Panoramic Annular Localizer: Tackling the Variation Challenges of Outdoor Localization Using Panoramic Annular Images and Active Deep Descriptors

Ruiqi Cheng[1], Kaiwei Wang[1], Shufei Lin[1], Weijian Hu[1], Kailun Yang[1],
Xiao Huang[1], Huabing Li[1], Dongming Sun[1] and Jian Bai[1]

***Abstract*— Visual localization is an attractive problem that estimates the camera localization from database images based on the query image. It is a crucial task for various applications, such as autonomous vehicles, assistive navigation and augmented reality. The challenging issues of the task lie in various appearance variations between query and database images, including illumination variations, dynamic object variations and viewpoint variations. In order to tackle those challenges, Panoramic Annular Localizer into which panoramic annular lens and robust deep image descriptors are incorporated is proposed in this paper. The panoramic annular images captured by the single camera are processed and fed into the NetVLAD network to form the active deep descriptor, and sequential matching is utilized to generate the localization result. The experiments carried on the public datasets and in the field illustrate the validation of the proposed system.**

I. INTRODUCTION

Localization is one of important research topics concerning autonomous vehicles [1], [2], robotics [3] and assistive navigation [4], [5]. Generally, GNSS (global navigation satellite system) is the straightforward way to localize the autonomous vehicles in the urban areas. However, localization tends to fail at those fade zones, such as the streets with high-rises, or under severe conditions, such as bad space weather. Fortunately, as the non-trivial sensing source, visual information is eligible for providing localization tasks with sufficient place fingerprint. The proliferation of computer vision has spurred the researchers to propose a great deal of vision-based localization solutions [1], [2], [6]–[10]. Given a query image, visual localization predicts the camera location by searching the best-matching database images featuring the largest similarity to that query image.

The most challenging and attractive part of visual localization is that the appearance variations between query and database images impact on the similarity measurement of images, so as to impede the robustness of the algorithm. Those appearance variations involve illumination variations, season variations, dynamic object variations and viewpoint variations. The variations of illumination, season and dynamic objects have been thoroughly researched by the research community. On the contrary, viewpoint variations are highly related to camera FOV (field of view) , and are tough to tackle merely using ordinary cameras. Therefore, the expansion of FOV is essential to overcome viewpoint variations between query images and database images.

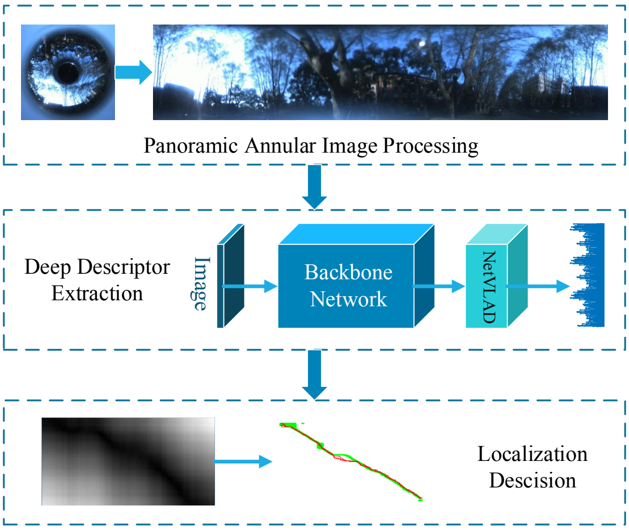

Panoramic Annular Image Processing
Deep Descriptor Extraction
Image
Backbone Network
NetVLAD
Localization Descision

Fig. 1. The schematic diagram of the proposed Panoramic Annular Localizer.

Based on our preliminary research [4], [5], [11], we propose a visual localization system PAL (panoramic annular localizer), which utilizes panoramic annular lens as the imaging sensor to capture the omnidirectional images of the surrounding environments and utilizes deep image descriptors to overcome various appearance changes. As shown in Fig. 1, the proposed PAL is composed of three phases: *panoramic image processing*, *deep descriptor extraction* and *localization decision*. The contributions of this paper are summarized as follows.

- In this paper, the panoramic annular lens is firstly integrated into outdoor visual localization system, which assists in the issues of viewpoint variations. Moreover, the car-mounted panoramic annular datasets captured in the real-world scenarios are released for the tasks of visual localization.
- Active image descriptors extracted from panoramic images are leveraged to measure the similarity between images featuring various appearance changes. The proposed active descriptor outperforms the passive descriptors derived from feature maps of convolutional

[1] All of the authors are with the State Key Laboratory of Modern Optical Instrumentation, Zhejiang University, Hangzhou, China. Corresponding author: Kaiwei Wang, wangkaiwei@zju.edu.cn

neural networks.

- The active deep descriptors are combined with the sequential matching scheme. The performance of the proposed PAL system is validated on public and self-captured panoramic datasets for outdoor visual localization.

II. Related work

In this section, the prevailing panoramic cameras and visual localization solutions based on those cameras are summarized.

A. Panoramic Cameras

Various imaging systems are eligible to capture panoramic or omnidirectional images, including *multi-camera systems*, *catadioptric cameras* and *panoramic annular cameras*. Those multi-camera systems, capture high-quality panoramic images, but the different exposure between different cameras may result in the illumination inconsistency in the panoramic image. A catadioptric camera usually constitute of an ordinary camera and a curved reflection mirror above the camera. Therefore, the mechanical structure is relatively complicated and the volume of camera is large compared with other types of panoramic cameras.

Compared with those cameras, the panoramic annular lens [12]–[16] features a simpler structure and lower cost. Through special optical design, the lateral light of lens is collected to the camera as much as possible by reflection twice within the panoramic annular lens. In view of the integrated package, the panoramic annular lens features flexible installation and is free from occlusion caused by camera's structure, especially compared with the catadioptric cameras. Apart from that, the annular images get rid of the upper and lower part of spherical images and focus more on the intermediate part images that contain important visual cues of places.

B. Image Features

Extracting robust features from images is the fundamental factor that impacts the performance of visual localization. The research community has focused on this topic for a long time. As the early research on vehicle visual localization, SeqSLAM [17] utilized the SAD (sum of absolute difference) of normorlized image patches to measure the similarity between query and database images, which is not robust against various appearance variations. Aggregating local features (such as SURF [18] and ORB [19]) into compact vectors, BoW (bag of words) places a vital role in place recognition [20], [21], especially becomes a popular place recognition approaches in SLAM (simultaneous localization and mapping) system. Unfortunately, in view of the limited description capability of hand-crafted local features, BoW performs badly under the conditions of illumination variations.

Apart from the hand-crafted descriptors mentioned above, the feature maps of prevailing CNN (Convolutional Neural Network) are also leveraged as powerful images descriptors [5], [22]. The "passive" deep descriptors cope with variation issues by the limited pretrained knowledge derived from vanilla classification tasks, which results in sub-optimal localization performance when compared with the active deep descriptors. There are several work [23]–[25] that falls into active image descriptors, which are trained specially to robustify the descriptor performance under the conditions of appearance variations.

C. Panoramic Visual Localization

In order to tackle the viewpoint variations, the research community has proposed various visual localization approaches to achieving robust place recognition. Murillo and Josecka proposed place recognition using GIST descriptors extracted from panoramic images [26]. As one of the earliest attempts on the task, the proposed approach achieved good performance on a large-scale dataset, but the issues of appearance variations are not attached importance to. Based on the NetVLAD module, the image retrieval approach of panoramic images proposed in [25] performed well on the street view dataset. However, the algorithm was not designed for the localization problems and not validated on other panoramic datasets. Oishi et al [27] proposed view-based robot localization and navigation, where panoramic images are one of the multi-modal data. The hand-crafted image features and the sliding window scheme were utilized for matching the panoramic images, which naturally behaves inferior when matching images with apparent variations.

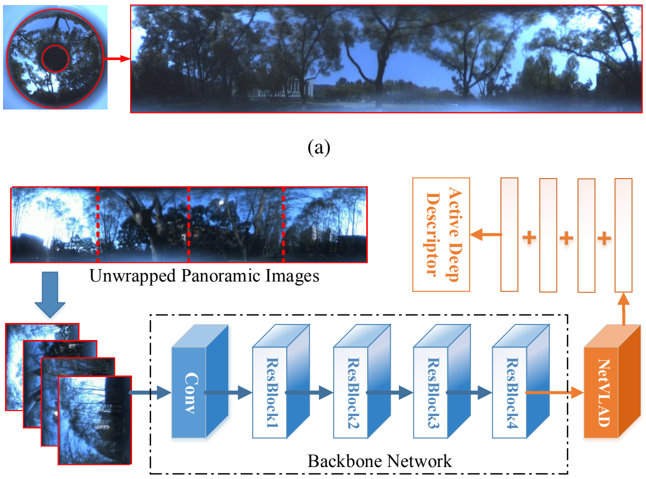
III. Methodology

In this section, we elaborate the proposed PAL system, a visual localization framework that is designed for the challenging variation issues of visual localization. Panoramic annular lens captures the omnidirectional images with 360° FOV in single frame, which is the effective camera to tackle the viewpoint variations in visual localization. Apart from that, efficient deep image features are utilized to extract the place fingerprint embedded in images, which boosts the robustness against appearance variations, such as illumination, season and dynamic object variations.

A. Preprocessing of panoramic annular images

As the name suggests, panoramic annular images [e.g. the left image of Fig. 2 (a)] are annular images that covers 360° field of view. In order to apply different kinds of feature extraction procedures to the panoramic images, the annular images are firstly unwrapped into the rectangular images, shown as the right image of Fig. 2 (a).

The camera model of panoramic annular lens is completely different from those of ordinary cameras where the pinhole model applies. Due to the wide FOV, the projection surface of the panoramic annular lens should be the curved surface (e.g. sphere), instead of the plane surface. Fortunately, the panoramic annular lens could be taken as the single-view wide-angle imaging system, and the related camera model has been studied by the research community [28]–[31]. In



(a)

(b)

Fig. 2. The unwrapping process and feature extraction of panoramic annular images. (a) Unwrapping the panoramic annular image into the rectangular image. (b) The extraction procedures of the active deep descriptor from the panoramic image.

this paper, OCamCalib [30] toolbox is utilized to calibrate the panoramic annular lens and to obtain the intrinsic parameters.

The unwrapping of the annular image is implemented complying with the following mapping, where the point (i, j) of the rectangular image corresponds to the point (x, y) of the annular image.

$$x = y_c + \rho \sin\theta \qquad y = x_c + \rho \cos\theta \tag{1}$$

$$\rho = R_{min} + \frac{R_{max} - R_{min}}{height} i \qquad \theta = \frac{2\pi j}{width} \tag{2}$$

The center of the annular image (x_c, y_c) is estimated by the calibration toolbox. Subsequently, the circular borders of the annular image are manually determined [see the double circles in the left image of Fig. 2 (a)], and R_{max} and R_{min} are thus obtained. The FOV ratio of the panoramic annular lens is utilized to determine the aspect ratio of the unwrapped rectangular image. The horizontal FOV of the panoramic annular lens is 360°, meanwhile the vertical FOV is determined by the projection model and the circular boarder of the annular image. In this paper, the panoramic annular lens features the vertical FOV of 75°, and the aspect ratio of unwrapped image is set to 4.8:1.

B. Active deep image descriptor

In this paper, we leverage NetVLAD [23] to describe the panoramic images effectively. The NetVLAD model is based on the backbone network, which is usually the pre-trained network for the classification task on the large-scale image datasets (e.g. ImageNet [32] or Places [33]). As shown in Fig. 2 (b), ResNet-18 [34] without fully connected layers is utilized as the base network of NetVLAD module.

As an analogue of the pooling layer of the CNN, NetVLAD pools the discriminative descriptor with place fingerprint from the preceding feature map. The pooling ability of NetVLAD is obtained from the training phase, when the images with diverse appearances (e.g. with different illumination, viewpoint and dynamic objects) but captured at the same place are leveraged as training data. Specifically, the triplet loss function [23] impels the descriptor of the query image to close to that of positive database images rather than that of negative images, which robustifies the adaptability of the descriptor under variations. Similar with the training procedures in [23], the dataset Pittsburgh-30k [23] with images of ordinary FOVs is utilized to in the training phase.

Having trained, the network is leveraged to extract active deep image descriptors from panoramic images. As shown in Fig. 2 (b), each unwrapped panoramic image is split into four parts along the horizontal direction. Subsequently, the four sub-images constituting a batch are fed into the proposed deep network to obtain four NetVLAD vectors. The active deep descriptor of the panoramic image is derived by adding (rather than concatenating) the four NetVLAD vectors, which is reasonable according to the principles of NetVLAD, meanwhile the final descriptor maintains compacted size.

C. Sequential localization decision

The extracted deep descriptors are leveraged to measure the similarity between images, thus to characterize the correspondence of query images and database images. Herein, we define D as the distance matrix, where the element $D_{i,j}$ is the cosine distance between the *i*-th query image and the *j*-th database image. Inspired by the offline cone-based searching proposed in [17], the online cone-based searching is executed upon the distance matrix D. As shown in Fig. 3, each query-database pair (i, j) within the distance matrix is associated with two symmetrical cone regions which is limited by sequential length n_q, maximal velocity v_{max} and minimal velocity v_{min}. The online searching algorithm only makes use of the "past" query images instead of the "future" query images.

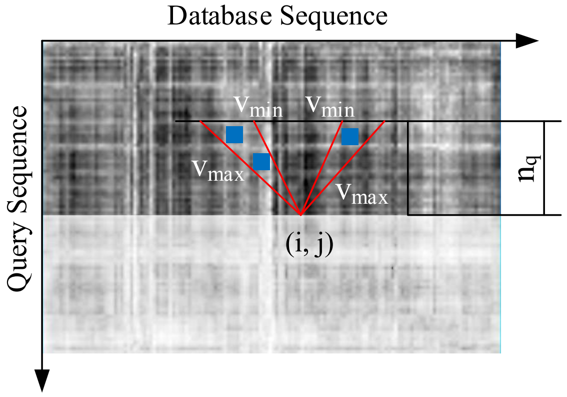


Fig. 3. Using cone-based sequential searching to score the matching correspondence of images.

Within the regions, the number of best-matching pairs n_{match} is counted. The query descriptor and its nearest neighbor among the database descriptors compose the best-matching pair. The score of the query-database pair (i, j) is defined as $s_{i,j} = n_{match}/n_q$. Naturally, all of the matching scores form into a score matrix S, where each query image corresponds to the best database image with the highest score. In order to get the final place recognition results, the matching score of the best query-database pair is evaluated through window uniqueness thresholding [17] to rule out the mismatched pairs.

IV. Experiments

In this section, both public and self-collected datasets are used to evaluate the proposed PAL system. Firstly, the performance of active deep descriptor was validated on MOLP dataset [10]. Secondly, the visual localization performance was tested in the field.

A. Comparison between passive and active deep descriptors

In the public panoramic dataset MOLP, four binocular cameras mounted on the vehicle were utilized to capture images of four different directions. In this paper, the summer night images of city roads captured in the reverse traversing direction are used to evaluate the performance of deep descriptors, including passive and active descriptors.

If the index difference between the place recognition result and the ground truth is not larger than the tolerance (set to 5 in this paper), the result is defined as a TP (true positive) result. Otherwise, the positive result is defined as a FP (false positive) result. Moreover, if the ground truth is not empty but the query does not match with any database image, the negative result is defined as a FN (false negative) result. The performance of localization is evaluated and analyzed in terms of F_1 score.

$$F_1 = 2 \times (P \times R)/(P + R), \tag{3}$$

where $R = TP/(TP + FN)$ and $P = TP/(TP + FP)$.

According to our previous research [5], the deep descriptors derived from GoogLeNet [35] pretrained on ImageNet [32] is the optimal choice on the task of visual localization among the prevailing networks. Therefore, the five optimal feature maps of convolutional or pooling layers in GoogLeNet (listed in Table I) are selected as one of the baseline of passive descriptors. Moreover, the conv3 and fc6 layer of AlexNet [36] pretrained on ImageNet are also used as another baseline of passive descriptors. Similar to the active descriptor extraction, the split images are fed into the network to obtain the feature maps of the designated layer, and those feature maps derived from sub-images are flatten and concatenated together to form into the descriptor. The reverse traversing direction of database and query sets is taken as a priori knowledge to concatenating the partial descriptors in the consistent order of sub-images.

Among active deep descriptors, the NetVLAD descriptors based on different backbone networks (i.e. AlexNet, VGG-16 [37] and ResNet-18) are compared to choose the optimal structure. The fully connected layers of the backbone networks are removed and the last convolutional layer is connected with the successive NetVLAD module. In this paper, all of the NetVLAD networks are trained on the public dataset Pittsburgh-30k with the default training parameters [23]. In order to compare various deep descriptors fairly, the input images fed into different networks are universally set to 224×224. Meanwhile, the split number of panoramic images are also compared in the experiment, where four-part split, two-part split and one-part split are tried. The brute force searching is utilized to find the nearest neighbor of the query descriptor as the best-matching results, which are evaluated with F_1 score.

TABLE I

The localization performance (F_1 score) of different split number on MOLP dataset (In.=Inception)

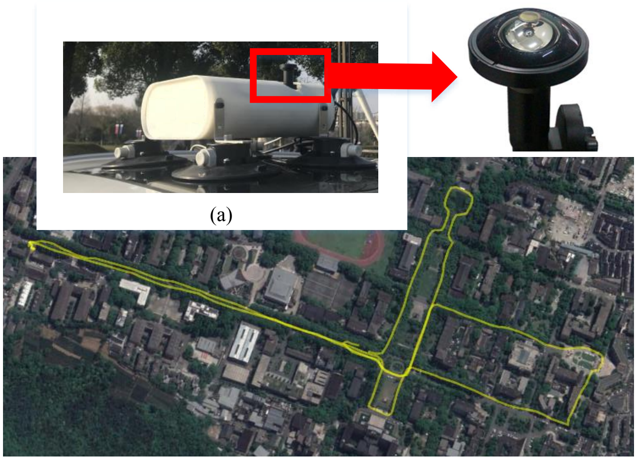
F_1		One-part	Two-part	Four-part
AlexNet	conv3	0.06	0.43	0.38
	fc6	0.03	0.18	0.11
GoogLeNet	In.3a/3×3	0.04	0.29	0.26
	In.3a/3×3_reduce	0.07	0.27	0.12
	In.3b/3×3_reduce	0.06	0.51	0.37
	In.3a/pool_proj	0.07	0.34	0.25
	In.5b/1×1	0.09	0.32	0.06
NetVLAD with AlexNet		0.19	0.29	0.28
NetVLAD with VGG-16		0.33	0.51	0.50
NetVLAD with ResNet-18		**0.41**	**0.51**	**0.54**

The experimental results are shown in Table I. The proposed active descriptors composed of NetVLAD network with ResNet-18 achieves the best performance among the listed descriptors. The feature maps of AlexNet and GoogLeNet behave better on the condition of two-part split, and are influenced substantially by the way of split. Comparatively, the descriptors derived from NetVLAD networks perform more stable among different split ways. Moreover, the superiority of the NetVLAD descriptors also lies in the combination way of sub-image descriptors. Due to the principle of netVLAD, the direct superposition of four descriptors does not require to know the traversing direction of datasets.

B. Performance on the real-world scenarios

The performance of PAL is validated on the Yuquan dataset, which was captured at the Yuquan Campus of Zhejiang University, China. The panoramic images were captured by the car-mounted panoramic annular lens on a three-kilometer route, as shown in Fig. 4. The database sequence were captured in the sunny afternoon, meanwhile the query sequences were captured both in the afternoon (subset-1) and at dusk (subset-2). Meanwhile, GNSS data were also collected.

It is worthwhile to note that the database sequence is not completely overlapped with the two query sequences. Those unseen query images matched with the database image are denoted as false results. The ratio of false results out of all positive results is FR (false rate). The PR (positive rate) of localization is defined as the ratio of matching results, the distance between which is within the tolerance ($50m$). In the experiment, the sequential matching parameters are set as $v_{min} = 0.4$, $v_{max} = 2.5$, $n_q = 10$. As shown in Table II, the proposed PAL surpasses OpenSeqSLAM2.0 [17]



(a)

(b)

Fig. 4. (a) The panoramic annular lens and the peripheral device are mounted on the top of the car. (b) The test route (denoted by yellow lines) covers around three kilometers.

on the real-world dataset. Some localization results of PAL system are shown in Fig. 5. It is concluded that the proposed approach is eligible to overcome the appearance variations (especially illumination variation). Moreover, the panoramic camera mounted on the roof of the car naturally reduces the impact of dynamic objects (like pedestrians and cars) on the performance of visual localization.

TABLE II

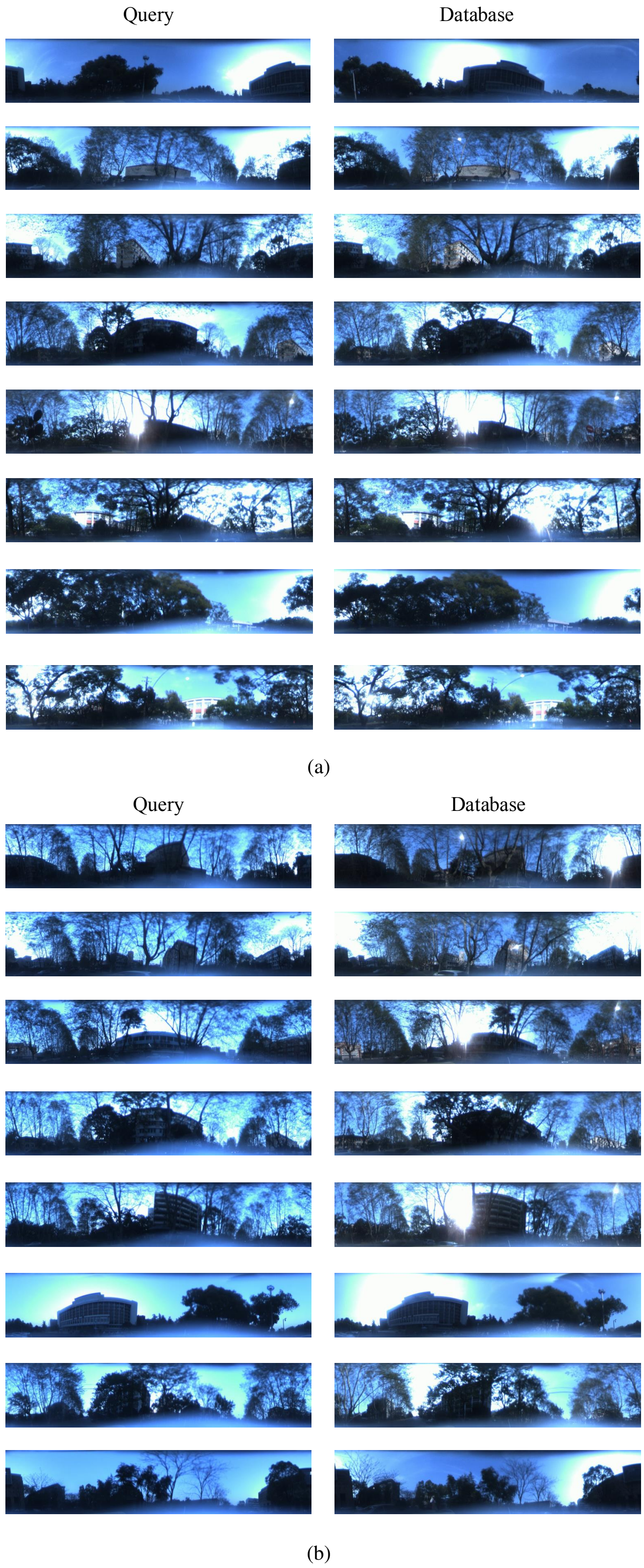
THE LOCALIZATION PERFORMANCE ON YUQUAN DATASET

Algorithms	Subset-1		Subset-2	
	FR	PR	FR	PR
OpenSeqSLAM2.0	17.15%	21.67%	23.81%	41.31%
PAL	**11.13%**	**32.89%**	**20.92%**	**45.24%**

For computational complexity, the network inference of the deep feature extraction phase consumes around $40ms$ on NVIDIA GPU (GTX-1060). Thereby, the active deep descriptor features not only robust description capability but also real-time performance. As for the phase of sequential matching, each image consumes around $13ms$ to get the decision of place recognition. In a brief, the proposed PAL system could perform in real time in the practical scenarios.

V. CONCLUSIONS

In order to tackling the variation issues of visual localization, PAL (panoramic annular localizer) is proposed in this paper. We incorporate panoramic annular lens and the active deep descriptor into the visual localization system. Firstly, the unwrapping of annular images is executed to prepare for the phase of descriptor extraction, where the split panoramic images are fed into the NetVLAD network based on ResNet-18. The descriptors obtained from sub-images are added together regardless of concatenation order, and the cone-based matching is leveraged to robustify the single-frame retrieval results. The experiment on MOLP dataset illustrate



(a)

(b)

Fig. 5. The visual localization results of (a) the afternoon test set and (b) the dusk test set.

that the proposed active descriptor outperforms the off-the-shelf deep descriptors. In the field test, the performance of the proposed system in the practical scenarios is validated.

The code and dataset related to the proposed PAL system are publicly available at https://github.com/chengricky/PAL. In the future, the scene understanding and pose estimation are planned to be researched based on this work.

ACKNOWLEDGMENT

This work was supported by the State Key Laboratory of Modern Optical Instrumentation.

REFERENCES

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

[2] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa, and E. Romera, "Are you able to perform a life-long visual topological localization?" *Autonomous Robots*, vol. 42, no. 3, pp. 665–685, Mar 2018.

[3] L. Tang, Y. Wang, X. Ding, H. Yin, R. Xiong, and S. Huang, "Topological local-metric framework for mobile robots navigation: a long term perspective," *Autonomous Robots*, vol. 43, no. 1, pp. 197–211, Jan 2019.

[4] R. Cheng, K. Wang, L. Lin, and K. Yang, "Visual localization of key positions for visually impaired people," in *2018 24th International Conference on Pattern Recognition (ICPR)*, Aug 2018, pp. 2893–2898.

[5] S. Lin, R. Cheng, K. Wang, and K. Yang, "Visual Localizer: Outdoor localization based on convnet descriptor and global optimization for visually impaired pedestrians," *Sensors*, vol. 18, no. 8, 2018.

[6] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, Feb 2016.

[7] P. Neubert and P. Protzel, "Beyond holistic descriptors, keypoints, and fixed patches: Multiscale superpixel grids for place recognition in changing environments," *IEEE Robotics and Automation Letters*, vol. 1, no. 1, pp. 484–491, Jan 2016.

[8] O. Vysotska and C. Stachniss, "Effective visual place recognition using multi-sequence maps," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1730–1736, April 2019.

[9] Y. Naiming, T. Kanji, F. Yichu, F. Xiaoxiao, I. Kazunori, and I. Yuuki, "Long-term vehicle localization using compressed visual experiences," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, Nov 2018, pp. 2203–2208.

[10] S. Siva and H. Zhang, "Omnidirectional multisensory perception fusion for long-term place recognition," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 1–9.

[11] R. Cheng, K. Wang, J. Bai, and Z. Xu, "OpenMPR: Recognize places using multimodal data for people with visual impairments," *Measurement Science and Technology*, 2019.

[12] S. Niu, J. Bai, X. yun Hou, and G. guang Yang, "Design of a panoramic annular lens with a long focal length," *Applied Optics*, vol. 46, no. 32, pp. 7850–7857, Nov 2007.

[13] Z. Huang, J. Bai, T. X. Lu, and X. Y. Hou, "Stray light analysis and suppression of panoramic annular lens," *Optics express*, vol. 21, no. 9, pp. 10 810–10 820, 2013.

[14] X. Zhou, J. Bai, C. Wang, X. Hou, and K. Wang, "Comparison of two panoramic front unit arrangements in design of a super wide angle panoramic annular lens," *Applied Optics*, vol. 55, no. 12, pp. 3219–3225, Apr 2016.

[15] Y. Luo, J. Bai, X. Zhou, X. Huang, Q. Liu, and Y. Yao, "Non-blind area pal system design based on dichroic filter," *Optics Express*, vol. 24, no. 5, pp. 4913–4923, Mar 2016.

[16] H. Xiao and B. Jian, "Analysis of the imaging performance of panoramic annular lens with conic conformal dome," in *AOPC 2015: Optical Design and Manufacturing Technologies*, vol. 9676. International Society for Optics and Photonics, 2015, p. 96760G.

[17] B. Talbot, S. Garg, and M. Milford, "OpenSeqSLAM2.0: An open source toolbox for visual place recognition under changing conditions," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2018, pp. 7758–7765.

[18] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 404–417.

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

[20] A. Glover, W. Maddern, M. Warren, S. Reid, M. Milford, and G. Wyeth, "OpenFABMAP: An open source toolbox for appearance-based loop closure detection," in *2012 IEEE International Conference on Robotics and Automation*, May 2012, pp. 4730–4735.

[21] R. Mur-Artal and J. D. Tardos, "ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, oct 2017.

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

[23] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: Cnn architecture for weakly supervised place recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1437–1451, June 2018.

[24] M. Lopez-Antequera, R. Gomez-Ojeda, N. Petkov, and J. Gonzalez-Jimenez, "Appearance-invariant place recognition by discriminatively training a convolutional neural network," *Pattern Recognition Letters*, vol. 92, pp. 89 – 95, 2017.

[25] A. Iscen, G. Tolias, Y. Avrithis, T. Furon, and O. Chum, "Panorama to panorama matching for location recognition," in *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, ser. ICMR '17. New York, NY, USA: ACM, 2017, pp. 392–396.

[26] A. C. Murillo and J. Kosecka, "Experiments in place recognition using gist panoramas," in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, Sep. 2009, pp. 2196–2203.

[27] S. Oishi, Y. Inoue, J. Miura, and S. Tanaka, "SeqSLAM++: View-based robot localization and navigation," *Robotics and Autonomous Systems*, vol. 112, pp. 13 – 21, 2019.

[28] C. Geyer and K. Daniilidis, "Paracatadioptric camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 687–695, May 2002.

[29] X. Ying and Z. Hu, "Can we consider central catadioptric cameras and fisheye cameras within a unified imaging model," in *Computer Vision - ECCV 2004*, T. Pajdla and J. Matas, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 442–455.

[30] D. Scaramuzza, A. Martinelli, and R. Siegwart, "A toolbox for easily calibrating omnidirectional cameras," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2006, pp. 5695–5701.

[31] C. Mei and P. Rives, "Single view point omnidirectional camera calibration from planar grids," in *Proceedings 2007 IEEE International Conference on Robotics and Automation*, April 2007, pp. 3945–3950.

[32] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 248–255.

[33] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.