Vision-based Place Recognition Using ConvNet Features and Temporal Correlation Between Consecutive Frames

Chu-Tak Li[1], Wan-Chi Siu[1], *Life-FIEEE*, and Daniel P.K. Lun[1], *SrMIEEE*

Abstract— **The most challenging part of vision-based place recognition is the wide variety in appearance of places. However temporal information between consecutive frames can be used to infer the next locations of a vehicle and obtain information about its ego-motion. Effective use of temporal information is useful to narrow the search ranges of the next locations, hence an efficient place recognition system can be accomplished. This paper presents a robust vision-based place recognition method, using the recent discriminative ConvNet features and proposes a flexible tubing strategy which groups consecutive frames based on their similarities. With the tubing strategy, effective pair searching can be achieved. We also suggest to add additional variations in the appearance of places to further enhance the variety of the training data and fine-tune an off-the-shelf, CALC, network model to obtain better generalization about its extracted features. Experimental results show that our proposed temporal correlation based recognition strategy with the fine-tuned model achieves the best (0.572) F1 score improvement over the original CALC model. The proposed place recognition method is also faster than the linear full search method by a factor of 2.15.**

I. Introduction

Visual Place Recognition (Visual Scene Recognition or Loop Closure Detection) is a crucial component to Visual Simultaneous Localization And Mapping (Visual SLAM) [1,2]. In the recent decade, there are various vision-based place recognition algorithms which take image (or frame) obtained from standard single frontal camera as input. Note that consecutive images (video sequence) contain a large amount of information, in which reliable place recognition can be achieved. The main function of visual place recognition is to judge whether the current capturing scene (query frame) has been previously visited or not and thereby we can obtain our location information in real time. Precise location information benefits navigation, localization, and visual SLAM systems as successful self-positioning is useful for reducing drift errors produced during navigation and mapping.

The simplest method is to compare the query frame with all the frames stored in a database using distance measure techniques like Euclidean distance or Cosine distance. If the most similar frame pair has the distance which is smaller than a pre-defined threshold, one can tell that the current scene has been seen before and the corresponding frame from the database will be reported. This is a single image-based method using single nearest neighbor search technique and relies much on the discrimination power of the frame descriptors. A Bag-of-Words (BoW) method, FAB-MAP [3], had been one of the most famous appearance-based SLAM systems in the last decade. Key points of frames are first detected and clustered to form a visual word dictionary. The dictionary is used to judge whether two frames come from the same scene. However, Milford and Wyeth [4] claimed that key points cannot be effectively recalled under extreme changes in surroundings and lighting. They proposed the first image sequence-based method, SeqSLAM, which employs sequences of images to boost the recognition performance. Later on, some researchers incorporate the temporal correlation given by the consecutive input frames into their proposed methods [5-10]. In this paper, temporal correlation is defined as the information obtained from the comparison of previous frames. For example, the frame differences between neighboring frames should be small when there is no motion.

Nowadays, features from Convolutional Neural Networks (ConvNet features) have been proven to be more generalized than conventional hand-designed features [11]. The most challenging part in place recognition is that images usually suffer from extreme changes in lighting conditions, viewpoints, appearance, and surroundings. This means that a general or even predictive image descriptor is required to robustly describe a scene. Theoretically, if a large amount of training data with a wide variety in the abovementioned changes are available, ConvNet features can provide promising generalization about the extracted features compared with hand-crafted features. [12-17] applied ConvNet features to place recognition and evaluated the corresponding performance. Their results also show the superior discrimination power of ConvNet features. Sünderhauf and his team [15] found that the ConveNet features from the convolutional layer 3 (conv3) of AlexNet [11] give the best performance but the conv3 features are high dimensional features ($\mathbb{R}^{384*13*13}\in \mathbb{R}^{64896}$) which result in slow pair searching if linear full search strategy is used. In addition, difficulties in gathering a large amount of labelled training data and high computational cost are two well-known barriers to CNN-based approaches. Hence, [12] proposed a shallow convolutional autoencoder network (CALC) to quicken the feature extraction and a way to generate training data without manual labelling. They proved the discrimination power of their features on small subsets of several large datasets.

In this paper, we present an efficient visual place recognition method with the emphasis on the temporal correlation based on our previous work [10,23]. Our work is equivalent to real-world localization as long as a geotagged database is concerned. The main contributions in this paper are as follows.

[1] Chu-Tak Li, Wan-Chi Siu, and Daniel P. K. Lun are with Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong (e-mail: chu-tak.li@connect.polyu.hk, enwcsiu@polyu.edu.hk, enpklun@polyu.edu.hk).

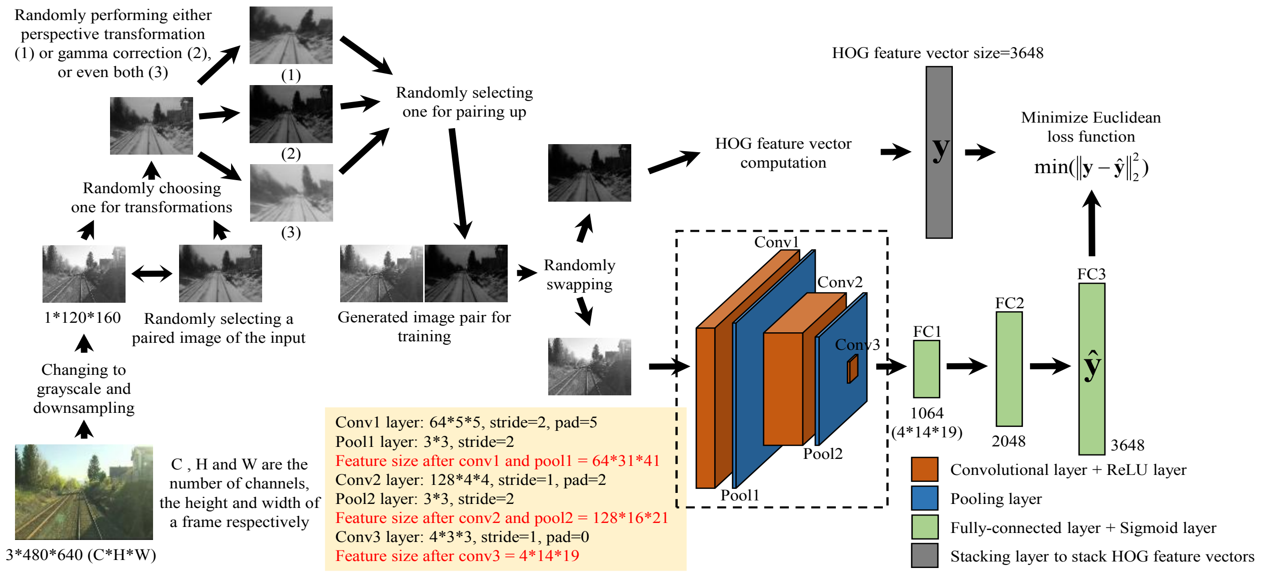

Randomly performing either perspective transformation (1) or gamma correction (2), or even both (3)
(1)
(2)
(3)
Randomly choosing one for transformations
Randomly selecting one for pairing up
1*120*160
Randomly selecting a paired image of the input
Generated image pair for training
Randomly swapping
Changing to grayscale and downsampling
3*480*640 (C*H*W)
C , H and W are the number of channels, the height and width of a frame respectively
Conv1 layer: 64*5*5, stride=2, pad=5
Pool1 layer: 3*3, stride=2
Feature size after conv1 and pool1 = 64*31*41
Conv2 layer: 128*4*4, stride=1, pad=2
Pool2 layer: 3*3, stride=2
Feature size after conv2 and pool2 = 128*16*21
Conv3 layer: 4*3*3, stride=1, pad=0
Feature size after conv3 = 4*14*19
HOG feature vector computation
HOG feature vector size=3648
y
Minimize Euclidean loss function
$\min(\|\mathbf{y}-\hat{\mathbf{y}}\|_2^2)$
Conv1
Conv2
Conv3
Pool1
Pool2
FC1
FC2
FC3
$\hat{\mathbf{y}}$
1064
(4*14*19)
2048
3648
Convolutional layer + ReLU layer
Pooling layer
Fully-connected layer + Sigmoid layer
Stacking layer to stack HOG feature vectors

Figure 1. The CALC model structure and its training procedure [12] with our modifications in the automatic training pair generation. (Information for readers. (1): Perspective transformed frame (2): Gamma corrected frame (3): Both perspective transformed and gamma corrected frame)

1) We propose an efficient place recognition strategy based on our previously proposed high confident initialization strategy [23] to further reduce the search range of each new coming frame.

2) We adopt the network model proposed in [12] and enhance the way of generating training data via enlarging the variety in the training data and fine-tuning the model on a subset of a place recognition dataset [18]. Hence, the discrimination power of the features from the fine-tuned model is generally improved;

Experimental work on three challenging datasets with various practical problems such as varying speeds, changes in appearance, viewpoints and illumination show the better generalization about the extracted features of the fine-tuned model and the advantages of using temporal correlation between consecutive query frames.

The paper is organized as follows. Section II provides a brief review on the architecture of the network to be used and how we modify the generation of the training data. Section III presents our proposed method and Section IV gives the experimental results. A conclusion is provided in Section V.

II. CONVOLUTIONAL AUTOENCODER NETWORK FOR VISUAL PLACE RECOGNITION

Merrill and Huang [12] have tried to remove the abovementioned barriers to CNN-based approaches by proposing a shallow convolutional autoencoder network for loop closure detection (CALC) and an automatic training data generation method. Because of the fast feature extraction stage and the exemption from manual labelling, they claimed that their method is lightweight and unsupervised. The denoising property of autoencoder network model has been studied in [19]. Autoencoder usually has several convolutional layers to achieve dimension reduction of the input. Extracted features are then used to reconstruct the initial input of the same size. With this process, the autoencoder learns to give an output which is a “denoised” version of the input. Hence, the autoencoder is able to identify the representative features to have satisfactory reconstruction.

For the training data generation [12], random perspective transformation to each input is adopted so as to pair up the input with its self-perspective transformed version for training without manual labelling. With this method, the extracted features from their proposed model are robust to changes in viewpoints. The training procedure of their proposed model with our modifications to the training data generation is shown in Fig.1. We adopt this training procedure[2] under the Caffe framework [20] to fine-tune the model for more desirable place recognition performance. Histograms of Oriented Gradients (HOG) [21,22] was chosen for their network training in which it benefits from (i) the robustness of changes in illumination as the local contrast normalization of HOG; (ii) the generalization about the changes in viewpoints as the employment of perspective transformed training data; (iii) the effective data compression of extracted features as the length of a HOG feature vector in their method is 3,648. Note that a simple $L2$ loss function is used, which aims to minimize the HOG feature vector, $\mathbf{y}$, and the ConvNet feature vector from the model, $\hat{\mathbf{y}}$. An image pair for training always represents the same scene so the two feature vectors should be close to each other. During the testing, only the convolutional layers (black dashed box in Fig.1) are used to extract the features. Hence, the feature size used for testing is 1,064 (4*14*19).

We suggest to add more variations in the frame pairs for training. We also employ random gamma correction and random selection of paired frames besides random perspective transformation. Equation (1) tells about the gamma correction and γ is randomly selected from 0.1 to 2.5.

$$F_c = \left(\frac{F}{255}\right)^{\gamma} \times 255 \tag{1}$$

[2] Source code and the pre-trained model are available online: https://github.com/rpng/calc. Please refer to it for the details.

where F is the input, F_c is the corrected frame. If $\gamma < 1$, a darker corrected frame is generated to simulate the changes in illumination. The opposite observation is made for $\gamma > 1$.

The process of creating a frame pair for training is as follows and is shown in Fig.1. First, all the frames input to the network is converted into grayscale and downsampled to 120*160 as the same in [12]. Second, a paired frame of the input is randomly selected from the dataset or the input itself is used directly as a pair (i.e. two same images as a pair). This is due to the fact that we have frames which capture the same place at different time slots. If we capture a place at 4 different time slots, we will have 4 possible frame pairs. We randomly pick one of them and generate the pair for training. Third, we randomly choose one frame from each pair and perform either the perspective transformation ((1) in Fig.1), or gamma correction ((2) in Fig.1), or even both ((3) in Fig.1). With this generation strategy, a wide variety of data can be achieved such as varying viewpoints, lighting and seasonal changes.

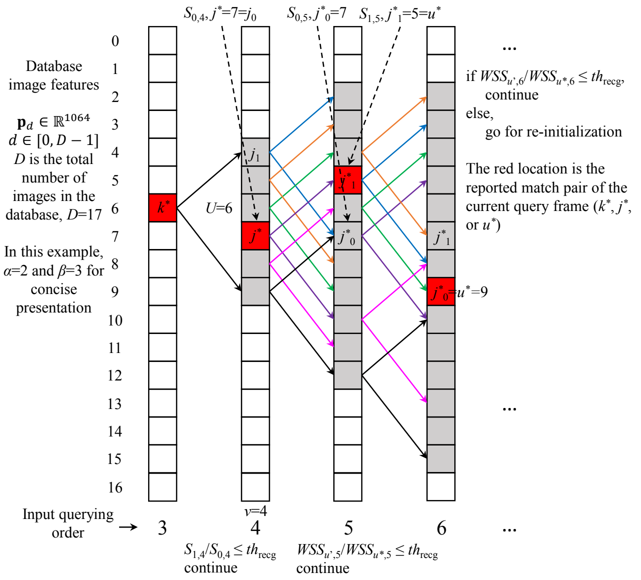
III. PROPOSED STRATEGY

To work on using temporal correlation for place recognition, we have to cope with speed changes in a sequence in practice. What is the optimal length of an image sequence section to be processed together or tube size? We define a tube of frames as the amount of temporal correlation which has to be taken into account so as to benefit from using it. This means that we have to find the starting point of a vehicle or the starting point of each tube of frames. A possible strategy is to assign a weight to each frame in comparison with the possible starting frame, and a weighted sum of some number (S) of consecutive frames can form similarity scores to define confidently the initial place [23]. Usually we can pick K nearby initial starting places. This means that there are K hypotheses about the initial location of the vehicle and each hypothesis has its search space. The hypothesis is represented by a weighted sum of S consecutive frames. Note that higher the score means better hypothesis. If the best hypothesis is larger than a pre-defined threshold, it will then be accepted as the best match to start the tube location. In this paper, we assume that the initialization has been done. This means that the starting position of a vehicle is known. Interested readers may refer to [10,23] for details.

A. Efficient Place Recognition Strategy

Once the initialization is done, we can confidently localize a vehicle and the search space for the coming query frames can be reduced based on the fact that the vehicle must travel along a route gradually without any sudden jump from one location to another location. Hence, linear full search is not necessary and the reduced search space helps to improve the place recognition accuracy. Fig.2 illustrates our effective place recognition strategy based on the initial match pair, k^*, which comes from our confident initialization.

For the 1st query frame after obtaining k^*, we perform the linear full search in a new search space to find the match pair using (2).



$S_{0,4}, j^*=7=j_0$
$S_{0,5}, j^*_0=7$
$S_{1,5}, j^*_1=5=u^*$
Database image features
$\mathbf{p}_d \in \mathbb{R}^{1064}$
$d \in [0, D-1]$
D is the total number of images in the database, D=17
In this example, α=2 and β=3 for concise presentation
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
k^*
U=6
j_1
j^*
j^*_1
j^*_0
j^*_1
$j^*_0=u^*=9$
...
if $WSS_{u',6}/WSS_{u^*,6} \le th_{\text{recg}}$, continue
else, go for re-initialization
The red location is the reported match pair of the current query frame (k^*, j^*, or u^*)
...
v=4
Input querying order →
3
4
5
6
...
$S_{1,4}/S_{0,4} \le th_{\text{recg}}$ continue
$WSS_{u',5}/WSS_{u^*,5} \le th_{\text{recg}}$ continue

Figure 2. A graphical illustration of our proposed efficient place recognition strategy

$$j^* = \underset{j\in[k^*-\alpha,k^*+\beta]}{\arg\max}\ C(\mathbf{q}_t,\mathbf{p}_j) = \underset{j\in[k^*-\alpha,k^*+\beta]}{\arg\max}\ (\mathbf{q}_t\cdot\mathbf{p}_j) \qquad (2)$$

where we use the Cosine similarity, $C(\mathbf{q},\mathbf{p})$, to perform similarity measure. $(\mathbf{q}\cdot\mathbf{p})$ is defined as the cosine of the angle difference between two normalized vectors $\mathbf{p}$ and $\mathbf{q}$. Here, $\mathbf{p}$ and $\mathbf{q}$ denote the normalized ConvNet feature vectors of the database frame and query frame given by the network model mentioned in Section II respectively. t is the current querying order. j^* is the single nearest neighbor to the current query frame among the new search space with the highest similarity score ($S_{0,t} = C(\mathbf{q}_t,\mathbf{p}_{j^*})$). If the ratio of $S_{0,t}$ to $S_{1,t}$ (the second highest similarity score found in the search space) is smaller than a threshold, th_{recg}, the confidence of j^* is at a satisfactory level and we will keep the tubing to make use of the temporal correlation between consecutive query frames. We believe that once we confidently localize the vehicle, fast recognition or tracking can be performed to improve the efficiency of the method until the confidence level of the output drops below a certain value. If the ratio is larger than th_{recg}, we will have to do a re-initialization. After finding j^*, all the similarity scores in the search space are denoted as $S_{u,t}$, $u \in [0,U\text{-}1]$ where U=α+β+1, and j_u denotes the location of uth neighbor in the search space which also defines the search space for the next query frame. From the next query frame, the weighted similarity score is calculated using (3,4).

$$j^*_u = \underset{j\in[j_u-\alpha,j_u+\beta]}{\arg\max}\ C(\mathbf{q}_t,\mathbf{p}_j) \qquad (3)$$

$$S_{u,t} = C(\mathbf{q}_t,\mathbf{p}_{j^*_u})\times(1-C(\mathbf{q}_t,\mathbf{q}_{t-1})) \qquad (4)$$

where α and β have been defined previously. Each j_u has its own search space and the corresponding single nearest neighbor is denoted as j^*_u. The weighted similarity score of the uth neighbor in time t is denoted as $S_{u,t}$. Note that only the U nearest neighbors are kept for decision making and j_u is updated continuously according to j^*_u. The match pair of the current query frame (u^*) is given by (5).

$$u^* = \underset{u \in [0, U-1]}{\arg\max}(S_{u,t} + \sum_{a=v}^{t-1} S_{u,a} \times C(\mathbf{q}_t, \mathbf{q}_a)) \qquad (5)$$

$$WSS_{u^*,t} = S_{u^*,t} + \sum_{a=v}^{t-1} S_{u^*,a} \times C(\mathbf{q}_t, \mathbf{q}_a) \qquad (6)$$

where the tube is reset after the initialization and v is the querying order in which j^* is computed. $S_{u,v}$ is calculated during the computation of j^* mentioned in above. Equation (6) shows the calculation of the weighted sum of scores of location u^* for the current query frame. Similarly, if the ratio of $WSS_{u*,t}$ to $WSS_{u',t}$ (the second highest weighted sum of scores found in the current search space) is smaller than a threshold, th_{recg}, the confidence of u^* still maintains at a satisfactory level and u^* is reported as the match pair for output. Otherwise, re-initialization will be activated.

IV. EXPERIMENTS

A. *Model Fine-tuning and Strategy Parameters*

We used parts of the Nordland dataset [18] to fine-tune the network model discussed in Section II. This dataset consists of 4 long rail sequences recorded at four seasons. We removed all the stop and tunnel parts for model fine-tuning. The sequences have been time-synchronized. This means that any frame in one sequence represents the same scene in the other sequences as long as the frame indexes are the same. The first 10,000 frames of each sequence were extracted and there were 30,820 frames for the fine-tuning (each sequence has 7,705 frames after the removal of stop and tunnel frames). For the parameters of our proposed place recognition strategy, α and β are set to 5 and 10 respectively. th_{recg} is pre-defined as 0.75.

B. *Datasets*

In our experiments, we compare our proposed method with several state-of-the-art approaches, namely SeqSLAM [4], ABLE-M [7,8], CALC [12], and AlexNet conv3 feature-based approach [11,15,24], on 3 challenging datasets. SeqSLAM employs downsampled grayscale normalized images as features for pair searching and assumes identical speed situation. ABLE-M is also a sequence-based approach which groups a different number of images as image sequences to form binary sequence codes for pair searching. We denote ABLE-M using different sequence lengths as ABLE-M l where l is the size of an image sequence, l=1,150 or 300. CALC is the model that we fine-tuned and used in our proposed method. The original CALC merely relies on the discrimination power of its features and adopts the simplest single nearest neighbor for pair searching. The AlexNet conv3 feature-based approach extracts the conv3 features from two AlexNet pre-trained on two datasets, ImageNet [11] and Places365 [24]. The former is for object classification and the latter is for scene classification. This approach also utilizes the single nearest neighbor search.

1) *UA Dataset*

Fig.3 shows a non-rail dataset and contains short sequences which were recorded on the campus of University of Alberta, Edmonton, Canada [25]. This dataset focuses on changes in illumination and has been time-synchronized. We extracted two sequences from this dataset, one daytime and one nighttime, both have 646 frames. The daytime sequence was used to construct the database.

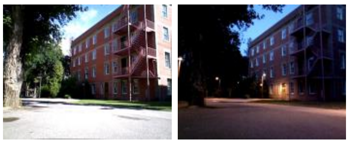


Figure 3. UA dataset (non-rail case with changing lighting conditions and viewpoints)

2) *Nordland Dataset*

We used the last 5,000 frames of each sequence for comparisons. We have ensured that there was no overlap with the training data and we did not remove the tunnel and stop frames for testing. The database was formed using the "Spring" sequence. Fig.4 shows the seasonal changes.

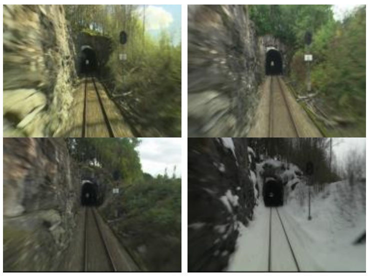


Figure 4. Nordland dataset (with extreme changes in appearance and seasonal changes)

3) *Light Rail Transit (LRT) Dataset*

LRT dataset was captured directly from a public transportation system in Hong Kong and is shown in Fig.5. There are 4 sequences of the same route, 3 daytimes and 1 nighttime. This dataset involves many practical difficulties such as varying speeds, extreme changes in lighting, and blurring. On average, each sequence has 2,566 frames. We used one of the daytime sequences to form the database.

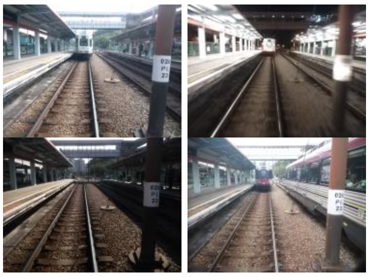


Figure 5. LRT dataset (with changes in illumination, varying speeds, moving objects, and blurring)

C. *Evaluation Metrics*

We have sorted all match pairs from different approaches based on their weighted similarity scores. Generally, a match pair with high score is more likely to be correct. For computing the precision, a pair is regarded as correct if its difference from the ground truth is less than 5 frames. We apply a set of 100 recall rates (0.01 to 1.00 with step 0.01) to the sorted scores and generate the corresponding set of precisions and recall rates. We used F1 score to evaluate various approaches for concise comparisons which is computed by (7),

TABLE I. OVERALL F1 SCORE COMPARISONS OF VARIOUS APPROACHES (**BOLD** FACE IS THE BEST; *ITALICS* IS THE SECOND BEST)

querying sequence	Ours*	Ours	CALC [12]	AlexNet, Places365[24]	AlexNet, ImageNet[11]	SeqSLAM [4]	Downsampled image [4]	ABLE-M 300 [7,8]	ABLE-M 150 [7,8]	ABLE-M 1 [7,8]
UA	**0.941**	0.889	0.369	0.419	0.582	0.654	0.155	*0.914*	0.899	0.476
Summer	0.760	0.637	0.478	0.492	0.381	0.832	0.094	**0.952**	*0.932*	0.450
Fall	0.827	0.761	0.586	0.557	0.534	0.855	0.122	**0.961**	*0.944*	0.579
Winter	0.590	0.478	0.308	0.494	0.152	**0.741**	0.052	*0.695*	0.540	0.054
LRT2	0.505	0.522	0.319	*0.596*	0.297	**0.658**	0.144	0.098	0.219	0.068
LRT3	0.742	0.736	0.721	**0.843**	*0.775*	0.420	0.113	0.213	0.338	0.340
LRT4	0.700	0.743	0.687	**0.821**	*0.751*	0.445	0.226	0.249	0.332	0.297
Average	**0.724**	*0.681*	0.495	0.603	0.496	0.658	0.129	0.583	0.601	0.323

$$F1 = 2 \times \frac{P \cdot R}{P + R} \tag{7}$$

where P is the precision defined as the ratio of the number of correct match pairs to the number of recalled match pairs. R is the recall rate [0.01,1.00] and is defined as the ratio of the number of recalled match pairs to all the query frames. Note that high F1 is attained if and only if both P and R are high. Hence, F1 can reflect the practicability of a method.

D. F1 Score Comparisons

Here, we summarize the performance of all the mentioned approaches in Section IV. *B*. and the F1 scores are listed in Table I. For each approach, each pair of precision and recall rate gives a F1 score and here we compare the maximum F1 score of each approach. Note that "Ours*" represents the method of the fine-tuned CALC with our modified training data generation method and proposed tubing strategy while "Ours" represents the fine-tuned network model without our recognition strategy. On UA dataset, "Ours*" obtains a 0.572 (=0.941-0.369) F1 improvement over the original CALC. On average, we have the highest F1 score (0.724) throughout the 3 challenging datasets. This means that our proposed method has better adaptability to general situations. For the discrimination power and generalization about the ConvNet features, "Ours" gives better performance (0.681) than the original CALC (0.495), AlexNet Places365 (0.603) and AlexNet ImageNet (0.496) on the three challenging datasets. This also reflects that the AlexNet pre-trained on the recognition-centric dataset performs better than that of the classification-centric dataset.

Apart from this, sequence-based methods like SeqSLAM and ABLE-M have good performance in long sequences for situations with identical speed, the Nordland dataset. ABLE-M 300 (0.869=(0.952+0.961+0.695)/3) clearly outperforms ABLE-M 150 (0.805=(0.932+0.944+0.540)/3) on the Nordland dataset and provides evidence that sequences with identical speed benefit from large tube size. However, the performance of these sequence-based methods drops drastically on the LRT dataset (0.56=(0.098+0.213+0.249)/3). This also gives evidence that large tube size could worsen the performance in situations with varying speeds. Hence, flexible tubing strategy is a must to get satisfactory performance in both situations with varying speeds and identical speed.

E. Time Cost Comparisons

Here, we show the time cost comparisons of all the methods, including our approach and conventional approaches, SeqSLAM and ABLE-M. Note that powerful computation resources and a large amount of labelled data are not always available in the reality. The CPU used is Intel Core^{TM} i7-6900k @ 3.2 GHz. The codes of SeqSLAM were written in MATLAB while ABLE-M and our tube strategy were written in C++. Hence, we divided the time reported by MATLAB by 10 for a possible fair comparison. We recorded the total processing time of each candidate video sequence and divided it by the total number of frames to obtain the average time cost per frame. This means that the time cost of our proposed initialization has been included in the evaluation. The time costs are shown in Table II in terms of milliseconds. "Ours" and the original CALC should have the same time cost as both employ single nearest neighbor search and have the same network structure. The two AlexNets pre-trained on two datasets should also have the same time cost.

TABLE II. OVERALL TIME COST COMPARISONS OF VARIOUS APPROACHES (MS, ONLY CPU IS USED)

Dataset	Ours*	Ours, CALC[12]	AlexNet [11,24]	SeqSLAM [4]	ABLE-M [7,8]
UA	49.6	67.0	288.4	2.0	0.6
Nordland	68.2	80.5	385.6	12.7	1.5
LRT	58.8	72.2	354.4	7.1	1.0
Average	58.9	73.2	342.8	7.3	1.0

The first observation is that conventional approaches are much faster than the ConvNet feature-based approaches as there is no heavy computation of convolutions. However, the discrimination power of ConvNet features superiorly outperforms the traditional hand-crafted features such as downsampled grayscale normalized image and binary descriptor as shown in Table I. The fact that SeqSLAM is sensitive to the length of sequences is also reflected here. On the Nordland dataset (5,000 frames), SeqSLAM requires 12.7 ms per frame while SeqSLAM only requires 2.0 ms per frame on the UA dataset (646 frames). For this reason, ABLE-M adopts image sequences to form binary sequence codes for efficient pair matching via the use of Hamming distance. Hence, ABLE-M is the fastest method among the comparing methods which only requires 1.0 ms per frame on average.

Among the ConvNet feature-based methods, "Ours*" is the fastest method because of the use of temporal correlation between consecutive query frames and the search space reduction. Note that the time cost of the feature extraction of the CALC model is 46.5 ms. This means that pair searching with our proposed tubing strategy costs only 12.4 ms (=58.9-46.5 ms) on average. Comparing with the linear full search (73.2-46.5=26.7 ms), our proposed tubing strategy is faster than it by a factor of 2.15. For the AlexNet conv3 feature-based approaches, on average 342.8 ms is needed to match a frame because of the slow feature extraction and slow matching of high dimensional feature vectors.

V. Conclusion

In this paper, we have proposed an efficient visual place recognition strategy via the temporal correlation with a sequence of consecutive query frames. We investigated training data generation methods to further increase the discrimination power of ConvNet features. Particularly, we have suggested to weight and sum the similarity scores based on the comparison of consecutive frames to obtain the final match pairs. A single nearest neighbor strategy of each possible match pair defines the search space of the coming query frames, hence efficient pair searching has been achieved. For future development, we will focus on real-world visual place recognition systems with online learning and mapping abilities. We believe that the training data generation method discussed in this paper provides a possible clue to design a better model for online place recognition.

Acknowledgment

This work was supported by the Hong Kong Polytechnic University(1-ZE1B) and consultancy work with Hong Kong Productivity Council on Video Analytics and ADAS. Mr. LI Chu-Tak would like to acknowledge the Research Studentship offered by the same university.

References

[1] Guillaume Bresson, Zayed Alsayed, Li Yu, and Sébastien Glaser, "Simultaneous Localization And Mapping: A Survey of Current Trends in Autonomous Driving," *IEEE Trans. on Intelligent Vehicles*, vol.2, no.3, pp. 194-220, Sept. 2017.

[2] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, Present, and Future of Simultaneous Localization And Mapping: Towards the Robust-Perception Age," *IEEE Trans. on Robotics*, vol.32, no.6, pp. 1309-1332, Dec. 2016.

[3] Mark Cummins and Paul Newman, "Highly Scalable Appearance-Only SLAM - FAB-MAP 2.0," *Proceedings*, *Conference on Robotics: Science and Systems* (*RSS*), Seattle, USA, pp. 39-46, Jun. 2009.

[4] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual Route-Based Navigation for Sunny Summer Days and Stormy Winter Nights," *Proceedings*, *IEEE International Conference on Robotics and Automation* (*ICRA*), Saint Paul, USA, pp. 1643-1649, May 2012.

[5] Jianliang Zhu, Yunfeng Ai, Bin Tian, Dongpu Cao, and Sebastian Scherer, "Visual Place Recognition in Long-term and Large-scale Environment based on CNN Feature," *IEEE Intelligent Vehicles Symposium* (*IV*), Changshu, Suzhou, China, pp. 1679-1685, Jun. 2018.

[6] Edward Pepperell, Peter I. Corke, and Michael J. Milford, "All-Environment Visual Place Recognition with SMART," *Proceedings*, *IEEE International Conference on Robotics and Automation* (*ICRA*), Hong Kong, China, pp. 1612-1618, May 2014.

[7] Roberto Arroyo, Pablo F. Alcantarilla, Luis M. Bergasa, and Eduardo Romera, "Towards Life-Long Visual Localization using an Efficient Matching of Binary Sequences from Images," *Proceedings*, *IEEE International Conference on Robotics and Automation* (*ICRA*), Seattle, Washington, pp. 6328-6335, May 2015.

[8] Roberto Arroyo, Pablo F. Alcantarilla, Luis M. Bergasa, and Eduardo Romera, "OpenABLE: An Open-source Toolbox for Application in Life-Long Visual Localization of Autonomous Vehicles," *Proceedings*, *IEEE International Conference on Intelligent Transportation Systems* (*ITSC*), Rio de Janeiro, Brazil, pp. 965-970, Nov. 2016.

[9] Sayem Mohammad Siam, and Hong Zhang, "Fast-SeqSLAM: A Fast Appearance Based Place Recognition Algorithm," *Proceedings*, *IEEE International Conference on Robotics and Automation* (*ICRA*), Singapore, pp. 5702-5708, May 2017.

[10] Chu-Tak Li, and Wan-Chi Siu, "Fast Monocular Vision-based Railway Localization for Situations with Varying Speeds," *Proceedings*, *APSIPA Annual Summit and Conference 2018* (*APSIPA-ASC 2018*), Hawaii, USA, pp. 2006-2013, Nov. 2018.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Proceedings*, *the 25th International Conference on Neural Information Processing Systems* (*NIPS*), Lake Tahoe, USA, pp. 1097-1105, Dec. 2012.

[12] Nate Merrill and Guoquan Huang, "Lightweight Unsupervised Deep Loop Closure," *Proceedings*, *Conference on Robotics: Science and Systems* (*RSS*), Pittsburgh, Pennsylvania, USA, pp. 1-9, Jun. 2018.

[13] Zetao Chen, Adam Jacobson, Niko Sunderhauf, Ben Upcroft, Lingqiao Liu, Chunhua Shen, Ian Reid, and Michael Milford, "Deep Learning Features at Scale for Visual Place Recognition," *Proceedings*, *IEEE International Conference on Robotics and Automation* (*ICRA*), Singapore, pp. 3223-3230, May 2017.

[14] Roberto Arroyo, Pablo F. Alcantarilla, Luis M. Bergasa, and Eduardo Romera, "Fusion and Binarization of CNN Features for Robust Topological Localization across Seasons," *Proceedings*, *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), Daejeon, South Korea, pp. 4656-4663, Oct. 2016.

[15] N. Sünderhauf, S. Shirzai, F. Dayoub, B. Upcroft, and M. J. Milford, "On the Performance of ConvNet Features for Place Recognition," *Proceedings*, *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), Hamburg, Germany, pp. 4297-4304, Sept. 2015.

[16] Tayyab Naseer, Wolfram Burgard, and Cyrill Stachniss, "Robust Visual Localization Across Seasons," *IEEE Trans. on Robotics*, vol.34, no.2, pp. 289-302, Apr. 2018.

[17] M. Shahid, T. Naseer, and W. Burgard, "DTLC: Deeply Trained Loop Closure Detections for Lifelong Visual SLAM," *Proceedings*, *Workshop on Visual Place Recognition*, *Conference on Robotics: Science and Systems* (*RSS*), Ann Arbor, USA, pp. 1-8, Jun. 2016.

[18] Niko Sünderhauf, Peer Neubert, and Peter Protzel, "Are We There Yet? Challenging SeqSLAM on a 3000 km Journey Across All Four Seasons," *Proceedings*, *Workshop on Long-Term Autonomy*, *IEEE International Conference on Robotics and Automation* (*ICRA*), Karlsruhe, Germany, pp. 102-115, May 2013.

[19] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol, "Extracting and Composing Robust Features with Denoising Autoencoders," *Proceedings*, *the 25th International Conference on Machine Learning*, New York, USA, pp. 1096-1103, Jul. 2008.

[20] Y Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrel, "Caffe: Convolutional Architecture for Fast Feature Embedding," *Proceedings*, *the 22nd ACM International Conference on Multimedia*, Orlando, Florida, USA, pp. 675-678, Nov. 2014.

[21] Navneet Dalal, and Bill Trigss, "Histograms of Oriented Gradients for Human Detection," *Proceedings*, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR*), San Diego, CA, USA, pp. 886-893, Jun. 2005.

[22] Chu-Tak Li, Wan-Chi Siu, and Daniel P.K. Lun, "Boosting the Performance of Scene Recognition via Offline Feature-Shifts and Search Window Weights," *Proceedings*, *IEEE International Conference on Digital Signal Processing* (*DSP*), Shanghai, China, pp. 1-5, Nov. 2018.

[23] Chu-Tak Li, Wan-Chi Siu, and Daniel P.K. Lun, "Semi-Supervised Deep Vision-based Localization using Temporal Correlation between Consecutive Frames," *Proceedings*, *IEEE International Conference on Image Processing* (*ICIP*), Taipei, Taiwan, pp. 1-5, Sept. 2019.

[24] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba, "Places: A 10 million Image Database for Scene Recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.40, no.6, pp. 1452-1464, Jul. 2017.

[25] Yi Hou, Hong Zhang, and Shilin Zhou, "Convolutional Neural Network-Based Image Representation for Visual Loop Closure Detection," *Proceedings*, *IEEE International Conference on Information and Automation* (*ICIA*), Lijiang, China, pp. 2238-2245, Aug. 2015.