features of these proposals by ConvNet. Regions extracted by a landmark detector in [5] are deep local features and robust to changes in viewpoint. These works need the environmental conditions with distinguished landmarks which is not always happen in real world.

Our approach follows the idea of using invariable features, but instead of selecting useful information, we delete the dynamic information, as cars, vehicles, pedestrians, without damaging context information. This is achieved through semantic segmentation to detect objects to remove and image inpainting to reconstruct the removed regions. To our best knowledge, this is the first work that tackles dynamic information removal and background information reconstruction in place recognition.

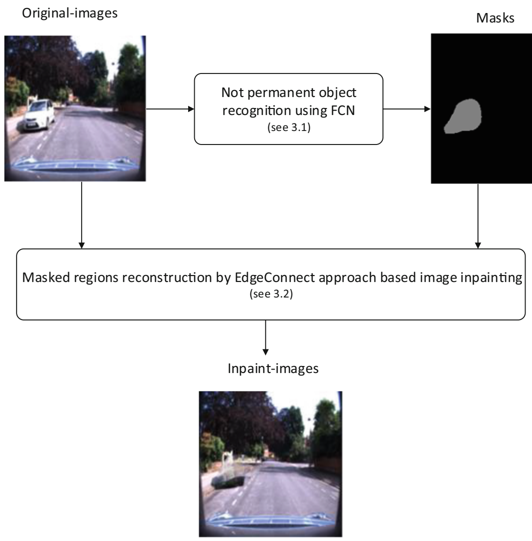

Original-images
Not permanent object recognition using FCN (see 3.1)
Masks
Masked regions reconstruction by EdgeConnect approach based image inpainting (see 3.2)
Inpaint-images

Fig. 2. Proposed system to decrease the impact of dynamic information of the image and reconstruction of its background.

3 Proposed System

Image-based localization aims to recognize the image in a local navigation dataset that represents the same place as a query image (the current image) based on their appearance.

To describe each image, a CNN model is used to extract image features. After comparing the similarity of a query image with each images in the reference sequence, the image which is the most similar to the query image is selected.

The proposed method that tries to decrease the impact of the dynamic objects in images is based on two-stages (Fig. 2): 1) objects detection and removing, and 2) image inpainting and reconstructing.

3.1 Proposed Dynamic Objects Removing

The first step of the proposed approach is to eliminate the objects of the image that are not permanent in the scene (as vehicles, pedestrians, $\cdots$). For that, FCN [8] is used to recognize these objects in image and make local predictions of these objects. Given an image of any size as input, this end-to-end fully convolutional network (FCN) provides an output image of the same size with dense prediction of each pixel. With this prediction, the position of each object can be easily marked. Without considering the position of the other objects, only the positions of pedestrians and vehicles are recognized and marked in this paper. These marked dynamic objects are not directly removed from the image, but are seen as missing regions in the image inpainting method that is chosen. The 3.2 part will explain this step in detail.

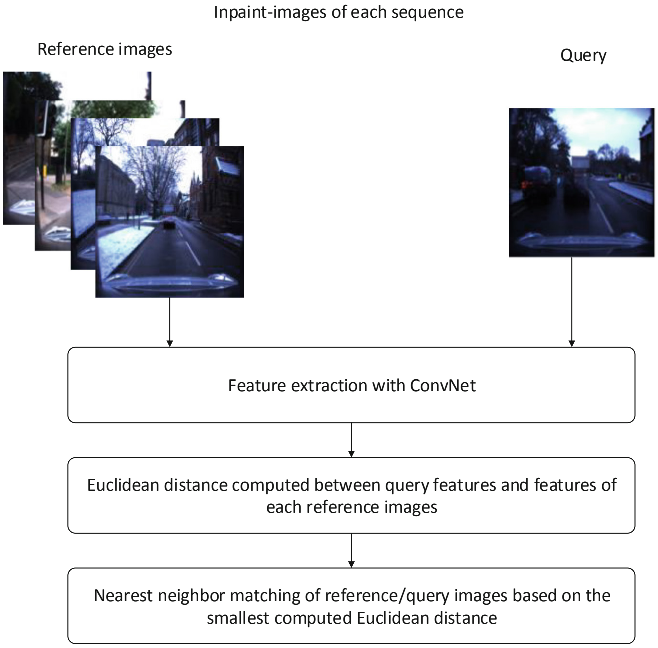

Inpaint-images of each sequence
Reference images
Query
Feature extraction with ConvNet
Euclidean distance computed between query features and features of each reference images
Nearest neighbor matching of reference/query images based on the smallest computed Euclidean distance

Fig. 3. The process of place recognition.

To get the output of FCN, a fine-tuned classification model is needed. In our work, the resnet101 Model which is trained on PASCAL VOC dataset is picked. The output image of FCN will be considered as an input mask image in image inpainting step.

3.2 Images Repairing with Image Inpainting Approach

Image inpainting means reproducing the missing regions of a ground truth image as if it has not being corrupted. In our paper, we choose a state-of-art image inpainting method: EdgeConnect [11]. Given an original image as the ground truth image and a mask image generated by FCN, the output of EdgeConnect is our expected inpainting-image. In the process of inpainting, EdgeConnect will see the dark part of mask image as background and recognize the rest part as missing regions which will be reconstructed later.

The reconstructing of EdgeConnect consists of two parts. First, it trains an edge generator to hallucinate edges in the missing regions. Then, it needs an image completion network to combine the color and texture information of the existing parts with the edges in the missing parts of the image. Combining the two parts of work together, an end-to-end trainable network is proposed to reconstruct missing regions with fine details. By solving two computer vision problems: Image-to-Edges and Edges-to-Image, EdgeConnect improves the inpainting quality of the corrupted images.

Table 1. Testing datasets description.

Datasets	Sequences	No. of frames	Variations in			Resolution	
			Traffic	Weather	Season	Original	Resized
St.Lucia	different time in a day	227	Mild	None	None	640 × 480	224 × 224
CMU	summer V.S fall	207	Severe	Moderate	Moderate	1024 × 768	224 × 224
Oxford Robotcar	summer V.S winter	282	Severe	Moderate	Severe	1280 × 960	224 × 224

3.3 Place Recognition Approach

By comparing the Euclidean distances between the feature vector of query image and of every reference images, a nearest matching reference image is found per query image. The place of this nearest neighbor reference image is then considered as the place of the query image. And the featurization process in this paper is to learn the deep feature of each image by a ConvNet. The whole place recognition process is shown in Fig. 3.

4 Experimental Setup

4.1 Datasets

In order to evaluate the proposed approach, three sequences are selected from three datasets which deal with mild, moderate and severe variations in environmental conditions, including illumination, traffic and pedestrians, weather and seasonal changes. Details about the datasets are described below and summarized in Table 1.

- St. Lucia Dataset[1]: The St.Lucia Dataset was captured in a suburb. The car pass through some routes several times.
- CMU Seasons Dataset[2]: The CMU Seasons dataset is a part of CMU Visual localization dataset created by *Badino et al.* [3]. We use the left mono images that represent the scenes of the same route in summer and fall. The summer images are used as references and fall images are used as queries.
- Oxford RobotCar Dataset: The Oxford RobotCar Dataset [10] is composed of over 100 different sequences of Oxford. We choose a fixed route and compare it between summer and winter.

4.2 Data Pre-processing

One fixed route in each dataset is selected. The sequence of each route is a subset of two traversals which were taken over different time. After considering

[1] https://wiki.qut.edu.au/display/cyphy/OpenRatSLAM+datasets.
[2] https://www.visuallocalization.net/datasets/.

Table 2. Place recognition precision comparison when using original-images and inpaint-images with four featurization ConvNets.

Precision (%)	VGG19		NIN_ImageNet		AlexNet		bvlc_GoogleNet	
	Original	Inpaint	Original	Inpaint	Original	Inpaint	Original	Inpaint
St.Lucia	98.24	**99.56**	97.80	**99.12**	98.68	**99.56**	99.56	99.56
CMU	84.54	**86.96**	92.75	**96.14**	87.92	**92.75**	87.44	**92.27**
Oxford RobotCar	90.78	**98.58**	86.88	**90.43**	88.65	**91.84**	96.10	**98.23**

one traversal as the query traversal, the images in the other traversal are selected one by one to make sure that each query image has a corresponding image, which was captured at the same place. Then, each query image and its corresponding image are renamed with the same number from 1 to the length of the query sequence according to their positions. Since the same number represents the same place and adjacent numbers represent adjacent places, it is easy to find out whether a query image is correctly matched or not by selecting the best matching images according to their numbers.

4.3 Evaluation Criteria and Baselines

Based on the evaluation of [13], a precision parameter to benchmark visual place recognition is defined as the percentage of query images whose predicted place is approximately same as its real place: the query image and its nearest reference image are seperated by no more than 3 images (about 5 m of their ground truth positions).

Table 2 shows the results obtained using deep feature matching on the original query and reference images to benchmark the proposed method.

The method to generate the deep feature vector of an image is applied using four different ConvNets: VGG19, NIN_ImageNet, AlexNet, bvlc_GoogleNet. The best results obtained using the proposed approach are based on layer con5_1 in VGG19, conv3 in NIN_ImageNet, conv4 in AlexNet and pool3/3×3_s2 in bvlc_GoogleNet.

The four ConvNets are detailed as below:

- VGG19. The model used here [15] is an improved version of VGG19 in the ILSVRC-2014 competition.
- NIN_ImageNet [7]. It is a small model for ImageNet, yet fast to train and has slightly better performance than AlexNet.
- AlexNet. AlexNet model is proposed in [6] and is trained on images from ImageNet.
- bvlc_GoogleNet. This model was trained by Sergio Guadarrama[3] and it is a replication of the work in [17].

[3] https://github.com/BVLC/caffe/tree/master/models/bvlc_googlenet.

5 Results

Table 2 shows the place recognition precision obtained using original-images and their corresponding inpaint-images for three different datasets through four ConvNets.

Results of Table 2 shows that no matter the traffic conditions and weather conditions are moderate or severe changing, deleting the cars and pedestrians and then repairing the background permits to improve the results whatever the network used (VGG19, NIN_ImageNet, AlexNet). When using bvlc_GoogleNet to generate features of images from St.Lucia dataset, removing the dynamic objects have no impact on the performance of place recognition.

As illustrated in Fig. 1, the FCN is effective in recognizing objects and EdgeConnect can repairing images, but there are also some challenges. FCN is not sensitive to details, so it can not recognize cars which are far away. For some objects, FCN outputs are too coarse and parts of the objects remain, that will result in those objects being reconstructed by EdgeConnect unexpectly. As for EdgeConnect, the inpainting quality may be not good when faces to the images with a large missing portion, and the reconstructed regions may suffer from artifacts.

Even though, the performance of place recognition can be improved without perfect inpainted images.

6 Conclusion and Future Work

This paper proposed a new feature extraction system for visual place recognition. Before putting the whole images into a ConvNet, they are submitted to semantic segmentation and image inpainting method to decrease the presence of dynamic information of images. This method improves the performance of place recognition in changing environments, especially in changing traffic conditions.

The robustness of features extracted by the proposed system relies on the performance of semantic segmentation and image inpainting. It is a hard task for image inpainting methods to inpaint the image with a large missing portion, so improving the image inpainting performance may further improve the performance of place recognition. This is the research direction of our following work.

Acknowledgment. The authors gratefully acknowledge financial support from China Scholarship Council, and useful discussions with Tao Yang.

References

1. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006). https://doi.org/10.1007/11744023_32

2. Chen, Z., Lam, O., Jacobson, A., Milford, M.: Convolutional neural network-based place recognition (2014). arXiv:1411.1509
3. Hernan Badino, D.H., Kanade, T.: The CMU visual localization data set (2011). http://3dvis.ri.cmu.edu/data-sets/localization
4. Johns, E., Yang, G.: Feature co-occurrence maps: appearance-based localisation throughout the day. In: IEEE International Conference on Robotics and Automation, pp. 3212–3218, May 2013
5. Kanji, T.: Self-localization from images with small overlap. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4497–4504 (2016)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. **25**, 1097–1105 (2012)
7. Lin, M., Chen, Q., Yan, S.: Network in network (2013). arXiv:1312.4400
8. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015
9. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the International Conference on Computer Vision (ICCV), p. 1150 (1999)
10. Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 year, 1000 km: the Oxford robotcar dataset. Int. J. Robot. Res. **36**(1), 3–15 (2017). https://doi.org/10.1177/0278364916679498
11. Nazeri, K., Ng, E., Joseph, T., Qureshi, F.Z., Ebrahimi, M.: EdgeConnect: Generative image inpainting with adversarial edge learning (2019). arXiv:1901.00212
12. Neubert, P., Sünderhauf, N., Protzel, P.: Superpixel-based appearance change prediction for long-term navigation across seasons. Robot. Auton. Syst. **69**, 15–27 (2015)
13. Olid, D., Fácil, J.M., Civera, J.: Single-view place recognition under seasonal changes (2018). arXiv:1808.06516
14. Qiao, Y., Cappelle, C., Ruichek, Y., Yang, T.: ConvNet and LSH-based visual localization using localized sequence matching. Sensors **19**(11) (2019)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv:1409.1556
16. Sünderhauf, N., et al.: Place recognition with ConvNet landmarks: viewpoint-robust, condition-robust, training-free. In: Robotics: Science and Systems, July 2015
17. Szegedy, C., et al.: Going deeper with convolutions. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015
18. Yang, T., Cappelle, C., Ruichek, Y., Bagdouri, M.: Multi-object tracking with discriminant correlation filter based deep learning tracker. Integr. Comput.-Aided Eng. **26**, 1–12 (2019). https://doi.org/10.3233/ICA-180596