CFVL: A Coarse-to-Fine Vehicle Localizer with Omnidirectional Perception across Severe Appearance Variations

Yicheng Fang[1], Kaiwei Wang[2], Ruiqi Cheng[1] and Kailun Yang[3]

***Abstract*— Visual localization in vehicle navigation remains a crucial image retrieval task to determine the best matched image. Developing an efficient algorithm to address the localization issues of vehicle is highly difficult, for severe appearance variations with vehicles moving around can bring about significant challenges and big obstacles. In this paper, we propose the CFVL framework which takes panoramas into use in the localizer and the system processes from coarse to fine, in order to attain more robust and stable descriptors. NetVALD descriptors based on explicit panorama construction, which are regarded robust to appearance changes, are extracted in the coarse stage, while Geodesc keypoint descriptors, which are believed to detect more detailed information, are utilized in the fine stage, so as to perceive the accurate localization. A comprehensive set of experiments is carried on several datasets with different appearances across seasonal cycling, illumination variations, diverse traversals, and so on, to verify the effectiveness of the coarse stage and fine stage in our system. Brute Force (BF) matching and Fundamental Matrix mapping are utilized to match and locate correct locations after coarse stage and after fine stage. The accuracy of the coarse matching and fine matching are verified separately. Our system is demonstrated to be with high location recall, generalization capacity across different environments.**

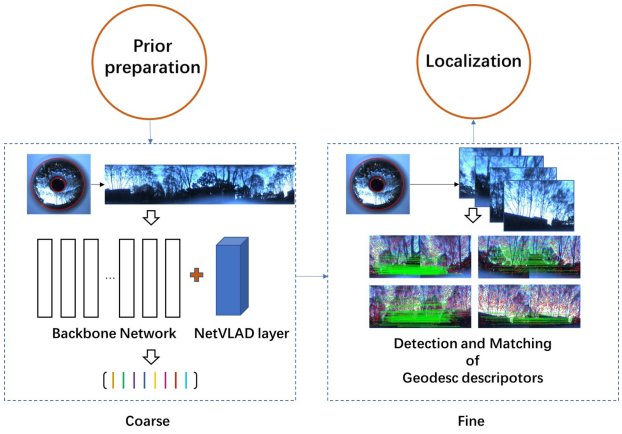


Fig. 1. Overview of the proposed CFVL framework. In the prior preparation phase, the panoramic annular images will be processed in explicit panoramas or plane images in two forms. In the localization phase, NetVLAD descriptors and Geodesc descriptors will be detected in the coarse stage and fine stage respectively. After that, Brute Force (BF) matching and Fundamental Matrix mapping will be conducted respectively.

I. INTRODUCTION

Visual localization has always been an unsolved and challenging problem in the field of vehicle navigation. Day-night cycling, seasonal variations, highly dynamic objects, viewpoint changing, and illumination variations can all be regarded as severe appearance variations that limit the robustness of navigation systems [1] [2] [3] [4]. Panoramas based visual localizition, as a state-of-art field of localization, has also been tried sporadically. Panoramas can greatly diminish the impact of changes in viewpoints. In addition, for constant attention from all directions is required, and omnidirectional perception is highly needed for autonomous vehicles, panoramas are also necessary [5] [6] [7].

In the literature, a cluster of visual localization algorithms based on panoramas has been presented [8] [9] [10] [11] [12], but some of the features they design may not be adaptive to large-scale driving scenarios that appear in vehicle navigation. Besides, the lack of open-source panoramic datasets with diverse scenes, and the low computation efficiency make it harder for researchers to leverage panoramic localization for intelligent vehicles applications.

Convolutional Neural Networks (CNN) descriptors have been extensively utilized in visual localization and achieved competent performances [13] [14] [15] [16] [17]. Recently, a state-of-art network called NetVLAD [18], which combines a backbone network and a NetVLAD layer, is confirmed to reach superior performances than common CNNs in visual place recognition and location retrieval tasks. Common active deep descriptors are usually extracted from models pre-trained by local images with only forward view of a scene, which are unreliable for omnidirectional perception of the whole surroundings due to the discrepancies in panoramas and conventional pin-hole images [5]. On the other hand, only active deep features are not accurate enough to yield the exact top1 result, but they allow to generate a coarse range for the right locations [2].

For these reasons, we propose CFVL, a Coarse-to-Fine Vehicle Localizer, which performs omnidirectional perception using panoramas explicitly, to tackle the challenges brought by severe appearance variations in vehicle navigation. Our approach can be divided into three parts, where the overview of the proposed CFVL framework is shown in Figure 1. The contributions of this paper are summarized as follows:

- This paper proposes a novel vehicle localization system: CFVL, which utilizes the panoramic images to assist localization under viewpoint changing during outdoor

[1]Y. Fang and R. Cheng are with State Key Laboratory of Mordern Optical Instrumentation, Zhejiang University, China {fangyicheng, rickycheng}@zju.edu.cn

[2]K. Wang is with National Optical Instrumentation Engineering Technology Research Center, Zhejiang University, China wangkaiwei@zju.edu.cn (Corresponding author)

[3]K. Yang is with Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Germany kailun.yang@kit.edu

driving, meanwhile obtains omnidirectional perception.
- CFVL combines NetVLAD descriptors and Geodesc [19] keypoint descriptors, taking advangate of appearance invariance for NetVLAD descriptors and capacity of obtaining detailed information for Geodesc descriptors, which form a Coarse-to-Fine vehicle localization method. It is demonstrated that the coarse stage is able to search a rough range for accurate localization, while the fine stage can provide much finer results.
- We collect a real-world panoramic dataset with perfect vehicle views——Chengyuan dataset, to facilitate the study of panoramic visual localizition, and to validate our CFVL system. Chengyuan dataset can be obtained at https://github.com/dachengzihhh/Chengyuan-dataset.

II. Related work

In this section, some related work including acquisition of wide-angle images, diverse descriptors for image retrieval and attempts on panoramic localization are reviewed.

A. Acquisition of wide-angle images

Sometimes large-angle views are needed to acquire omnidirectional information of the surroundings in the fields of panoramic monitoring, pipeline detection, machine vision and autonomous vehicle. Generally, two ways are widely employed to obtain wide-angle images, one of them is by image stitching [20], with superiorities of high image resolution and equipment simplification. Another is single-sensor gaze imaging technology, which reaches a perfect real-time performance, where fish-eye lens [21] and panoramic annular lens (PAL) [22] are two applications of it, as no latency will be incurred in synchronization and data fusion.

The panoramas gained by image stitching has limitations of poor real-time performance and stitching errors. Fish-eye lens can get over these difficulties, however, the images taken by them have a strong negative distortion in the edge region, and the large areas of skyward contents in the images are regarded useless for visual localization. Luckily, PAL, whose illumination of the phase surface is uniform, and the images taken by whom don't have serious negative distortion on the edge, thus can perfectly satisfy our requirement in vehicle localization tasks. PAL consists of a PAL block, a set of Relay Lens (RL), and a camera for imaging. After light enters, it is reflected three times by the PAL block to reduce the incident angle, and then the image is captured on the camera through the RL system with positive optical focus which contributes again to the reduction of incident angle, the basic structure and the imaging optical path of the system of PAL are shown in Figure 2.

B. Diverse descriptors for image retrieval

In the field of visual localization, adaptive descriptors are utilized when the circumstances change. Sometimes holistic images can be presented by local features [23] [24] while sometimes by global descriptors [25] [26]. GIST [26] [27] is a kind of handcrafted global descriptors, which is extracted

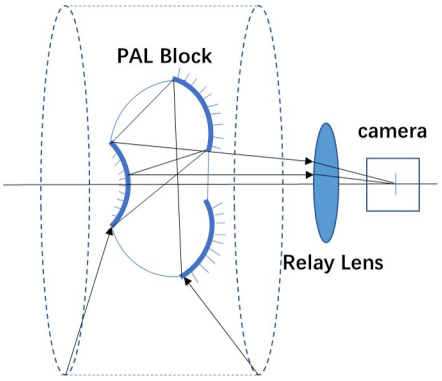


Fig. 2. Basic structure and the imaging optical path of PAL.

from holistic images, shows sensitivity to the viewpoint changing. Similarly, SeqSLAM [25], as a holistic algorithm based on sequence matching, is also found not robust to viewpoint changes, although it performs remarkably well under appearance changes. However, SIFT [24], SURF [28] local descriptors have demonstrated viewpoint invariance, but degrades the robustness on the appearance changing conditions. Our previous work has proved that local descriptors combined with holistic descriptors method can cope with the visual localization problems against environment changing [2].

Extensive Convolutional Neural Networks descriptors are prevailing in perceiving the accurate localization in recent days [13] [14] [15] [16]. One application of CNN descriptors is to utilize a pre-trained network model to obtain a knowledge from classification tasks. S. Lin et al. [3] evaluated different layer features derived from five prevailing ConvNets (AlexNet, VGGNet, GoogLeNet, SqueezeNet, MobileNet) on their robustness against various environmental changes. It is demonstrated that GoogLeNet has overwhelming advantages over other ConvNets. However, as the information extracted from pre-trained network is limited, more active and adaptive CNN descriptors are required. Then, the more robust NetVLAD network was proposed, by combining a backbone network and a NetVLAD layer, where the NetVALD descriptors can effectively improve the ability to express images of the same category and enhance the capability of image retrieval [18].

C. Attempts on panoramic localization

I. Ahmet et al. [29] proposed a panorama to panorama matching method based on NetVLAD descriptors for location recognition. By testing through implicit construction of a panorama in the descriptor space and explicit construction of a panorama in the image space respectively on the "street view" imagery, it comes out that a single NetVLAD descriptor is preferable than aggregating individual views into a vector in most of the situations for visual localization tasks. R. Cheng et.al. [1] also presented a Panoramic Annular Localizer based on panoramic annular images and active deep descriptors. It is demonstrated that active deep descriptors, especially NetVLAD descriptors, obtain a superior performance than some passive methods. The

proposed CFVL utilizes explicit panoramas to train active NetVLAD descriptors, instead of common local forward-facing images, and attaches a fine matching procedure to enhance matching ability. We will compare their matching results with our CFVL on the Yuquan dataset [1] in the following experiments.

III. Methodology

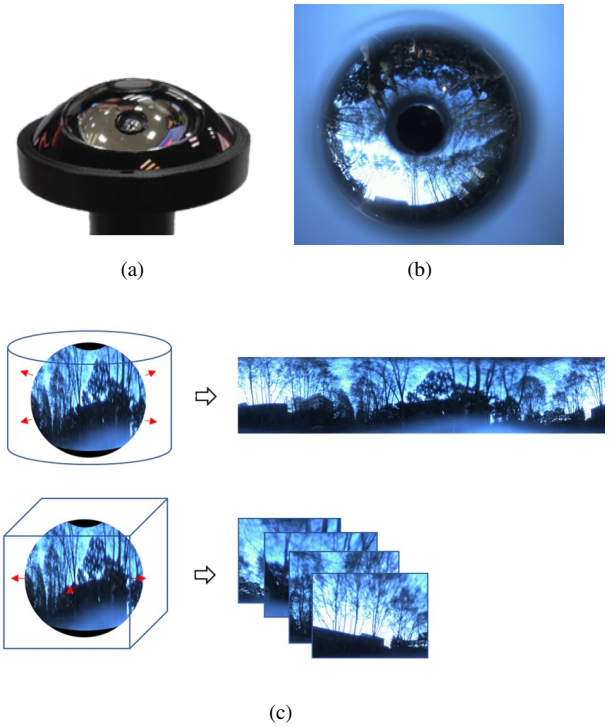
In this section, methodology of our CFVL framework will be interpreted in three dominating parts: prior preparation of panoramic annular images, descriptors extraction on both coarse and fine stages, as well as localization.

A. Prior preparation: panoramic annular image processing in two forms

Because the panoramic annular images are not consistent with human visual sense, processing them in which form is of great significance. Figure 3(a) shows a PAL and Figure 3(b) shows a panoramic annular image imaged by PAL. To eliminate distortion in the panoramas, we need to project the view of PAL into a unit spherical surface first, to ensure that each pixel is of the same distance from the original point. Then the projection on the spherical surface will be processed in two forms, as shown in Figure 3(c). The upper method projects the unit spherical surface onto a cylinder, and then unfolds the cylinder. In this way, a complete explicit panorama is produced. For the NetVLAD network is trained on the explicit panoramas, the test images would better be explicit panoramas unfolded in the first form, so that feature consistency can be guaranteed. Another processing form is projecting the spherical surface onto a cube, as shown in Figure 3(c) (the bottom method), in which four plane images will be obtained from one panoramic annular image. In the Geodesc descriptor matching procedure, the Fundamental matrix [30] of query images and database images will be computed, where the planarity of the images must be confirmed so that solid geometry principle can be satisfied. The second processing way can perfectly suits this strict principle.

B. Descriptors extraction on both coarse and fine stages

- Coarse stage: NetVLAD global descriptors
  The NetVLAD network architecture includes a standard CNN and a NetVLAD layer. By training such network in a backward propagating way, the model will obtain the learning ability, to recognize location information from different appearances provided by images. Triple loss is designed to impel the query images to find out which are the most similar images from the dataset, to distinguish positives and negatives. In this way, once we feed a panorama into the NetVLAD network, a robust holistic descriptor will be extracted to tackle the challenges of appearance changing.
- Fine stage: Geodesc keypoint descriptors
  Geodesc descriptors are state-of-art deep-learning based keypoint descriptors, which offer a novel batch constructed method that simulates the pixel-wise matching and effectively samples useful data for the learning process. The model is pre-trained on the Hpatch dataset [31], so it is convenient to deploy with any other images in a straightforward way. Geodesc descriptors are more adaptive to image content, compared to SIFT descriptors. Figure 4 shows an example of matching result comparison between SIFT descriptors and Geodesc descriptors on two similar ordinary images with slightly different viewpoint, where more matching pairs and fewer mismatches are detected when Geodesc descriptors are utilized. In this sense, it is worth considering to implement Geodesc descriptors to ensure that we have sufficient points for the final localization.



(a) (b) (c)

Fig. 3. (a) PAL; (b) A panoramic annular image imaged by PAL; (c) Two panoramic annular image processing forms, the upper method projects the unit spherical surface onto a cylinder, while the bottom method projects the unit spherical surface onto a cube.

C. Localization

Given a query panorama, Brute Force (BF) [32] matching helps NetVLAD descriptors find out a rough accurate localization result, by computing the Euclidean distance [33] between query image and database image as follows:

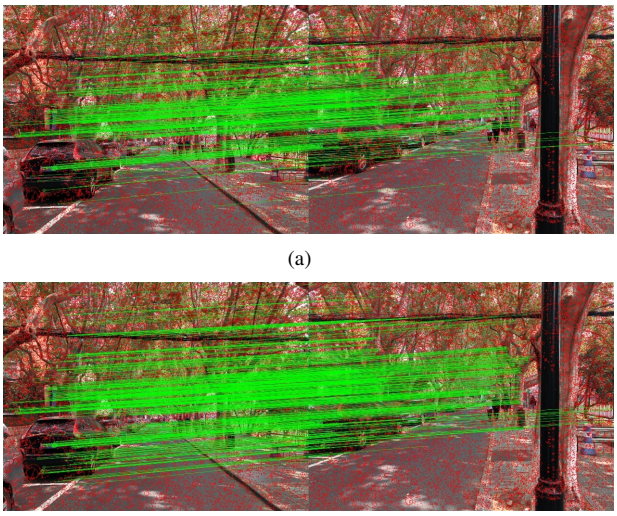
$$\begin{aligned} d(x,y) &= \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2} \\ &= \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \end{aligned} \tag{1}$$

where

$$x = (x_1, x_2, \ldots, x_n) \tag{2}$$

represents the feature extracted from query image;

$$y = (y_1, y_2, \ldots, y_n) \tag{3}$$



(a)

(b)

Fig. 4. Comparison of matching result between (a) SIFT descriptors and (b) Geodesc descriptors.

represents the feature extracted from database image; and $d(x, y)$ represents the Euclidean distance between the query image and the database image.

Apart from the rough localization proposed above, Geodesc descriptors help to determine a finer result. It is able to find out a mapping relationship based on Fundamental Matrix, between keypoints from query images and database images, on the condition that the images are all regarded as planes. Fundamental Matrix can be represented as follows:

$$q_1 \mathbf{F}^{\mathrm{T}} q_2 = 0 \tag{4}$$

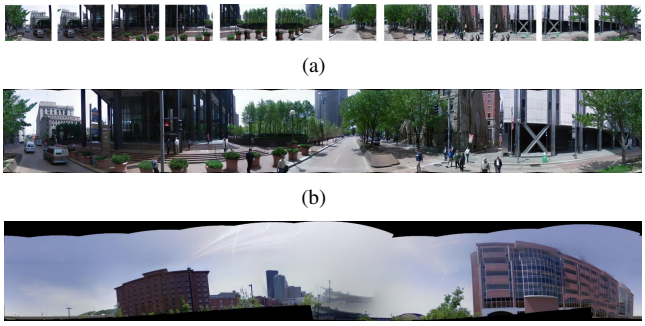
where q_1, q_2 represent pixel coordinates of two images; F represents the Fundamental Matrix. However, panoramas can not be regarded as planes any more, because they are processed from panoramic annular images as cylinders other than planes, which can be a huge problem for panoramic matching, thus panoramic annular image projection onto a cube will also be taken into consideration in our experiments.

IV. Experiments

A. Panoramic stitching of Pittsburgh dataset

It is found that Pittsburgh dataset [34] can perceive more scenarios due to its image diversity. Each image from the Pittsburgh dataset is associated with a GPS location and a total of 24 images are associated with the same GPS location. The 24 images, which perceive 360° omnidirectional information of a location, are collected from different perspectives and there are overlapped areas between each two adjacent images.

Out of expectation to construct explicit panoramas in order to acquire training data and validation data, we adopt the existing stitching method [20]. In addition, because we need only scenes from the lower perspective that is suitable for intelligent vehicle applications, only the bottom 12 images are stitched. This also helps to save memory and improve computation efficiency. Figure 5(a) shows random bottom 12 images in one location from Pittsburgh dataset, and Figure 5(b) shows the stitched result. The existence of faint black blur on the edge is reasonable after stitching. Not all the images are stitched perfectly and successfully, for the features between adjacent images are too similar or too few. Figure 5(c) shows a mis-stitched result. We transform images from Pitt250k train subset and Pitt250k val subset to the stitched panoramas according to the image sequence numbers, and the mis-stitched images are discarded. Finally 3632 database and 296 query panoramas are included in the Pitt250k train subset, while 3207 database and 294 query panoramas are contained in the Pitt250 val subset. We train NetVLAD network on Pitts250k train subset.



(a)

(b)

(c)

Fig. 5. (a) Random bottom 12 images in one location from Pittsburgh dataset; (b) The stitched panorama of (a); (c) A mis-stitched result.

B. Validation of coarse stage: whether rough range can be selected?

To validate whether the coarse stage can select a rough range for a accurate location, query explicit panoramas are unfolded in the first form (as shown in Figure 3(c), the upper approach), and then are fed into AlextNet [35], VGG16 [36] and ResNet18 [37] with NetVLAD network respectively. This way, we obtain three kinds of NetVLAD descriptors, which will be utilized to determine top20, or in other words, a rough range of accurate localization.

TABLE I
Localization results of coarse stage on Pitt250 val subset

Model	*recall@1*	*recall@5*	*recall@10*	*recall@20*
C:AlexNet+NetVLAD	0.7585	0.9354	0.9226	0.9762
C:VGG16+NetVLAD	0.9286	0.9864	0.9898	0.9898
C:ResNet18+NetVLAD	0.9456	0.9864	0.9932	0.9932

[a]C means coarse stage.

Firstly we verify the feasibility of the NetVLAD models trained by explicit panoramas, by experimenting on stitched Pitt250k val subset. Recall@TopN means the precision when only if one of the N results is in the range of ground truth, it is regarded as a accurate localization. Matching results are shown in Table 1, we can easily see that the models perform a remarkable performance on not only rough range, but even on fine range. The best performed model ResNet18 with NetVLAD on rough range reaches a high recall@20 of 0.9932. What's more, the finest top1 result gets to 0.9456, even the top 12 images and mis-stitched images

are discarded, which shows a great success of NetVLAD models trained by explicit panoramas. However, here the train subset and val subset are from the same dataset, the versatility and generalization capacity of the models should also be taken into consideration by verifying the performance in unseen domains.

TABLE II
LOCALIZATION RESULTS OF COARSE STAGE ON MOLP DATASET-BACKWARD AS QUERY, FORWARD AS DATABASE

Model	Query: Backward, Database: Forward			
	recall@1	*recall@5*	*recall@10*	*recall@20*
C:AlexNet+NetVLAD	0.2525	0.5170	0.6453	0.7976
C:ResNet18+NetVLAD	0.3006	0.6253	0.7635	0.8657
C:VGG16+NetVLAD	0.4008	0.6613	0.7495	0.8417

[a]C means coarse stage.

Secondly the NetVLAD models are tested on the public explicit panoramic MOLP dataset [9] across reverse traversing directions, to further validate the generalization capability for large-scale images and discriminability for appearance changes. MOLP dataset is captured by four binocular cameras in different seasons, in driving perspectives. We evaluate our coarse stage on summer night subset which includes forward and backward routes from city, and the tolerance is set to 5 images before and after the ground truth. The matching results are shown in Table 2. The best recall@20 reaches 0.8657, it is believed that the rough range can be relatively accurate to include correct locations. It proves that our NetVLAD models trained by explicit panoramas can learn something irrelevant with day-night appearances, although the train data have no night information, which is helpful to identify the location features from environments.

TABLE III
LOCALIZATION RESULTS OF COARSE STAGE ON YUQUAN DATASET-AFTERNOON2 AS QUERY, AFTERNOON1 AS DATABASE

Model	Query: Afternoon2, Database: Afternoon1			
	recall@1	*recall@5*	*recall@10*	*recall@20*
C:ResNet18+NetVLAD	0.6388	0.7975	0.8343	0.8697
C:AlexNet+NetVLAD	0.7323	0.8470	0.8966	0.9334
C:VGG16+NetVLAD	0.8173	0.8924	0.9292	0.9589
	Query: Dusk, Database: Afternoon1			
	recall@1	*recall@5*	*recall@10*	*recall@20*
C:ResNet18+NetVLAD	0.3556	0.6038	0.7068	0.7881
C:AlexNet+NetVLAD	0.3788	0.5181	0.6313	0.7605
C:VGG16+NetVLAD	0.5893	0.8200	0.9057	0.9057

[a]C means coarse stage.

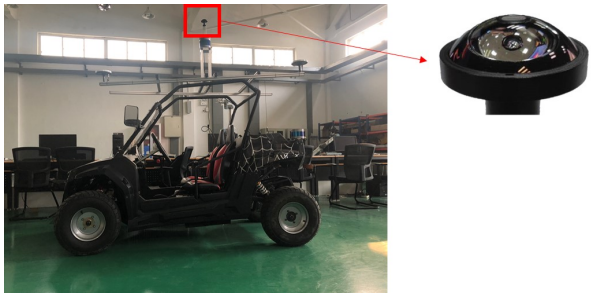
The third validation is conducted on Yuquan dataset from real-world vehicle scenarios overcoming different illumination and traverses. Yuquan dataset is collected with PAL on a three-kilometer route in Zhejiang University. The subset Afternoon1 and subset Afternoon2 are captured both on sunny afternoon but from different traverses and another subset Dusk is captured at dusk. Setting Afternoon1 subset as database and Afternoon2 subset as query allows to validate performances of coarse stage against different traverses while setting Afternoon1 subset as database and Dusk subset as query facilitates the comparison of our system under different illumination conditions. The matching results are shown in Table 3. In our experiments, ground truth is annotated based on GPS information and the tolerance distance between two images is set to 50 meters. The NetVLAD models also succeed in harsh real-world scenarios, the best performed VGG16 with NetVLAD achieves a recall@20 of 0.9589 when Afternoon2 is set as query, and achieves a recall@20 0f 0.9057 when Dusk is set as query.

We can also easily compare the top1 results on Yuquan datset with the Panoramic Annular Localizer proposed by R. Cheng et al. [1] under the same conditions. Our top1 best matching results, 0.8173 when the query is Afternoon2 and 0.5893 when the query is Dusk, outstripping their approach which feeds panoramas directly into NetVLAD with ResNet18 trained by local images method, whose positive rate is 0.4524 and 0.3289 respectively. In this sense, our model superiority trained by explicit panoramas is proved. However, the pity is that we can't easily get access to the calibration results of the datasets above, so the processing the panoramas into planes is unattainable, which can't satisfy the matching conditions of the fine stage, but only the coarse stage can already reach an excellent performance under challenges.

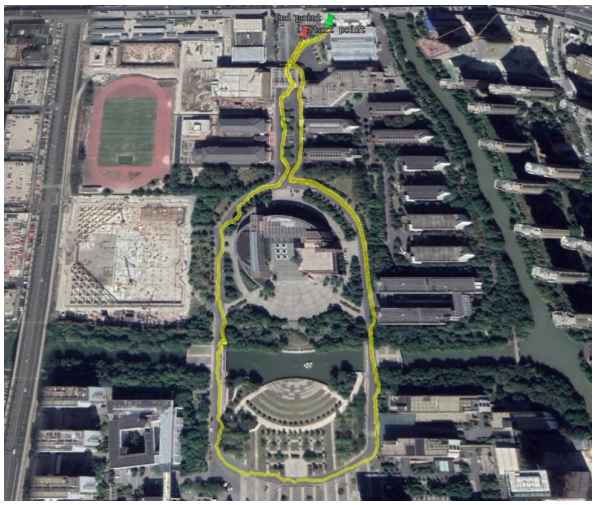
C. Validation on Chengyuan dataset: from coarse to fine

We collect Chengyuan dataset on Chengyuan campus in Gongshu District in Hangzhou, China to facilitate the study of vehicle localizer, where the fully electric instrumented vehicle shown in Figure 6(a) is utilized. The setting of the acquisition program enables the real-time output of the two forms of unfolded images shown as Figure 3(c), and the requirements of coarse stage and fine stage can be met simultaneously. The image acquisition equipment and route are shown in Figure 6. This dataset covers the variations of summer and winter, as well as sunny afternoon and cloudy morning. Among them, subset1 with scenarios of winter sunny afternoon is set as database, subset2 with summer views and subset3 with winter cloudy morning scenarios are set as query respectively.

In coarse stage, NetVLAD descriptors will be obtained by the above method, which will be utilized to determine a rough range. Only the top10 database images selected by the coarse stage need to conduct the following fine stage. In the fine stage, four plane images will be unfolded by the second processing method of Figure 3(c) from one panoramic annular image. The corresponding part of the four plane images between database and query images will be extracted Geodesc descriptors to compute Fundamental Matrix as mapping relationship, respectively. Inliners and outliners will be distinguished by Fundamental Matrix, and the total inliner number of the four parts will help to determine the final top1 result, or in other words, finer result. Figure 7 draws line charts of the matching results. The coarse matching results are drawn as dashed line, while the results of combination of

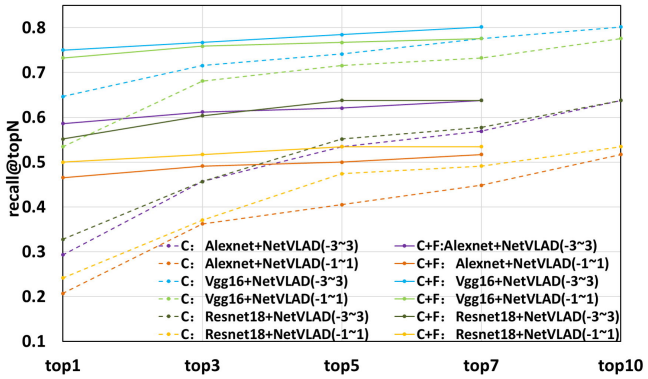


(a)

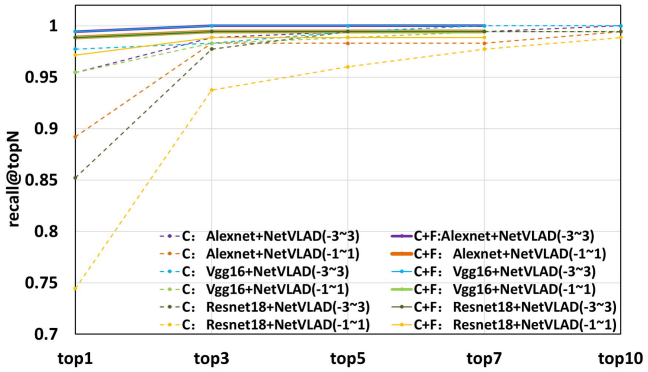


(b)

Fig. 6. (a) Image acquisition equipment; (b) Traveling route (the route is denoted in yellow, and the red point refers to start point, the green point refers to end point.)



(a)



(b)

Fig. 7. Line chart of matching results on Chengyuan dataset. (a) Subset2 as query and subset1 as database; (b) Subset3 as query and subset1 as query. (C means coarse stage, C+F means combination of coarse stage and fine stage.)

coarse matching and fine matching are given in solid lines, results of 1 or 3 images before and after the ground truth setting as accurate localizations are both evaluated. As shown in Figure 7, the challenges brought by season changes seem to be more serious than those from weather changes, for the recall in Figure 7(b) is not as accurate as Figure 7(a). We can also see that after combination of coarse matching and fine matching, the finer results improve dramatically than only through the coarse matching. In this sense, our CFVL obtains a great success verified by both numerical and qualitative results. Figure 8 displays some correct visual localization results on Chengyuan dataset.

V. CONCLUSIONS

In this paper, we propose a conceptually simple coarse-to-fine vehicle localizer: CFVL, which can perceive omnidirectional information of surroundings during vehicle traveling, against severe appearance variations, such as illumination changing, day-night cycling, traverse variations and so on. The system processes from coarse stage to fine stage, coarse stage strengthens the ability of recognizing different appearances, while the fine stage improves the ability to learn detail information from images. CFVL reaches excellent performances on both coarse stage and fine stage. After conducting a series experiments, it is demonstrated that the coarse stage can select a rough range of accurate locations while the fine stage helps provide finer results. Compared to results from others, our system reaches superior accuracy, even only through the coarse stage, by training NetVLAD network with explicit panoramas other than training with locally viewed small images. Fine matching plays an important role in producing finer results.

ACKNOWLEDGMENT

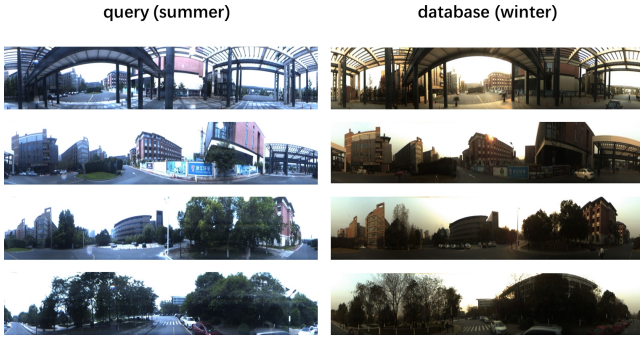
The fully electric instrumented vehicle utilized to capture Chengyuan dataset is provided by Center for Robotics and Intelligent Manufacturing Engineering, College of Engineers, Zhejiang University.

REFERENCES

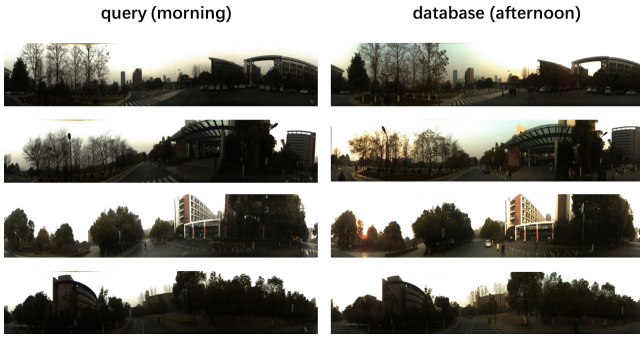
[1] R. Cheng, K. Wang, S. Lin, W. Hu, K. Yang, X. Huang, H. Li, D. Sun, and J. Bai, "Panoramic annular localizer: Tackling the variation challenges of outdoor localization using panoramic annular images and active deep descriptors," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, Oct 2019, pp. 920–925.

[2] Y. Fang, K. Wang, R. Cheng, K. Yang, and J. Bai, "Visual place recognition based on multilevel descriptors for the visually impaired people," in *Target and Background Signatures V*, vol. 11158, International Society for Optics and Photonics. SPIE, 2019, pp. 21 – 32.

[3] S. Lin, K. Wang, R. Cheng, and K. Yang, "Visual localizer: Outdoor localization based on convnet descriptor and global optimization for visually impaired pedestrians." *Sensors (Basel)*, vol. 18, no. 8, pp. 1424–8220, 2018.



(a)



(b)

Fig. 8. The visual localization results. (a) Subset2 as query and subset1 as database; (b) Subset3 as query and subset1 as database. (The left column are the query images, and the right column are the corresponding database images.

[4] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015.

[5] K. Yang, X. Hu, H. Chen, K. Xiang, K. Wang, and R. Stiefelhagen, "Ds-pass: Detail-sensitive panoramic annular semantic segmentation through swaftnet for surrounding sensing," *arXiv preprint arXiv:1909.07721*, 2019.

[6] L. Yu, C. Joly, G. Bresson, and F. Moutarde, "Monocular Urban Localization using Street View," in *14th International Conference on Control, Automation, Robotics and Vision (ICARCV'2016)*, Phuket, Thailand, Nov. 2016. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01425639

[7] M. Meilland, A. I. Comport, and P. Rives, "Dense omnidirectional RGB-D mapping of large scale outdoor environments for real-time localisation and autonomous navigation," *Journal of Field Robotics*, vol. 32, no. 4, pp. 474–503, June 2015.

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

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

[10] Z. Liao, J. Shi, X. Qi, X. Zhang, W. Wang, Y. He, R. Wei, and X. Liu, "Coarse-to-fine visual localization using semantic compact map," *ArXiv*, vol. abs/1910.04936, 2019.

[11] A. Holliday and G. Dudek, "Scale-robust localization using general object landmarks," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2018, pp. 1688–1694.

[12] 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.

[13] Y. Hou, H. Zhang, and S. Zhou, "Convolutional neural network-based image representation for visual loop closure detection," in *2015 IEEE International Conference on Information and Automation*, Aug 2015, pp. 2238–2245.

[14] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *CoRR*, vol. abs/1312.6229, 2013.

[15] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," *ArXiv*, vol. abs/1310.1531, 2013.

[16] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: An astounding baseline for recognition," *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 512–519, 2014.

[17] Y. Hou, H. Zhang, and S. Zhou, "Bocnf: efficient image matching with bag of convnet features for scalable and robust visual place recognition," *Autonomous Robots*, vol. 42, no. 6, pp. 1169–1185, Aug 2018.

[18] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[19] Z. Luo, T. Shen, L. Zhou, S. Zhu, R. Zhang, Y. Yao, T. Fang, and L. Quan, "Geodesc: Learning local descriptors by integrating geometry constraints," in *Computer Vision – ECCV 2018*. Cham: Springer International Publishing, 2018, pp. 170–185.

[20] M. Zhu, W. Wang, B. Zhu, and J. Huang, "A fast image stitching algorithm via multiple-constraint corner matching," 2013.

[21] J. Kannala and S. S. Brandt, "A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1335–1340, Aug 2006.

[22] D. L. Lehner, A. G. Richter, D. R. Matthys, and J. A. Gilbert, "Characterization of the panoramic annular lens," *Experimental Mechanics*, vol. 36, no. 4, pp. 333–338, Dec 1996.

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

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

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

[26] N. Sünderhauf and P. Protzel, "Brief-gist - closing the loop by simple means," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sep. 2011, pp. 1234–1241.

[27] G. Singh, "Visual loop closing using gist descriptors in manhattan world," in *Omnidirectional Robot Vision workshop, held with IEEE ICRA*, 2010.

[28] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *ECCV*, 2006.

[29] 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: Association for Computing Machinery, 2017, p. 392–396.

[30] Z. Zhang and C. Loop, "Estimating the fundamental matrix by transforming image points in projective space," *Computer Vision and Image Understanding*, vol. 82, no. 2, pp. 174 – 180, 2001.

[31] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk, "Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors," in *CVPR*, 2017.

[32] J. H. Friedman, F. Baskett, and L. J. Shustek, "An algorithm for finding nearest neighbors," *IEEE Transactions on Computers*, vol. C-24, no. 10, pp. 1000–1006, Oct 1975.

[33] P.-E. Danielsson, "Euclidean distance mapping," 1980.

[34] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *CVPR*, 2013.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[37] 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.