Visual place recognition based on multi-level descriptors for the visually impaired people

Yicheng Fang, Kaiwei Wang*, Ruiqi Cheng, Kailun Yang and Jian Bai
State Key Laboratory of Modern Optical Instrumentation, Zhejiang University, 38# Zheda Road, Hangzhou 310027, China

ABSTRACT

The Visually Impaired People (VIP) have the difficulty in perceiving the accurate localization in their daily life. Developing an efficient algorithm to address the localization issues of the VIP is crucial. Visual Place Recognition (VPR) refers to using the image retrieval algorithms to determine the location of a query image in the database, which is promising to help the VIP solve their localization problems. However, the accuracy of VPR is directly affected by the changes of scene appearances such as illumination, seasons and viewpoints. Therefore, finding a method to extract robust image descriptors under the changes of scene appearance is one of the most critical tasks in current VPR research. In this paper, we propose a VPR approach to assist the localization and navigation of visually impaired pedestrians. The core of our proposal is a combination of multi-level descriptors by using appropriate descriptors: the whole image, local regions and keypoints, aimed to enhance the robustness of VPR. The matching procedure between query images and database images includes three steps. Firstly, we obtain the Convolutional Neural Networks (CNN) features of the whole images from a pre-trained GoogLeNet, and the Euclidean distances between the query images and the database images are computed to determine the top 10 good matches. Secondly, local salient regions are detected from the top-10 best-matched images with Non-Maximum Suppression (NMS) to control the number of bounding boxes. Thirdly, we detect the SIFT keypoints and extract the geodesc descriptors of the keypoints, from the local salient region, and determine the top 1 among the top 10 good matches. In order to verify our approach, a comprehensive set of experiments has been conducted on dataset with challenging environmental changes, such as the GardensPointWalking dataset.

Keywords: visually impaired people; place recognition; image matching; descriptors

1. INTRODUCTION

Visual Place Recognition (VPR) has always been an unsolved problem in the field of robotics, unmanned system, self-driving, and especially the navigation for the Visually Impaired People (VIP), which aims to cope with significant changes induced by day-night cycles and viewpoint variations[1, 2]. Thus, we try to find one method that can effectively help the VIP locate their position, to come over the challenging changes of appearances and viewpoints, by treating their location as a query and find the best match in the database.

Holistic features based methods like GIST[3-5], extract a single descriptor from the whole image and then score the matching degree between the query images and the database images, to determine the best matches, but come at a cost of sensitivity to viewpoint changes. Local region detection is extracted as the salient region through object proposal methods, which shows impressive performance against the viewpoint changes. Recent researches have shown dramatic performance on how deep-learning feature based approach can be applied in the domain of VPR, as both holistic images and local images can obtain CNN features, from a pre-trained ConvNet[6, 7]. Moreover, keypoints are often a significant component of an effective VPR system, which can be detected from the whole image or local regions. Specifically, some newly proposed deep-learning based methods used for description of keypoints show better performance than conventional keypoint features in image matching, which can be taken into consideration for VPR.

In this paper, we propose a multi-level visual localization algorithm, which combines the keypoints, local region and the whole image with adaptive descriptors. The procedures of multi-level descriptors based VPR are described as follows, where the flow diagram of our approach is shown in Figure 1.

1) We obtain the CNN features of the whole image from a pre-trained GoogLeNet, where the inception_3a/3x3_reduce layer and the inception_3a/3x3 layer are selected to provide the features. A single descriptor is extracted from one image, and the Euclidean distance between the query images and the database images are computed to determine the top 10 good matches.

2) Local salient regions are detected from the top-10 best-matched images by using the method BING[8], which can discard the relatively unimportant areas for VPR like sky and ground, in order to reduce the computation complexity. Besides, the coefficient of Non-Maximum Suppression (NMS)[9] decides the size and number of the detected areas.

3) We detect the SIFT keypoints and extract the geodesc[10] descriptors of the keypoints, from the aforementioned local salient regions. After that, we compute the Fundamental Matrix or Homography Matrix between the query local salient regions and database local salient regions, to distinguish the outliers and inliers, where the number of inliers determines which is the top 1 among the top 10 good matches.

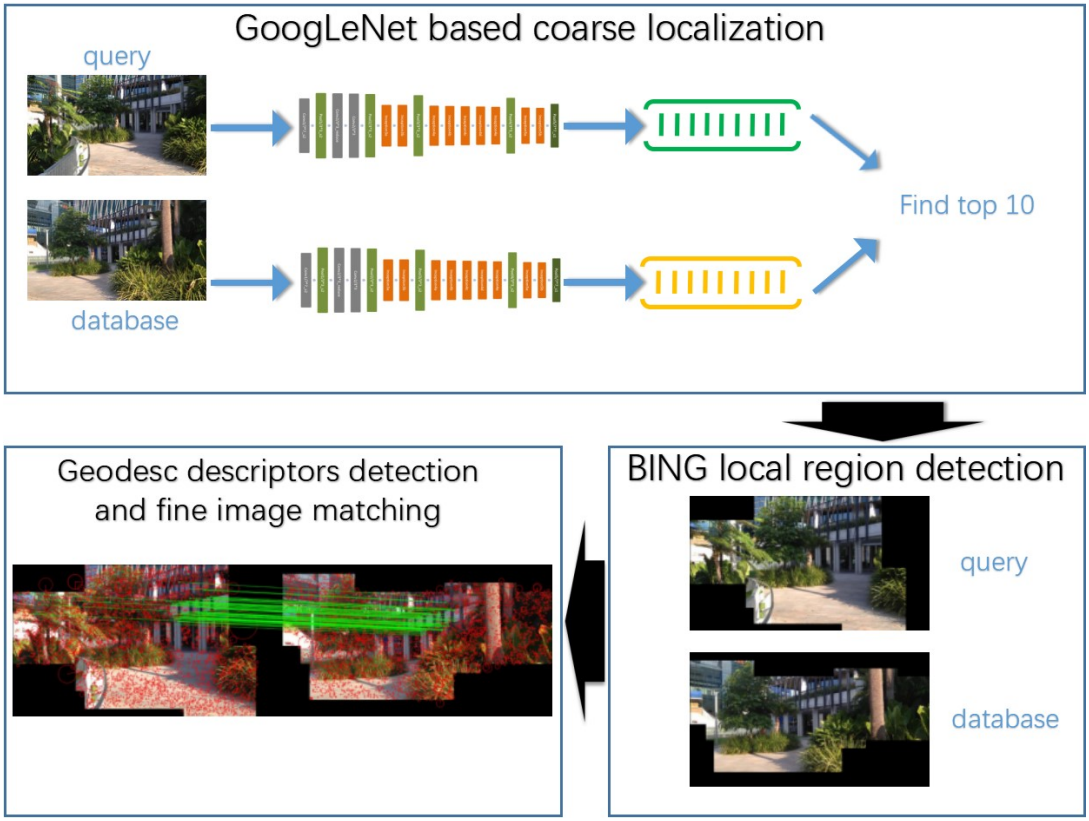


Figure 1. The procedures of multi-level descriptors based VPR

To test our approach, we carry out a series of experiments on GardensWalkingPoint dataset[11], and mainly pay attention to the results on conditions with viewpoints change. We evaluate the matching results by computing the matching precision. Our multi-level descriptors based method are proved to have a superior performance on the walking navigation dataset suitable for the VIP.

The paper proceeds as follows. Section 1 does a brief introduction of our proposed multi-level descriptors based approach and corresponding experiments and results. Section 2 provides an overview of some related work. Our method will be described in detail in Section 3. A series of experiments will be conducted in Section 4. Section 5 will give the conclusion.

2. RALATED WORK

The VIP have the difficulty in perceiving the accurate localization in their daily life. Efficient algorithms to address the localization issues of the VIP are required. Extensive approaches have been presented in the field of VPR. For all the approaches in VPR, there are several dominating ideas: holistic features based methods, local features or local region based methods, CNN descriptors based methods. In changing environments, adaptive features are utilized, where sometimes local features can present the holistic image while sometimes an image is represented with its global statistics.

SeqSLAM[12] is a holistic method based on sequence matching, it performs remarkably well under severe appearance changes, like season variations, day-night cycles, and weather changes, but it shows sensitivity to viewpoint changes. Another holistic algorithm named GIST[3], which is based on handcrafted global descriptors, uses the responses of an image to a Gabor filter bank. GIST is also found not robust to viewpoints change, where VPR faces severe challenges.

In terms of local feature based algorithms, SIFT[13], SURF[14] detectors have demonstrated a significant degree of viewpoint invariance, but only show a limited degree of appearance invariance. FAB-MAP[15] is the first work that combines SURF features with BoW[6] encoding methodology for VPR. Due to the invariance properties of SURF, FAB-MAP can get over challenging viewpoint changes. However, it fails in conditions with severe appearance changes, and it does not generalize well to new environments without learning a new site-specific vocabulary[16]. Region based methods within an image can achieve some of the advantages of both keypoints and the whole image, to address both viewpoint and environmental changes by site-specific training[17]. Recent research have also demonstrated how generic deep learning-based features trained for object recognition can be successfully applied in the domain of place recognition[18, 19] . Sünderhauf, N. et al. have presented a novel place recognition system that builds on object detection methods (such as Edge Boxes[20] and BING[8]) and convolutional visual features, by utilizing a generic ConvNet pre-trained on a large image classification dataset [21].

Recent advances in deep-learning or CNN based method in VPR have demonstrated that this method can be a potential solution to overcoming the weaknesses of handcrafted features. It has been shown that CNN features, when used as a global descriptor for an entire image, have a better discrimination capacity than state-of-the-art handcrafted global features[10, 22-26]. The availability of pre-trained network models makes it easy to experiment with such approaches for different tasks in VPR. Shufei, L., Ruiqi, C. et al. evaluated different layers 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 against other ConvNets because of its appearance invariance and viewpoint invariance as well as the modest computational complexity[7].

In the following section, we will investigate how our multi-level descriptors based method works, to combine three levels of descriptors: the whole image, local regions and keypoints, in order to enhance the robustness to viewpoint and environment change. After that, our experiment validates the effectiveness of the proposal by observing how it performs in the pedestrian dataset suitable for the VIP.

3. METHODOLOGY

3.1 System overview

Figure 2 shows the assistant device Intoer which integrates a mobile computing platform, a bone-conduction headphone and a RGB-Depth sensor. The sensor detects 3D scenes and obtains real-time RGB-D images, while the portable computer processes the images to predict traversable areas[27, 28], then the bone-conduction headphone transforms the image information into sounds to transmit it to the VIP. The detected scenes by the sensor can be seen as query images with viewpoint and appearance changes. Aiming to assist them to perceive the accurate localization, a series of query images will be matched to a series of pre-recorded database images in order to find the best matches. In this way, the best matched query image and database image are regarded as the same position.

Our algorithm conducts the place recognition from coarse to fine, and extracts multi-level features and descriptors. GoogLeNet extracts holistic CNN features from the whole images, to determine the best-matched top 10 images corresponding to the one query image from database, which can be seen as the coarse stage. While in the fine stage, the BING method is utilized to detect local regions and the deep-learning based geodesc features will be extracted from the local region, to achieve more sophisticated localization. The CNN based method, BING, and geodesc descriptors will be introduced as follows.

Figure 2. Assistant device Intoer for the VIP.

3.2 GoogLeNet based coarse localization

Our previous work has demonstrated that GoogLeNet is found to feature the highest robustness against environmental changes, by comprehensively comparing the performance of five prevailing ConvNets (AlexNet, VGGNets, GoogLeNet, SqueezeNet and MobileNet)[7]. By concatenating two compressed convolutional layers of GoogLeNet, an image can be represented efficiently with only thousands of bytes. We take the output of two layers including inception_3a/3*3_reduce and inception_3a/3*3 from GoogLeNet to represent the input images. A query image and a database image can respectively acquire a tensor with same dimension. The architecture of GoogLeNet and the Inception module is shown in Figure 3 and Figure 4. The layers we utilized are marked below.

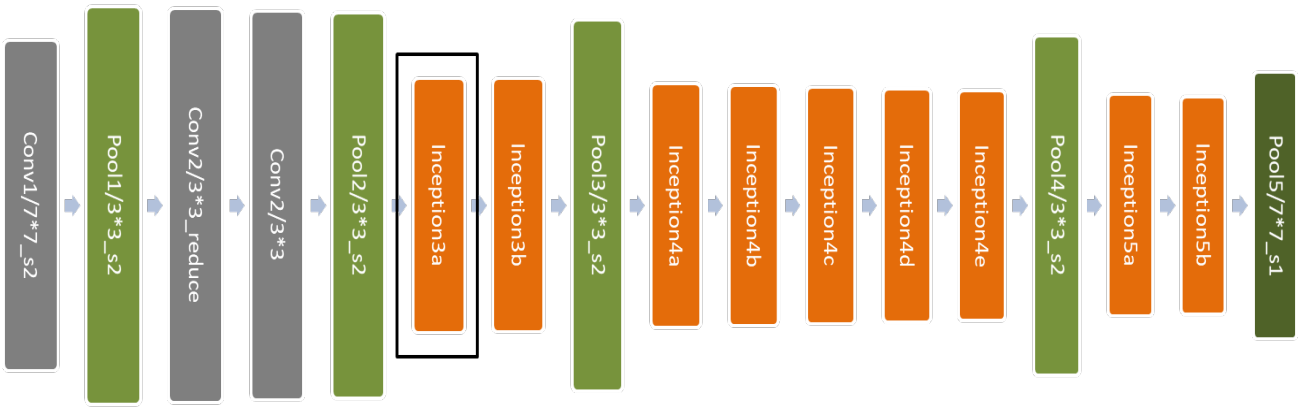


Figure 3. The architecture of GoogLeNet.

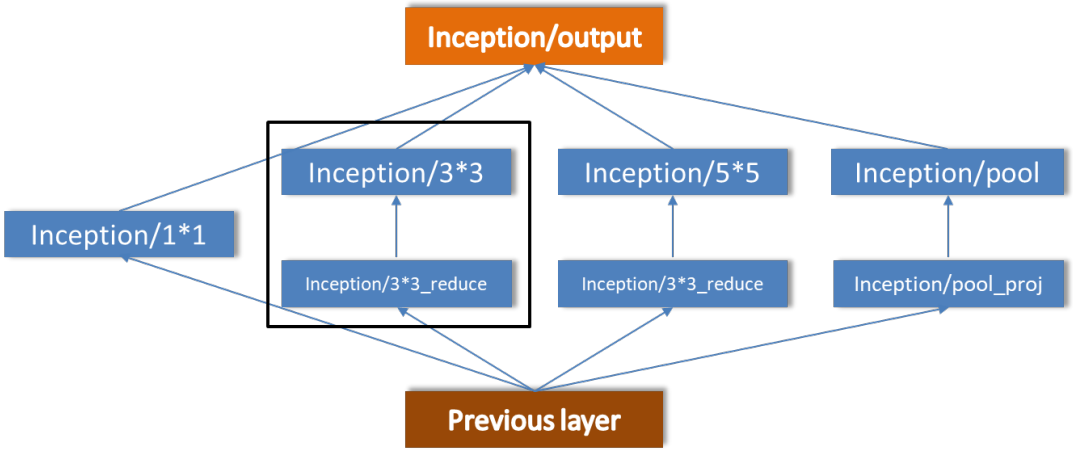


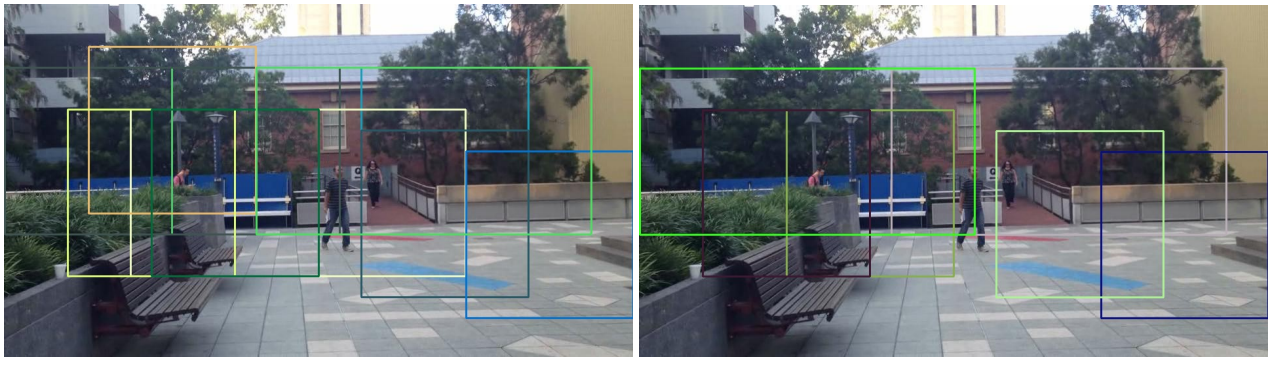
Figure 4. The architecture of Inception module.

3.3 Object detection: BING

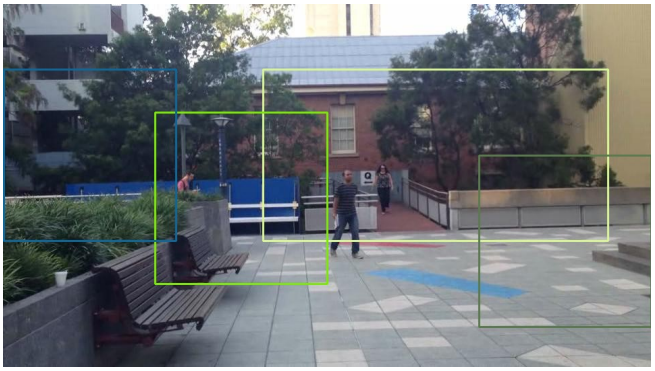
Researches from cognitive psychology and neuro-biology suggest that humans have a strong ability to perceive objects before identifying them. Based on the human reaction time that is observed and the biological signal transmission time that is estimated, human attention theories hypothesize that the human vision system processes only parts of an image in detail, while leaving others nearly unprocessed. This further suggests that before identifying objects, there are simple mechanisms in the human vision system to select possible object locations. In the field of VPR, in order to reduce the useless scenes in the environments, and just leave the significant and salient areas that need to be considered, training a scene measure which is generic across diverse environments is necessary.

A binarized normed gradients for objectness estimation method called BING can help identify salient areas using objectness scores, which reflects how likely an image window covers an object. As a salient object detection model, BING tries to detect the most eye-catching scene, and then segments the detected scene from the whole image. For simple images, such salient object segmentation can be applied in image scene analysis. Additionally, it can be used as a tool to process large number of images benefiting from its high efficiency. The method BING generates a series of landmarks that are both stable and repeatable against severe viewpoint and appearance changes, which is very suitable for VPR.

As shown in Figure 5(a), we apply the BING method on one image from the GardensPointWalking dataset and see how likely it can detect the salient scenes. It turns out that the redundant parts of the image, like sky and ground that are little-contributing for VPR, are set aside and the significant scenes are grabbed by the bounding boxes. However, too many overlapped areas would result in unnecessary calculation and influence the computational efficiency. Thus, the Non-Maximum Suppression (NMS) method is used to suppress the redundant overlapped areas after BING detection. The goal of the NMS method is to keep the optimal boxes, that is, to suppress the bound boxes whose overlapped part is not a maximum, and to search for the local maximum. There will be a suppressing threshold. When we adjust the threshold to 0.35, the detected result is shown in Figure 5(b), while Figure 5(c) shows the result after adjusting the threshold to 0.2.



(a) (b)

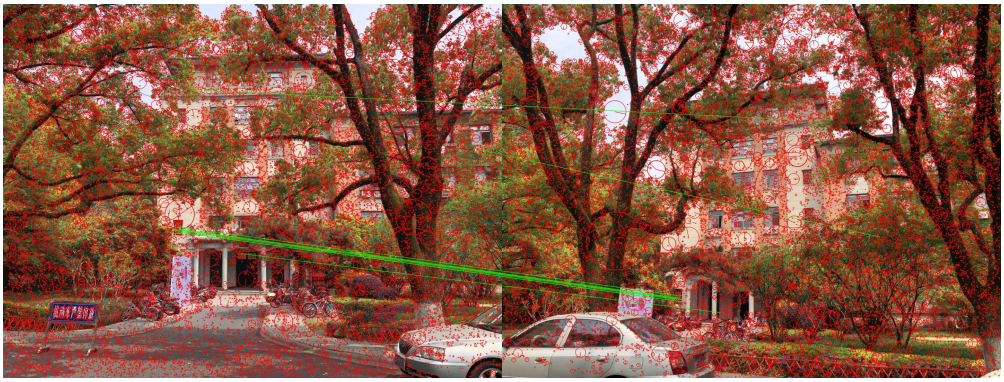


(c)

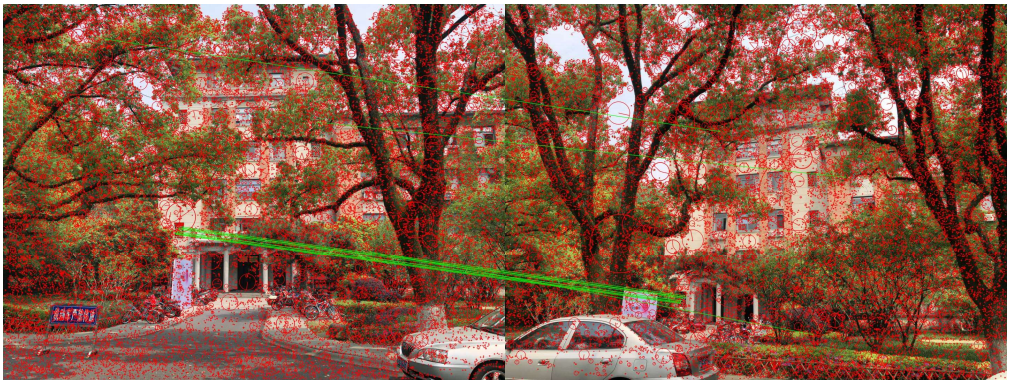
Figure 5. a) BING detection result; b) Non-Maximum Suppression (NMS) for reducing overlapped areas (the threshold is 0.35); c) Adjust NMS threshold to 0.2

3.4 Geodesc descriptors

Some kinds of traditional handcrafted keypoint descriptors like SIFT, SURF and so on have already been used in VPR[29], while deep-learning based keypoint descriptors, which are more adaptive to image content, are rarely used. Geodesc[10] is a state-of-the-art deep-learning based method to describe the keypoints from the images. This approach offered a novel batch constructed method that simulates the pixel-wise matching and effectively samples useful data for the learning process. It used the L2-Net directly and test on the Hpatch[30] dataset, whose images all have same 32*32 size and can be approximatively regarded as keypoints.



(a)



(b)

Figure 6. (a) Matching result of SIFT based method on a pair of pictures with viewpoint changes taken in front of the No.3 Building, Zhejiang university; (b) Matching result of geodesc based method on the same pair of pictures.

The procedures are as follows: Firstly, the SIFT keypoints rather than the SIFT descriptors are detected from the images; Secondly, the keypoint areas are cropped into patches of 32*32 to facilitate geodesc features extraction; Thirdly, we extract geodesc feathers from the patches. If two images are prepared to be matched, the geodesc descriptors will help to identify the similar parts from the images and select good matches between the two images. An example image pair, which is taken with viewpoint changes, is shown in Figure 6 by comparing the matching results between SIFT and geodesc descriptors.

As we can see from the Figure 6, it turns out that, compared to the SIFT approach, geodesc descriptor-based method can find more good matching pairs with few mismatches. Geodesc based approach outperforms the SIFT in the viewpoint-changing conditions and shows impressive results on image matching. Thus, we integrate it in our VPR system. By extracting geodesc descriptors from the images after the BING local region detection process, we only detect the keypoints and geodesc descriptors from the local areas. For the uselessness of the redundant areas, the operation of blacking out these areas is needed. Figure 7 shows the operation of combining BING local region approach with geodesc method in image matching on GardensWalkingpoint dataset. This stage can be regarded as fine matching after GoogLeNet-based coarse localization.

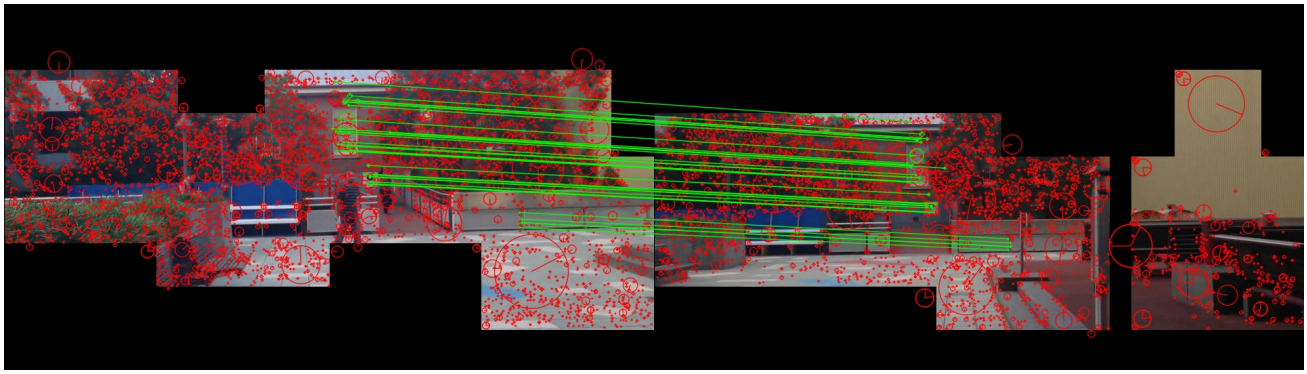
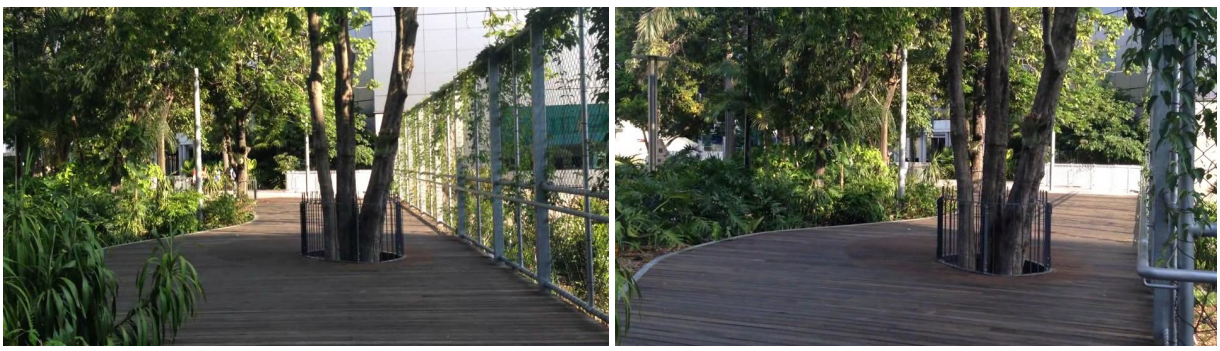


Figure 7. The operation of combining BING local region approach with geodesc method in image matching on GardensWalkingpoint dataset.

4. EXPERIMENTS

4.1 GardensWalkingPoint dataset

We measure the performance of our VPR system on GardensWalkingPoint dataset[11], a vision dataset of a single route through the Gardens Point Campus, Queensland University of Technology, Brisbane, Australia. The visual data was collected with a forward-facing, hand-held iPhone 5. The route was traversed three times, twice during the day and once at night, where each part includes 200 pictures. Those pictures with the same sequence number are located at the same position. One of the day route is traversed on the left-hand side of the path and the other day route and the night route are traversed on the right-hand side of the path, to capture both pose and condition change.



(a) (b)

Figure 8. a) Image087 from the day-left subset; b) Image087 from the day-right subset of GardensWalkingPoint dataset.

We only utilize the day-right part and day-left part to test our system in viewpoint-changing conditions. Figure 8 shows an arbitrarily selected pair of images with the same ordinal number of the dataset, and the obvious change of perspective can be seen. We use day-left as the query image and day-right as the database image.

4.2 GoogLeNet based VPR

After extracting the features from the two layers of GoogLeNet, matching procedure should be considered. Euclidean distance[31] between the two tensors will be computed as follows:

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

where $x=(x_1,x_2,\cdots,x_n)$ represents the feature extracted from query image; $y=(y_1,y_2,\cdots,y_n)$ represents the feature extracted from database image; and $d(x,y)$ represents the Euclidean distance between the query image and the database image.

It is obvious that the lower Euclidean distance is, the more similar the two images are. Thus, for a query image, we can determine the top 10 most similar ones, the image with the least Euclidean distance can regarded as the best match, or namely top 1. When test the approach on GardensWalkingPoint dataset, the day-left images are regarded as the query images, while the day-right images are taken as the database images. We draw the corresponding matching figure between query images and best top 1 matched database image, as shown in Figure 9. In Figure 9, the red points show the corresponding matching relationship between query images and database images. If the tendency of red points is close to a diagonal line, the images are well-matched. However, as shown in figure 9, there are some existing mismatches, indicating that the top 1 selected by this method has obvious errors. Therefore, this method is not suitable for fine matching, but only appropriate to select a rough range for VPR.

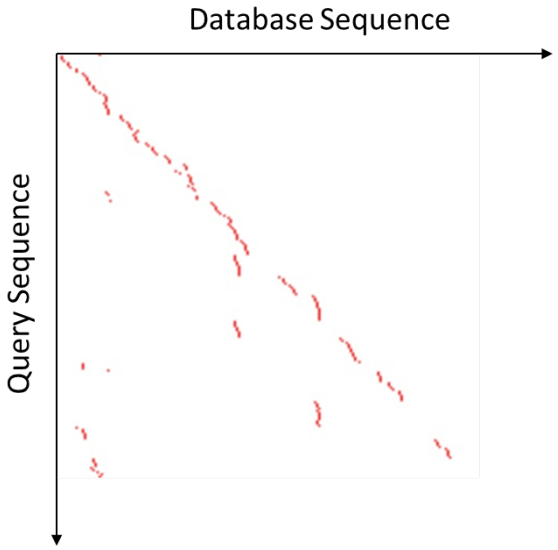


Figure 9. The corresponding matrix between query images and the best matched images.

For a query image, we regard the database image which has the same sequence number with the query image as ground truth, and 5 images before and 5 images after the ground truth can all be seen as correct matching results. At the same time, the range of 3 images before and after the ground truth is also considered. The precision of top 1 and top 10 can be also respectively computed. An image is considered as a correct match even if only one image of the best ten appears in the ground truth range. When the range is -5~5, the top 1 precision is 0.59, while the top 10 precision reaches 0.91. It is further confirmed that GoogLeNet based VPR is only suitable to filter for a rough range, but not suitable for fine matching. The detailed matching results of GoogLeNet based VPR precision is shown in Table 1.

Table 1. The detail matching results of GoogLeNet based VPR precision

precision	Top 1	Top 3	Top 5	Top 7	Top 10
GG(-3~3)	0.535	0.700	0.765	0.800	0.870
GG(-5~5)	0.590	0.735	0.820	0.850	0.910

4.3 Fundamental Matrix vs Homography Matrix

GoogLeNet based coarse localization helps us find the top 10 database images for a query image, and then it comes to the stage to find finer matches utilizing our multi-level descriptors based method. After detecting local regions and extract geodesc descriptors from local region, we take Fundamental Matrix and Homography Matrix into consideration respectively to find a mapping relation between keypoints from query images and keypoints from database images, to distinguish inliers and outliers, and to conduct further matches.

For Fundamental Matrix[32], if given a pair of images which are taken from different viewpoints, a random point on the first image has a counterpart to an epipolar line on the second image, and the epipolar line which passes through the center ray of the first image is the projection of the point. We can see the mapping of two images from point to epopolar line follows Fundamental Matrix. Fundamental Matrix can be represented as follows:

$$q_1^T F q_2 = 0, \tag{2}$$

where q1, q2 represent pixel coordinates of two images; F represents the Fundamental Matrix.

For Homography Matrix[33], the Homogeneous coordinate of a point-pair from two images with different perspectives can be expressed using a projective transformation. It is suitable for the case where two images are on the same plane. Usually, two images can be hypothesized on the same plane when they are far apart. Homography Matrix can be computed as follows:

$$q_1^T H q_2 = 0, \tag{3}$$

where q1, q2 represent pixel coordinates of two images; H represents the Homography Matrix.

The two matrixes can both help us find out the corresponding relationship between a query image and a database image with different principles. Keypoints following the corresponding principle are considered as inliers, otherwise it is regarded as outliers. Then RANSAC algorithm helps us eliminate mismatches. Obviously, the more inliers are, the more similar between the query image and database image are. In this way, we can find out the finer matches from the top 10 after GoogLeNet based coarse localization. The detailed matching results of Fundamental Matrix based VPR precision is shown in Table 2, and the detailed matching results of Homography Matrix based VPR precision is shown in Table 3. Range of -5~5 and range of -3~3 for the ground truth are also considered.

Table 2. The detail matching results of Fundamental Matrix based VPR precision

precision	Top 1	Top 3	Top 5	Top 7
F(-3~3)	0.815	0.870	0.870	0.870
F(-5~5)	0.895	0.905	0.905	0.910

Table 3. The detail matching results of Homography Matrix based VPR precision

precision	Top 1	Top 3	Top 5	Top 7
H(-3~3)	0.790	0.860	0.865	0.870
H(-5~5)	0.885	0.900	0.910	0.910

Additionally, we draw the precision broken line graph of the matching results of GoogLeNet based localization, Fundamental Matrix based method, and Homography Matrix based method as shown in Figure 10.

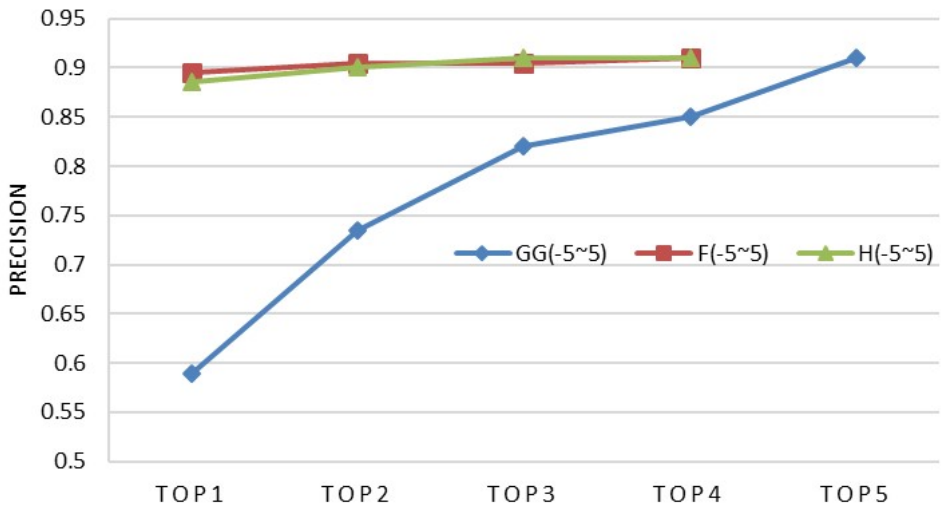


Figure 10. The precision broken line graph of matching results of GoogLeNet based localization (GG), Fundamental Matrix based method (F), and Homography Matrix based method (H).

By comparing Table1, Table2, Table 3 and considering Figure 10, we can find out that our multi-level descriptors based VPR gets a superior performance than GoogLeNet based localization. The precision of the top 1 is significantly improved compared to GoogLeNet based localization, which demonstrates that our method can recognize positions accurately even with difference viewpoints in the dataset suitable for blind pedestrian navigation. Moreover, Fundamental Matrix based method is slightly more accurate than the Homography Matrix based approach in our VPR framework because it is more reasonable and general to use Fundamental Matrix. We suppose the keypoints from two images are on the same plane on the Homography Matrix based condition, which actually are not. Therefore, Homography Matrix based approach is not as precise as Fundamental Matrix based method, but still can work and gets a decent performance, which further indicates the robustness of the proposed algorithm. In conclusion, our multi-level descriptors based VPR method is proved to be effective, and robust to viewpoint changes.

5. CONCLUSION

In this paper, a multi-level descriptors based VPR approach is proposed, to help the VIP perceive the accurate localization in the changing environments, especially in the severe viewpoint-changing conditions. The proposed approach involves three levels of descriptors: the whole image, local region and keypoints. In the whole image level, or namely the coarse localization stage, top 10 best matches are selected based on GoogLeNet features. In the local region level, local regions are extracted from the whole images by using the image detection approach BING. In the keypoints level, keypoints and geodesc features are extracted only from the local region. We compared the GoogLeNet based localization method, and our multi-level descriptors based method, to validate our method achieves good precision of VPR. Furthermore, the Fundamental Matrix performs slightly better than Homography Matrix in the last matching procedure. In the future, we aim to integrate the approach in the wearable devices of walking navigation for the VIP and conduct user studies.

REFERENCES

[1] Cheng, R., Wang, K., Lin, L. and Yang, K., "Visual Localization of Key Positions for Visually Impaired People," 2893-2898 (2018).

[2] Cheng, R., Wang, K., Lin, S., Hu, W. and Bai, J., "Panoramic Annular Localizer: Tackling the Variation Challenges of Outdoor Localization Using Panoramic Annular Images and Active Deep Descriptors," arXiv preprint arXiv:1905.05425 (2019).
[3] Niko Sünderhauf, P. P., "BRIEF-Gist – Closing the Loop by Simple Means," IEEE/RSJ International Conference on Intelligent Robots & Systems (2011).
[4] Oliva, A. and Torralba, A., "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope," International Journal of Computer Vision 42 3, 145-175 (2001).
[5] Singh, G. and Kosecka, J., "Visual loop closing using gist descriptors in manhattan world," ICRA Omnidirectional Vision Workshop (2010).
[6] Hou, Y., Zhang, H. and Zhou, S. J. a. R., "BoCNF: efficient image matching with Bag of ConvNet features for scalable and robust visual place recognition," 42 6, 1169-1185 (2018).
[7] Lin, S., Cheng, R., Wang, K. and Yang, K., "Visual localizer: Outdoor localization based on convnet descriptor and global optimization for visually impaired pedestrians," Sensors 18 8, 2476 (2018).
[8] Cheng, M. M., Zhang, Z., Lin, W. Y. and Torr, P., "BING: Binarized Normed Gradients for Objectness Estimation at 300fps," Computer Vision & Pattern Recognition (2014).
[9] Neubeck, A. and Gool, L. J. V., "Efficient Non-Maximum Suppression," International Conference on Pattern Recognition (2006).
[10] Luo, Z., Shen, T., Zhou, L., Zhu, S., Zhang, R., Yao, Y., Fang, T. and Quan, L., "GeoDesc: Learning Local Descriptors by Integrating Geometry Constraints," in Computer Vision – ECCV 2018, 170-185 (2018).
[11] Sünderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B., Milford, M., "On the performance of Convnet features for place recognition," in Proceedings of the IEEE International Conference on Intelligent Robots and Systems, 4297-4304 (2015).
[12] Milford, M. J. and Wyeth, G. F., "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," 2012 IEEE International Conference on Robotics and Automation, 1643-1649 (2012).
[13] Lowe, D. G., "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision 60 2, 91-110 (2004).
[14] Bay, H., Tuytelaars, T. and Gool, L. V., "SURF: Speeded up robust features," European Conference on Computer Vision (2006).
[15] Cummins, M. and Newman, P., "FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance," The International Journal of Robotics Research 27 6, 647-665 (2008).
[16] Espinace, P., Soto, A., Kollar, T. F. and Roy, N., "Indoor Scene Recognition Through Object Detection," IEEE International Conference on Robotics & Automation (2010).
[17] Mcmanus, C., Upcroft, B. and Newmann, P., "Scene signatures : localised and point-less features for localisation," in Robotics: Science and Systems X, University of California, Berkeley, CA, Journal Issue (2014).
[18] Sünderhauf, N., Dayoub, F., Shirazi, S., Upcroft, B. and Milford, M., "On the performance of ConvNet features for place recognition," IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 4297-4304 (2015).
[19] Chen, Z., Lam, O., Jacobson, A. and Milford, M., "Convolutional neural network-based place recognition," arXiv preprint arXiv:1411.1509 (2014).
[20] Zitnick, C. L. and Dollar, P., "Edge Boxes: Locating Object Proposals from Edges," ECCV (2014).
[21] Sünderhauf, N., Shirazi, S., Jacobson, A., Dayoub, F., Pepperell, E., Upcroft, B. and Milford, M., "Place Recognition with ConvNet Landmarks: Viewpoint-Robust, Condition-Robust, Training-Free," Robotics: Science and Systems (2015).
[22] Hou, Y., Zhang, H. and Zhou, S., "Convolutional Neural Network-Based Image Representation for Visual Loop Closure Detection," IEEE International Conference on Information & Automation (2015).
[23] Krizhevsky, A., Sutskever, I. and Hinton, G., "ImageNet Classification with Deep Convolutional Neural Networks," International Conference on Neural Information Processing Systems (2012).
[24] Sermanet, P., Eigen, D., Zhang, X., Mathieu, M. L., Fergus, R. and Lecun, Y., "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks," CoRR abs/1312.6229 (2013).
[25] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E. and Darrell, T., "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition," arXiv preprint arXiv:1310.1531 (2013).
[26] Razavian, A. S., Azizpour, H., Sullivan, J. and Carlsson, S., "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition," IEEE Conference on Computer Vision and Pattern Recognition Workshops 512-519 (2014).

[27] Yang, K., Wang, K., Hu, W. and Bai, J., "Expanding the Detection of Traversable Area with RealSense for the Visually Impaired," Sensors 16 11, 1954 (2016).
[28] Yang, K., Wang, K., Bergada, L.M., Romera, E., Hu, W., Sun, D., Sun, J., Cheng, R., Chen, T. and López, E., "Unifying Terrain Awareness for the Visually Impaired through Real-Time Semantic Segmentation," Sensors 18 5, 1506 (2018).
[29] Kameda, Y. and Ohta, Y., "Image Retrieval of First-Person Vision for Pedestrian Navigation in Urban Area," International Conference on Pattern Recognition (2010).
[30] Balntas, V., Lenc, K., Vedaldi, A. and Mikolajcz3, K., "HPatches: A benchmark and evaluation of handcrafted and learned local descriptors," Computer Vision & Pattern Recognition (2017).
[31] Danielsson, P. E. J. C. G. and Processing, I., "Euclidean distance mapping," Computer Graphics and Image Processing 14 3, 227-248 (1980).
[32] Zhengyou, Z., Charles, L., " Estimating the Fundamental Matrix by Transforming Image Points in Projective Space," Computer Vision and Image Understanding 82 2,174-180 (2001).
[33] Detone, D., Malisiewicz, T. and Rabinovich, A., "Deep Image Homography Estimation," arXir preprint arXir:1606.03798 (2016).