PL-VSCN: Patch-level vision similarity compares network for image matching

Xiong You¹ | Qin Li¹ | Ke Li¹ | Anzhu Yu¹ | Shuhui Bu²

¹Department of Geography Information System, Information Engineering University, Zhengzhou, China
²Department of Artificial Intelligence, Northwestern Polytechnical University, No.127 Youyi Road, Xi'an, China

Correspondence
Qin Li, Information Engineering University, No.66 Longhai Road, 450000, Zhengzhou, China. Email: leequer120419@163.com

Abstract
Image matching plays an important role in various computer vision tasks, such as image retrieval and loop closure detection in Simultaneous Localization and Mapping. The authors propose a discriminative patch-based image matching method that converts the problem of whole image matching to that of local patch matching. To construct the patch representation, the Patch-Level Vision Similarity Compare Network (PL-VSCN) is proposed to produce the patch feature. In the image matching process, local patches that potentially contain objects within images are initially detected, and the discriminative feature of each patch is extracted based on the pre-trained PL-VSCN. Then, the similarities between the patch pairs are calculated to construct the similarity matrix, and the corresponding patch pairs are detected based on the mutual matching mechanism on the similarity matrix. Experimental results indicate that the proposed PL-VSCN can generate the discriminative patch feature, which can accurately match the patch pairs with the corresponding content and distinguish those with non-corresponding content. In addition, the comparison experiments demonstrate that the proposed image matching method outperforms existing approaches on most datasets and effectively completes the image matching task.

INTRODUCTION

Image matching is critical in many vision applications such as appearance-based navigation [1,2], place recognition [3,4] and loop closure detection in Simultaneous Localization and Mapping (SLAM) [5,6]. Feature representation and similarity measures are two basic steps in the process of image matching [7]. The image features are initially extracted to construct image representation, and the similarity measurement model is then constructed to produce the similarity score, which is ultimately utilized to predict whether the image pair is matched. Generally, the main methods of image matching can be split into two categories. The first category involves constructing the image descriptor based on the hand-crafted features such as the SIFT [8,9], SURF [10,11] and ORB [12]. As the features only depict the local region pixels around the key points, all the features within an image should be further gathered based on the aggregating models, such as the Bag-of-Words (BoW) [13,14], the Vector of Locally Aggregated Descriptors (VLAD) [15] and the Fisher Vector (FV) model [16–18]. Furthermore, the similarity score of the image pair is obtained by calculating the distance between the image descriptors. The other category involves completing the image matching based on deep learning and particularly Convolutional Neural Networks (CNNs) [19], which has achieved excellent performance in many vision applications. Deep learning-based methods effectively fuse feature extraction and similarity measures into an end-to-end network wherein the input of the network corresponds to the image pair, and the output is the similarity score. The CNNs gradually mine the intrinsic essence in images and construct the discriminative features, which could effectively match image pairs with the corresponding content and distinguish the non-matched pairs. Thus, the deep learning-based methods can achieve excellent performance in image matching.

Specifically, the core of determining whether or not two images are matched depends on whether the image pair contains the corresponding content, and the similarity score reflects the quantity of the corresponding content wherein a higher similarity score implies that the image pair exhibits more corresponding content. Given that the real-world scenario changes constantly, the matched image pairs can significantly...
vary in appearance. Figure 1 demonstrates a matched image pair and the corresponding content, enclosed by red bounding boxes, in respective images. Although the images in Figure 1 are matched with each other, the corresponding contents are significantly less than the non-corresponding content.

The non-corresponding content, which typically occupies the main part of the whole image, inevitably disturbs the calculation process of image matching, and plays a negative role in predicting the image pair as matched.

We aim to solve the challenging problem and complete the whole image matching task based on matching the local patch within an image pair. Instead of representing the whole image directly, we describe the local patch and match the corresponding patch pairs between images. Figure 2 shows the flow chart of the proposed image matching method. In the process, the local patches are initially detected in the whole images, and the discriminative feature of each patch is constructed based on the well-trained PL-VSCN. Subsequently, the similarity values of the patch pairs are calculated to constitute the similarity matrix based on which the corresponding patch pairs are detected.

The Edge Boxes algorithm [20,21] is adopted to detect the image patches, and the detected patch regions generally cover the meaningful objects within images. With respect to each detected patch, it contains the significant object target, the covered image regions are called as the relevant contents, and the image contents, out of the patch regions, are called as the irrelevant contents, which generally cover the background of the image, such as the sky and the pavement. In the process of local patch matching, only the covered object patches are involved, which effectively eliminates the interference of the irrelevant contents, thus the local patch matching is significantly easier than the whole image matching.

The experimental results indicate that the propose PL-VSCN could produce the discriminative features for local patches and the patch-based image matching method could achieve better performance than those matching the whole image directly.

The relevant researches in the paper provide three main contributions as follows:

- **Discriminative feature extraction based on comparison mechanism:** In order to match the corresponding patch pairs and distinguish the non-correspondences, the PL-VSCN is trained based on comparison mechanism, and the training process brings the corresponding patch pairs nearer to each other and takes the non-correspondences further from each other.

- **Complementary convolution information integration:** In order to fully describe the patch information, the normal convolution and dilated convolution are employed to construct deep convolutional networks and build the patch representation, respectively. And the Fusion-Net is designed to further integrate the complementary features, the integrated feature is more discriminative than the single convolution operation based feature.

- **Mutual matching based on similarity matrix:** To match the corresponding patches between an image pair, the similarity matrix, which consists of the similarity scores of the patch pairs, is first constructed, then the mutual matching mechanism is employed to detect the corresponding patch pairs, and the matching results are the optimal scheme for both the images, which specifically reduces the random matching errors.

## 2 RELATED WORKS

Recently, image matching has been explored by considerable researchers, many of which focus on the deep leaning based methods. Some researchers employ the CNNs to extract the image feature, the CNNs have achieved great success in image classification, it could effectively describe the semantic information in images, thus the CNNs based image representation is assumed to be discriminative. Niko et al. [22] deploy the pre-trained AlexNet [19] to construct a holistic image descriptor. Each individual layer from the network can generate the image feature, and the matching performances of different layer features are comparatively analysed in the research, which provides a meaningful reference for relevant exploration. Chen et al. [23] adopt the pre-trained Overfeat [24] to construct the deep hierarchical feature, each dimension of the output feature is separately obtained from the corresponding layer in the network. Different from the single layer based image feature, the deep hierarchical feature involves all the layers in the network, it could describe the image information from different levels. Inspired by the VLAD model [15], Relja et al. [25] introduce the NetVLAD to construct the whole image representation. The trainable CNN structure is first employed to extract the image feature, then the NetVLAD layer is

![Figure 1](image1.png)  
**Figure 1** Matched image pair and corresponding patch pair between images. The images in (a) and (b) match each other, regions marked by red boxes contain the corresponding content, and the non-corresponding content (out of the regions) occupies the main part of the images. (a) Image 1. (b) Image 2
adopted to generate the VLAD vector for image. The NetVLAD has achieved great success in place recognition. Instead of constructing the holistic image representation, Li et al. [7] generate the local patch feature. The images are first divided into the regular patches, and the image similarity is evaluated according to the patch similarities. Compared with representing the whole image directly, the local patch-based method can achieve greater robustness in place recognition. Silvia et al. [26] adopt the Edge Boxes algorithm [20,21] to detect objects within images, then the CNN feature of each object patch is constructed based on the pre-trained AlexNet [19], and the covisibility graph is adopted to represent the spatial relations between the objects. The CNN feature and spatial relation are effectively integrated in the similarity measurement model, and the method can effectively complete the loop closure detection task in SLAM.

In addition, lots of researchers focus on the Siamese networks [27] that receive the image pair as input and produce a similarity score to predict matching results. MatchNet is an end-to-end network of local patch matching and was proposed by Han et al. [28]. The whole architecture can be divided into two parts, the deep convolutional network that extracts features from patches and a network with three fully connected layers that computes a similarity between the extracted features. As the processes of feature extraction and similarity measurement are integrated together, the flexibility is greatly limited in some vision applications. Simoserra et al. [29] adopted the Siamese networks to construct the patch representation. In the process of training, the Euclidean distance is utilized to construct the loss function. Additionally, the hard samples, including the corresponding samples with significant changes and the non-corresponding samples with similar looking, are effectively mined during the training process, and the best proportion of the hard sample in training data is explored to train the discriminative model. As the feature extraction network just consists of three convolutional layers, which is insufficient to describe the intrinsic essence of the patch, the discriminative capability of the patch feature is limited. Zagoruyko et al. [30] explored to construct the loss function based on the patches comparison, which effectively solves the problem of lacking labelled training data. The matching performance of various Siamese structures is evaluated in the research, and it provides meaningful references for the relevant research. The PN-NET, which receives three patches as input, was introduced by Balntas et al. [31]. The loss function is constructed based on three patch features, which specifically add the constraints in the training process. Consequently, the training efficiency is significantly improved, and the features produced by the network are comparatively discriminative. Melékhov et al. [32] adopted the deep convolutional network to generate the patch descriptors, and the patch matching task was completed based on the Euclidean distance between the patch descriptors. Additionally, some training techniques, including the histogram equalization operation on input patch, and batch normalization operation on convolutional output, are explored to improve the matching performance.

Some researchers [33,34] directly adopt the Siamese networks to conduct the whole image matching task, and the image descriptors that are constructed based on the deep convolutional networks exhibit strong generalization capability. However, the matched image pairs generally exhibit a significant amount of non-corresponding content, this can severely disturb the matching process, and significantly limit the matching performance.

The paper proposes an image matching method based on local patch matching, which can effectively eliminate the negative interference of irrelevant content. The PL-VSCN is constructed to produce the discriminative feature for local patches, and the feature can effectively match
the corresponding patch pairs and distinguish the non-correspondences. In the whole image matching process, the mutual matching mechanism is adopted to detect the corresponding patch pairs between images, and a number threshold for corresponding patch pairs is set, the image pair can be predicted as matched if the number of the corresponding pairs exceeds the threshold. The experimental results demonstrate that the proposed image matching method can achieve better performance when compared with the existing methods.

3 | METHODS

To convert the whole image matching to the local patch matching, we initially detect the local object patches within images, and the PL-VSCN is then constructed to generate the discriminative feature for the detected patches. Furthermore, we construct a similarity matrix, which consists of similarity values of the patch pairs, to detect the corresponding patch pairs based on mutual matching mechanism.

3.1 | Patch detection

In order to perform the patch-based image matching task, the first step involves detecting local patches, which should cover the meaningful objects within images. We adopt the Edge Boxes algorithm [20] to construct the initial patches, in the algorithm, the significant pixel points are firstly detected to construct the edge image, as depicted in Figure 3b, and the Non-Maximum Suppression (NMS) operation is performed on the edge image to make the detected points sparse. The adjacent points which are almost in a straight line in the edge image, are clustered as an edge group. If the angle between two edge groups approximately equals the average direction of them, the two edge groups are similar with each other, and they come from the same object. The similar edge groups are further clustered based on the edge clustering method [21], and the edges of the same objects are grouped together, the object patches can be specified according to the bounding boxes of the clustered edge groups.

Figure 3c demonstrates the bounding boxes of the detected patches, which are seriously overlapped. We perform the NMS operation on the detected patches, which could effectively eliminate the overlapped patches, and save the significant patches meanwhile. The remaining patches have less overlapped contents, and they generally cover the meaningful objects in images. The NMS algorithm sets the relevant parameter to define the maximum overlap ratio, and the parameter is assigned 0.4 in our proposed system, which means that the overlapped area between two patches is less than 0.4 times the minimal patch area, then the remaining patches would have less overlapped contents.

In the practice of patch detection, some detected patches are very large, even the whole image is detected as a patch in some cases, which is unreasonable. What’s more, as some large patches occupy the main part of whole images, some meaningful object patches may be covered by the large patches, the object patches may be eliminated in the NMS operation step, and the large patches could conflict with the meaningful object patches in the patch detection process. As the meaningful object patches are more significant than the large patches in our local patch-based image matching method, in order to save the meaningful object patches, it is advisable to eliminate the large patches.

In addition, the detected patches would be resized to 64 × 64 to satisfy the input of PL-VSCN in the subsequent steps. The patches with excessively small or large sizes, and those with evident variation in height and width are regarded as the unqualified patches, because they would suffer severe

![Image](https://via.placeholder.com/50x102)
![Image](https://via.placeholder.com/545x297)

**FIGURE 3** Process of patch detection. The test image is shown in (a), and its edge image is depicted in (b), the detected patches based on Edge Boxes algorithm are demonstrated in (c), the patches in (c) are seriously overlapped, the remaining patches after NMS operation and size constraints are depicted in (d). (a) Test image. (b) Edge image; (c) Initial patches detected based on Edge Boxes algorithm, (d) Remaining patches after NMS operation and size constraints.
deformation in the resizing operation, and it’s essential to delete those unqualified patches from the current patches set. To achieve that, we define the constraints of patch size, as listed in Equations (1) and (2), where \( W \) and \( H \) denote the width and height of the detected patches, respectively. All the detected patches are traversed, and the sizes of them are strictly checked. With respect to each patch, if the width and height cannot follow the constraints in Equations (1) and (2), the patch is deleted from the current patches set. Consequently, the remaining patches are all qualified, their sizes are moderate, and their shapes are comparatively square.

\[
64 \times 64 < W \times H < 256 \times 256 \quad (1)
\]
\[
0.5 < W/H < 2.0 \quad (2)
\]

Figure 3d shows the bounding boxes of the detected patches based on the proposed method. The detected patches have less overlapped regions, and nearly all the patches exhibit moderate size and comparatively square shape, which provides good foundation to construct the discriminative feature for image patches in the subsequent steps.

### 3.2 | PL-VSCN

In order to effectively match the corresponding patches and distinguish the non-correspondences, the PL-VSCN is constructed to produce the discriminative features for the local patches. In the feature space, the corresponding patches get close to each other, and the non-corresponding patches are far from each other, the corresponding pairs can be clearly separated from the non-correspondences according to the feature distance.

#### 3.2.1 | Architecture

The PL-VSCN is constructed based on the comparison mechanism, and it includes two identical towers with the shared weights, as shown in Figure 4. The input of the network is the patch pairs with the regular size of \( 64 \times 64 \). The training label denotes the matching ground truth (if the patch pair is matched, the label is 1, otherwise the label is -1).

Each tower of the PL-VSCN consists of three parts, namely NC-Net, DC-Net and Fusion-Net, as shown in Figure 5. With respect to the input patch, we initially employ the DC-Net and NC-Net to produce the patch feature, and Fusion-Net is then designed to further integrate the outputs from DC-Net and NC-Net. The parameter details of each tower in PL-VSCN are shown in Table 1.

**NC-Net**

The architecture is the normal deep convolutional network with seven convolutional layers and a fully connected layer. The network describes the patch from different levels via the convolution operation on the input patch from layer to layer.

As the abstract semantic information hidden in the image patch is effectively represented, the feature constructed based on the essential representation is quite discriminative.

**DC-Net**

Different from NC-Net, DC-Net employs dilated convolution to construct the patch feature in each convolution layer. Figure 6 demonstrates the process of the dilated convolution and normal convolution wherein dilated convolution increases the reception field. Additionally, for each unit in the upper layer, the associated pixels in the lower layer differ between the dilated convolution and normal convolution with the exception of the convolution centre, hence, the DC-Net-based feature is considerably complementary with NC-Net.

**Fusion-Net**

The Fusion-Net consists of two fully connected layers and is designed to fuse the outputs of NC-Net and DC-Net. The features produced by DC-Net and NC-Net are considerably complementary with less redundancy, thus the integrated feature is assumed as more discriminative. In addition, we perform the normalization operation on the output descriptor to ensure that the length of patch feature is one.

#### 3.2.2 | Loss function

In order to match the corresponding patch pair and distinguish the non-correspondences, the objective of constructing the patch feature involves minimizing the feature distance of the corresponding pairs and maximizing non-correspondences to the maximum possible extent.

We employ the cosine distance between the patch features to evaluate the similarity of the patch pair. As the length of patch feature is 1, the dot product between the patch features simply corresponds to the cosine distance, and the high cosine value implies that the angle between the patch descriptors is small and similarity score between the patch pair is high. The values in each dimension of the output feature range from
-1 to 1, thus the value interval of the similarity score (i.e. cosine value between the output features) corresponds to \([-1,1]\). We define the loss function as

\[
\Gamma = -\frac{1}{2N} \sum_{i=1}^{N} \left\{ (1 + L_i) \cdot \log \left( 1 + \frac{D_i^1 \cdot D_i^3}{2} \right) + (1 - L_i) \cdot \log \left( 1 - \frac{D_i^1 \cdot D_i^3}{2} \right) \right\},
\]

where \(D_i^1\) and \(D_i^3\) denote the output features of the patch pair, and \(D_i^1 \cdot D_i^3\) corresponds to the patch similarity of the patch pair. \(L\) represents the training label (the label of the corresponding pair is one and that of the non-correspondence is -1), and \(N\) is the sample number in each training batch. The training process makes the feature distance of the corresponding pairs tend to one and those of the non-correspondences tend to -1.

### 3.3 Image matching based on similarity matrix

To predict whether an image pair is matched or not, the local patches are first detected within each image, then the discriminative feature of each patch is constructed based on the PL-VSCN. The patch features of each image are stacked to form the feature matrices, denoted by \(F_1 \in \mathbb{R}^{M \times 128}\) and \(F_2 \in \mathbb{R}^{N \times 128}\) respectively, where \(M\) and \(N\) represent the numbers of the detected patches in respective images. The similarity matrix is constructed based on the feature matrices and is given as follows:

\[
S = F_1 \cdot F_2^T
\]

where \(F_2^T\) denotes the transpose of \(F_2\), \(S\) is the similarity matrix, and the element \(s_{ij}\) in \(S\) represents the similarity score (i.e. cosine value) between the \(i^{th}\) patch in the first image and \(j^{th}\) patch in the second image.

The cosine value varies slightly around 0, and we convert the cosine value into the angle between the patch descriptors, denoted by \(s_{A} = arccos(S)\), to represent the patch similarity. In addition, the \(p^{th}\) patch in the first image is denoted by \(P_i^1\), the \(q^{th}\) patch in the second image is \(P_j^2\), and the similarity level between \(P_i^1\) and \(P_j^2\) is represented by \(s_{pq}\), the element on the \(p^{th}\) row and \(q^{th}\) column of \(S_A\). We define the following requirements, listed in Equations (5)(6)(7), where \(R_p\) denotes the elements on the \(p^{th}\) row of \(S_A\), \(R_p = \{s_{pj}, j = 1, 2, \ldots, N\}\), \(C_q\) represents the elements on the \(q^{th}\) column of \(S_A\), \(C_q = \{s_{qp}, i = 1, 2, \ldots, M\}\), \(Se_{min}(C_q)\) and \(Se_{min}(R_p)\) are the secondary minima of \(C_q\) and \(R_p\), respectively.

\[
s_{pq} < \pi/6 \quad (5)
\]

\[
s_{pq} = min(C_q), \frac{Se_{min}(C_q)}{s_{pq}} > 1.2 \quad (6)
\]

![Diagram](image-url)
TABLE 1 Layer parameters of each tower in PL-VSCN. The output dimension is given by height × width × channel. DC-Net and NC-Net exhibit the same architecture with seven convolutional layers and a fully connected layer, and the difference lies in the type of convolution operation, the normal convolution and dilated convolution are employed in NC-Net and DC-Net, respectively. Specifically, C(3) denotes the convolutional layer with the filter size of 3 × 3, stride is 1, the padding model is ‘SAME’ and the dilatation rate is one in DC-convolutional layer with the filter size of 3.

| Layer       | Operation         | Output Dim |
|-------------|-------------------|------------|
| DC-Net/NC-Net |                   |            |
| Conv1       | C(3) + T + MP(2)  | 32 × 32 × 32 |
| Conv2       | C(3) + T + MP(2)  | 16 × 16 × 64 |
| Conv3       | C(3) + T          | 16 × 16 × 64 |
| Conv4       | C(3) + T          | 16 × 16 × 128 |
| Conv5       | C(3) + T + MP(2)  | 8 × 8 × 128  |
| Conv6       | C(3) + T          | 8 × 8 × 256  |
| Conv7       | C(3) + T + MP(2)  | 4 × 4 × 256  |
| Fc          | F + T + L2_norm   | 4096 → 128  |
| Fusion-Net  |                   |            |
| Fc1         | F + T             | 256 → 128   |
| Fc2         | F + T + L2_norm   | 128 → 128   |

FIGURE 6 Dilated convolution and normal convolution. Red blocks and lines denote the process of the dilated convolution, and green represents the normal convolution. With respect to each unit in the upper layer, the associated pixels in the lower layer are different between these two convolution methods with the exception of the convolution centre.

\[ s_{pq} = \min(R_p), \quad \frac{S_{e.min}(R_p)}{s_{pq}} > 1.2 \] (7)

If \( s_{pq} \) satisfies the aforementioned requirements, \( s_{pq} \) is much less than other values in \( R_p \) and \( C_p \), which indicates that \( P^*_p \) is similar with \( P^*_q \), and the similarity level is much higher than those between \( P^*_q \) and any other patches in the second image as well as those between \( P^*_q \) and any other patches in the first image, thus \( P^*_q \) and \( P^*_q \) are the corresponding patch pair.

The image matching task can be completed based on local patch matching, and the image pair can be predicted as matched if a corresponding patch pair is matched between two images. However, as the real word scenario is quite complex, some different objects have the similar appearance when mapped into the images, and the proposed PL-VSCN is incapable to distinguish these object patches with similar-looking. Then the corresponding patch pairs could be incorrectly matched between two non-matched images, and the non-matched pairs would be incorrectly predicted as matched.

In the practice of image matching, the number of the corresponding patch pairs is set as the metric to predict whether or not the image pair is matched, the more corresponding patch pairs, the higher confidence level of predicting the image pair as matched. We set the number threshold of the corresponding patch pairs, denoted by \( T_N \), and the image pair can be predicted as matched if the number of corresponding patch pairs exceeds \( T_N \).

To achieve good matching performance, the value of \( T_N \) should be assigned according to the concrete contents of the test image. If there are many similar objects in the image contents, the value of \( T_N \) should be high, then the image pair is still regarded as non-matched if only few corresponding patches are matched. If the scene content is simple, the recommended value of \( T_N \) is a universal one.

4 | EXPERIMENTS

In order to verify the discriminative capability of the patch feature produced by the proposed PL-VSCN, we employ the well-trained PL-VSCN model to generate patch features and perform the patch matching experiment. Also, the comparison experiment is conducted to prove that the proposed PL-VSCN outperforms other networks.

Furthermore, in order to validate the feasibility of our image matching method, in the second experiment, we conduct the image matching task on public datasets and perform comparisons between the existing methods and our discriminative patch-based method.

4.1 | Experiments about the patch matching

In the first experiment, we would find the answer to the following two questions: (1) whether the proposed PL-VSCN can match the corresponding patch pairs and distinguish the non-correspondences; (2) whether the proposed PL-VSCN can achieve better matching performance when compared with the existing networks.

We employ the Multi-View Stereo dataset (MVS) [35] to train the PL-VSCN, and the dataset contains approximately 1.5 M image patches and 500k 3D points. All the patches are grey scale with the size of 64 × 64, and each patch is associated with a specific 3D point. Two patches can comprise a training
sample, and the patch pair is corresponding (i.e. positive sample) if the patches observe the same 3D point, otherwise the patch pair is non-corresponding (i.e. negative sample) as shown in Figure 7.

In addition, the patches in the dataset come from different scenes, it can be divided into three subsets: the Statue of Liberty (LY), Notre Dame (ND), and Half Dome in Yosemite (YO). We employ each subset to construct the training samples, test samples, and validate samples separately. Different subsets are joined to form the training data, including LY + YO (test on ND), LY + ND (test on YO), YO + ND (test on LY), and LY + YO + ND (test on LY + YO + ND), the test samples are not included in the training samples.

The whole training process consists of 51 epochs, all the training samples are divided into 1000 batches, and each batch contains 400 samples wherein the number of positive samples is equal to the negative samples (with respect to the subsets combination of LY + YO + ND, the batch number is 600).

The Stochastic Gradient Descent (SGD) is adopted to optimize the network in the training process. To achieve an optimal model with high efficiency, the learning rate decays dynamically with the increase of iterations, as given in Equation (8), where $\text{init}_l$ denotes the initial learning rate (0.01), and $\text{cur}_\text{iter}$ represents the quantity of current iterations (total iterations is $51 \times 1000 \approx 5 \times 10^5$).

$$\text{cur}_l = \text{init}_l \cdot 0.9^{\frac{\text{cur}_\text{iter}}{5 \times 10^5}} \tag{8}$$

The training samples are employed to optimize the PL-VSCN, the model, which could achieve minimal loss on the validate samples, is saved, and the saved model is called as the well-trained PL-VSCN. In the patch matching practice, the input of the well-trained PL-VSCN is the image patch, the output is just the normalized feature, and the dot production between two patch features is just the patch similarity.

Figure 8 demonstrates the similarities of the test patch pairs based on the untrained and well-trained models. The initial similarities of the positive and negative samples have no difference on the whole, as shown in Figure 8a. Figure 8b demonstrates that the similarities of the positive samples significantly exceed those of the negative samples, which indicates that the well-trained PL-VSCN can effectively separates the positive samples from negative samples, and the training process make the corresponding pairs (i.e. positive samples) close to each other (high cosine value) and the non-correspondences (i.e. negative samples) far from each other (low cosine value).

In the patch matching experiment, the test samples are adopted to evaluate the matching performance of the well-trained PL-VSCN. Each sample contains two patches, we adopt the well-trained PL-VSCN to construct the patch features, and the similarity of the patch pair (sample similarity) is calculated based on the dot production between the patch features. The similarity threshold is set to calculate the precision and recall (PR) of the test samples, as depicted in Equations (9) and (10), where $N_p$ is the number of position samples, $N_T$ is the number of test samples which are predicted as matched (the similarities are higher than the threshold), $N_{TP}$ is the number of positive samples which are correctly predicted as matched (the similarities are higher than the threshold).

$$P = \frac{N_{TP}}{N_T} \tag{9}$$

$$R = \frac{N_{TP}}{N_p} \tag{10}$$

By changing the similarity threshold, we obtain a set of PR characteristics, the PR curve is drawn, and the area under the PR curve (AUC) is precisely calculated to quantitatively evaluate the matching performance, the higher the AUC value, the better matching performance the well-trained model.

In order to verify the matching performance of the proposed PL-VSCN, we compare our models with other networks, including the followings:

**CNN3 [29]:** The Siamese network consists of two identical towers, and each tower contains three convolutional layers, the similarity of the patch pair is evaluated based on the $L_2$ distance between the patch descriptors.

**CNN7 [32]:** The architecture adopts the deep convolutional network to produce discriminative feature for the input patch, and the Euclidean distance between the patch features is interpreted as the patch similarity.

**MatchNet [28]:** The MatchNet is an end-to-end network that receives the patch pair as input and directly generates the similarity score of the patch pair. The architecture consists of the feature network that constructs the patch representation and the metric network that generates the similarity of the patch pair.

![Figure 7](image-url) Some patch pairs. (a) Positive samples, (b) negative samples
In addition, we separately train the DC-Net and NC-Net to produce the patch feature and perform the patch matching task in the comparison experiment. Furthermore, we manually combine the features from DC-Net and NC-Net to generate the integrated feature as defined in Equation (11), where \( F_D \) and \( F_N \) denote the output features of DC-Net and NC-Net, respectively, \( \overline{F} \) denotes the manually integrated feature, and \( \text{Norm}(\cdot) \) is the normalization operation. We adopt the manually integrated feature to conduct the patch matching task and term the method as ‘Manually-Fusion’ for simplicity.

\[
\overline{F} = \text{Norm}([F_D, F_N])
\]  

(11)

We implement the aforementioned methods and employ the same training sets to learn the top-performing models. The comparison of PR curves between the mentioned networks is shown in Figure 9, and the AUC values of those PR curves are calculated and listed in Table 2. In addition, matching efficiency is another metric to evaluate the matching performance, the total times of the testing process based on various models are precisely recorded, and the average times of processing each sample are calculated and listed in Table 3. The experimental results indicate the following conclusions:

The CNN3 and CNN7 could achieve high efficiency, but matching results in Table 2 demonstrate that the models perform poorly on the test samples. The MatchNet can achieve high accuracy on all datasets, but the matching efficiency is very low, and the matching performance is still slightly inferior to the proposed PL-VSCN.

The PL-VSCN can be regarded as the combination of DC-Net and NC-Net, the architecture of PL-VSCN is more complex, and the matching efficiency is consequently limited, as depicted in Table 3. As the DC-Net and NC-Net describe the patch information from different levels, the corresponding features are mutually complementary. The PL-VSCN integrates the features from DC-Net and NC-Net, it could completely describe the patch information and produce the patch feature with high discriminative capability, the experimental results in Table 2 demonstrate that the PL-VSCN is superior to DC-Net and NC-Net.

Although the Manually-Fusion approach exhibits good performance by manually integrating the features from DC-Net and NC-Net, the matching performance cannot reach the level of PL-VSCN. Specifically, the PL-VSCN adopts FusionNet to optimize the fusion model in the training process, it could effectively eliminate the human interference and automatically achieve an optimal combination. The experimental result indicates that Fusion-Net is more competent to integrate the complementary features from DC-Net and NC-Net.

The proposed PL-VSCN evidently outperforms other models, especially CNN3 and CNN7, it is more capable of producing the discriminative feature despite of the complex architecture, and it can effectively match the positive samples and distinguish the negative samples.
TABLE 2  Area under the PR curves (AUC) of the comparison networks on different set combinations. Bold numbers denote the optimal results, and the proposed PL-VSCN achieves optimal matching performance on all the set combinations.

| Training Sets | Testing Set/sets | CNN3 | CNN7 | MatchNet | DC-Net | NC-Net | PL-VSCN | Manually-Fusion |
|---------------|-----------------|------|------|----------|--------|--------|---------|----------------|
| LY + YO       | ND              | 0.8385 | 0.8527 | 0.9614 | 0.9469 | 0.9550 | 0.9668 | 0.9601 |
| LY + ND       | YO              | 0.7817 | 0.7565 | 0.9474 | 0.9391 | 0.9417 | 0.9656 | 0.9541 |
| YO + ND       | LY              | 0.8626 | 0.8158 | 0.9547 | 0.9387 | 0.9512 | 0.9880 | 0.9518 |
| LY + YO + ND  | LY + YO + ND    | 0.8509 | 0.8414 | 0.9715 | 0.9607 | 0.9629 | 0.9731 | 0.9632 |

TABLE 3  Average times of processing each sample based on the comparison models.

| Methods          | Elapsed Time (ms) |
|------------------|-------------------|
| CNN3             | 0.799             |
| CNN7             | 1.248             |
| MatchNet         | 1.760             |
| DC-Net           | 1.232             |
| NC-Net           | 1.244             |
| PL-VSCN          | 1.327             |
| Manually-Fusion  | 2.476             |

4.2 | Experiments about the image matching

In order to verify the validity of the proposed image matching method in practical applications, we adopt the discriminative patch-based method to perform the image matching task in the second experiment, and we make comparison between the proposed method and other three approaches, which employ the deep convolutional network to directly represent the whole image.

We adopt the public image datasets [36] that contain five subsets (i.e., London Eye, San Marco, Tate Modern, Times Square and Trafalgar), to generate the test image pairs. The matched image pairs (i.e., positive pairs) are directly provided in each subset, and the non-matched pairs (i.e., negative pairs) are manually constructed. Two images that do not exhibit any corresponding content are randomly selected to constitute a negative pair.

With respect to a test pair, we initially detect the object patches in the images and then construct the patch features based on the proposed PL-VSCN. The similarities of the patch pairs are calculated to constitute the similarity matrix, then the corresponding patch pairs are detected based on the mutual matching mechanism. We record the number of the corresponding patch pairs and predict whether or not the image pair is matched.

We randomly select 1000 positive pairs and 1000 negative pairs from each subset and adopt the proposed method to conduct the image matching task on these test pairs. The matching results on five subsets are shown in Figures 10–14, where the red dots denote the positive pairs and the blue dots represent the negative pairs.

The matching results in Figures 10–14 demonstrate that the positive pairs generally possess more corresponding patch pairs on all subsets. This indicates that the proposed method can predict whether or not the image pair is matched according to the number of corresponding patch pairs, and the positive pairs can be separated from the negative pairs by setting the number threshold of corresponding patch pairs ($T_N$).

In addition, the numbers of corresponding patch pairs in different subsets vary greatly, and the corresponding patch pairs of the London Eye and San Marco subsets significantly exceed those of the Tate Modern subset. This is because the scene contents of London Eye and San Marco subsets are more complicated with lots of objects, consequently, more object patches are detected.

In order to compare the proposed method with other algorithms, image matching experiments based on four approaches are performed on the five test sets, and the performance metrics of precision and recall are recorded at different matching thresholds. The comparison approaches are detailed as follows:

**AlexNet [19]:** The network achieves great success in image classification, and it could effectively describe the abstract semantics in an image. Thus, the AlexNet based image representation is assumed as discriminative to match image pairs with the corresponding content. In order to produce the image descriptor based on the pre-trained model, the output of Fc7 layer, which is a vector with 4096 dimensions, is set as the image feature. Furthermore, image similarity is evaluated based on the Euclidean distance between the image features.

**HybirdNet [37]:** Although HybridNet has the same architecture with AlexNet, the approach adopts more training data to optimize the model, and relevant research [38] indicates that the pre-trained model achieves excellent performance in image retrieval on the Oxford Building dataset [39]. In addition, it exhibits strong generalization capability and generates the discriminative feature for other images. Similar to AlexNet, we adopt the output of Fc7 layer to construct image representation and measure image similarity based on the Euclidean distance.

**sHybirdNet [34]:** sHybirdNet is a Siamese network with two identical HybridNet towers. In the training stage, the input of the network is the image pair, each HybridNet tower produces the image descriptor, which is still the 4096-dimensional feature from the output of Fc7 layer, and the Euclidean distance between the image descriptors is interpreted as the
matching score. The weights of the structure are initialized based on the pre-trained HybridNet model and further fine-tuned based on the training data. In our comparison experiment, four subsets are joined to train the network, and the remaining subset is set as the validate and test data.

Proposed method: The proposed method is adopted to detect corresponding patch pairs in each test image pair. In the process, a set of PR characteristics are obtained by changing the number threshold ($T_N$). If the number of the corresponding patch pairs exceeds $T_N$, then the image pair is predicted as matched.

The comparison of PR curves based on the aforementioned methods is shown in Figures 15–19. Different from other approaches, the PR curves of the proposed method, drawn by the red dotted lines, are discrete. The values of $T_N$ can only be non-negative integers, and only several points on the PR curves, marked by asterisks, are meaningful.

The experimental results, shown in Figures 15–19, demonstrate that the proposed method can achieve better performance on most of the datasets. Compared with the whole image matching, the local patch matching can
effectively eliminate the interference of non-corresponding content and achieve high matching accuracy, thus, the patch-based image matching method can achieve great success.

In addition, Figure 15 indicates that the proposed method performs comparatively poor on the Tate Modern subset, this is because relatively few patches are detected in the patch detection stage, as shown in Figure 10, and the corresponding contents in some positive samples are not synchronously detected as patches, consequently, the corresponding patch pair between images cannot be found and the proposed method would fail to match the positive pairs.

The matching performances of the proposed method vary significantly on different datasets, which indicates that the matching results are closely related to the concrete contents and spatial layouts of images. And the proposed method is more competent to match the complicated images which contain lots of objects, and abundant object patches can be detected to predict whether or not the image pair is matched.
5 | CONCLUSION

The article proposes the discriminative patch-based image matching method aimed at the challenging problem in image matching, wherein the non-corresponding content between the matched image pairs generally occupies the main part of the whole image and severely disturbs the image matching process. Instead of directly matching the whole image, we convert the image matching task into the local patch matching, which is less challenging. In the method, the object patches within respective images are first detected, and the discriminative feature of each patch is constructed based on the proposed PL-VSCN. Subsequently, patch similarities are calculated to constitute the similarity matrix and the corresponding patch pairs are detected based on the mutual matching mechanism on the similarity matrix.

The core of the proposed image matching method involves extracting the discriminative feature for the local patch, which can match the corresponding patch pairs and distinguish the non-correspondences. To achieve that, we train the PL-VSCN based on the comparison mechanism, and comparison experiments demonstrate that the proposed PL-VSCN outperforms the existing methods. Furthermore, we compare the proposed image matching method with other approaches, and the experimental results indicate that the patch-based image matching method achieves better matching performance on most of the test datasets.

Nevertheless, the proposed method confronts a new issue that the corresponding content between the image pair may not be synchronously detected as patches in the patch detection stage, in which case the proposed method is invalidated. In order to solve this problem, future research will focus on constructing the spatial adjacent relationship between the patches. Both the similarity score and spatial adjacent between the patches will be set as the criteria to determine whether or not the image pair is matched. As the patch-based representation could precisely describe the image information, relevant researches, involving the patch detection, patch feature extraction and patch adjacent representation, are promising in many vision tasks, and it is meaningful to further optimize the patch-based image matching method.

REFERENCES

1. Dudek, G., Jugessur, D.: Robust place recognition using local appearance based methods. IEEE International Conference on Robotics and Automation, San Francisco, USA, vol. 2, pp. 1030–1035 (2000)
2. Blžiki, M., et al.: Appearance-based landmark selection for visual localization. J. Field Robot. 36(1), 1–7 (2019)
3. Garg, S., Suennderhauf, N., Milford, M.: Lost? appearance-invariant place recognition for opposite viewpoints using visual semantics (2018). arXiv preprint arXiv:1804.05526
4. Ouenghi, S., et al.: Visual odometry and place recognition fusion for vehicle position tracking in urban environments. Sensors. 18(6), 930–939 (2018)
5. Cummins, M.J., Newman, P.M.: Fab-map: appearance-based place recognition and mapping using a learned visual vocabulary model. In: Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, pp. 3–10 (2010)
6. Cummins, M., Newman, P.: FAB-MAP: probabilistic localization and mapping in the space of appearance. Int. J. Robot Res. 27(6), 647–665 (2008)
7. Li, Q., et al.: Place recognition based on deep feature and adaptive weighting of similarity matrix. Neurocomputing. 199(C), 11418C127 (2016)
8. Lowe, D.G.: Object recognition from local scale-invariant features. IEEE Int. Conf. Computer Vision (ICCV), Kerkyra, Greece, vol. 99, pp. 1150–1157 (1999)
9. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60(2), 91IC110 (2014)
10. Bay, H., Tuytelaars, T., Van, Gool, L.: Surf: speeded up robust features. In: European Conference on Computer Vision, Berlin, Heidelberg, pp. 404–417 (2006)
11. Bay, H., et al.: Speeded-up robust features (SURF). Comput Vis Image Underst. 110(3), 346–359 (2008)
12. Rublee, E., et al.: ORB: an efficient alternative to SIFT or SURF. In: 2011 IEEE Int. Conf. Computer Vision (ICCV), Barcelona, Spain, vol. 11, pp. 2–10 (2011)
13. Liu, J.: Image retrieval based on bag-of-words model (2013). arXiv preprint arXiv:1304.5168
14. Sarwar, A., et al.: A novel method for content-based image retrieval to improve the effectiveness of the bag-of-words model using a support vector machine. J. Inf. Sci. 45(1), 117–135 (2019)
15. Jigou, H., et al.: Aggregating local descriptors into a compact image representation. In: IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, USA, pp. 3304–3311 (2010)
16. Jaakkola, T., Haussler, D.: Exploiting generative models in discriminative classifiers. Adv. Neural Inf. Process. Syst. 11(11), 487–493 (1999)
17. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: 2007 IEEE Conf. Computer Vision and Pattern Recognition, Minneapolis, USA, pp. 1–8 (2007)
18. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: European Conference on Computer Vision, Berlin, Heidelberg, pp. 143–156 (2010)
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process Syst., 1097–1105 (2012)
20. Zitnick, C.L., Dollár, P.: Edge boxes: locating object proposals from edges. European Conference on Computer Vision, pp. 391–405. Springer, Cham (2014)
21. Dollár, P., Zitnick, C.L.: Structured forests for fast edge detection. In: 2013 IEEE Conf. Computer Vision and Pattern Recognition, Sydney, Australia, pp. 1841–1848 (2013)
22. Sünderhauf, N., et al.: On the performance of convnets features for place recognition. In: IEEE/RSJ international conference on intelligent robots and systems (IROS), Hamburg, Germany, pp. 4297–4304 (2015)
23. Chen, Z., et al.: Convolutional neural network-based place recognition (2014). arXiv preprint arXiv:1411.1509
24. Sermanet, P., et al.: Overfeat: integrated recognition, localization and detection using convolutional networks (2013). arXiv preprint arXiv:1312.6229
25. Arandjelovic, R., et al.: NetVLAD: CNN architecture for weakly supervised place recognition. IEEE Trans. Pattern Anal. Mach. Intell. 40(6), 1437–1451 (2018)
26. Casianelli, S., et al.: Robust visual semi-semantic loop closure detection by a covisibility graph and CNN features. Robot Autonom. Syst. 92, S58–66 (2017)
27. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Conf. Computer Vision and Pattern Recognition, San Diego, USA, pp. 539–546 (2005)
28. Han, X., et al.: Matchnet: unifying feature and metric learning for patch-based matching. In: 2015 IEEE Conf. Computer Vision and Pattern Recognition, Boston, USA, pp. 3279–3286 (2015)
29. Simo-Serra, E., et al.: Discriminative learning of deep convolutional feature point descriptors. In: 2015 2011 IEEE Int. Conf. ComputerVision (ICCV), Santiago, Chile, pp. 118–126 (2015)
30. Zagoruyko, S., Komodakis, N.: Learning to compare image patches via convolutional neural networks. In: 2015 IEEE Conf. Computer Vision and Pattern Recognition, Boston, USA, pp. 4353–4361 (2015)
31. Bhatnag, V., et al.: PN-Net: conjoined triple deep network for learning local image descriptors, (2016). arXiv preprint arXiv:1601.05030
32. Melekhov, I., Kannala, J., Rahtu, E.: Image patch matching using convolutional descriptors with euclidean distance. In: Asian Conference on Computer Vision, pp. 638–653. Springer, Cham (2016)
33. Lin, T.Y., et al.: Learning deep representations for ground-to-aerial geolocalization. In: 2015 IEEE Conf. Computer Vision and Pattern Recognition, Boston, USA, pp. 5007–5015 (2015)
34. Melekhov, I., Kannala, J., Rahtu, E.: Siamese network features for image matching’ 23rd International Conference on Pattern Recognition (ICPR), pp. 378–383. Cancun, Mexico (2016)
35. Brown, M., Hua, G., Winder, S.: Discriminative learning of local image descriptors. IEEE Trans. Pattern Anal. Mach. Intell. 33(1), 43–57 (2010)
36. Cao, S., Snavely, N.: Learning to match images in large-scale collections. In: European Conference on Computer Vision, pp. 259–270. Springer, Berlin, Heidelberg (2012)
37. Zhou, B., et al.: Learning deep features for scene recognition using places database. In: Adv. Neural Inform. Process. Syst, pp. 487–495 (2014)
38. Chandrasekhar, V., et al.: A practical guide to CNNs and Fisher vectors for image instance retrieval. Signal Process. 128(C), 426–439 (2016)
39. Philbin, J., et al.: Object retrieval with large vocabularies and fast spatial matching. In: IEEE Conf. Computer Vision and Pattern Recognition, Minneapolis, USA, pp. 1–8 (2007)

How to cite this article: You X, Li Q, Li K, Yu A, Bu S. PL-VSCN: Patch-level vision similarity compares network for image matching. IET Comput. Vis. 2021;15:122–135. https://doi.org/10.1049/cvi2.12018