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

A R T I C L E   I N F O

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

A B S T R A C T

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).

© 2016 Published by Elsevier B.V.

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 (FABMAP) [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
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.
2. Related work

In recent years, the image similarity measurement for place recognition has been explored by a lot of researches, many of which focus on the feature extraction and the construction of image descriptor to ensure the robustness of the similarity result.

A method to localize a vehicle with the visual imagery or range information was explored by Badino et al. [23]. In the method, a map by navigating the route was constructed using a GPS-equipped vehicle and the database of simple visual and 3D features were built firstly, the current location was determined by matching the current image features with those in database. Roy and Mukherjee [2] proposed a global descriptor for the image. Firstly, the low level image feature, such as color histogram, color coherence, was extracted, then the Sobel edge feature was added to get better performance, finally the Manhattan distance was adopted to determine the similarity between images. Aimed at the problems of perceptual aliasing in image matching tasks, Wang and Zhang [24] introduced a novel method to select a proper image descriptor in a specific application. The performance of the image descriptor was measured according to its power to distinguish the different objects and match the same ones in the image. Liu and Siegwart [25] adopted the uncalibrated omnidirectional camera to extract the geometric information and color features, based on which the adaptive descriptor of image was generated. The proposed method effectively solved the problem of simultaneous localization in topological regions, and achieved high computational efficiency.

Even though global descriptors are widely used in many recognition applications, the shortcoming that global descriptors ignore the spatial relation of objects in images still exists. Considering the great discriminative, many researchers adopt the local descriptors in place recognition.

Cummins and Newman [14] proposed a complete probabilistic framework for place recognition, which was even applicable in visually similar environments. Firstly, the Speed-Up Robust Feature (SURF) was extracted from the test video, and the appearance based navigation was considered as a recursive Bayesian filtering problem based on the probabilistic on top of the BoW representation. What's more, several principled approaches were explored to achieve high robustness in special environments, a bail-out strategy was proposed to improve the efficiency, and the method was evidently superior to the standard TF-IDF ranking. Han et al. [26] adopted the sparse coding representations as feature to describe the salient patch in an image, and the representations were constructed by learning many training patches from relevant dataset. The method can be widely used in image retrieval, segmentation and compression. Valgren and Lilienthal [27] proposed a straightforward method to solve the problem of appearance-based topological localization by comparing single image pairs directly using local image features. In the method, the suitable feature for the particular task was determined firstly, then the location algorithm was optimized according to the accuracy and the epipolar constraint was also introduced to further improve the performance. Hua and Hasegawa [28] proposed a novel method to extract the image feature according to the geometric structure. The extracted feature can keep high robustness in dynamic environments, and the method showed great advantages on accuracy and stability in outdoor SLAM. Fukumoto et al. [29] proposed an effective method to localize a car based on the on-vehicle video. The method adopted the Affine SIFT to extract the robust feature from Temporal Height Image (THI), and the image was represented based on the BoW model. The experimental results demonstrated that the method can effectively match the same place image without using any location devices.

Different from above mentioned methods, we combine the local and global feature to describe the image. The image is divided into patches firstly, then each patch is described as a whole. And the image is ultimately described by a matrix composed of each patch descriptor. To obtain more representative description, we adopt the CNNs to extract hierarchical feature for representing each patch property. In addition, the final similarity is computed by a novel comparing mechanism on similarity matrix, which achieves better robustness and accuracy, rather than the direct patch similarity comparison.

3. Image description matrix

In computer vision tasks, strong representations are vital for good performance. Recent researches show that good intrinsic representations are hierarchical. Deep learning based feature extractors provide a powerful framework to learn such hierarchical features. Moreover, convolution operation is close to the mechanism of human eyes’ capturing features, which makes the networks highly invariant to translation, scaling, tilt, or other deformations.

In this work, the core of image similarity measurement is to get the description vector or matrix, which can represent the essence of image. First, we segment the image into superpixels to represent the objects precisely, and meanwhile the feature of superpixel is calculated by the CNNs. Second, each image is divided into $4 \times 4$ patches with the same size, and each patch descriptor is built based on the covered superpixels. The image description matrix is constructed on the basis of the patch descriptors.

3.1. Superpixel segmentation

Recently, many researchers have explored to extracted spatial features based on segmented regions, which can effectively describe structure information of objects in the image [30]. Superpixel is a small region in an image, which contains a series of neighboring pixels with closed color, brightness, or texture [31]. Generally, the superpixel can effectively improve the ability of anti-noise, keep the original structural information, and retain the edges of objects in image. For an image, a single pixel is meaningless, and
people get the image information from the area composed of many pixels. Thus, it is essential to perform the superpixel segmentation to describe the image effectively. In our study we divide the image into superpixels to better represent the objects in image. Because the number of pixels is far more than that of superpixel in an image, the computational efficiency is also greatly improved.

Recently, many researchers have explored different superpixel segmentation methods [32–37]. Felzenszwalb and Huttenlocher [32] proposed a graph-based segmentation method, which is extremely fast. Comaniciu and Meer [36] proposed to use mean shift based segmentation, which is robust to local variations. However, these methods produce superpixels with irregular sizes and shapes which tend to straddle multiple objects. Shi and Malik [37] proposed to use Normalized Cut (NCut) for superpixel segmentation. NCut has the nice property of producing superpixels with similar sizes and compact shapes, which is expected for vision algorithms. Nevertheless, the computational performance of NCut is relatively low. Li and Chen [34] adopted the Linear Spectral Clustering (LSC) to produce compact and uniform superpixels with low computation. Liu et al. [33] proposed an entropy rate superpixel segmentation method. Achanta et al. [35] adopted the Simple Linear Iterative Clustering (SLIC) to complete the process. The method of SLIC can efficiently generate compact, nearly uniform superpixels and the simplicity of the method makes it extremely easy to implement and use.

We evaluate various superpixel segmentation methods, and find that the SLIC method [35] can preserve the boundaries of adjacent objects well, which is more reasonable for scene similarity measure. In addition, it has high efficiency to generate superpixels. Therefore, we adopt the method to segment images into regions for subsequent processing.

### 3.2. Superpixel feature

Recently, CNNs have shown remarkable advantages in various image and vision applications [18–20]. It is trained with multiple stages, where the upper layer’s inputs are the outputs of the lower layer. In addition, the pooling can make the upper layer cover larger range, therefore, CNNs can generate hierarchical feature from input data. The input and the output of each stage are sets of arrays called feature maps.

In our research, the original color images are used as the input of neural networks, thus each feature map is regarded as a two-dimensional array containing color channels of the input images. After one stage, the output feature map is treated as further abstraction of input feature map. Generally, each stage is composed of three parts: convolution operation, non-linearity transformation, and feature pooling. And the CNNs with $L$ layers can be described as a sequence of linear transformations (*operator), interspersed with non-linear symmetric squashing units like sigmoid function or tanh function (non-linear operator), and pooling operations (pool operator). The output of the $l$-th stage, represented by $F_l$, can be defined as

$$F_l = \text{pool} \left( \tanh(W_l \ast F_{l-1} + b_l) \right),$$  \hspace{1cm} (1)

where $l \in 1, \ldots, L$, $b_l$ is the bias parameter of the $l$th layer, $W_l$ is the convolutional kernel. In addition, the source image can be
regarded as the initial feature map $F_0$. Consequently, the whole networks can be constructed by stacking each stage layer upon layer.

Once the output feature maps of all layers have been generated, we adopt the bilateral interpolation to upsample them into the same size of input image and then concatenate them to produce a three dimensional arrays $F \in \mathbb{R}^{N \times H \times W}$ in which the three dimensions are the height of images $H$, the width of images $W$, and the number of feature maps $N$, then the arrays $F$ can be seen as hierarchical descriptors, which is depicted in formula (2), where $\text{up}$ operator is the process of upsampling, $\text{up}(F_l) \in \mathbb{R}^{N_l \times H_l \times W_l}$. $N_l$ denotes the numbers of feature maps in $l$th layer,

$$F = \{\text{up}(F_1), \text{up}(F_2), \ldots, \text{up}(F_l)\}. \quad (2)$$

For any area in an image, its descriptor can be produced by the arrays $F$, and each dimension of the descriptor is described by the corresponding area in a feature map.

Even though strong features can be generated by making full use of the outputs of all layers, the information in some feature maps is redundant, which will reduce the computational efficiency. Therefore, in our method we just select the feature maps of several convolutional layers to build the superpixel descriptors to improve the efficiency and ensure the quality of the features at the same time.

In order to generate appropriate features, the architectures of the CNNs should have wide layer range, in which case, the information in low layers and high ones is evidently different, then we can make full use of the complementary information to produce high representative features. Furthermore, the numbers of the feature maps in some layers should be suitable to ensure the feature with proper dimension. We use MatConvNet [38] and public available pre-trained model to extract hierarchical features, which save a lot of work in our research. The model adopted in our method is “imagenet-vgg-f”, which contains 21 layers, and the architecture of the model can easily produce high representative features with proper dimension. The feature maps in 5th layer contain much detail information, as shown in Fig. 3, and those in 13th and 16th layers contain semantic information, as shown in Figs. 4 and 5, where the first 64 feature maps in the two layers are depicted respectively. The combination of the complementary information can produce feature with high quality. In addition, the three layers have 64, 256, 256 feature maps respectively, and the dimension of the descriptor constructed by these layers is just proper to ensure the high efficiency. Therefore, we just adopt these layers to generate superpixel feature to balance the computation efficiency and feature quality.

The feature maps are the results of the convolution operation, those in low layers are the input of the high layers, and the outputs of the high layers are the further abstraction of the input feature maps. The feature maps in middle layers can be considered as the combination of those in low layers, which may not be easily understood through visualizing of feature maps, and the process makes the non-linear information in low layers more separable, for example, the gray value of some pixels in feature maps have no evident difference in low layers, but those in high layers are discriminating.

In fact, the structure is just the imitation of the human cognition, and the feature maps in higher layers are more abstract, which are more capable to represent the original semantics in image. For example, the feature maps in the highest layer can be consider as a vector, each dimension of which represents the probability of a particular category that the image belongs to. And the feature extracted from the semantic information is more representative, which can be more capable to reflect the intrinsic information in image.

Each superpixel, composed of a series of neighboring pixels with closed characteristic, can be described on the basis of the CNNs, the process is illustrated in Fig. 2.

Because the lower layer’s feature maps just represent the edge or rapid change, we use the information entropy of the superpixel as the feature to improve the representation capability. The low entropy indicates that the salient locations concentratively exist in the superpixel area, and the high one indicates the salient locations are dispersed all over the area [39]. In the calculation, the
value range is divided into a series of bins at regular intervals, and the calculation of superpixel entropy feature is formulated as
\[ p_i = n_i / n_t, \]
\[ H = -\sum_{i=1}^{N_t} p_i \log_2 p_i, \]
where \( n_t \) denotes number of pixels belong to ith bin, \( n_t \) is the total number of the area pixels, \( p_i \) is the probability of each bin, \( N_t \) represents the number of bins, and \( H \) denotes the information entropy of the area pixels.

The other two layers' outputs have abundant semantic information, as a consequence, we adopt the average value to describe the area pixels directly. Finally, the feature of superpixel can be represented by a vector \( \mathbf{F} \) with 576 dimensions, which consists of the entropy feature of 64 dimensions and averaged feature of last two convolutional layers.

3.3. Image patch descriptor

In order to describe the whole image and make the computing similarity with spatial information, the image is split into 4 \times 4 patches with the same size. The descriptor of each patch is calculated by the weighted average of covered superpixel's feature, where the weight represents the proportion of superpixel area in the patch. The weight of ith superpixel is defined as
\[ w_i^P = \frac{N_i^P}{N^P}, \]
where \( N_i^P \) is the pixel number of the ith superpixel in the patch, and \( N^P \) denotes the pixel number of the patch.

The weighted average is adopted to obtain patch descriptor, and the calculation is formulated as
\[ \mathbf{F}_i^P = \sum_{i=1}^{N^P} w_i^P \mathbf{F}^P_i, \]
where \( N^P \) denotes the number of superpixels in the patch, \( w_i^P \) denotes the weight of the ith superpixel, \( \mathbf{F}^P_i \) represents the descriptor of ith superpixel, and \( \mathbf{F}_i^P \) is the patch descriptor. According to the step above, we build a 576 dimensional descriptor for each patch after normalizing it. As a consequence, the image can be represented by a 16 \times 576 dimensional description matrix composed of the 16 patch descriptors.

3.4. Similarity matrix

Each row vector of the image description matrix is the corresponding patch descriptor, the similarity of two images can be reflected by the similarities of corresponding patches they contain. The patch similarity can be described by calculating the cosine of the two patch descriptors, as defined as
\[ S_{ij}^P = \cos \langle \mathbf{F}_i^P, \mathbf{F}_j^P \rangle = \frac{\mathbf{F}_i^P \cdot \mathbf{F}_j^P}{||\mathbf{F}_i^P|| \cdot ||\mathbf{F}_j^P||}, \]
where \( S_{ij}^P \) represents the similarity score of patch i and j, and the larger the value is, the more similar the two patches are. Particularly, if the two patches are identical, \( S_{ij}^P \) is 1. As the patch descriptors have been normalized with the length of 1, the dot product of the patch descriptors is just the patch similarity.

Actually, the dot product between the first image description matrix and the transpose of the second is just the 16 \times 16 dimensional similarity matrix, denoted by \( S \). The element on the ith row and jth column in similarity matrix represents the similarity between the ith patch in the first image and the jth patch in the second. Fig. 6 shows two sets of image pairs and their similarity matrices.

The diagonal elements in similarity matrix is \( D = \{d_1, d_2, ..., d_{16}, d_1 = \text{Sp}\}, \) and the off-diagonal elements are \( R = \{(r \in S) \land (r \notin D)\} \).

Generally speaking, if two images origin from the same place, the 16 values in \( D \) are the maximums on the corresponding rows and columns, the situation is depicted in Fig. 6(a1) and (b1), the image pair comes from the same place, and most diagonal elements in similarity matrix have the max values on corresponding rows and columns.

4. Adaptive weighting image similarity

The place recognition is based on measuring the similarities between images. The traditional methods just use the global or local descriptor based image vector to measure the image similarity. However, these approaches usually incorrectly output the similarity score of the same place images with local changes and the different place images with similar looking.

In this work, we propose a novel comparing mechanism to obtain better performance. We determine the probability that image pair comes from the same place by comparing the values in \( D \) with those in \( R \), and obtain the weight of each patch similarity by detecting the abnormal values in \( D \). According to the similarity matrix, we define the overall similarity as
\[ S = \xi \sum_{i=1}^{16} w_i^P S_{ii}^P, \]
where \( S \) is the similarity of image pair, \( \xi \) is the probability that two images origin from the same place, \( S_{ii}^P \) denotes the ith patch similarity, \( w_i^P \) represents the weight of the corresponding patch similarity with the requirement of \( \sum_{i=1}^{16} w_i^P = 1 \).

4.1. Direct method

After obtaining the patch similarities of image pair, the image similarity can be simply computed through averaging the 16 patch similarities. The similarity is formulated as
\[ S = \frac{1}{16} \sum_{i=1}^{16} S_{ii}^P. \]
In fact, it is just a particular case of our proposed formula, where \( \xi = 1 \) and \( w_i^P = 1/16 \). We call this method as the direct method, which is easily to be understood and implemented.

However, some problems inevitably exist: first, as the environment is changing all the time and the images are captured at different time, it is quite possible that the same place images are locally changed, which will lead to a low similarity if evaluated by the direct method. The situation is illustrated in Fig. 6(a1) and (b1). The the 7th and 11th patches are changed, the relevant values in \( D \) are distinctly smaller than other diagonal values in similarity matrix. Second, if the different place images are similar in vision, the diagonal values in similarity matrix are generally large, in which case the similarity score of the direct method will be high. This case is depicted in Fig. 6(a2) and (b2), where the diagonal values are very large, but the image pair comes from different places. The above problems mentioned will reduce the robustness of the direct method and even lead to wrong recognition. To boost the robustness of the image similarity calculation, we propose a method based on analyzing the data difference in similarity matrix to determine the image similarity.
4.2. Probability of the same place

Aiming at the two problems existing in the direct method, we propose an adaptive weighting method based on similarity matrix. The different place images with similar-looking is distinguished by evaluating the probability that image pair origins from the same place. In addition, in order to solve the problem that two images capturing the same scene but some objects get changed, which will lead to a low similarity, we assign an adaptive weight for each patch to alleviate the effect of the changed patches.

In the similarity matrix of two images, if the values in \( D \) are distinctly larger than those in \( R \), we can infer that the images are from the same place with high probability. If the values in \( D \) and \( R \) have no distinct difference, the images may origin from different places, even though the values in \( D \) may be very large sometimes. Generally, elements in \( R \) can be considered as the similarities of two irrelevant patches, the element values has the probabilistic distribution as

\[
R \sim N(\mu_r, \delta^2),
\]

where \( m \) denotes the element number in \( R \), the average of values in \( D \) is \( \mu_d = \frac{1}{m} \sum_{r \in R} r \). We define the event denoted by \( E \): for the similarity matrix of two images, the values in \( D \) are distinctly larger than those in \( R \). \( E = 1 \) indicates that the event is true while \( E = 0 \) indicates the false. We can conclude that the probability that two images origin from the same scene, denoted by \( \xi \)

\[
\xi = P(E = 1) = P(\mu_d > \mu_r).
\]

We use hypothesis testing to obtain the value of \( \xi \). The hypothesis testing is defined as

\[
E_0 : \mu_d \leq \mu_r,
\]

\[
E_1 : \mu_d > \mu_r.
\]

The normal distribution, \( z_\alpha \), is the value of the test statistic, and the \( \alpha \) is the significance level to reject the null hypothesis. The larger the value of \( \mu_d \) is than that of \( \mu_r \), the larger the value of \( z_\alpha \) is, and the smaller the value of \( \alpha \) is, the situation is depicted in Fig. 7.

We define the variable \( p_a = 1 - \alpha \), where \( p_a \) can reflect the level that \( \mu_d \) is larger than \( \mu_r \). Generally, the larger the value of \( p_a \) is, the more distinct that values in \( D \) are larger than those in \( R \). Obviously, if the value of \( \mu_d \) is larger than that of \( \mu_r \), \( p_a \in [0.5, 1) \). In order to get a comparable result, the domain is extended to the interval of \([0, 1]\), and \( \xi \) is represented as

\[
\xi = \begin{cases} 
0 & : \mu_d \leq \mu_r, \\
2(p_a - 0.5) & : 0.5 < p_a \leq 0.75, \\
1 - 2(1 - p_a) & : p_a > 0.75.
\end{cases}
\]

Consequently, the probability that images origin from the same place is determined by comparing the data difference in \( D \) and \( R \). For the different place images with similar-looking, the entries in \( D \) are large, but those in \( R \) are also very large, as shown in Fig. 6(b2). The entries in \( D \) and \( R \) have no distinct difference and the value of \( \xi \) evaluated by the proposed method will be relatively small.
4.3. Image similarity computation

Through analyzing the 16 elements in D, we find that the outliers may exist. In the similarity matrix, the values in D are larger than those in R on the whole, but some individual values in D have distinct differences with other diagonal values and they are more closed to the values in R. This is illustrated in Fig. 6(b1), where most of the diagonal values are larger than the off-diagonal, but the 7th and the 11th are closed to the values in R. We can infer with high probability that the corresponding patches have been changed.

In order to improve the robustness, the weights of the changed patches should be small when we adopt the formula (8) to calculate the image similarity. Based on the above discoveries, the adaptive weighting can be modeled as: the larger the values in D is than those in R, the larger the corresponding weights are; and vice versa. The level that the ith element in D is larger than those in R is defined as

\[ x_i = \begin{cases} \frac{d_i - \mu_r}{\delta} & : d_i > \mu_r, \\ 0 & : d_i \leq \mu_r, \end{cases} \quad (18) \]

where \( d_i \) is the ith element in D. Consequently, the adaptive weight of the ith patch similarity is defined as

\[ w_i^p = \frac{x_i}{\sum_{j=1}^{16} x_j} \quad (19) \]

In this way, we can solve the first problem in the direct method. For the changed patches, the corresponding weights will be small according to the formulas (18) and (19). In this case, the method can alleviate the effect of the abnormal data and consequently improve the robustness.

5. Experiments

In order to verify the robustness and validity, and prove that the proposed method outperforms other methods, in the first group experiments, comparisons of four methods are conducted. The methods include image histogram based image similarity, the method of BoW in FAB-MAP [14], the proposed direct method, and proposed adaptive weighting method. To demonstrate the comparison, we select a series of typical image pairs, including the same place images with local changes and the different place images with similar-looking.

Furthermore, to validate the reliability of the proposed adaptive weighting method in practical applications, in the second experiment, we use the method to solve the loop closure detection in single-camera SLAM system.

In our experiments, we adopt the SLIC to segment the image into superpixels. And several experiments are conducted to find the proper region size which ensures the high computation efficiency and great experimental performance at the same time. What’s more, we rewrite and optimize the source codes of CNNs and superpixel segmentation using C++, and perform the computation on GPU to improve the efficiency.\(^1\) In addition, we measure the running time in the experiment on a computer with Xeon 3.2 GHz CPU and 16 G memory.

5.1. Experiments about the robustness

We design the first experiment to check the robustness of our proposed adaptive weighting method and make analyses on the efficiency of the four methods. We adopt the mentioned methods to recognize the particular image pairs, including the same place images with local changes and the different place images with the similar-looking. What’s more, the elapsed time of describing the images based on those methods is also recorded respectively.

5.1.1. Experiment design

Two sets of image pairs from the same place are shown in Figs. 8 and 9. The image in Fig. 8(a) is locally different with the one in Fig. 8(b) because of the arrival of a box, and the location of a box is changed in Fig. 9(a) and (b). What’s more, the image pairs shown in Figs. 10–12 origin from the different places. Especially, the image pairs in Figs. 11 and 12 are very similar in vision with dense grasses and brick walls, respectively.

In addition, we adopt the four methods separately, to describe the images in the new college dataset [40]. The elapsed time of the process is precisely recorded, and the average time of dealing with each image is calculated.

5.1.2. Experiment analysis

The similarity scores of image pairs calculated by four different methods are listed in Table 2, and the average time of describing the images in dataset based on four method is summarized in Table 3. Through the experimental results, we can obtain the following discoveries.

For the image histogram based method, the similarity of the image pair in Fig. 8 is incorrectly low. The appearance of the box greatly affects the performance, which indicates that this method lacks robustness. Because the histogram describes the image as a whole, and ignores the spatial information, the result of the image pair in Fig. 10 is also incorrectly high. In addition, the high similarities of the image pairs in Figs. 11 and 12 indicate that the method fails to recognize the similar-looking images. When adopting the method of the BoW in FAB-MAP, although the similarity scores of the locally changed images are lower than those of other methods, the results are acceptable. However, the similarities of the similar-looking image pairs are still incorrectly high.

The proposed direct method can get reasonable similarity results for the locally changed images, and the similarities of the images in Figs. 8 and 9 are 0.9434 and 0.9655, which can be considered as right, this is mainly because the changed areas between two images are too small and most diagonal values in similarity matrix are very large. However, the direct method does not achieve good performances on the different place image pairs, especially, the similarities of the similar-looking images in Figs. 11 and 12 are 0.9885 and 0.9249, which are considered as obviously wrong.

Compared with other methods, the proposed adaptive weight method can accurately determine the similarities of all the image pairs. The image pairs in Figs. 8 and 9 origin from the same place, the similarities are very high despite that the images are locally changed. And the similarities of the different place images are low, even though the image pairs in Figs. 11 and 12 are similar in vision.

In addition, Table 3 demonstrates that the elapsed time of the four methods has no evident difference, and the image descriptor can be constructed with high efficiency.

Although these methods have no difference in efficiency, the proposed adaptive weighting method obviously outperform other three methods in robustness, not only the locally changed but also the similar-looking images can be effectively recognized.

The experiment results demonstrate that the proposed adaptive weighting method can achieve good performance, however, as

\(^1\) The source codes are available at http://www.adv-ci.com/blog/source/pi-cnn and http://www.adv-ci.com/blog/source/pi-slic.
the real world is complicated, the proposed method may be failed in some special situations. For example, the similarity score of the different place image pair in Fig. 14 is 0.7363, which is incorrectly high. And it is mainly because the selected image pair has the similar layout in content. The two images contain many similar objects, and the corresponding objects generally lie in the same positions in individual image, such as the desk, the computer screen and the flower pot, in which case the corresponding image patches may contain the similar objects, then the differences between the two images can not be effectively distinguished.

5.2. Experiment on loop closure detection

Loop closure detection is a hard and essential problem for SLAM, which is solved by finding the similar images from the previous. We design the second experiment to adopt the proposed adaptive weighting method to solve the loop closure detection. We use new college dataset [40] in our experiments and firstly choose the first 100 images in the dataset to simplify the experiments which already form the loop. In order to avoid the incorrect similarities between images with neighboring sequences, only similarities of images have enough intervals are calculated.

![Fig. 8](image_url)

![Fig. 9](image_url)

![Fig. 10](image_url)

5.2. Experiment on loop closure detection

Loop closure detection is a hard and essential problem for SLAM, which is solved by finding the similar images from the previous. We design the second experiment to adopt the proposed adaptive weighting method to solve the loop closure detection. We use new college dataset [40] in our experiments and firstly choose the first 100 images in the dataset to simplify the experiments which already form the loop. In order to avoid the incorrect similarities between images with neighboring sequences, only similarities of images have enough intervals are calculated.
5.2.1. Experiment design

Adopting the proposed adaptive weighting method to get the similarity arrays of the last 50 images, we choose the elements whose values are larger than the threshold (the threshold is set to be 0.85 in the experiment). The image pairs with high similarity are summarized in Table 2, where the \( n \)th image is denoted by No.\( n \), for example No.95 denotes the 95th image.

As shown in Table 4, we find a series of neighboring images, which are similar with the first four images, thereby the loop closure can be detected. The positions of the test images are drawn in Fig. 13.

From Fig. 13(c), we know that the loop closure is formed at the 95th images, and Table 4 shows that the 95th image has the high similarity with the 1st image, which indicates that the loop closure is successfully detected. We pick out the 1st and the 95th images from the dataset, shown in Fig. 15.

The similarities between the 95th image and the first 45 images are shown in Fig. 16, and the similarities between the 1st image and the last 50 images are shown in Fig. 17.

From Fig. 13(c), we know that the loop closure is formed at the 95th images, and Table 4 shows that the 95th image has the high similarity with the 1st image, which indicates that the loop closure is successfully detected. We pick out the 1st and the 95th images from the dataset, shown in Fig. 15.

The similarities between the 95th image and the first 45 images are shown in Fig. 16, and the similarities between the 1st image and the last 50 images are shown in Fig. 17.

From Fig. 13(c), we know that the loop closure is formed at the 95th images, and Table 4 shows that the 95th image has the high similarity with the 1st image, which indicates that the loop closure is successfully detected. We pick out the 1st and the 95th images from the dataset, shown in Fig. 15.

The similarities between the 95th image and the first 45 images are shown in Fig. 16, and the similarities between the 1st image and the last 50 images are shown in Fig. 17.

From Fig. 13(c), we know that the loop closure is formed at the 95th images, and Table 4 shows that the 95th image has the high similarity with the 1st image, which indicates that the loop closure is successfully detected. We pick out the 1st and the 95th images from the dataset, shown in Fig. 15.

The similarities between the 95th image and the first 45 images are shown in Fig. 16, and the similarities between the 1st image and the last 50 images are shown in Fig. 17.

From Fig. 13(c), we know that the loop closure is formed at the 95th images, and Table 4 shows that the 95th image has the high similarity with the 1st image, which indicates that the loop closure is successfully detected. We pick out the 1st and the 95th images from the dataset, shown in Fig. 15.

The similarities between the 95th image and the first 45 images are shown in Fig. 16, and the similarities between the 1st image and the last 50 images are shown in Fig. 17.

From Fig. 13(c), we know that the loop closure is formed at the 95th images, and Table 4 shows that the 95th image has the high similarity with the 1st image, which indicates that the loop closure is successfully detected. We pick out the 1st and the 95th images from the dataset, shown in Fig. 15.
judging the existence of the high similarities between the current image and the previous, whose values are larger than the threshold. Experimental result is shown in Fig. 18. In the figure, the revisited locations are labeled by the red circles, and the corresponding location images, which are similar with the revisited ones, are labeled by green circles, according to which we can detect the formation of the loop closure. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

From Fig. 18, we can know that most revisited locations are detected. To evaluate the detection performance, the precisions at different thresholds are calculated. For a set of image pair captured at two locations, we define that the image pair comes from the same place if the distance of the locations is less than the maximum permissible, and the maximum permissible is called the error range. The detection precision is defined as:

\[ Pr = \frac{n_{tp}}{n_{to}} \]  

where \( n_{tp} \) represents the total number of the same place image pairs we find out using the proposed method at a specific threshold, \( n_{to} \) is the number of those detected image pairs, which come from the same place actually within the specific error range. We calculate the precisions at different error ranges, and the results are listed in Table 5, which shows that high accuracy can be achieved in the range of permitted errors.

5.2.2. Statistics on the elapsed time

The elapsed time of each stage in processing the 992 images, whose sizes are all 640 × 480, is precisely recorded. For an image, we calculate the average time of each stage, and the results are listed in Table 6, where the steps of superpixel segmentation and feature maps creation are the most time-consuming.

In the process of loop closure detection, we need to calculate the similarities between the current image and the previous ones. The elapsed time mainly consists of two parts, one is to construct the current image description matrix, and the other is to compare all previous images and find the similar ones. The elapsed time of the current image can be modeled as:

\[ T_i = (T_{sm} + T_n)(i - 1) + (T_{seg} + T_f + T_{sup} + T_{dm}) \]  

where \( i \) is the image sequence number, \( T_i \) represents the elapsed time of the current image in loop closure detection. Generally, the similar images of the current frame can be detected within several milliseconds, which indicates that the loop closure detection can be finished in real time.

5.2.3. Experiment analysis

From the experiments above, we can verify that the proposed adaptive weighting method can effectively solve the loop closure detection by similarity comparison. And the experimental results demonstrate that the method is reliable and robust, the similarities of the same place images are significantly higher than those of the different place images whose values are generally smaller than 0.5. And the revisited locations can be effectively detected in real time, which indicates that the method can be applied to practical SLAM.
6. Conclusion

Aiming at solving the common problems in place recognition that the same place images get changed locally and the different place images are similar in vision, the paper proposes an algorithm of image similarity measurement for place recognition. Firstly we apply superpixel segmentation on image and generate the superpixel descriptors according to the feature maps calculated by

![Figure 16. The similarities between the 95th image and the first 45 images.](image)

![Figure 17. The similarities between the 1st image and the last 50 images.](image)

![Figure 18. The detection of the previously seen location for all images in dataset. The revisited locations are labeled by the red circles, and the corresponding location images, similar with revisited ones, are labeled by green circles.](image)

**Table 5**

<table>
<thead>
<tr>
<th>Error range (m)</th>
<th>Threshold = 0.85</th>
<th>Threshold = 0.9</th>
<th>Threshold = 0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.3987</td>
<td>0.5347</td>
<td>0.7636</td>
</tr>
<tr>
<td>4</td>
<td>0.5391</td>
<td>0.6779</td>
<td>0.8727</td>
</tr>
<tr>
<td>5</td>
<td>0.6427</td>
<td>0.7634</td>
<td>1.0</td>
</tr>
<tr>
<td>6</td>
<td>0.7097</td>
<td>0.8211</td>
<td>1.0</td>
</tr>
<tr>
<td>7</td>
<td>0.7520</td>
<td>0.8529</td>
<td>1.0</td>
</tr>
<tr>
<td>8</td>
<td>0.7998</td>
<td>0.8956</td>
<td>1.0</td>
</tr>
<tr>
<td>9</td>
<td>0.8309</td>
<td>0.9225</td>
<td>1.0</td>
</tr>
<tr>
<td>10</td>
<td>0.8533</td>
<td>0.9384</td>
<td>1.0</td>
</tr>
</tbody>
</table>

**Table 6**

<table>
<thead>
<tr>
<th>Stage</th>
<th>Elapsed time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image segmentation ($T_{seg}$)</td>
<td>23.8</td>
</tr>
<tr>
<td>Feature map creation ($T_f$)</td>
<td>11.5</td>
</tr>
<tr>
<td>Superpixel descriptor construction ($T_{sup}$)</td>
<td>3.5</td>
</tr>
<tr>
<td>Image description matrix construction ($T_{dm}$)</td>
<td>1.4</td>
</tr>
<tr>
<td>Similarity matrix construction ($T_{sm}$)</td>
<td>0.042</td>
</tr>
<tr>
<td>Image similarity evaluation ($T_{is}$)</td>
<td>0.456</td>
</tr>
</tbody>
</table>

**Figure 15.** The 1st and the 95th images in the dataset come from the same place.
CNNs. Secondly, the image is divided into 4 × 4 patches, each patch descriptor is built according to the superpixels the patch contains, and the image description matrix is constructed based on the patch descriptors. Then the similarity matrix is generated by the dot product between the first image description matrix and the transpose of the second one. We determine the probability that the image pair origins from the same place by comparing the diagonal values with the off-diagonal, and evaluate the final similarity based on the data difference in similarity matrix rather than the patch similarity values, which greatly improve the robustness of the algorithm.

The experimental results show that the proposed adaptive weighting method is more robust and outperforms other three methods. It assigns high similarity scores for the same place images despite the local changes and low scores for the different place images despite the similar-looking. Furthermore, the method is adopted to solve the loop closure detection where the results demonstrate that it can effectively solve the practical problems in SLAM.

In our proposed adaptive weighting method, the deep feature represents the essence of image, and the analysis on the similarity matrix accurately describes the similarity between images. Although the method achieves better performance, some detailed information between images is still inevitably ignored. Currently, the image is just divided into 4 × 4 patches, which may cause low spatial description. In order to obtain a more precise description, the image can be divided into more patches to lessen the negative effect. In addition, if the same place images have evident difference in scale, our methods may be failed, and the issue is also our focus of future work, which can be solved by extracting the multi-scale features and exploring the corresponding similarity measure method. In fact, further research on exploring the similarity matrix is promising in many vision tasks. For example, the positions of the changed objects in image can be detected by analyzing the abnormal values in similarity matrix, and the values in similarity matrix are centered on the diagonal on the whole if the image pair origins from the same place. In general, the similarity matrix can be considered as a compact and effective representation, which reflects the relations between images, and lots of potential information in similarity matrix still needs to be mined to make full use of it in different vision applications.

Acknowledgments

Our research is partially supported by grants from National Natural Science Foundation of China (41201390, 61202185, and 61573284), the Project of Science and Technology Innovation of Henan Province (310201401-(JCQ01009, JCQ01012)), the Fundamental Research Funds for the Central Universities (310214011-(JQ01009, JQ01012)), and the Open Projects Program of National Laboratory of Pattern Recognition (NLPR).

References


Qin Li received the BS degree from Zhengzhou Institute of Surveying and Mapping, China in 2013. Currently, he is working toward the MS degree at Zhengzhou Institute of Surveying and Mapping, China. His research interests include image processing and computer vision.

Ke Li received the PhD degree in Zhengzhou Institute of Surveying and Mapping, China in 2008. He is an associate professor at Zhengzhou Institute of Surveying and Mapping, China. His research includes GIS, image processing, and computer vision. He has published over 15 journal and conference papers in the related areas.

Xiong You received the PhD degree in Zhengzhou Institute of Surveying and Mapping, China in 1994. He is a professor at Zhengzhou Institute of Surveying and Mapping. His research includes Virtual Geographic Environment (VGE), and Geographic Information System (GIS).

Shuhui Bu received the MSc and PhD degrees in the College of Systems and Information Engineering from University of Tsukuba, Japan in 2006 and 2009. He was an assistant professor (2009–2011) at Kyoto University, Japan. Currently, he is an associate professor at Northwestern Polytechnical University, China. His research interests are concentrated on computer vision and robotics, including SLAM, 3D shape analysis, image processing, pattern recognition, 3D reconstruction, and related fields. He has published approximately 40 papers in major international journals and conferences, including the IEEE TMM, IEEE TBME, TVC, C&G, ACM MM, ICPR, CGI, SMI, etc.

Zhenbao Liu received the PhD degree in computer science from the College of Systems and Information Engineering, University of Tsukuba, Japan, in 2009. He is an associate professor at Northwestern Polytechnical University, Xi’an. He was a visiting scholar in the GrUVi Lab of Simon Fraser University in 2012. His research interests include 3D shape analysis, matching, retrieval and segmentation.