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

Appearance-based Loop Closure Detection via Bidirectional Manifold Representation Consensus

Kaining Zhang, Zizhuo Li, and Jiayi Ma

***Abstract*— Loop closure detection (LCD), which aims to deal with the drift emerging when robots travel around the route, plays a key role in a simultaneous localization and mapping system. Unlike most current methods which focus on seeking an appropriate representation of images, we propose a novel two-stage pipeline dominated by the estimation of spatial geometric relationship. When a query image occurs, we select candidates on-line according to the similarity of global semantic features in the first stage, and then conduct robust geometric confirmation to verify true loop-closing pairs in the second stage. To this end, a robust feature matching algorithm, termed as bidirectional manifold representation consensus (BMRC), is proposed. In particular, we utilize manifold representation to construct local neighborhood structures of feature points and formulate the matching problem into an optimization model, enabling linearithmic time complexity via a closed-form solution. Furthermore, we propose a dynamic place partition strategy based on BMRC to segment image streams with similar content into a place, which can mine more valid candidate frames, improving the recall rate of the whole system. Extensive experiments on several publicly available datasets reveal that BMRC has a good performance in the general feature matching task and the proposed pipeline outperforms the current state-of-the-art approaches in the LCD task.**

I. Introduction

Simultaneous localization and mapping (SLAM) system refers to realize self-orientation and map-building. Due to the existence of cumulative error generated by odometry and optimization, it would be challenging to achieve accurate localization and construct drift-free map. To address this issue, loop closure detection (LCD) engine is considered, which is responsible for identifying revisited trajectory regions during the traveling route and providing supplementary information regarding with the measurements' arrangement in the 3D space, for example, it would fuse the current pose with an earlier one to rectify incremental pose drift. Therefore, LCD constitutes a crucial component of SLAM systems, enabling a significant improvement of the ultimate performance.

In the literature, LCD is typically divided into two steps: selecting candidate frames and identifying loop-closing pairs. The first procedure is analogous to the image retrieval task, which aims to select the most similar images from a database. Clearly, an appropriate image representation is essential and majority approaches are based on bag-of-words (BoW) [1]. These methods first search local features and descriptors based on handcrafted techniques such as SIFT [2], SURF [3] and ORB [4], followed by relying on visual words generated via quantizing the space of local descriptors, coupled to TF-IDF scoring. BoW model usually ignores the geometric structure, and its expressive ability would degenerate in case of complex scenes. In addition, the performance relies on the size of codebook, and sometimes a new codebook is required when new regions occur.

This work was supported by the National Natural Science Foundation of China (No. 61773295). The authors are with the Electronic Information School, Wuhan University, Wuhan 430072, China (J. Ma is the corresponding author. E-mail: jyma2010@gmail.com).

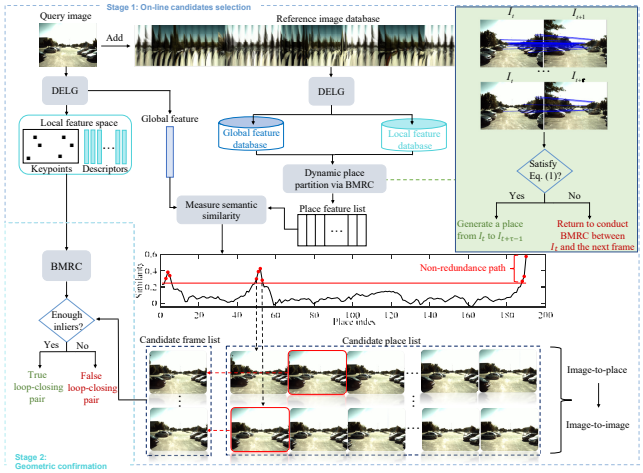

Fig. 1: A schematic diagram of the whole LCD system.

In general, selecting candidate frames is devoted to acquiring high recall, while the second step of LCD, also known as geometric confirmation, tries to improve precision. It is substantially achieved by first constructing putative matches between an image pair and then removing mismatches. Nearly all state-of-the-art LCD methods adopt RANSAC [5], *i.e.*, acquiring the smallest possible outlier-free subset to estimate a parametric model by resampling based on a hypothesize-and-verify strategy, as the default setting. It performs well when geometric constraints are parametric, but degenerates when non-rigid deformation and dominated outliers exist, or the geometric model is difficult to obtain. Therefore, when complex scenes occur, both the accuracy and efficiency of RANSAC would degrade severely, cutting down the performance of LCD unavoidably.

Contributions. There are three main contributions in this work. Firstly, a simple yet efficient robust feature matching method, called bidirectional manifold representation consensus (BMRC), is proposed. It mainly considers the intrinsic property of the potential inliers' neighborhoods, reconstructing the bidirectional local topological structures via manifold representation, which enables to establish reliable correspondences in scenes where RANSAC may fail, *i.e.*, scenes with non-rigid deformation or dominated outliers. Meanwhile, a

closed-form solution with linearithmic time complexity is derived for our BMRC, enabling it to achieve real time. Secondly, based on BMRC, a dynamic place partition (DPP) strategy is designed according to the sequential nature of spatial data in the robotics domain. It segments consecutive images with similar content into a place, which can achieve image-to-place association and avoid adjacent images to compete for the identification of candidate. Thirdly, combining BMRC and DPP, a novel two-stage semantic-to-geometric pipeline for LCD is proposed. It focuses on mining spatial geometric relationship between different scenes and can achieve state-of-the-art performance in the LCD task.

II. Related Work

A. Loop Closure Detection

State-of-the-art LCD approaches can mainly be divided into two categories according to the procedure of codebook construction, where methods training the codebook off-line are placed in the first category. For example, FAB-MAP 2.0 [6] utilizes a pre-trained codebook followed by a probabilistic model including a Chow Liu tree. It can perform reliable results over a 1000 KM trajectory. Later, to accelerate LCD procedure, an off-line codebook has been built via discretizing a binary descriptor space, which achieves one order of magnitude faster than previous approaches [7].

Off-line approaches are popular for its generality, but readily affected by the quality of codebooks. To address this issue, on-line methods, or incremental methods, which belong to the second category, have led to an increasing interest during recent years. Angeli *et al*. [8] extended the BoW model to incremental conditions using local shape and color information and utilized Bayesian filtering to determine loop-closing pairs. Relying on agglomerative clustering, a scalable and automatic BoW technique is proposed in [9], which results in a satisfying performance even at small vocabulary sizes. Later, IBuILD [10] aims to build an incremental binary vocabulary based on feature tracking between consecutive images. Similarly, Zhang *et al*. [11] learned codewords from a pair of matched features from two adjacent frames, ensuring temporally-derived perspective invariance to camera motion. Except for applying on-line technique, the concept of dynamic islands has been introduced to save computational time [12]. Recently, many researchers pay attention to tackle the LCD task through voting schemes directly on descriptors' space. Typically, Tsintotas *et al*. [13] first defined a place using a dynamic segmentation of incoming image stream, and then adopted a nearest neighbor voting scheme on the descriptors' space to select the most appropriate match. Subsequently, they designed an on-line feature named "Tracked Word" for consecutive images through a tracking technique coupled with a guided-feature-detection technique, followed by casting votes to the corresponding instances [14].

B. Image Feature Matching

RANSAC is very popular in the existing LCD methods for loop closure verification. It describes geometric constraints of correspondences via low-DoF parametric models, such as fundamental matrix, homography or affine transformation. To address a more general feature matching problem, several methods based on non-parametric interpolation by applying slow-and-smooth prior have been proposed during the past decade [15], [16], such as ICF [17] and VFC [18]. However, their applicability on real-time tasks would be limited due to their cubic complexities. Later, to improve the efficiency, several locality consistency assumption-based methods have been proposed. For example, Bian *et al*. [19] encapsulated motion smoothness as the statistical likelihood of a certain number of matches in a region. Ma *et al*. [20] maintained the local neighborhood structures of potential true matches and formulated the problem into a mathematical model, realizing feature matching in linearithmic time and linear space complexities. Jiang *et al*. [21] cast feature matching into a spatial clustering problem with outliers, which enables the method to achieve quasi-linear time complexity.

III. Methodology

A new pipeline for LCD is implemented as outlined in Fig. 1. The procedure is mainly divided into two stages: on-line candidates selection and robust geometric confirmation. For better understanding, we present the symbols with regard to our approach in Table I.

A. On-line Candidate Frames Selection

1) Feature Extraction via DELG: Conventional LCD systems typically rely on extracting handcrafted local features, followed by BoW-related strategies to generate signatures/embeddings for each image. Compared with local ones, global descriptors aim at transferring semantic information and excel at delivering high image retrieval performance with compact representations. However, its results always have a poor discrimination, thus requiring further geometric verification via accurate image matching. That means additional local feature extraction is indispensable, which may cause a decline in efficiency both in time and storage space.

To address this issue, we adopt the latest state-of-the-art technique DELG [22] to jointly extract global and local features of an image. Specifically, a deep feature map $\mathcal{M}_D$ with $C_D = 2,048$ channels is acquired from the *conv5* output of ResNet-101 (R101), followed by GeM pooling and a whitening fully-connected layer, and finally produces a *global feature* of size $C_F = 2,048$. Regarding *local features*, an attention module is adopted. Concretely, the shallower feature map $\mathcal{M}_S$ of the *conv4* output of R101 is transferred to a small convolutional network and an attention score map $\mathcal{M}_A$ would be produced, indicating the probability of each location to be a keypoint. Later, according to $\mathcal{M}_S$, a small convolutional autoencoder module is applied to produce descriptors with dimensionality $C_L = 128$ of each location. In this paper, we at most choose locations with top-υ_{max} attention scores as keypoints to address subsequent tasks. Under the GPU technique, $110.3ms$ is needed to conduct feature extraction for an image with resolution $1,226 \times 370$.

TABLE I: Parameter illustration.

Parameter	Description	Value
α, β	Place formulation threshold	30, 0.5
υ_{max}	Maximum number of features	500
υ_{min}	Minimum number of features	20
σ	Candidate place threshold	0.2
T_r	Time in non-redundance path (s)	10
K	Number of neighbors in BMRC	13
η_m	Parameters of the filter	[0.2, 0.5, 0.5]
λ	Inlier threshold	0.17
θ	Loop closure event threshold	\

2) Dynamic Place Partition: To improve recall rates of the LCD system, *i.e.*, identifying more valid candidate frames in the first stage, we dynamically segment image streams with similar content into places via BMRC (which will be introduced later) according to the sequential nature of spatial data in the robotics domain. Denote the image captured at timestamp t as I_t, and its global descriptor extracted from DELG model as $\mathbf{g}_t$. A place P_n including τ_n images would generate if it meets the condition:

$$|G\left(I_t, I_{t+\tau_n}\right)| < \alpha \quad \text{or} \quad \frac{|G\left(I_t, I_{t+\tau_n}\right)|}{|G\left(I_t, I_{t+1}\right)|} < \beta, \tag{1}$$

where $G\left(I_t, I_{t+\tau_n}\right)$ indicates the inlier set preserved after BMRC and $|\cdot|$ means the cardinality of a set. Note that to avoid failure in feature matching, images containing less than $\upsilon_{\min}$ local features are automatically segmented into the current place. Compared with segmenting fixed-length sequence as a place, our dynamic place partition (DPP) strategy can deal with turning points and sight changes during the travelling path. Subsequently, the semantic information of place P_n would be produced based on the sum of global features of images belonging to the same place:

$$\mathbf{p}_n = \tfrac{1}{\tau_n} \textstyle\sum_{i=t}^{t+\tau_n-1} \mathbf{g}_i. \tag{2}$$

When a query image I_t comes to the pipeline, we identify candidate places according to the similarity defined as:

$$s(I_t, P_n) = -\log 10 \left\|\mathbf{g}_t - \mathbf{p}_n\right\|_2, \tag{3}$$

where $\|\cdot\|_2$ indicates L_2 norm. Only when $s(I_t, P_n) > \sigma$ would P_n be regarded as a candidate place. Subsequently, similarity between I_t and any image in the candidate place would be calculated, and the most similar one would be identified as the candidate frame, participating in the geometric confirmation. To avoid images acquired close in time to compete among them, $\xi = f \cdot T_r$ neighboring images obtained before the query would not be considered whether they meet the condition of being a candidate, where f stands for the frame rate and T_r is the consuming time in non-redundance path. Meanwhile, the query image I_t would be added to the reference image database, followed by judging whether a new place should be created according to Eq. (1). That is to say, the first stage of our pipeline totally operates on-line and transforms from image-to-place to image-to-image matching.

B. Geometric Confirmation via BMRC

When candidate frames are identified, geometric confirmation would be implemented via feature matching. Combining the local features in DELG and the distance ratio method [2], we can acquire a set of $N < \upsilon_{max}$ putative matches $\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ with $\mathbf{x}_i$ and $\mathbf{y}_i$ being the coordinates of extracted features in the query I and the candidate frame $I^{'}$. Specifically, for each feature in I, we search the two nearest neighbors in descriptor space in $I^{'}$, and subsequently compute their distance ratio. Only when passed distance ratio test, the putative match can be kept. Then we would use the new proposed BMRC to filter out mismatches in the putative set.

1) Neighbor Manifold Representation: In general, the local neighborhood structure among feature points may not tend to change freely due to the physical constraints in a small region around a point, even when non-rigid deformation occurs in the scene [23]. Based on this observation, we consider representing the neighborhoods of feature points mathematically, followed by removing outliers according to their similarity. To this end, we introduce an efficient strategy similar to locally linear embedding (LLE) [24], which is proposed as a nonlinear dimensionality reduction method to preserve the local neighborhood structure in a low-dimensional manifold. For $(\mathbf{x}_i, \mathbf{y}_i)$, based on Euclidean pixel distance, we first search the K nearest neighbors for $\mathbf{x}_i$ in I and denote them as $\mathcal{N}_{\mathbf{x}_i}$. Meanwhile, we denote the corresponding features of $\mathcal{N}_{\mathbf{x}_i}$ in $I^{'}$ as $\mathcal{C}_{\mathbf{y}_i}$. Based on $\mathcal{N}_{\mathbf{x}_i}$, neighbor topological structures of $\mathbf{x}_i$ could be represented via an $1 \times K$ weight vector. Considering all points in I with $i \in [1, N]$, the topological structures could be indicated by an $N \times K$ weight matrix $\mathbf{W}^{\mathcal{N}_\mathbf{x}}$ under a constraint that the rows of it sum to one, *i.e.*, $\forall i, \sum_{j=1}^{K} \mathbf{W}_{ij}^{\mathcal{N}_\mathbf{x}} = 1$. Thus the overall reconstruction error could be measured by the cost function defined as:

$$\epsilon(\mathbf{W}^{\mathcal{N}_\mathbf{x}}) = \textstyle\sum_{i=1}^{N} \left\| \mathbf{x}_i - \sum_{j=1}^{K} \mathbf{W}_{ij}^{\mathcal{N}_\mathbf{x}} \mathbf{x}_j \right\|^2. \tag{4}$$

The optimal solution can be obtained by addressing a least squares problem. Similarly, $\mathbf{W}^{\mathcal{C}_\mathbf{y}}$ could be acquired based on $\mathcal{C}_{\mathbf{y}_i}$. Thus outliers could be identified through the distance of the topological structure, which is measured by:

$$Dist^{\mathbf{x}\to\mathbf{y}}(\mathbf{x}_i, \mathbf{y}_i) = \left\| \mathbf{W}_{i,:}^{\mathcal{N}_\mathbf{x}} - \mathbf{W}_{i,:}^{\mathcal{C}_\mathbf{y}} \right\|^2. \tag{5}$$

The above manifold representation can be defined in a bidirectional manner to further promote the accuracy:

$$Dist^{\mathbf{y}\to\mathbf{x}}(\mathbf{x}_i, \mathbf{y}_i) = \left\| \mathbf{W}_{i,:}^{\mathcal{N}_\mathbf{y}} - \mathbf{W}_{i,:}^{\mathcal{C}_\mathbf{x}} \right\|^2, \tag{6}$$

where $\mathcal{N}_{\mathbf{y}_i}$ refers to the K nearest neighbors for $\mathbf{y}_i$ in $I^{'}$ and $\mathcal{C}_{\mathbf{x}_i}$ records the corresponding features of $\mathcal{N}_{\mathbf{y}_i}$ in I. Based on $\mathcal{N}_{\mathbf{y}_i}$ and $\mathcal{C}_{\mathbf{x}_i}$, $\mathbf{W}^{\mathcal{N}_\mathbf{y}}$ and $\mathbf{W}^{\mathcal{C}_\mathbf{x}}$ could be defined accordingly.

2) Iterative Filtering Strategy: As Fig. 2 shows, ideally, $\mathcal{N}_\mathbf{x}$ and $\mathcal{C}_\mathbf{y}$ with respect to the two corresponding points in an inlier should have similar topological structures, while for an outlier they will be of great difference. Therefore, according to Eqs. (5) and (6), it is theoretically easy to distinguish outliers from the putative set $\mathcal{S}$. However, due to the existence of contaminated data, there would be several outliers in an inlier's neighborhood, causing the distances

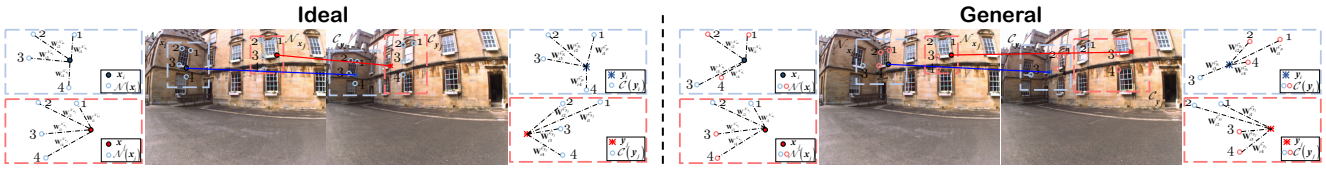

Fig. 2: Illustration of the topological structure in ideal and general situations, respectively. Red: outlier; blue: inlier.

defined in Eqs. (5) and (6) become large even for an inlier. To address this issue, we introduce an element-based iterative filtering strategy to acquire a relatively clean set of matches for neighbor manifold representation. Clearly, inliers would have more common elements in their neighborhoods, *i.e.*, $n_i = |\mathcal{N}_{\mathbf{x}_i} \cap \mathcal{N}_{\mathbf{y}_i}|$ tends to have a large value. Hence we define a confidence based on $\mathcal{N}_{\mathbf{x}_i}$ and $\mathcal{N}_{\mathbf{y}_i}$ as:

$$Ratio(i) = n_i/K. \tag{7}$$

Later, a set of thresholds $\{\eta_m\}_{m=1}^{M}$ is introduced to gradually filter unreliable matches which cannot meet the condition $Ratio > \eta_m$. The reliable set $\mathcal{U}_m$ composed of preserved matches in each iteration could be denoted as:

$$\mathcal{U}_m = \{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{S} | Ratio(i) > \eta_m, i \in [1, N]\}. \tag{8}$$

Note that $\mathcal{U}_0$ is defined by the original putative set $\mathcal{S}$, and in m-th iteration, $\mathcal{N}_{\mathbf{x}_i}$ and $\mathcal{N}_{\mathbf{y}_i}$ are obtained based on $\mathcal{U}_{m-1}$, followed by constructing $\mathcal{U}_m$ according to Eqs. (7) and (8). By using this iterative filtering, the final set $\mathcal{U}_M$ is typically clean enough for neighbor manifold representation.

3) Problem Formulation: Denoting $\mathcal{I}$ the unknown inlier set, its optimal solution is written as:

$$\mathcal{I}^* = \arg\min_{\mathcal{I}} C(\mathcal{I}; S, \mathcal{U}_M, \lambda), \tag{9}$$

with the cost function C defined as:

$$C(\mathcal{I}; S, \mathcal{U}_M, \lambda) = \textstyle\sum_{i \in \mathcal{I}} Dist(\mathbf{x}_i, \mathbf{y}_i) + \lambda(N - |\mathcal{I}|), \tag{10}$$

where $Dist(\mathbf{x}_i, \mathbf{y}_i)$ is expressed as:

$$Dist(\mathbf{x}_i, \mathbf{y}_i) = \frac{1}{2}\big(Dist^{\mathbf{x}\to\mathbf{y}}(\mathbf{x}_i, \mathbf{y}_i) + Dist^{\mathbf{y}\to\mathbf{x}}(\mathbf{x}_i, \mathbf{y}_i)\big). \tag{11}$$

All neighborhoods are constructed through $\mathcal{U}_M$. The first term in Eq. (10) aims to penalize any match which does not have a similar local topological structure, while the second one discourages the outliers, and parameter $\lambda > 0$ achieves a trade-off between the two terms.

4) A Closed-form Solution: To optimize the objective function in Eq. (10), we introduce an $N \times 1$ binary vector $\mathbf{p}$, where $p_i \in \{0, 1\}$ indicates the match correctness of the i-th correspondence $(\mathbf{x}_i, \mathbf{y}_i)$. Specifically, $p_i = 1$ represents an inlier, and an outlier otherwise. Therefore, the cost function in Eq. (10) can be rewritten as:

$$C(\mathbf{p}; S, \mathcal{U}_M, \lambda) = \textstyle\sum_{i=1}^{N} p_i Dist(\mathbf{x}_i, \mathbf{y}_i) + \lambda(N - \sum_{i=1}^{N} p_i). \tag{12}$$

By merging the items related to p_i, we obtain:

$$C(\mathbf{p}; S, \mathcal{U}_M, \lambda) = \textstyle\sum_{i=1}^{N} p_i\big(Dist(\mathbf{x}_i, \mathbf{y}_i) - \lambda\big) + \lambda N. \tag{13}$$

Clearly, $Dist(\mathbf{x}_i, \mathbf{y}_i)$ can be identified in advance as long as the putative set is provided due to the fact that the local topological structure is fixed. Hence the only unknown variable is p_i. Additionally, $Dist(\mathbf{x}_i, \mathbf{y}_i) > \lambda$ would lead to a positive term, resulting in an increase of the cost, and vice versa. Therefore, to minimize the cost, the optimal solution of $\mathbf{p}$ can be determined by:

$$p_i = \begin{cases} 0, & Dist(\mathbf{x}_i, \mathbf{y}_i) > \lambda, \\ 1, & Dist(\mathbf{x}_i, \mathbf{y}_i) \le \lambda, \end{cases} \quad i = 1, \ldots, N. \tag{14}$$

Finally, the optimal solution can be denoted as:

$$\mathcal{I}^* = \{i | p_i = 1, i = 1, 2, \ldots, N\}. \tag{15}$$

Subsequently, candidate images satisfying $|\mathcal{I}^*| > \theta$ with the query image are identified as loop closure events. As our method exploits the consensus of the bidirectional neighbor manifold representation, we name it BMRC.

5) Computational Complexity: According to K-D tree [25], the time complexity of searching K nearest neighbors is about $O((K+N)\log N)$. Therefore, the time complexity of constructing $\{\mathcal{N}_{\mathbf{x}_i}^K\}_{i=1}^N$, $\{\mathcal{N}_{\mathbf{y}_i}^K\}_{i=1}^N$ along with the filtering procedure in Eq. (8) can be denoted as $O((MK+MN+K+N)\log N)$. Additionally, each row of $\mathbf{W}^{\mathcal{N}_\mathbf{x}}$ can be solved separately with $O(K^3)$ time complexity, thus obtaining $Dist(\mathbf{x}_i, \mathbf{y}_i)$ in Eq. (11) for all $i \in [1, N]$ requires $O(NK^3)$ time complexity. Therefore, the total time complexity of our BMRC can be recorded as $O((MK+MN+K+N)\log N + NK^3)$. The space complexity of BMRC is $O(MKN)$ due to the memory requirement for storing $\mathcal{N}_{\mathbf{x}_i}$, $\mathcal{C}_{\mathbf{y}_i}$, $\mathcal{N}_{\mathbf{y}_i}$ and $\mathcal{C}_{\mathbf{x}_i}$ for all $i \in [1, N]$, as well as the neighborhoods constructed during the procedure of iterative filtering. Generally, M and K are far smaller than N, then the time and space complexities of our method can be simplified as $O(N \log N)$ and $O(N)$ respectively. That is to say, our BMRC has linearithmic time complexity and linear space complexity in regard to the scale of the given putative set, which is efficient for solving large-scale or real-time vision-based tasks.

IV. Experiments

A. Experimental Setup

All parameter settings have been summarized in Table I, and we fix them throughout the experiments. Concretely, we run our system on an Intel(R) Core(TM) i9-9920X CPU @ 3.50GHZ machine with three TITAN RTX GPUs. We select four publicly-available LCD datasets to evaluate the performance of our system, and detailed information is shown in Table II. Specifically, we adopt images captured by left cameras in CC and NC, and meanwhile the frames of NC are resampled to 1Hz, due to the robot's low velocity and high camera frequency. Ground truth (GT) is provided in the form

TABLE II: Properties of LCD datasets.

Dataset	Characteristic	# Images	Frame rate	Image size
KITTI 00 (K00) [26]	Long-term trajectory (Outdoor)	4541	10	1241×376
CityCentre (CC) [27]	Variation in speed (Outdoor)	2474	7<	640×480
NewCollege (NC) [28]	Considerable loop closure examples (Outdoor)	52480	20	512×384
Lip6Indoor (L6I) [8]	Strong perceptual aliasing examples (Indoor)	388	1	240×192

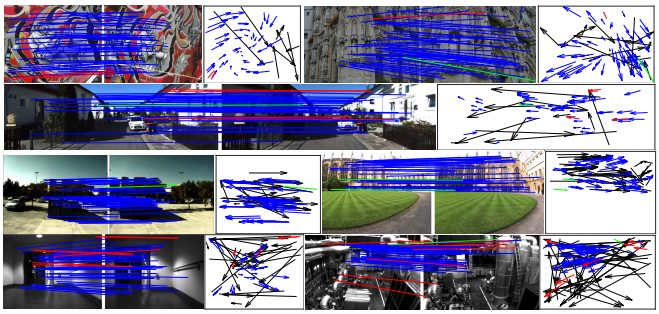

Fig. 3: Feature matching results of BMRC on seven image pairs. For visibility, at most 50 randomly selected matches are presented. In the motion field, the head and tail of each arrow correspond to the spatial positions of two pairwise feature points. From left to right, top to bottom, the inlier percentages are 83.25%, 74.89%, 74.41%, 69.18%, 43.16%, 52.68% and 40.85%. (Blue: true positive, red: false positive, green: false negative, black: true negative.)

of a binary matrix, where $GT_{ij} = 1$ means a loop closure event occurs between I_i and I_j and $GT_{ij} = 0$ indicates otherwise. GT in K00, CC and NC are labeled manually in [7] based on dataset's odometry information, while that of L6I is established through the GT data in [8]. As we apply an image-to-place strategy, new GT, denoted as $\widetilde{GT}$, would be created according to the initial one. Concretely, we consider $\widetilde{GT}_{ij} = 1$ if there is at least one loop-closing camera pose existing between place j and image i.

B. Performance Evaluation on Feature Matching

We first validate BMRC on addressing the general feature matching problem. To this end, two matching datasets are adopted. The first one is composed of 30 true loop-closing pairs selected from the abovementioned LCD datasets (*LCD*), mainly including scenes with complex transformations and repetitive structures, and the second one constitutes 51 image pairs extracted from *VGG* [18], dominating by scenes with large scale changes, local non-rigid deformations and low-overlapped areas. The average inlier ratio of the two dataset is 49.67% (*LCD*) and 88.10% (*VGG*) respectively, thus there are more outliers in *LCD*. Fig. 3 presents intuitive results of BMRC on seven typical image pairs, where the precision, recall and F-scores are (96.56%, 99.18%, 97.86%), (98.87%, 98.79%, 98.83%), (91.54%, 97.35%, 94.36%), (99.49%, 98.01%, 98.75%), (82.20%, 96.04%, 88.58%), (84.92%, 99.07%, 91.45%) and (86.73%, 97.70%, 91.89%), respectively. We can see that majority of correct matches are identified successfully. Even in indoor scenes where texture is deficient or strong perceptual aliasing examples exist, BMRC can still achieve good performance.

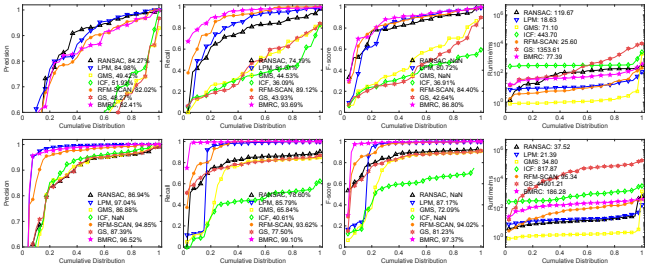

Fig. 4: Quantitative feature matching results on *LCD* (top) and *VGG* (bottom). From left to right: precision, recall, F-score and runtime (ms) with respect to the cumulative distribution. The coordinate (x, y) on the curves means that there are $100 * x$ percents of image pairs which have precisions, recalls or runtime no more than y.

To provide a comprehensive quantitative evaluation of BMRC, six state-of-the-art feature matching methods including RANSAC [5], LPM [20], GMS [19], ICF [17], RFM-SCAN [21] and GS [29] are adopted for comparison. F-score is defined as the ratio of $2 \cdot Precision \cdot Recall$ and $Precision + Recall$, thus it is a comprehensive evaluation. As Fig. 4 shows, BMRC outperforms the state-of-the-art approaches in the two challenging datasets due to its high F-scores in the third column. By contrast, because RANSAC searches the optimal parametric model relying on only a part of the whole true correspondences that obey some specific geometric constraints, its performance becomes poor in *LCD* due to the existence of dominated outliers and complex patterns. As for GS, GMS and ICF, they do not show satisfying results in these challenging situations, because GS cannot automatically estimate the factor for affinity matrix and it is not affine-invariant, GMS requires that the putative correspondences should be in large scale, while ICF cannot accurately recover the corresponding function when there is large depth discontinuity or motion inconsistency in the scene. RFM-SCAN and LPM obtain better results compared with other comparative methods, but they tend to preserve false correspondences located around the repetitive patterns, which occurs frequently in a typical LCD scene.

The last column of Fig. 4 illustrates the runtime statistics of the seven methods. Clearly, GS spends a lot of time seeking a maximum inlier set with subgraph isomorphism theory, limiting its applicability in the LCD task. RANSAC is fast in the *VGG* dataset, but turns slow in the LCD dataset. By contrast, LPM, GMS, RFM-SCAN and our BMRC are all efficient in these challenging datasets, hence considering the real-time requirement, we would embed these four methods, as well as the baseline RANSAC, into the proposed LCD pipeline to further validate the advantage of BMRC in not only general feature matching, but also the LCD task.

C. Results on Loop Closure Detection

We adopt the pipeline presented in Fig. 1 to conduct the LCD experiments and generate PR-curves according to different θ, and results are shown in Fig. 5. Precision (P) is defined as the number of true positive LCDs over the total systems identifications. Recall (R) is the ratio between

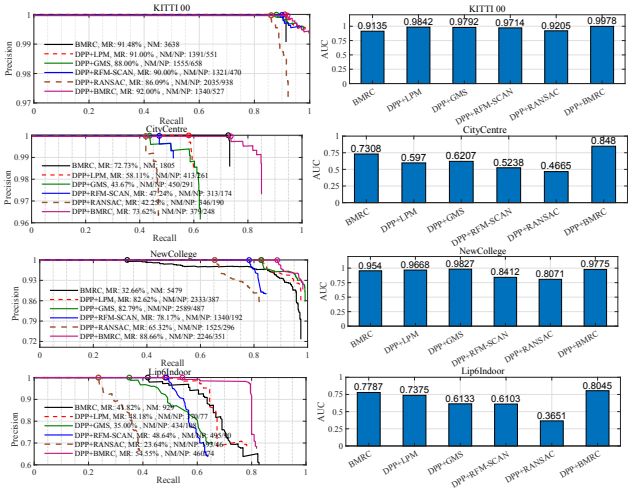

Fig. 5: Precision-recall (PR) curves and area under the curve (AUC) of four LCD datasets. (DPP: dynamic place partition. MR: maximum recall rate at 100% precision. NM: number of matching during the stage of geometric confirmation. NP: number of places.)

the true positive detections and the actual total of loop closure events defined by the GT. Note that to illustrate the effectiveness of DPP, an image-to-image procedure, denoted as BMRC in Fig. 5, is additionally conducted for comparison. Concretely, when a query image I_q occurs, its global feature $\mathbf{g}_q$ would be acquired via DELG, followed by calculating the similarity score $s(I_q, I_r)$ between it and any individual I_r in the reference image database according to Eq. (3). Then at most 10 images satisfying $s(I_q, I_r) > \sigma$ constitute the candidate list, entering the stage of geometric confirmation via BMRC. The tolerance used for the image-to-image strategy is set to 10 neighboring locations [14].

Performance analysis. According to Fig. 5 we can conclude that the proposed LCD system embedded with BMRC performs the best on the four datasets. Specifically, K00 travels a long-term path at a steady rate and includes adequate visual information, thus different feature matching methods have similar results. In contrast, CC has a variable velocity and loop-closing pairs always have large scale and depth changes, which requires the robustness of the feature matching method. Thus pipelines embedded with BMRC substantially improve the performance. Note that in this dataset, due to low correlation between consecutive images, adopting DPP or not has a similar result. As for NC, the robot moves slowly and sometimes stands still, resulting in high similarity among consecutive images. Thus the simple image-to-image strategy would unavoidably miss valid candidates, causing low recall rates on the overall performance. L6I is an indoor dataset equipped with deficient texture and strong perceptual aliasing examples (*i.e.*, several distinct places look similar). Thus it is challenging for appearance-based methods to obtain good results. But our LCD pipeline embedded with BMRC can still achieve the best compared with others.

Time analysis. When a query image comes to the pipeline, our operation is mainly composed of feature extraction via

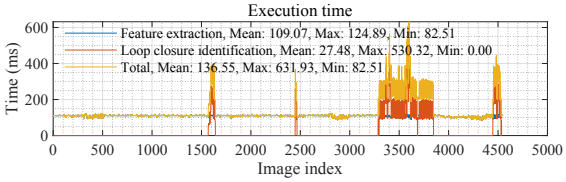

Fig. 6: Execution time for each frame in KITTI 00.

TABLE III: Comparative results of MR. Bold indicates the best.

Approaches \ Datasets	KITTI 00	CityCentre	NewCollege	Lip6Indoor
SeqSLAM [30] (2012)	83.54	42.80	55.12	20.91
Bampis *et al.* [31] (2016)	81.54	68.49	77.55	45.69
PREVIeW [32] (2018)	97.30	71.14	**92.74**	52.22
Tsintotas *et al.* [14] (2019)	**97.7**	\	83.00	\
Proposed	92.00	**73.62**	88.66	**54.55**

DELG, selecting candidate frames according to similarity scores and geometric checking via BMRC. Note that DDP is excluded due to its off-line property. As K00 is the longest dataset included, we evaluate our execution time on it and show the results in Fig. 6. We can see the total time of handling a query image is 136.55 ms in average, and 631.93 ms/frame is the maximum execution time when several loop-closing relationships exist. Thus the proposed pipeline can satisfy the real-time requirement in the LCD task.

Comparative results. Table III shows the ultimate results of the proposed algorithm and comparative algorithms as reported in the corresponding articles, including SeqSLAM [30], Bampis *et al.* [31], PREVIeW [32] and Tsintotas *et al.* [14]. Clearly, our approach has overall the best performance on two of the four datasets. As the result of our system is mainly dominated by feature matching, the superiorities of it would be shown on datasets with more challenging scenes, *i.e.*, CC and L6I. However, K00 and NC are composed of relatively simple scenes, where false negatives mainly appear in situations recorded at the same locations with different directions. Thus our framework performs unfavorable against others, but still achieves a high level.

V. Conclusion

In this paper, we propose a novel two-stage pipeline for the LCD task, *i.e.*, firstly selecting candidate frames and then identifying loop-closing pairs via geometric confirmation. To this end, an efficient feature matching approach termed as BMRC is proposed. It can establish reliable correspondences between an image pair in real time. Based on BMRC, a dynamic place partition strategy, which is an incremental operation when the robot travels the path, is further proposed to segment images captured close in time and content into a place, aiming at mining more valid candidate frames in the first stage and improving the recall rates of the system. Extensive experiments prove the validation of BMRC in general feature matching and reveal that the proposed LCD pipeline can obtain good results in real time compared with other state-of-the-art approaches.

REFERENCES

[1] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proceedings of the International Conference on Computer Vision*, 2003, pp. 1470–1477.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[3] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Proceedings of the European Conference on Computer Vision*, 2006, pp. 404–417.

[4] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *Proceedings of the International Conference on Computer Vision*, 2011, pp. 2564–2571.

[5] 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.

[6] M. Cummins and P. Newman, "Appearance-only slam at large scale with fab-map 2.0," *The International Journal of Robotics Research*, vol. 30, no. 9, pp. 1100–1123, 2011.

[7] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, 2012.

[8] A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer, "Fast and incremental method for loop-closure detection using bags of visual words," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1027–1037, 2008.

[9] T. Nicosevici and R. Garcia, "Automatic visual bag-of-words for online robot navigation and mapping," *IEEE Transactions on Robotics*, vol. 28, no. 4, pp. 886–898, 2012.

[10] S. Khan and D. Wollherr, "Ibuild: Incremental bag of binary words for appearance based loop closure detection," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2015, pp. 5441–5447.

[11] G. Zhang, M. J. Lilly, and P. A. Vela, "Learning binary features online from motion dynamics for incremental loop-closure detection and place recognition," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2016, pp. 765–772.

[12] E. Garcia-Fidalgo and A. Ortiz, "ibow-lcd: An appearance-based loop-closure detection approach using incremental bags of binary words," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3051–3057, 2018.

[13] 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.

[14] ——, "Probabilistic appearance-based place recognition through bag of tracked words," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1737–1744, 2019.

[15] 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.

[16] 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.

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

[18] J. Ma, J. Zhao, J. Tian, A. L. Yuille, and Z. Tu, "Robust point matching via vector field consensus," *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1706–1721, 2014.

[19] 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.

[20] 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.

[21] 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, 2020.

[22] B. Cao, A. Araujo, and J. Sim, "Unifying deep local and global features for image search," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 726–743.

[23] Y. Zheng and D. Doermann, "Robust point matching for nonrigid shapes by preserving local neighborhood structures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 643–649, 2006.

[24] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[25] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[26] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.

[27] M. Cummins and P. Newman, "Fab-map: Probabilistic localization and mapping in the space of appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

[28] M. Smith, I. Baldwin, W. Churchill, R. Paul, and P. Newman, "The new college vision and laser data set," *The International Journal of Robotics Research*, vol. 28, no. 5, pp. 595–599, 2009.

[29] H. Liu and S. Yan, "Common visual pattern discovery via spatially coherent correspondences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1609–1616.

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

[31] L. Bampis, A. Amanatiadis, and A. Gasteratos, "Encoding the description of image sequences: A two-layered pipeline for loop closure detection," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2016, pp. 4530–4536.

[32] ——, "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.