2019 International Conference on Robotics and Automation (ICRA)
Palais des congres de Montreal, Montreal, Canada, May 20-24, 2019

MRS-VPR: a multi-resolution sampling based global visual place recognition method

Peng Yin[1], Rangaprasad Arun Srivatsan[3], Yin Chen[4], Xueqian Li[3], Hongda Zhang[1], Lingyun Xu[1], Lu Li[3], Zhenzhong Jia[3], Jianmin Ji[2,*], Yuqing He[1,*]

***Abstract*—Place recognition and loop closure detection are challenging for long-term visual navigation tasks. SeqSLAM is considered to be one of the most successful approaches to achieve long-term localization under varying environmental conditions and changing viewpoints. SeqSLAM uses a brute-force sequential matching method, which is computationally intensive. In this work, we introduce a multi-resolution sampling-based global visual place recognition method (MRS-VPR), which can significantly improve the matching efficiency and accuracy in sequential matching. The novelty of this method lies in the coarse-to-fine searching pipeline and a particle filter-based global sampling scheme, that can balance the matching efficiency and accuracy in the long-term navigation task. Moreover, our model works much better than SeqSLAM when the testing sequence is over a much smaller time scale than the reference sequence. Our experiments demonstrate that MRS-VPR is efficient in locating short temporary trajectories within long-term reference ones without compromising on the accuracy compared to SeqSLAM.**

I. INTRODUCTION

In mobile robotic systems, simultaneous localization and mapping (SLAM) is a process of constructing and updating the map of an unknown environment while performing localization [1]. Visual place recognition (VPR) plays a vital role in finding reliable loop closures, helping SLAM to optimize the global localization and mapping. To improve the robustness against varying conditions, Milford *et al.* proposed a sequence matching method, SeqSLAM [2]. Given a set of M reference frames and a set of N testing frames, SeqSLAM can detect the potential matches based on feature similarities between the frames, with a computation complexity of $O(MN)$ using a brute-force searching method. As a result, sequential matching-based VPR is impractical in real robot navigation tasks because of two main challenges: (1) the exhaustive sequence searching method is computationally expensive with the stored frame sequence growing boundlessly, and (2) down-sampled frame sequence introduces uncertainty in matching process, when the length of testing frame sequence is too small.

To overcome these challenges, we apply a multi-resolution sampling (MRS) based method to improve the sequential matching efficiency. In lower resolution level, the sequence matching of each particle can be evaluated quickly, resulting in fast convergence of the distribution of particles. This property helps the particles obtain good initial estimation at the beginning. As see in Fig 1, the computation complexity for each particle at the lower resolution level is smaller than at the higher resolution level. While the number of particles decreases at the denser resolution level, the overall efficiency of the dense frame sequence match is not compromised. Therefore, we balance the matching efficiency and accuracy on higher resolution level.

The main contributions of this paper are listed as follows,

- Based on the sequential matching-based place recognition method, we propose a multi-resolution sampling scheme to balance the matching accuracy and searching efficiency. Our method is combined with a coarse-to-fine searching approach and a particle filtering scheme. This method is faster and more accurate than the original SeqSLAM method. It has wide potential for real world long-term robotic navigation tasks.
- We present a theoretical basis of our MRS-based method in the sequence matching. We also compare the improvement in performance of the MRS-VPR to the original SeqSLAM method. In the experiment part, we investigate the matching efficiency and accuracy of our proposed method, and discuss the key parameters within the MRS-VPR framework.

The rest of this paper is organized as follows. In Section II, we briefly describe the recent developments in VPR methods. In Section III, we introduce our method. Section IV demonstrates the performance of our method, details the experiment designs, and evaluates results, and conclusions are presented in Section V. The linked video[1] gives a better visualize results of the proposed method.

This paper was supported by the National Natural Science Foundation of China (No. 61573386, No. 91748130, U1608253) and Guangdong Province Science and Technology Plan projects (No. 2017B010110011).

P. Yin, L. Xu, H. Zhang and Y. He are with the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, University of Chinese Academy of Sciences, Beijing. (yinpeng, xulingyun, zhanghongda, heyuqing@sia.cn) J. Ji is with the School of Computer Science and Technology, University of Science and Technology of China, Hefei Anhui. (jianmin@ustc.edu.cn) R.A. Srivatsan, X. Li and L. Li are with the Biorobotics Lab, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. (arangapr, xueqianl, lilu12@andrew.cmu.edu) Y. Chen is with the School of Computer Science, University of Beijing University of Posts and Telecommunications, Beijing. (chenyin@bupt.edu.cn)

[1] https://youtu.be/2Pu_4kDFtMI

978-1-5386-6027-0/19/$31.00 ©2019 IEEE

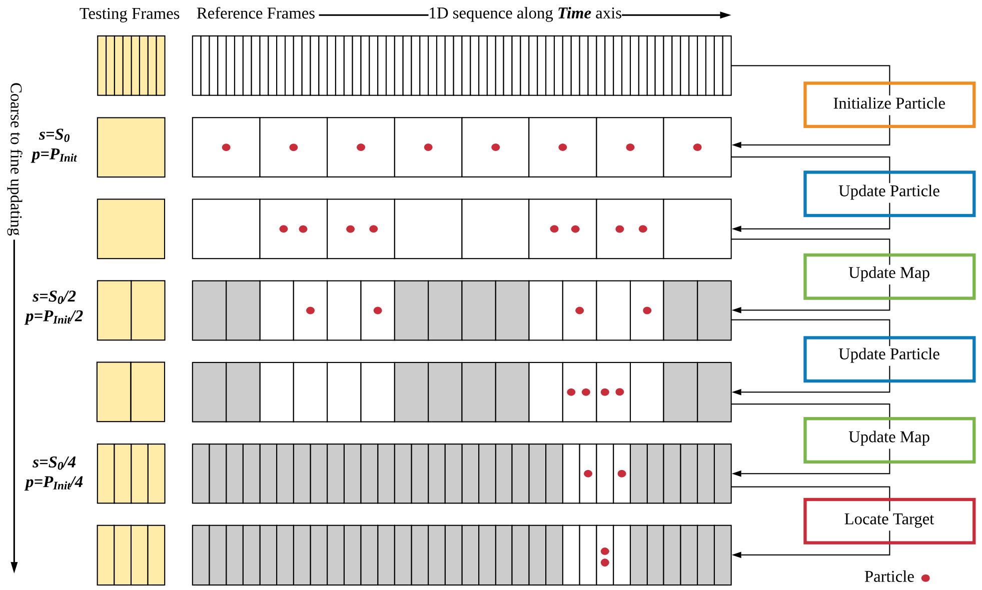

Fig. 1: The Multi-Resolution Sampling method. s is the frame sampling interval, and p is the number of particles. Row 1 shows the testing frames (yellow) and reference frames (white). In row 2, both sequential frames are down-sampled to the lowest resolution level, and initial particles are uniformly sampled in the reference sequence, each particle represents a potential matching trajectory. By iterative updating the sequence resolution level, particles try to find the best match under highest resolution level.

II. Related Work

Appearance changing under variant conditions leads to an unstable place recognition in SLAM frameworks. Traditional V-SLAM methods use BoW [3] (vector of local handcrafted features) as the image descriptor, or rely on prior 3D maps for online matching [4], or use hierarchical BoW [5].

FABMAP [6] uses Bayesian filtering to achieve long-term place recognition over 1000 km [6]. However, FABMAP cannot handle scenarios with variant changes in environmental conditions. Another family of appearance-based place recognition method, SeqSLAM [2, 7], uses a series of frame sequence to improve the robustness under variant environments. Lowry *el al.* [8] assumed the differences caused by geometry features are relatively smaller than the differences caused by season-to-season appearance changes. They developed a Principal Component Analysis (PCA) [9] approach to remove the season-related features, and extract remaining features as the season-invariant descriptions.

Sequential matching-based methods are not practical in real world applications due to their computational complexity. To improve the robustness in sequence matching, Naseer *et al.* [7] proposed an minimum cost flow-based data association, which could deal with non-matching image sequences that result from temporal occlusions or from visiting new places. Vysotska *et al.* [10], improved the work of Naseer *et al.* with GPS priors.

Even though there is a rich literature that focus on dealing with varying conditional problems [8, 7, 10], very few works focus on improving the efficiency and accuracy of searching in long-term place recognition tasks [11, 12]. Recently, with the development of deep learning for computer vision, Porav *et al.* [13] improved feature robustness against variant conditions by extracting reliable convolution layer features.

More recently, Sayem [11] proposed a Fast-SeqSLAM method, which improved the searching efficiency by utilizing an approximate nearest neighbor (ANN) as the initial estimate for potential matches. Since ANN in Fast-SeqSLAM still relies on single image feature similarities, the initial search efficiency may decrease when the original matching frame sequence is of a relatively long-time scale. Liu and Zhang [12] applied a particle filter to improve the matching efficiency, where each particle represented a potential subset of the frame sequence [14]. Rather than evaluating the whole frame sequence, they predicted the weights of multiple particles based on frame sequence similarities and the robot motion. However, both the methods described above require a good estimation of the initial matched location.

III. Proposed Method

Our work avoids the brute-force searching scheme in the traditional sequential matching methods by introducing a multi-resolution sampling approach, which combines a coarse-to-fine searching scheme and a particle filter method. Each particle represents a potential frame sequence in reference frames. As shown in Fig. 1 and Algorithm 1, our method can be divided into following steps:

1) Set the initial resolution level. down-sample trajectories according to the current resolution level and the initial particles; (line $5 \sim 6$)
2) Update the particle status based on their evaluation results; (line $9 \sim 18$)

Algorithm 1: MRS-VPR

```
   Input  : $M$ = Reference Frames, $N$ = Testing Frames
   Output: Predicted reference index
1  begin
2      $s = S_0$;
3      for $i \leftarrow 1$ to $l_{max}$ do
4          /* Step1 Map Updating */;
5          $m \longleftarrow$ M(s), $n \longleftarrow$ N(s), $s = s/2$;
6          Generate initial particles $P_{init}$ according to
             Eq 1;
7          /* Step2 Particle Updating */;
8          while $M_{cover} >= 50\%$ do
9              foreach $p_j$ in $P$ do
10                 $t_M, t_N$ = Extract($m, n, p_j.index$);
11                 $value, new\_index$ = Evaluate($t_M, t_N$);
12                 $p_j.weight = p_j.weight * value$;
13                 $p_j.index$ = new_index;
14             end
15             Particles weighting normalization according
                 to Eq 6;
16             effectiveness = Evaluate particles efficiency
                 according to Eq 7;
17             if effectiveness < threshold_effect then
18                 Particles Resampling;
19             end
20             Calculate map coverage according to Eq 8;
21         end
22         /* Step3 Map Updating */;
23         Update sequence frames, and particles' status;
24     end
25     Sort particles according to the particle weight;
26     return Best particle index
27 end
```

3) Update the current resolution level and particle indexes. If the map resolution reaches the maximum level, go to step 4; else, go to step 2; (line 20 ~ 21)
4) Sort particles by their weightings, and predict the best particle. (line 24 ~ 25)

A. Particle Initialization

In the particle initialization step, by setting the frame skipping interval as $s = S_0$, we can down-sample both the reference and the testing frame sequence into lowest map resolution level. At the lowest level, particles are sampled uniformly along the whole frame sequence. Also, the initial number of particles P_{init} satisfies the following equation,

$$P_{init} = \frac{M}{N}\tau, \tag{1}$$

where M and N are the frame sequence length of reference frames and testing frames respectively; τ is the hyper parameter, which determines the overlaps between one potential neighbor frame sequence. The overlap rate of the neighbor frame sequence can be calculated by

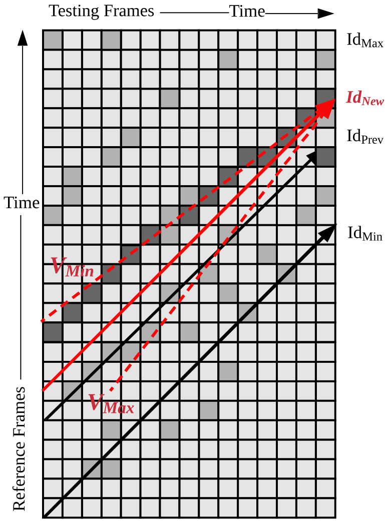

Fig. 2: The sequence matching mechanism for the testing and local reference frames.

$$Overlaps = \frac{\tau - 1}{\tau}. \tag{2}$$

High overlaps will improve searching robustness, while reducing the matching time. In the experiment part, we will analyze the influence of the τ configuration in the place recognition task.

Initial particles are uniformly sampled from the reference frame sequence, as shown in the second row of Fig. 1. Therefore, the entire particle sets have the following format,

$$P = \{p_t^{[1]}, p_t^{[2]}, p_t^{[3]}, ..., p_t^{[M]}\} \tag{3}$$
$$p_t^i = [index_t^i, weight_t^i],$$

where $index_t^i$ and $weight_t^i$ represent the end index and particle weight of the predicted reference frame sequence respectively.

B. Particle Evaluation

For each particle, we use the sequential matching to evaluate their measurements. The goal for sequential matching is to estimate the frame sequence similarity with a given testing and local reference trajectories, and then update the predicted particle status. As shown in Algorithm 2, we use the pre-processed image data as the feature description. The feature distance is defined as the sum of absolute feature differences, similar to the original SeqSLAM [2]. Although there are other feature extraction methods, this paper only focuses on the efficiency and accuracy in sequence matching. Thus, we only use the basic feature extraction method.

With the extracted features, the difference matrix can be calculated based on the feature distance between reference and testing trajectories. As shown in Fig. 2, x-axis represents

```
Algorithm 2: Evaluation
Input  : $T_M$ = Reference Frames, $T_N$ = Testing
         Frames
Output: Matching Value and refined index
begin
    $d_M$ = descriptor($T_M$), $d_N$ = descriptor($T_N$);
    $D$ = GetDifferenceMatrix($d_M, d_N$);
    Values = [ ];
    foreach $j$ from $(Id_{prev} - Id_{shift})$ to
      $(Id_{prev} + Id_{shift})$ do
        foreach $v$ in $[0.8, 0.9, ..., 1.2]$ do
            value = score($j, v$),   From Eq. 4 ;
        end
        Values.add(min values);
    end
    best_score = min Values;
    best_index = $Id_{min} + arg\min_{id}$ Values;
    return best_score, best_index
end
```

current testing frames t_N, and y-axis serves as potential reference frames t_M; the color of matrix cell represents the feature similarities, where darker colors imply relative place descriptions are more similar. Id_{prev} is the index of previous reference prediction, Id_{new} is the index of new predicted reference. Thus Id_{new} is searched within $(Id_{prev} - Id_{shift}, Id_{prev} + Id_{shift})$.

To assign weights from the current frame sequence match, we retrieve different trajectories for each potential reference index. At each end index, we apply different speed proportional constants $\frac{V_{test}}{V_{ref}} \in (0.8, 1.2)$ between testing and reference frames since the speed varies along the frame sequence. Thus the frame sequence similarity score can be evaluated by

$$score(j, v) = \sum_{t=1}^{L_{test}} D(t, j - v(L_{test} - t)), \tag{4}$$
$$Id_{new} = arg\min_{id} score,$$

where $D \in \mathbb{R}^{L_{test} \times [L_{test} + 2 \cdot Id_{shift}]}$ is the difference matrix, and L_{test} is the length of testing frames. Finally, the new index of the particle is updated according to the smallest frame sequence difference.

C. Particle Filtering

In the particle filtering step, for each particle (potential frame sequence), we use the sequence matching scheme in Sec. III-B to evaluate the frame sequence score based on Eq. 4. Then the new particle weighting $\hat{\omega}_k^i$ is obtained by,

$$\omega_k^i = \omega_{k-1}^i \times \frac{1}{1 + e^{-score_i}}, \tag{5}$$

After updating all particles, we normalize the weights of the particles by

$$\omega_k^i = \frac{\hat{\omega}_k^i}{\sum \hat{\omega}_k^i}, \tag{6}$$

and calculate the effectiveness of particles $\hat{N}_{eff}$ as

$$\hat{N}_{eff} = \frac{1}{\sum \left(\omega_k^i\right)^2}. \tag{7}$$

On the same map resolution level, particles are re-sampled around effective particles when the particle efficiency $\hat{N}_{eff}$ is smaller than a given threshold $thres_{particle}$. As shown in the third row of Fig. 1, finally, the particles will converge to potential matching targets.

D. Map Updating

Since computation complexity of sequence matching grows with the resolution level, we need to restrict particle sampling area to guarantee the matching efficiency. After particles reach the stabilized status on current map level, we compute the map coverage according to valid particles. We define a map coverage rate M_{cover}, which indicates the current particle convergence level

$$M_{cover} = \frac{M_{cur}}{M_{prev}}, \tag{8}$$

where M_{cur} and M_{prev} are current map coverage and previous map coverage separately. If the convergence rate M_{cover} is shrunk below a given threshold (in our experiments, this threshold is arbitrarily chosen to be 50%), we update both sequence frames into higher resolution level.

E. Speedup Analysis

In this subsection, we investigate the computation complexity of our MRS-VPR method by comparing it with the original SeqSLAM. For SeqSLAM, the computation complexity is $O(MN)$, where N and M are the number of frames of testing and reference frame sequence, respectively. For the MRS-VPR, we first generate P_{init} initial particles on the whole reference sequence space. Then for map resolution level i, the computation complexity is $O\left(\frac{P_{init}}{2^i} N_i\right)$, where N_i is the testing frame index on the i^{th} resolution level. $N_i = \frac{N}{2^{l_{max} - i}}$, while l_{max} is the maximum resolution level. The computation complexity ratio between the original *SeqSLAM* method and our proposed *MRS-VPR* is,

$$\mathcal{C}_{\frac{Seq}{MRS}} = \frac{MN}{\sum_{i=0}^{l_{max}} \frac{P_{init}}{2^i} \cdot N_i} = \frac{N}{\tau} \cdot \frac{2^{l_{max}}}{l_{max}},$$

where we substitute for P_{init} from Eq. 1. For example, if we set $l_{max} = 3$ and $\tau = 2.0$, the computation complexity ratio will be $1.33N$.

IV. Experiments

In this section, we investigate the performance of our method in the long-term global place recognition task.

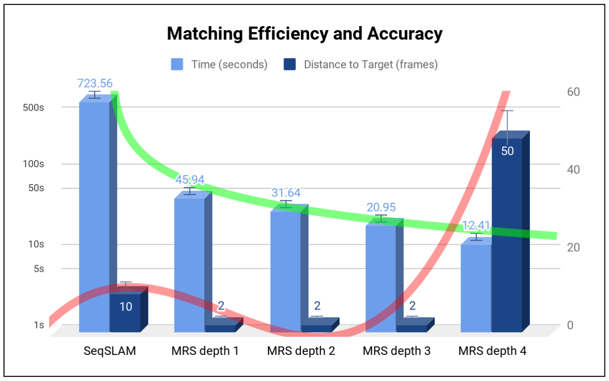

Fig. 3: The efficiency and accuracy of place recognition results on the same *Nordland* dataset [15]. When the MAS map depth $l_{max} \leq 3$, the matching efficiency grows with the map resolution depth without reducing the matching efficiency. But when $l_{max} = 4$, the matching error grows since testing frames are too sparse.

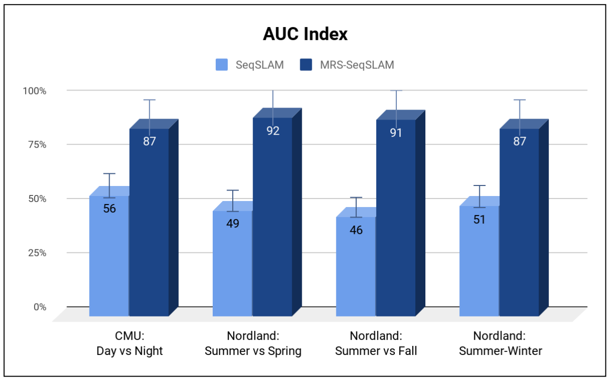

Fig. 4: The accuracy of place recognition results. We use AUC (Area Under precision-recall curve) to indicate the matching performance. For *CMU* dataset, we test the place recognition performance under day-night condition. For *Nordland* datasetwe test the matching results using summer versus spring, fall and winter conditions.

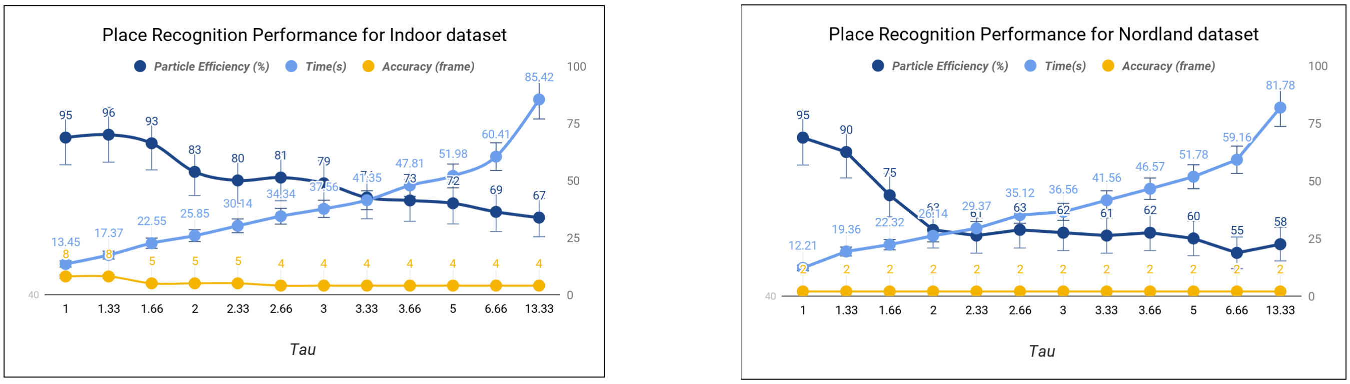

Fig. 5: The place recognition performance of the MRS-VPR method in *CMU* and *Nordland* dataset, with the same MRS depth level $l_{max} = 3$.

TABLE I: Datasets for place recognition task.

Dataset	*CMU*	*Nordland*
Reference (frames)	9000	9000
Testing (frames)	300	300
Conditions	Day/Night	Four seasons
Viewpoints	Not Fixed	Fixed
Dynamics Objects	Yes	No

A. Datasets

We use two datasets to test our method: the *Nordland* dataset [15], which is a 728 km long train ride in northern Norway, covering the same route in four different seasons; a *CMU* day-night dataset, which is a 1 km indoor sequence generated from a phone based camera with variant dynamics and viewpoint differences. We manually collected the second dataset, since we could not find large indoor VPR datasets containing day-night conditions. Table. I shows the details about two datasets. We observe that the main differences between *Nordland* and *CMU* dataset are viewpoints (fixed or not) and the existence of dynamics objects. While the *Nordland* dataset is collected by mounting a camera in the cab of a train, the *CMU* dataset is collected with a hand-held mobile phone camera. With the *CMU* dataset, it is hard to guarantee a stable viewpoint along the route. In addition, *CMU* dataset has lots of dynamic objects in the indoor environment. Thus, it is even harder to find potential matches in *CMU* dataset, compared to the *Nordland* dataset.

B. Accuracy & Efficiency Analysis

To inspect the accuracy and efficiency of different algorithms, we leverage short-term testing frame sequence to match the relative long-term reference frame sequence. For both *CMU* and *Nordland* datasets, the proportion between reference and testing frames number is 30, as shown in Table I.

One important parameter in our method is the map depth. Fig 4 shows the matching efficiency and accuracy of different methods in the *CMU* datasets [2].

We see that, when the map resolution depth $l_{max} = \{1, 2, 3\}$, the matching efficiency tends to improve with the

[2] Video link: https://youtu.be/dS028yXKNlw

growth of the map resolution depth, but the matching error is held at 2 frames. In contrast, the original sequential matching method took $723.56s$, and the final error between predicted match and the best alignment matches is 10 frames. However, when $l_{max} = 4$, the testing frames at the lowest level only have $\frac{300}{2^4} \sim 19$ frames. If the testing sequence is too small, the robustness of sequential matching against changing viewpoints and illuminations is lost; and particles at the lowest resolution level cannot provide good estimations at the beginning. We also repeated the experiment on the *Nordland* datasets. We empirically observe that $l_{max} = 3$ is a suitable map depth for both the datasets.

Another important parameter in our method is τ, which determines the initial number of particles. We investigate the matching performance under variant τ settings. As observed in Fig 5, with the increasing of τ, the particle effectiveness index $\hat{N}_{eff}$ decreases. This means that there will be more particles converging to the potential optimal index. But when τ is less than 1.0, the overlaps between two particles reduces to 0, and the particles are less likely to converge to optimal positions. In addition, the matching time also increases with τ. In order to balance both efficiency and accuracy, we set τ within $(1.5, 2.5)$, depending on the requirement of efficiency. In our experiment, the default τ value is 2.0.

Fig 3 shows the area under curve (AUC) index; higher the AUC index indicate more accurate matching results. Compared to traditional sequential matching method, our method is more stable under varying conditions; all the AUC indexes are above 80%. This indicates that the coarse-to-fine searching scheme can improve the initial estimation for the best sequence matching. In summary, our method has the potential to balance the efficiency and accuracy in the long-term place recognition under variant environmental conditions.

V. Conclusions

In this paper, we propose a multi-resolution particle filter-based sequence matching method. Our framework leverages coarse-to-fine searching methods to improve the robustness of place recognition, when the testing sequence is much smaller than the reference sequence (e.g. overlap ratio $\frac{M}{N} > 30$). The experiments on *Nordland* and *CMU* datasets show that our MRS-VPR framework outperforms the appearance based sequence matching method SeqSLAM in the long-term place recognition task. For the future work, we plan to combine the proposed place recognition method with topological maps to construct more robust reference maps for real robot long-term navigation task.

References

[1] A. J. Davison, “Real-time simultaneous localisation and mapping with a single camera,” in *Proceedings Ninth IEEE International Conference on Computer Vision*, Oct 2003, pp. 1403–1410 vol.2.

[2] M. J. Milford and G. F. Wyeth, “SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights,” in *IEEE International Conference on Robotics and Automation*, May 2012, pp. 1643–1649.

[3] H. Jgou, F. Perronnin, M. Douze, J. Snchez, P. Prez, and C. Schmid, “Aggregating local image descriptors into compact codes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1704–1716, Sept 2012.

[4] G. Pascoe, W. Maddern, A. D. Stewart, and P. Newman, “Farlap: Fast robust localisation using appearance priors,” in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 6366–6373.

[5] E. Garcia-Fidalgo and A. Ortiz, “Hierarchical place recognition for topological mapping,” *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1061–1074, 2017.

[6] M. Nowakowski, C. Joly, S. Dalibard, N. Garcia, and F. Moutarde, “Topological localization using Wi-Fi and vision merged into FABMAP framework,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sept 2017, pp. 3339–3344.

[7] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, “Robust visual robot localization across seasons using network flows,” in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[8] S. Lowry and M. Milford, “Change removal: Robust online learning for changing appearance and changing viewpoint,” *ICRA15 WS VPRiCE*, 2015.

[9] S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[10] O. Vysotska, T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, “Efficient and effective matching of image sequences under substantial appearance changes exploiting GPS priors,” in *ICRA*. IEEE, 2015, pp. 2774–2779.

[11] S. M. Siam and H. Zhang, “Fast-SeqSLAM: A fast appearance based place recognition algorithm,” in *IEEE International Conference on Robotics and Automation (ICRA)*, May 2017, pp. 5702–5708.

[12] Y. Liu and H. Zhang, “Towards improving the efficiency of sequence-based SLAM,” in *IEEE International Conference on Mechatronics and Automation*, Aug 2013, pp. 1261–1266.

[13] H. Porav, W. Maddern, and P. Newman, “Adversarial training for adverse conditions: Robust metric localisation using appearance transfer,” *arXiv preprint arXiv:1803.03341*, 2018.

[14] R. Van Der Merwe, A. Doucet, N. De Freitas, and E. A. Wan, “The unscented particle filter,” in *Advances in neural information processing systems*, 2001, pp. 584–590.

[15] N. Sünderhauf, P. Neubert, and P. Protzel, “Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons,” in *Proc. of Workshop on Long-Term Autonomy, ICRA*. Citeseer, 2013.