Neurocomputing 199 (2016) 114–127

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Place recognition based on deep feature and adaptive weighting of similarity matrix

Qin Li [a], Ke Li [a,*], Xiong You [a], Shuhui Bu [b], Zhenbao Liu [b]

[a] Zhengzhou Institute of Surveying and Mapping, China
[b] Northwestern Polytechnical University, China

ARTICLE INFO

Article history:
Received 12 October 2015
Received in revised form
22 February 2016
Accepted 12 March 2016
Communicated by Jungong Han
Available online 31 March 2016

Keywords:
Convolutional Neural Networks (CNNs)
Image description matrix
Similarity matrix
Place recognition

ABSTRACT

Effective features and similarity measures are two key points to achieve good performance in place recognition. In this paper we propose an image similarity measurement method based on deep learning and similarity matrix analyzing, which can be used for place recognition and infrastructure-free navigation. In order to obtain high representative feature, Convolutional Neural Networks (CNNs) are adopted to extract hierarchical information of objects in the image. In the method, the image is divided into patches, then the similarity matrix is constructed according to the patch similarities. The overall image similarity is determined by a proposed adaptive weighting scheme based on analyzing the data difference in the similarity matrix. Experimental results show that the proposed method is more robust than the existing methods, and it can effectively distinguish the different place images with similar-looking and the same place images with local changes. Furthermore, the proposed method has the capability to effectively solve the loop closure detection in Simultaneous Locations and Mapping (SLAM).

1. Introduction

Image similarity measurement is a core technique to recognize the place through measuring the image similarity. In Simultaneous Locations and Mapping (SLAM), loop closure detection is a hard problem solved by checking the similarities among candidate frames. In robot autonomous navigation, when a robot revisits a previously seen location, it is often necessary to determine robot's position just by the robot's internal sensors, termed “appearance-based navigation” [1], because the external infrastructures may be invalid in some environments such as indoors, near the tall building, and underground cavern. We can adopt the image similarity measurement to find the same place with the robot's first being there.

The key to measure the image similarity is to build a vector or matrix which can describe the distinct characteristic of image and identify it from others [2]. Generally, the methods to build the image descriptor can be divided into two categories: one is to describe the image as a whole, such as the color histogram [3], color coherence vector [2], and Gist [4,5]. Generally, the global feature suffers from poor generalization capability because of the lack of thorough understanding of the biological mechanisms [6]. The image histogram is a global feature describing the image in a holistic way, which can be easily calculated and understood. It is popular to adopt the histogram to describe image. But the histogram does not consider the spatial information of the objects in an image, images with different appearances may have the similar histograms [7]. What's more, the performance can be evidently affected by many factors, such as the resolution, illumination, and the arrival or departure of the objects, therefore, the image histogram lacks robustness.

The other is to represent the image based on the local descriptors, like Scale Invariant Feature Transform (SIFT) [8], Speed-Up Robust Feature (SURF) [9,10], which describe the salient patches around key points within the image. The local descriptors are generally quantized into visual words, then the image can be expressed as a vector of words ultimately. The method is termed Bag-of-Words (BoW) [11], which has achieved excellent performances on many vision applications, such as the object recognition, content-based image retrieval (CBIR), and image classification and annotation [12,13]. In Fast Appearance Based Mapping (FAB-MAP) [14–16], a recent technology used to solve the loop closure detection in SLAM, the BoW model is adopted to build the descriptor of each video frame. Although the local features have achieved great performance in some vision applications, they just depict the part information of objects in images, which may cause inconsistent performances in different tasks. And the local feature is still insufficiently powerful to describe the spatial and structural information of objects in the image [17], which greatly limits its

* Corresponding author.
E-mail addresses: leequer120419@163.com (Q. Li), Like19771223@163.com (K. Li), youarexion@163.com (X. You), bushuhui@nwpu.edu.cn (S. Bu), liuzhenbao@nwpu.edu.cn (Z. Liu).

http://dx.doi.org/10.1016/j.neucom.2016.03.029
0925-2312/

capability. In addition, the BoW method ignores the spatial information of image, it cannot describe the spatial relationships of objects in an image, thus the performance may be affected if the image content is changed.

There exist several problems need to be solved in place recognition [14]. Firstly, the real world is changing all the time. Two images of the same place, captured at different time, may be locally changed. For example, some new objects appear in the scene, some old objects disappear and the positions of some objects in the scene get changed, in which case, the similarity of the same place images may be low if evaluated by the traditional methods. Secondly, and more challengingly, different place images may be similar in vision, because the world is visually repetitive, such as the brick walls and dense grasses, in which case the similarity evaluated by the traditional methods may be high. All these existing problems will lead to incorrect recognition sometime.

Aiming at the challenging problems, we use Convolutional Neural Networks (CNNs) [18–22] to extract hierarchical feature which has more representative capability. Furthermore, a strategy which integrates global and local image information is proposed, and the similarity of image pair is calculated by statistics analysis on the data difference in similarity matrix, rather than just comparing individual patch or image similarity values.

The flow of determining the similarity of image pair is shown in Fig. 1.

Because the deep feature and the analysis on similarity matrix are applied in computing the similarity, compared to other methods, there are three main contributions as follows:

- *Deep feature extraction*: The CNNs use the pooling to make the upper layer cover larger region, therefore, it can generate hierarchical feature from an input image, which has strong representative capability. And the convolution operation is similar with the mechanism of human eyes, thus, the deep feature is highly invariant to translation, scaling and other deformation. As a consequence, the deep learning based descriptor can effectively represents the essence of image, and improves the performance in measuring the image similarity.
- *Spatial comparison*: To describe the spatial information, the image is divided into patches, based on which, the image description matrix is constructed. Therefore, more detailed information can be reflected, and the differences between images can be precisely described.
- *Analysis on the similarity matrix*: To improve the robustness of the measuring method, a novel adaptive weighting of similarity matrix is proposed. The probability of the same place and the weight of each patch similarity are determined by a comparing mechanism on similarity matrix, in which case, the performance of similarity measurement is greatly improved.

In order to verify the robustness of the proposed method, and validate the reliability in practical application, comprehensive experiments are conducted. The results demonstrate that the proposed method can achieve good performance, listed as follows:

- *Effectiveness*: The proposed method is more robust than others, not only the same place images with local change but also the different place images with similar-looking can be reliably recognized. What's more, it can find the revisited location with high accuracy, and effectively solve the problem of loop closure detection in SLAM.
- *Efficiency*: The proposed method can describe the image with high efficiency, it can find the similar images from the dataset within several milliseconds and finish the process of loop closure detection in real time.

In order to improve the readability of the whole paper, the major mathematical symbols, used in the later content, and their exact meaning are summarized in Table 1.

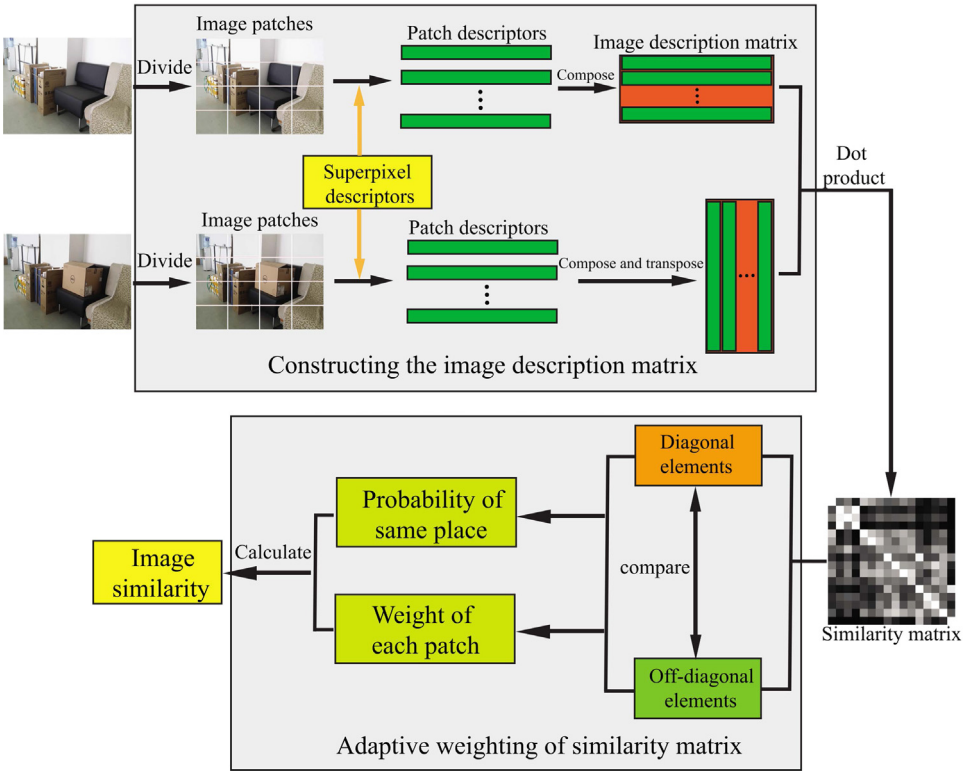

Fig. 1. The flow of determining the similarity of image pair.