2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

Visual Place Recognition via Local Affine Preserving Matching

Xinyu Ye and Jiayi Ma

***Abstract*— Visual Place Recognition (VPR) is a crucial component for long-term mobile robot autonomy. In this paper, we exploit a coarse-to-fine paradigm to recognize places. In particular, we first select candidate frames for each query image, and then check the spatial geometric relationship between the query and its candidate frames to determine the final place match. In the coarse match stage, we employ the deep learning network to extract global features that encode semantic information of images, then by comparing the similarity between features to obtain a candidate list of the query place. In the fine match stage, we propose an effective and efficient feature matching algorithm for real-time geometrical verification of candidate places, termed as local affine preserving matching (LAP). Extensive experimental results demonstrate that our LAP can significantly promote the VPR performance, and the proposed overall VPR method can achieve much better performance over the current state-of-the-art approaches.**

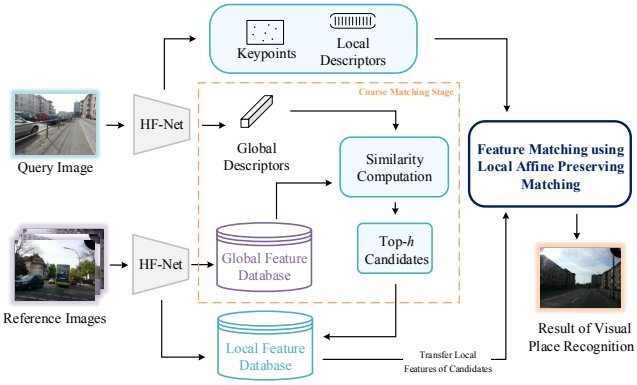

Fig. 1. Overview of the proposed visual place recognition method via local affine preserving matching.

I. Introduction

In the field of autonomous mobile robotics, numerous applications require that robots can recognize previously visited places. When the robot's primary sensors are cameras, the ability to remember pre-visited places is researched as Visual Place Recognition (VPR). However, the VPR problem remains a challenging task since the significant variations in the appearance of places [1], [2]. The variations include illumination change, seasonal change, viewpoint change, or dynamic object occlusion. Moreover, a VPR system also needs to consider the constraints of storage and computation requirements [3].

To address the VPR problem, a variety of methods have been developed over the past decades. The bag-of-words (BoW) framework is a classical approach for VPR, which quantizes local descriptors into visual words to represent images, then employs the obtained compact representations to perform fast queries on image databases. However, the BoW framework is sensitive to perceptual aliasing due to ignoring spatial position information of image features. Moreover, conventional methods are principally based on local and global handcrafted feature descriptors. The global descriptor is computationally efficient but sensitive to environmental changes such as viewpoint variations and partial occlusion, while local descriptors are less sensitive to scale and viewpoint changes but computationally expensive.

In recent years, with the development of deep learning, more and more researchers focus on exploiting convolutional neural networks (CNN) to represent images in VPR tasks [4], [5]. Their results demonstrate that CNN-based descriptors can encode rich semantic information and are more robust to some conditional variations than handcrafted features. Most of the CNN-based VPR methods directly match places by computing the similarity of CNN-based descriptors. This way also limits the VPR performance because it still does not consider the spatial geometric relationship for images. As both BoW-based and CNN-based VPR methods lack the exploration of the spatial geometric relationship, some approaches [6], [7] use RANSAC [8] to conduct a geometrical consistency check. Nevertheless, the RANSAC-alike method relies on a parametric model, thus the performance is degraded in the presence of non-rigid transformation and other complex transformations between image pairs. Moreover, its runtime is unfriendly to real-time tasks like VPR and grows exponentially with an increased outlier rate.

To overcome the above issues, we exploit a coarse-to-fine paradigm to recognize places. We consider the VPR problem as an on-line image retrieval task, where the current input image is regarded as a query image, while pre-visited images are considered as reference images. The overview of our VPR approach is shown in Figure 1. In the coarse matching stage, we exploit a CNN named HF-Net [9] to extract global descriptors, which can encode rich semantic information. The query image's candidate frames are obtained by computing the descriptor similarity between the query and database images. In this way, we can achieve a reliable coarse place match due to the robustness of the CNN-based features. In

This work was supported in part by the National Natural Science Foundation of China (61773295), in part by the Key Research and Development Program of Hubei Province (2020BAB113), and in part by the Natural Science Fund of Hubei Province (2019CFA037).

X. Ye is with the School of Electronic Information and Electrical Engineering, Shanghai Jiaotong University, Shanghai 200240, China (E-mail: xinyu_ye@sjtu.edu.cn).

J. Ma is with the Electronic Information School, Wuhan University, Wuhan 430072, China (J. Ma is the corresponding author. E-mail: jyma2010@gmail.com).

the fine matching stage, we aim to mine the spatial geometric relationship between the query image and its candidate to improve VPR performance as much as possible. To this end, we present a novel feature matching algorithm to do a geometric consistency check for each candidate. The proposed feature matching algorithm employs local features extracted by HF-Net to establish accurate feature correspondences. In the end, if the candidate frame passes the check, then it is regarded as the final place match result.

The major contributions of this paper include the following two aspects. On the one hand, we propose a new feature matching algorithm called local affine preserving matching (LAP) for real-time geometrical verification. The key idea of LAP is to utilize the property of affine-invariant for preserving the neighborhood structures among feature points. The method is able to handle complex transformations in the VPR tasks and obtain accurate correspondences in only a few milliseconds. On the other hand, by making full use of the information of global and local features extracted by the deep learning network, the proposed coarse-to-fine VPR framework can consider both visual appearances and spatial geometrical relationships and emphasizes the latter of importance in the VPR tasks. Extensive experiments reveal that our method can significantly promote the VPR performance and outperforms the current state-of-the-art.

II. Literature Review

A. Image Representation for Visual Place Recognition

Robust image representation is essential for VPR approaches to deal with complicated and changeable environments. The classic VPR method [10] used the Bag of Words (BoW) scheme [11] to encode local features as a more compact representation and efficiently retrieved database images by combining the invert index technique. This place description is viewpoint invariant but ignores the geometric and spatial structure so that not robust to complex scenarios. Recently, different off-the-shelf CNNs [12], [13] have been applied to VPR and shown promising performance. In particular, Sunderhauf *et al.* [14] evaluated the VPR performance for the response of each layer of AlexNet [15]. Nevertheless, these CNN-based global image representations are not robust enough to viewpoint variations and partial occlusion. Subsequently, Arandjelovic *et al.* [12] trained an end-to-end CNN architecture for VPR. Khaliq *et al.* [16] identified salient regions of convolutional layer responses and used vector of locally aggregated descriptors (VLAD) [17] to encode the regional features. Although these methods combine CNNs with regional features, they still do not consider the spatial relationship between local features, which is beneficial to improve the robustness of place matching [1]. To this end, in this paper, CNN-based global and local features are utilized, and the latter is devoted to mining the geometric structure relationship.

B. Geometric Verification

For a VPR system, we can verify the place matches by the geometrical and temporal consistency checking. This paper focuses on using geometrical information of image feature points to enhance the robustness of VPR systems. Common geometric verification methods in VPR adopt feature matching algorithms to establish reliable correspondences. By determining whether a fundamental matrix can be computed or there are enough inliers between the query image and candidate frame, the system can decide whether to regard the candidate frame as the final place match result. RANSAC [8] is widely utilized in the geometrical consistency check. It is a kind of resampling-based feature matching method that relies on a pre-defined parameter model, leading to less effectiveness in addressing complex transformations and dominating outliers. Several variants of RANSAC have been proposed to improve the algorithm, such as SCRAMSAC [18], GC-RANSAC [19] and MAGSAC++ [20]. Besides, researchers have designed several effective geometric verification methods for VPR. For example, Yue *et al.* [21] proposed a graph verification method, which builds an undirected triangular graph for candidate images to graph match with the current image. Camara *et al.* [22] encoded spatial information in the form of CNN features to verify candidate place matches. Recently, several locality consistency assumption-based methods [23], [24], [25], [26], [27] are proposed for image matching and achieved encouraging performance in both accurate and time efficiency. In this paper, we present a new feature matching method also based on the locality consistency assumption for better geometric verification in VPR, and extensive experiments prove noticeable improvements over the state-of-the-art.

III. Approach Overview

Figure 1 illustrates an overview of the proposed VPR method. We use a state-of-the-art technique HF-Net [9] to extract global and local descriptors of each image. HF-Net employs a MobileNet [13] backbone as the encoder, then computes the global descriptor by a NetVLAD [12] layer, and adopts the SuperPoint [28] architecture to obtain the exact location of keypoints and local descriptor vectors. HF-Net trained jointly these three models with multitasking distillation from different off-the-shelf teacher networks. On the off-line stage, the features of reference images are stacked to form two databases: global and local feature databases. On the on-line stage, a coarse-to-fine strategy is utilized to find the place match of the query place. We first select top k candidates as the coarse match by ranking the global feature similarity between the query image and reference images, then determine the final place match by verifying the geometric consistency between the query image and its candidates.

IV. Local Affine Preserving Matching

In the first stage of our VPR approach, we can obtain candidate frames of each query image only by matching global features. Although global features extracted by HF-Net can capture strong semantic information, they ignore spatial geometric relationships. Thereby, other places similar in appearance to the query place compete as candidates of

the query place. To distinguish true place match from false place match in the candidate list, we present an effective and efficient feature matching algorithm to mining geometric relationships between the query and its candidate frames. Our feature matching algorithm aims to exploit local features extracted by HF-Net to establish accurate feature correspondences between the query image and the candidate frame. If a place match, *i.e.*, a query image and one of its candidate frames have enough feature correspondences, we consider it as a true place match.

A. Problem Formulation

After obtaining local feature descriptors and their spatial positions in the image pairs, we utilize a similarity constraint to construct a set of putative correspondences S. The similarity constraint requires that points can only match points with similar local descriptors. Thereby we can achieve the putative set $S = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$, where $\mathbf{x}_i$ and $\mathbf{y}_i$ are vectors denoting the spatial positions of feature points in the query image and the candidate frame image, respectively. However, due to ambiguities in the similarity constraint, S is typically contaminated by unavoidable noise and outliers. Therefore, we aim to remove the false matches in the putative set S and find an optimal inlier set $\mathcal{I}^*$. Namely, we formulate the mismatch removal problem into an optimization model:

$$\mathcal{I}^* = \arg\min_{\mathcal{I}} C(\mathcal{I}; S, \lambda), \tag{1}$$

with the cost function C defined as:

$$C(\mathcal{I}; S, \lambda) = \sum_{i \in \mathcal{I}} c(\mathbf{x}_i, \mathbf{y}_i) + \lambda(N - |\mathcal{I}|), \tag{2}$$

where $|\cdot|$ denotes the cardinality of a set, $c(\mathbf{x}_i, \mathbf{y}_i)$ represents the outlier's confidence, and a larger value indicates a higher probability that $(\mathbf{x}_i, \mathbf{y}_i)$ is a false match. The second term of the cost function is used to discourage the outliers, and parameter $\lambda > 0$ controls the trade-off between the two terms. In the following, we will focus on designing a geometric constraint to define $c(\mathbf{x}_i, \mathbf{y}_i)$, which makes outliers have a larger value while inliers have a smaller value. Ideally, the optimal solution should realize zero penalty, *i.e.*, the first term of C should be zero.

In the place recognition tasks, there are not only rigid transformations, *e.g.*, scale and rotation change, but also non-rigid transformations, *e.g.*, viewpoint change. After these transformations, the absolute positions of feature points would be changed, but the topological structure of neighborhoods among feature points may not modify freely due to physical constraints. Moreover, non-rigid transformations cannot be modeled globally, but an affine or homography transformation in general can well model the topological structures of local regions inside the image. In other words, the local topological structures of inliers could preserve affine-invariant, while outliers could not preserve. Therefore, we aim to exploit local affine-invariant to measure the probability of outliers, *i.e.*, $c(\mathbf{x}_i, \mathbf{y}_i)$.

Fig. 2. Schematic illustration of the motion consistency. The inliers (in blue) and outliers (in red) are shown in the left and right parts, respectively. The head and tail of each arrow in the motion field correspond to the positions of two corresponding feature points in the image pair.

B. Neighborhood Construction Using Motion Consistency

For an inlier, the local affine-invariant can be preserved well when inliers construct its local topological structure. However, if we directly search the K-nearest neighbors for each feature point and regard them as neighborhood, then the neighborhood of inliers would be contaminated by outliers, leading to bad local affine-invariant preservation. Therefore, we use motion consistency to build a reliable neighborhood for each feature point.

Figure 2 shows the result of feature matching on the image pair captured by the same place. As we can see, the correct matches (blue displacement vectors) tend to have coherent motions and neighboring points sharing similar motions, while the false matches (red displacement vectors) tend to be randomly distributed across the image domain. Based on this observation, we define the motion consensus similarity:

$$\mu_{ij} = \frac{\min\{|\mathbf{v}_i|, |\mathbf{v}_j|\}}{\max\{|\mathbf{v}_i|, |\mathbf{v}_j|\}} \cdot \frac{(\mathbf{v}_i, \mathbf{v}_j)}{|\mathbf{v}_i| \cdot |\mathbf{v}_j|}, \tag{3}$$

where $\mathbf{v}_{(\cdot)}$ is the displacement vector associated with point pair $(\mathbf{x}_{(\cdot)}, \mathbf{y}_{(\cdot)})$, $\mu_{ij} \in [-1, 1]$ and $(\cdot, \cdot)$ denoting the inner product. The larger the value, the higher the consensus.

We first search the M-nearest neighbors for each feature point $\mathbf{x}_i$, then compute the motion consensus similarity between vector $\mathbf{v}_i$ and its neighbor displacement vectors, finally select the top K neighbors with the greatest similarity scores as neighborhood $\mathcal{N}_{\mathbf{x}_i}$. Meanwhile, the corresponding feature points of $\mathcal{N}_{\mathbf{x}_i}$ in the paired image is denoted as $\mathcal{N}_{\mathbf{y}_i}$.

C. Reconstruction Error Calculation Based on Local Affine Invariance

For K ($K > 3$) neighbors of each feature point, we choose randomly 3 neighbors to construct a minimum topological unit, hence there have

$$U = C_K^3 = \frac{K!}{6(K-3)!} \tag{4}$$

topological units in the neighborhood of $\mathbf{x}_i$.

Figure 3 illustrates the change of one topological uint after the affine transformation. According to the theory of multiple view geometry in computer vision, the affine transformation has three important invariants: parallel lines, ratios of lengths of parallel line segments, and ratios of areas. In this paper, we use the invariance of the ratio of areas to represent local affine invariance. As shown in Figure 3, given a feature point $\mathbf{x}_i$ and its neighbors $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$, they can form three triangles. After affine transformation, they are converted to $\mathbf{y}_i$, $\mathbf{y}_1$, $\mathbf{y}_2$ and $\mathbf{y}_3$, respectively. Moreover, the areas of these

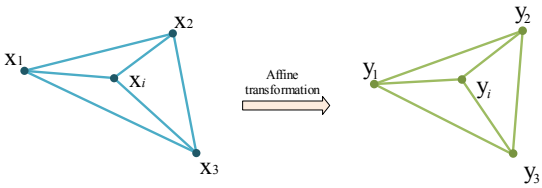

Fig. 3. The affine-invariant of topological structures. Points $\mathbf{x}_i$, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ form three triangles $\Delta\mathbf{x}_i\mathbf{x}_1\mathbf{x}_2$, $\Delta\mathbf{x}_i\mathbf{x}_2\mathbf{x}_3$ and $\Delta\mathbf{x}_i\mathbf{x}_3\mathbf{x}_1$ which correspond to $\Delta\mathbf{y}_i\mathbf{y}_1\mathbf{y}_2$, $\Delta\mathbf{y}_i\mathbf{y}_2\mathbf{y}_3$ and $\Delta\mathbf{y}_i\mathbf{y}_3\mathbf{y}_1$ after transformation.

triangles satisfy the following equation based on the affine invariant:

$$\frac{S_{\Delta\mathbf{x}_i\mathbf{x}_1\mathbf{x}_2}}{S_{\Delta\mathbf{y}_i\mathbf{y}_1\mathbf{y}_2}} = \frac{S_{\Delta\mathbf{x}_i\mathbf{x}_2\mathbf{x}_3}}{S_{\Delta\mathbf{y}_i\mathbf{y}_2\mathbf{y}_3}} = \frac{S_{\Delta\mathbf{y}_i\mathbf{y}_3\mathbf{y}_1}}{S_{\Delta\mathbf{y}_i\mathbf{y}_3\mathbf{y}_1}}. \tag{5}$$

Then we determine the three ratios that are the ratio of any two of the three triangles for $\mathbf{x}_i$, *i.e.*,

$$r_1 = \frac{S_{\Delta\mathbf{x}_i\mathbf{x}_1\mathbf{x}_2}}{S_{\Delta\mathbf{x}_i\mathbf{x}_2\mathbf{x}_3}}, \quad r_2 = \frac{S_{\Delta\mathbf{x}_i\mathbf{x}_2\mathbf{x}_3}}{S_{\Delta\mathbf{x}_i\mathbf{x}_3\mathbf{x}_1}}, \quad r_3 = \frac{S_{\Delta\mathbf{x}_i\mathbf{x}_3\mathbf{x}_1}}{S_{\Delta\mathbf{x}_i\mathbf{x}_1\mathbf{x}_2}}. \tag{6}$$

Similarly, the three ratios for $\mathbf{y}_i$ are represented as $\{r'_m\}_{m=1}^3$. By using a simple transform for Eq. (5), we have $\{r_m = r'_m\}_{m=1}^3$. That is to say, ideally, if a minimum topological unit satisfies the affine transformation, then r_m would be very close to r'_m, vice versa. Based on this fact, we define $c(\mathbf{x}_i, \mathbf{y}_i)$ in the cost function as

$$c(\mathbf{x}_i, \mathbf{y}_i) = \sum_{u=1}^{\alpha U} \sum_{m=1}^{3} \left(1 - \exp\left(|r_{ium} - r'_{ium}|\right)\right), \tag{7}$$

where $\alpha \in (0, 1)$, r_{ium} and r'_{ium} represents the m-th ratio in the u-th minimum topological unit of $\mathbf{x}_i$ and $\mathbf{y}_i$, respectively. Although exploiting motion consistency can obtain reliable neighborhoods, there still exist a few outliers to perturb the topological structure of inliers, leading to a large penalty. To further mitigate the effects of this situation, we sort U reconstruction errors, *i.e.*, $|r_{ium} - r'_{ium}|$, then select the smallest αU topological units to participate in the calculation. In this way, we can ensure the neighborhood topological structure of inliers as little as possible affected by outliers and achieve the low $c(\mathbf{x}_i, \mathbf{y}_i)$ values as much as possible. At the same time, due to the neighborhood topological structure of outliers that cannot be preserved after transformation, they also have no local affine invariance, thereby causing the large $c(\mathbf{x}_i, \mathbf{y}_i)$ values.

D. A Closed-form Solution

To optimize the objective function, we introduce an $N \times 1$ binary vector $\mathbf{p}$ to associate the putative set S, where $p_i = 1$ indicates an inlier, and an outlier otherwise. Accordingly, Eq. (2) is equivalent to the following minimization problem:

$$C(\mathbf{p}; S, \lambda, \tau) = \sum_{i=1}^{N} p_i c(\mathbf{x}_i, \mathbf{y}_i) + \lambda(N - \sum_{i=1}^{N} p_i). \tag{8}$$

We can rewrite the form of Eq. (8) by merging the terms related to p_i and obtain:

$$C(\mathbf{p}; S, \lambda, \tau) = \sum_{i=1}^{N} p_i(c_i - \lambda) + \lambda N, \tag{9}$$

Algorithm 1: The LAP Algorithm

Input: putative set $\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, parameters M, K, α, λ

Output: inlier set $\mathcal{I}^*$

1 Construct reliable neighborhood $\{\mathcal{N}_{\mathbf{x}_i}\}_{i=1}^N$ and $\{\mathcal{N}_{\mathbf{y}_i}\}_{i=1}^N$ based on motion consistency;
2 Calculate cost $\{c_i\}_{i=1}^N$ using Eq. (7) based on local affine invariance;
3 Determine $\mathcal{I}^*$ using Eqs. (10) and (11).

where c_i is short for $c(\mathbf{x}_i, \mathbf{y}_i)$. Due to the neighborhood relationship among the feature points is fixed, $\{c_i\}_{i=1}^N$ can be estimated in advance. Hence, the only unknown variable in Eq. (9) is p_i. Any putative match with c_i smaller than λ will lead to a negative term of Eq. (9), and vice versa. Consequently, the optimal solution of $\mathbf{p}$ is determined as:

$$p_i = \begin{cases} 1, & c_i \le \lambda \\ 0, & c_i > \lambda \end{cases}, \quad i = 1, \ldots, N\,. \tag{10}$$

Finally, the optimal inlier set $\mathcal{I}^*$ is represented as:

$$\mathcal{I}^* = \{i | p_i = 1, i = 1, \ldots, N\}\,. \tag{11}$$

We name the proposed feature matching method as LAP (local affine preserving matching), and summarize it in Algorithm 1. In this paper, we use LAP to verify the geometric relationship between the query image and its candidate frames.

E. Computational Complexity

As the putative set S has N feature point correspondences, the time complexity of searching M nearest neighbors for each feature point is about $O((M+N)\log N)$ by using the K-D tree. The total time complexity of construct reliable neighborhoods is less than $O(MN + M^2)$. The calculation of cost c_i includes computing and sorting U reconstruction errors, resulting in that the time complexity is less than $O(UN + U^2)$. Therefore, the total time complexity of our LAP is close to $O((M+N)\log N + (M+U)N + M^2 + U^2)$. As M and U is a constant and less than N, the time complexity of LAP can be simply written as $O(N \log N)$. That is to say, LAP has linearithmic complexity about the element number of the given putative set, which is significant for the VPR task.

V. Experimental Results

A. Implementation Details

Table I summarizes the parameter settings of the proposed VPR approach. In the experiment, we run the HF-Net on an Intel(R) Core(TM) i7-5930K CPU @ 3.50GHz machine, with four GeForce GTX TITAN X GPU. Moreover, we select the top 500 reliable features for VPR to check candidate image pairs, and the distance of Non-Maximum-Suppression (NMS) is set to 4 pixels.

TABLE I

PARAMETER SETTINGS.

Parameter	Description	Value
M	Size of the initial neighborhood	20
K	Size of the reliable neighborhood	10
α	Ratio of reliable topological units	0.5
λ	Threshold of distinguishing inliers and outliers	0.55
h	Number of candidate frames	3

TABLE II

BENCHMARK PLACE RECOGNITION DATASETS.

Dataset	Traverse		Variation	
	Ref.	Query	Viewpoint	Condition
Gardens Point	200	200	Adequate	Significant
Berlin Halenseestrasse	67	157	Significant	Moderate
Berlin A100	85	81	Adequate	Adequate
Synthesized Nordland	1415	1415	Moderate	Strong

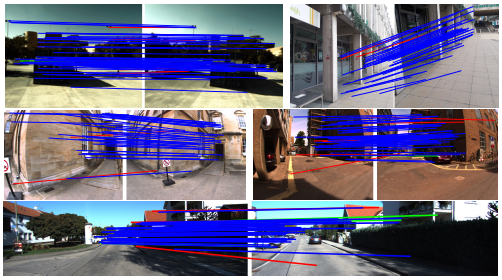

Fig. 4. Feature matching results of LAP on 5 representative image pairs. The ratio of outliers in the 5 image pairs are 88.33%, 63.64%, 88.89%, 64.56% and 50.72%, respectively. The head and tail of each arrow correspond to the positions of feature points in two images (blue = true positive, green = false negative, red = false positive).

B. Datasets

We use four representative datasets to evaluate the VPR performance more comprehensively, with details summarized in Table II. The Gardens Point dataset contains three different traverses for the same route, and we choose the traverses recorded from the left side of the walkways at day-time as reference images and the traverses recorded from the right side of the walkways at night-time as query images. Both the Berlin Halenseestrasse and Berlin A100 datasets have strong viewpoint variations. Berlin Halenseestrasse contains day-time traverses taken by a car user and a bicycle user, while Berlin A100 contains two traverses recorded from different cars in the day-time. The Synthesized Nordland dataset contains two traverses collected by a camera mounted on a train in summer and winter. The frame-to-frame ground-truth files of these datasets are available [22], and we use a ground truth tolerance of ± 3 frames.

C. Results on Feature Matching

Before evaluating the overall VPR performance, we first verify the effectiveness and efficiency of our LAP matching. Figure 4 illustrates the qualitative feature matching results on five image pairs. These image pairs covered several challenging situations for VPR, such as viewpoint variation, non-rigid transformation, massive repetitive structures, and scale variations. We manually build ground truth by checking each putative match in each image pair. The matching performance is characterized by precision and recall. In this case, precision is defined as the percentage of true inliers among preserved matches by a matching algorithm, and the recall is the ratio of the preserved true inliers among the whole ground-truth inlier set. With our LAP, the precision and recall of these image pairs are (89.66%, 98.11%), (96.08%, 100%), (95.23%, 100%), (89.29%, 98.04%), (91.43%, 91.43%), respectively. We can see that most of the correct matches are detected successfully while only preserving a few false matches. The result demonstrates that our LAP has great potential in tackling complex scenes in VPR tasks.

Figure 5 shows the quantitative comparisons of feature matching using our LAP and five state-of-the-art feature matching methods, including RANSAC [8], ICF [30], GMS [24], MR-RPM [31], and LMR [32] on two feature matching datasets. The first dataset consists of 24 challenging image pairs from place recognition datasets, where the ground-truth correspondences are established by manually checking the correctness of each putative correspondence in each image pair, and we called it *VPR-FM*. Another dataset is called *DTU* [33], a publicly available dataset, which contains the images of many different scenes taken from 49 or 64 positions. From the result, we can find that the precision of our LAP ranks second in the *VPR-FM* and third in the *DTU*, but it can achieve satisfying recall and F-score, where F-score is the harmonic mean of precision and recall and equals to $2 \cdot Precision \times Recall/(Precision + Recall)$. For *VPR-FM* dataset, our LAP achieves the best F-score, and its runtime is more than one order of magnitude faster than RANSAC. For *DTU* dataset, the F-score of LMR is slightly higher than LAP, but LAP is almost four times faster than LMR in time performance. That is to say, LAP can achieve the best trade-off on precision, recall and runtime on these two datasets. In addition, Figure 6 illustrates the result of the ablation experiments with different parameter settings.

D. Results on Visual Place Recognition

To demonstrate the effectiveness of our VPR system and analysis the effect of the number of candidates, we illustrate the precision-recall curves (PR curves) of four VPR schemes on four datasets in Figure 7. The first scheme compares the similarity between the global descriptors extracted by HF-Net and selects the top two reference images that are the most similar to the query image to conduct a ratio test. Only if it passes the test, the reference image is regarded as the current place match result. The remaining schemes follow the proposed VPR framework, *i.e.*, we first select respectively 1, 3, 5 candidates for a query image based on the similarity between the global descriptors, then use LAP to do geometric verification. If the inliers number between the candidate and the query image is greater than the pre-defined threshold, the candidate can be considered the place match result. By setting different thresholds, we can obtain the PR curve. In the VPR tasks, there have three critical indexes for the PR curve, such as the area under the curve (AUC), the maximum recall rate when the precision rate is 100%

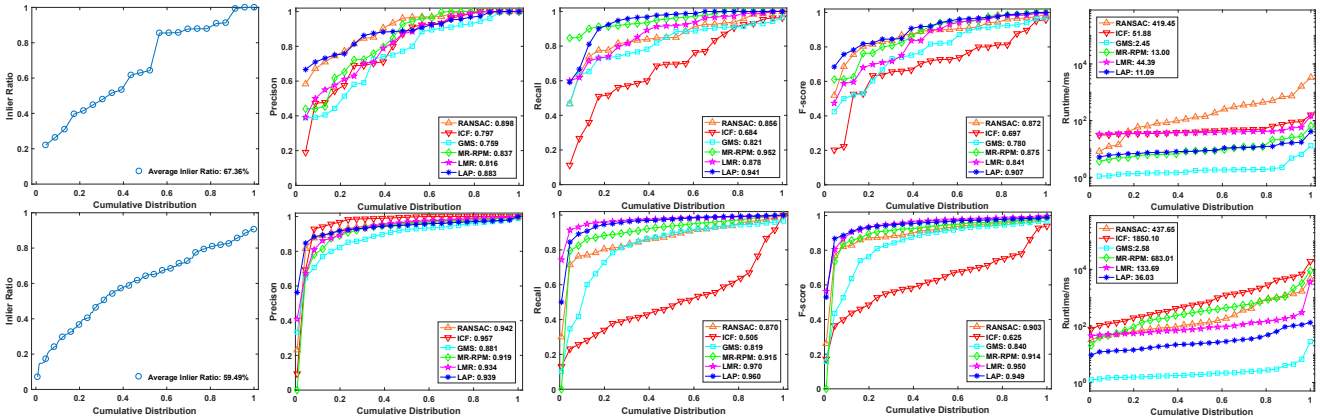

Fig. 5. Quantitative feature matching results of RANSAC, ICF, GMS, MR-RPM, LMR and LAP on the two datasets, such as *VPR-FM* (top) and *DTU* (bottom). From left to right: Initial inlier ratio, precision, recall, F-score and runtime with respect to the cumulative distribution. A point on the curve with coordinate (x, y) denotes that there are $100 \times x$ percents of image pairs with initial inlier ratio, precision, recall, F-score or runtime no more than y.

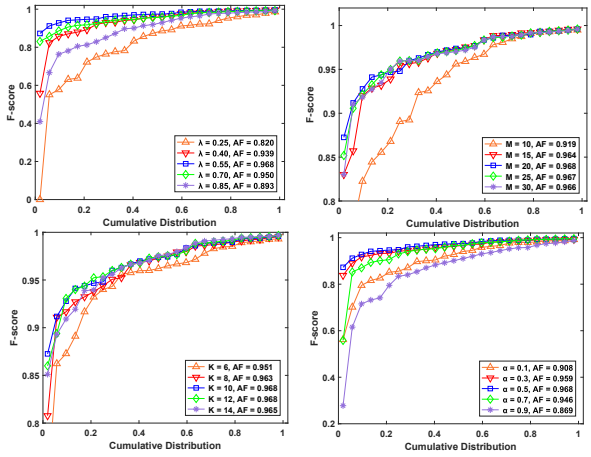

Fig. 6. F-score with respect to the cumulative distribution with different parameter settings on the *DAISY* dataset [29]. λ controls the threshold of distinguishing inliers and outliers. The larger the parameter M is, the larger the neighborhood range to be considered. The higher the value of K and $alpha$, the greater the number of minimum topological units. Excessive or insufficient minimum topological units are not beneficial to the performance.

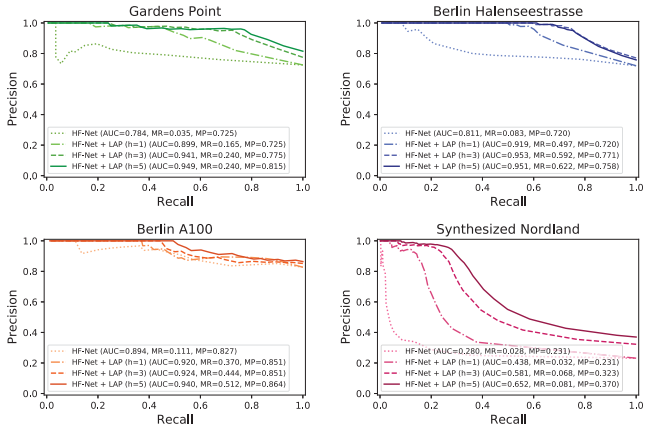

Fig. 7. Precision-recall curves of four schemes for VPR.

(MR), the maximum precision rate when the recall rate is 100% (MP). From the result, we see that feature matching can significantly improve VPR performance, especially for

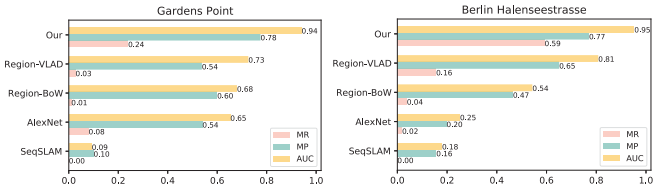

Fig. 8. Quantitative comparison of AUC, MR, MP with four state-of-the-art methods on two datasets.

the MR index, which is of great significance when VPR is applied to SLAM systems. Besides, as the number of candidates becomes increase, we see that the three indexes increase, and the performance is approximate on $h = 3$ and $h = 5$. Therefore, we choose $h = 3$ as a trade-off between time and performance.

Figure 8 compares the AUC, MR and MP results of our VPR method with SeqSLAM [34], AlexNet [14], Region-BoW [35] and Region-VLAD [16]. From the results, we see that the proposed VPR approach notably outperforms the other comparison methods in all three indexes. Moreover, we can achieve the promising results that AUC is 0.94 and 0.95, respectively, although these datasets contain strong viewpoint variation and illuminate variation.

VI. Conclusion

In this paper, we propose a coarse-to-fine framework for visual place recognition. The proposed approach makes full use of the semantic information and the spatial geometric relationship in VPR data. Instead of focusing on image representation, we aim to emphasize the importance of geometric relationships in the VPR tasks. To this end, we present an effective and efficient feature matching algorithm named LAP to conduct a geometric consistency check for candidates. LAP exploits the property of affine-invariant for preserving the neighborhood structures among feature points. It can significantly improve VPR performance while keeping real-time performance. Qualitative and quantitative comparative experiments have demonstrated that our method can handle various challenges and outperform the state-of-the-art on four representative datasets.

REFERENCES

[1] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2015.

[2] M. Zaffar, S. Ehsan, M. Milford, and K. McDonald-Maier, "CoHOG: A light-weight, compute-efficient, and training-free visual place recognition technique for changing environments," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1835–1842, 2020.

[3] S. Garg and M. Milford, "Fast, compact and highly scalable visual place recognition through sequence-based matching of overloaded representations," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2020, pp. 3341–3348.

[4] A. Khaliq, S. Ehsan, Z. Chen, M. Milford, and K. McDonald-Maier, "A holistic visual place recognition approach using lightweight cnns for significant viewpoint and appearance changes," *IEEE Transactions on Robotics*, vol. 36, no. 2, pp. 561–569, 2020.

[5] N. Merrill and G. Huang, "CALC2. 0: Combining appearance, semantic and geometric information for robust and efficient visual loop closure," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2019, pp. 4554–4561.

[6] K. A. Tsintotas, L. Bampis, and A. Gasteratos, "Assigning visual words to places for loop closure detection," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2018, pp. 5979–5985.

[7] L. Bampis, A. Amanatiadis, and A. Gasteratos, "Fast loop-closure detection using visual-word-vectors from image sequences," *The International Journal of Robotics Research*, vol. 37, no. 1, pp. 62–82, 2018.

[8] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[9] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 716–12 725.

[10] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: a versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[11] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2003, pp. 1–8.

[12] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5297–5307.

[13] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[14] N. Sunderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015, pp. 4297–4304.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[16] A. Khaliq, S. Ehsan, Z. Chen, M. Milford, and K. McDonald-Maier, "A holistic visual place recognition approach using lightweight cnns for significant viewpoint and appearance changes," *IEEE Transactions on Robotics*, vol. 36, no. 2, pp. 561–569, 2019.

[17] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proceedings of the IEEE computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3304–3311.

[18] T. Sattler, B. Leibe, and L. Kobbelt, "SCRAMSAC: Improving ransac's efficiency with a spatial consistency filter," in *Proceedings of the IEEE International Conference on Computer Vision*, 2009, pp. 2090–2097.

[19] D. Barath and J. Matas, "Graph-cut RANSAC," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[20] D. Barath, J. Noskova, M. Ivashechkin, and J. Matas, "MAGSAC++, a fast, reliable and accurate robust estimator," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1304–1312.

[21] H. Yue, J. Miao, Y. Yu, W. Chen, and C. Wen, "Robust loop closure detection based on bag of superpoints and graph verification," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2019, pp. 3787–3793.

[22] L. G. Camara and L. Přeučil, "Spatio-semantic convnet-based visual place recognition," in *Proceedings of the European Conference on Mobile Robots*, 2019, pp. 1–8.

[23] J. Ma, J. Zhao, J. Jiang, H. Zhou, and X. Guo, "Locality preserving matching," *International Journal of Computer Vision*, vol. 127, no. 5, pp. 512–531, 2019.

[24] J.-W. Bian, W.-Y. Lin, Y. Liu, L. Zhang, S. K. Yeung, M.-M. Cheng, and I. Reid, "GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence," *International Journal of Computer Vision*, vol. 128, no. 6, pp. 1580–1593, 2020.

[25] X. Jiang, J. Ma, J. Jiang, and X. Guo, "Robust feature matching using spatial clustering with heavy outliers," *IEEE Transactions on Image Processing*, vol. 29, pp. 736–746, 2019.

[26] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image matching from handcrafted to deep features: A survey," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 23–79, 2021.

[27] X. Jiang, J. Ma, G. Xiao, Z. Shao, and X. Guo, "A review of multimodal image matching: Methods and applications," *Information Fusion*, vol. 73, pp. 22–71, 2021.

[28] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 224–236.

[29] E. Tola, V. Lepetit, and P. Fua, "Daisy: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 815–830, 2009.

[30] X. Li and Z. Hu, "Rejecting mismatches by correspondence function," *International Journal of Computer Vision*, vol. 89, no. 1, pp. 1–17, 2010.

[31] J. Ma, J. Wu, J. Zhao, J. Jiang, H. Zhou, and Q. Z. Sheng, "Nonrigid point set registration with robust transformation learning under manifold regularization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 12, pp. 3584–3597, 2019.

[32] J. Ma, X. Jiang, J. Jiang, J. Zhao, and X. Guo, "LMR: Learning a two-class classifier for mismatch removal," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 4045–4059, 2019.

[33] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, "Large-scale data for multiple-view stereopsis," *International Journal of Computer Vision*, vol. 120, no. 2, pp. 153–168, 2016.

[34] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *Proceedings of theIEEE International Conference on Robotics and Automation*, 2012, pp. 1643–1649.

[35] Z. Chen, F. Maffra, I. Sa, and M. Chli, "Only look once, mining distinctive landmarks from convnet for visual place recognition," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017, pp. 9–16.