Hierarchical Visual Localization for Visually Impaired People Using Multimodal Images

Ruiqi Cheng^a, Weijian Hu^a, Hao Chen^a, Yicheng Fang^a, Kaiwei Wang^{a,*}, Zhijie Xu^b, Jian Bai^a

^aNational Engineering Research Center of Optical Instrumentation, Zhejiang University, Hangzhou, China.

^bSchool of Computing and Engineering, University of Huddersfield, Huddersfield, United Kingdom.

Abstract

Localization is one of the crucial issues in assistive technology for visually impaired people. In this paper, we propose a novel hierarchical visual localization pipeline based on the wearable assistive navigation device for visually impaired people. The proposed pipeline involves the deep descriptor network, 2D-3D geometric verification and online sequence matching. Images in different modalities (RGB, Infrared and Depth) are fed into Dual Desc network to generate robust attentive global descriptors and local features. The global descriptors are leveraged to retrieve the coarse candidates of query images. The 2D local features, as well as 3D sparse point cloud, are used in geometric verification to select the optimal results from the retrieved candidates. Finally, sequence matching robustifies the localization results by synthesizing the verified results of successive frames. The proposed unified descriptor network Dual Desc surpasses the state-of-the-art NetVLAD and its variant on the task of image description. Validated on the real-world dataset captured by the wearable assistive device, the proposed visual localization utilizes multimodal images to overcome the disadvantages of RGB images and robustifies the localization performance by deep descriptor network and hierarchical pipeline. In the challenging scenarios of the Yuquan dataset, the proposed method achieves the F_1 score of 0.77 and the mean localization error of 2.75, which is satisfactory in practical use.

Keywords:

Deep Image Descriptor, Place Recognition, Geometric Validation, Sequence Matching, RGB-D-IR Image

1. Introduction

It is estimated that around 253 million people are suffering from impaired vision all over the world (Bourne et al., 2017). Due to the lack of visual sensory information, the visually impaired people are frequently vulnerable to various hazards in their daily life. Specifically, traveling in unfamiliar outdoor environments has been one of the sore points for many visually impaired people. Fortunately, with the development of intelligent systems, the electronic traveling assistance for the visually impaired people aroused more and more attention from the research community. The burgeoning assistive technology still needs further development in multiple fields, among which localization is one of the most crucial demands of visually impaired people. As a prevailing approach to locating, the positioning technology based on GNSS (global navigation satellite system) has become more and more precise. However, the localization precision of low-cost portable GNSS devices is insufficient for the demands of visually impaired people, in view that the localization error is up to decades of meters under some severe space weather conditions. Analog to the fact that ordinary people tend to recognize places by visual cues, visual images could be utilized for localization in assistive navigation.

Leveraging images to localize is known as *visual localization* or *place recognition*, which is to retrieve the corresponding image (or estimate the camera pose) of a given query image from a priori database (Lowry et al., 2016). Generally speaking, visual localization confronts two challenges in practical use: the visual variations of the same place and the aliasing of different places. Visual variations refer to the different sorts of visual changes between query images and the corresponding database images, including illuminance changes, viewpoint changes, and dynamic object changes, etc. The place aliasing refers to two distinct places sharing similar visual appearance, which results in the failure of place recognition. The research communities have proposed various solutions (Piasco et al., 2018) to resolve those two issues, and those solutions have been applied in many scenarios including augmented reality and autonomous vehicles.

The visual variations in the scenario of assistive navigation are rather severe, because the image sensors mounted on wearable assistive devices are prone to more motion blur, more viewpoint changes, and more dynamic object changes. Therefore, the visual localization system for visual assistance should be more robust against different kinds of appearance variations, and we have previously studied several visual localization algorithms by following this line. Based on multimodal images, i.e. RGB-Depth-Infrared (RGB-D-IR) images, we utilized multiple training-free descriptors in image retrieval, so as to achieve visual localization at the separate positions designated by the user (Cheng et al., 2018). After that, we in-

*Corresponding author: wangkaiwei@zju.edu.cn

Email addresses: rickycheng@zju.edu.cn (Ruiqi Cheng), huweijian@zju.edu.cn (Weijian Hu), haoc@zju.edu.cn (Hao Chen), fangyicheng@zju.edu.cn (Yicheng Fang), z.xu@hud.ac.uk (Zhijie Xu), bai@zju.edu.cn (Jian Bai)

Figure 1: The schematics of the proposed assistive localization pipeline of multimodal visual localization.

vestigated the place recognition performance of the image descriptors extracted from different deep convolutional networks and leveraged the network flow method to optimize retrieval-based localization results (Lin et al., 2018). Assembling both algorithms, we proposed a real-time multimodal place recognition algorithm OpenMPR (Cheng et al., 2019), which fused multiple descriptors with optimized weights in the stage of sequence matching. The experimental results on wearable devices validated that OpenMPR achieved acceptable localization performance in the scenario of assistive navigation. Our preliminary work illustrates that both the multimodal images and deep network descriptors benefit the task of visual localization on the problems of visual variations. Nevertheless, the multiple descriptors enlarge the system complexity and extend the computational latency, thus more powerful and compact image descriptors are desired for place recognition. Moreover, local image features are not extracted in those algorithms, and depth images are treated like vanilla gray-scale images during descriptor extraction. That is to say, the geometric similarity between images can not be fully utilized for place recognition in the preceding research.

To overcome those shortcomings, we propose a novel visual localization pipeline, whose schematics is shown in Figure 1. In the pipeline, the unified deep network *Dual Desc* extracts both local and global descriptors from multimodal images simultaneously. Specifically, the attentive global descriptors are extracted from RGB (color) images, and the 2D local descriptors are extracted from near-infrared images. Moreover, the 3D local descriptors (sparse point cloud) are extracted from depth images. In the offline phase, those local and global descriptors are extracted from a priori database, which prepares for online image matching. In the online phase, the global descriptors extracted from query images are utilized to retrieve the top- K nearest database candidates. The trained attentive global descriptors derived from the deep network features a robust description capability to recognize the same-place images with visual variations. The geometric verification based on in-

frared local features and depth point cloud select the verified database images from the retrieved candidates. Herein, 2D-3D joint geometric verification mitigates the place aliasing of the localization results by leveraging the detailed scene geometry. Subsequently, the result of visual localization is obtained by online sequence matching, which robustifies the performance using the localization continuity of successive frames. In summary, the contributions of this paper lie in the twofold points.

- The unified descriptor network *Dual Desc* extracts both attentive global and 2D-3D local descriptors from RGB-D-IR images, which leverages the different advantages of the image modalities to generate the robust image description for visual localization with visual variations.
- The proposed hierarchical visual localization pipeline involving image retrieval, geometric verification, and sequence matching makes full use of scene semantics and scene geometry of images and is validated to be superior in the scenario of assistive localization.

The remaining parts of the paper are as follows. Section 2 reviews the state-of-the-art assistive localization systems and visual localization algorithms. In section 3, the pipeline of assistive localization involving the stages of image retrieval, geometric verification, and sequence matching are presented. In section 4, we demonstrate the proposed descriptor network *Dual Desc* and its implementation details. Subsequently, we present the experiment results of assistive localization in section 6. Finally, the conclusions are addressed in section 7.

2. Related Work

In this section, assistive localization systems and visual localization algorithms are reviewed.

2.1. Assistive Localization

Localization is an important component of electronic assistance devices, so the research community has been studying

various localization systems. Gomes et al. (2018) proposed a hybrid indoor positioning system using various data sources, including WiFi, Bluetooth beacons, and inertial sensors. Nevertheless, the solution is not possible to be deployed in large-scale outdoor environments. Brilhault et al. (2011) proposed an assistive positioning fusing GNSS, GIS (geographical information system), and the vision-based landmark detection algorithm. However, the landmark detection relies on the tagged landmarks of GIS to obtain the localization results, which requires a large amount of manual annotation and limits the application range of the assistive system.

The research community has proposed various assistive localization solutions, but most of the assistive localization work (Al-Khalifa and Al-Razgan, 2016; Kim et al., 2016; Gleason et al., 2018) focused on the indoor scenarios, and those methods require the building floor plan or the beacons installed in the building. The assistive localization in outdoor environments, which is more challenging meanwhile not trivial for the visually impaired people, still remains to be exploited.

2.2. Visual Localization

Current visual localization methods fall into two groups, retrieval-based localization (Garg et al., 2019; Bai et al., 2018) and pose-based localization (Sarlin et al., 2019; Taira et al., 2018). The retrieval-based methods obtain the best-matching database image of the query image and take that database image as the localization result. Comparatively, the pose-based methods yield the relative pose of the query image to the environment. In this paper, we focus on the localization method based on image retrieval, which is suitable for assistive navigation in large-scale and dynamic environments because it does not require the precise metric maps of database.

2.2.1. Image Description

The robust descriptor extracting effective place fingerprints from images is crucial for retrieval-based visual localization. As presented in Table 1, some work utilized traditional image descriptors to resolve place recognition issues. Unluckily, those artificial descriptors usually feature limited robustness against the appearance variations. Therefore, those descriptors have to accompany with image pre-processing or complicated matching strategies to ensure the performance of visual localization.

Deep neural networks present superior performance on image descriptor extraction. The off-the-shelf layers of the deep network are taken as the image descriptors (Sünderhauf et al., 2015), but the localization performance is largely limited by the pre-trained network. The fine-tuned descriptors learned on place recognition dataset, such as NetVLAD (Arandjelović et al., 2018), CNN Image Retrieval (Radenović et al., 2016) and DIR (Revaud et al., 2019), naturally attain better performance than the off-the-shelf deep descriptors or artificial descriptors. As an effective feature pooling method for place recognition, the differential pooling module NetVLAD (Arandjelović et al., 2018) is analogue to VLAD (Jégou et al., 2010) and is trained by triplet ranking loss on the dataset with various appearance variations. Proceeding to the pioneering work, many variants

of NetVLAD were also developed by the research community. Zhu et al. (2018) proposed an attention-based pyramid aggregation network to replace the NetVLAD module. Yu et al. (2020) fed NetVLAD with spatial pyramid images and trained the network with the weighted triplet loss.

All of the solutions above are not designed for the scenario of assistive navigation, and the issues of assistive visual localization are not studied and resolved. In this paper, we propose a novel descriptor network *Dual Desc*, which extracts both global and local descriptors simultaneously from color and infrared multimodal images. Furthermore, the learnable global descriptor of *Dual Desc* introduces an attention mechanism that considers the distribution of local features in images during descriptor pooling. Meanwhile, data augmentation methods simulating the practical imaging situations of wearable cameras are applied during metric learning.

2.2.2. Descriptor Matching

After the image description, image retrieval is executed to obtain the result of visual localization. However, merely with the retrieval results derived from global image descriptors, the localization results are still vulnerable to appearance variations and especially place aliasing.

To address the issue, sequence matching algorithms were put forward by synthesizing the localization results of successive frames. The classical sequence matching solution is SeqSLAM (Milford and Wyeth, 2012), where the matching score of a query-database pair is proportional to counting the best retrievals within adjacent regions of that pair. Leverage sequential information to improve the localization performance, Naseer et al. (2018) exploited network flows to maintain several possible trajectories hypotheses in parallel. The drawback of those matching strategies is that they are not likely to be executed online, which is disadvantageous for assistive navigation.

Apart from sequence searching, geometric verification is another approach to refine the retrieval nearest neighbors and determine the final visual localization result. The vanilla method is to select the final result by the inlier number of the estimated Homography (or Fundamental) matrix derived from the detected local features of the image pair.

3. System Overview

The RGB-D-IR multimodal images captured by the wearable camera embedded on the assistive device are utilized for assistive localization.

3.1. Hardware Configurations

In our previous work, the functionalities of obstacle avoidance (Yang et al., 2017) and scene perception (Yang et al., 2018) have been integrated into the assistive system Intoer (KrVision, 2019). The commercialized Intoer features a pair of bone-conduction earphones, a portable computer, and different sensors, including Intel RealSense Camera (Keselman et al., 2017), an inertial measurement unit, and a customized GNSS receiver. The hardware configurations of the portable computer are Intel

Table 1: The Related Work of Traditional Descriptor-based Place Recognition

Work	Descriptor
Routed Roads (Pepperell et al., 2016)	SAD (sum of absolute differences)
OpenABLE (Arroyo et al., 2018)	LDB (Yang and Cheng, 2014)
ORB-SLAM2 (Mur-Artal and Tardós, 2017)	BoW (Galvez-López and Tardos, 2012)
Qiao et al.’s (Qiao et al., 2017)	GIST (Oliva and Torralba, 2001) and CSLBP (Heikkilä et al., 2006)

Figure 2: The assistive device Intoeer is used to capture multimodal data.

Atom x7-Z8700 CPU @ 1.60GHz and 4GB RAM. In this paper, we also leverage the off-the-shelf assistive device to implement the proposed visual localization pipeline. Moreover, the localization results can be used by other assistive algorithms requiring precise positions.

As shown in Figure 2, the RGB-D-IR multimodal images are collected by RealSense, which is a representative product among various RGB-D-IR cameras. Although the RGB image contains rich chromatic information, it also suffers from motion blur, because the RGB camera is with relatively long exposure time and the rolling shutter. Fortunately, the infrared modality features a better performance on imaging stability thanks to the global shutter, though it only remains monochromatic information. Thereby, RGB images featuring abundant visual cues are used to extract global descriptors, meanwhile, depth images featuring robust image details are used to extract 2D local descriptors. Besides, the dense depth images derived from the stereo infrared images provide the information of scene geometry. The bone-conduction earphones are designed for potential human-machine interaction, which is not used in this paper yet.

3.2. Yuquan Multimodal Dataset

As we previously proposed in OpenMPR, the real-world multimodal dataset available at (Cheng, 2019) was captured within the Yuquan campus of Zhejiang University. As shown in Table 2, each subset is composed of one query sequence and one database sequence. For each sequence, there are multimodal image tuples with GNSS data, and the spatial interval of adjacent images is around one meter. All of the images are with the size of 320×240 . Due to the inaccuracy of GNSS data, the ground truth is labeled manually as a best-matching database

Table 2: The Characteristics of Multimodal Yuquan Dataset

Subset	# Query	# Database	Changes
1	97	142	v / o
2	233	273	v / o / i

v = viewpoint variations, o = dynamic objects, and i = illumination variations.

image or no database image according to visual similarity. In the dataset, the index difference between the localization result and the ground truth is used to evaluate algorithm performance. If the index difference is below the tolerance of 5 images, the localization result is taken as a correct result.

The collected multimodal images feature apparent viewpoint variations between query and database sequence. Apart from that, all of the images also present dynamic object changes between query and database. Moreover, the illumination changes exist in Subset-2, where database and query images were captured in the afternoon and at dusk respectively. Furthermore, quite a few images were captured at sidewalks with trees, and they are visually similar to each other, which results in the place aliasing of place recognition. In summary, Yuquan multimodal dataset captured in real-world scenarios benefits the robustness evaluation of assistive visual localization.

4. Deep Descriptor Network: *Dual Desc*

Extracting robust descriptors is a key part of visual localization to overcome the appearance variation issues. In this paper, we leverage the unified network *Dual Desc* to extract both local and global descriptors simultaneously. As shown in Figure 3, the network is composed of a backbone, which provides with basic feature description capability, and two branches to extract global descriptors *G* (*NetAVLAD*) and local descriptor *L* (*Dense Desc*). The backbone network can be one of the prevailing convolutional neural networks, which are truncated to get rid of fully connected layers. The intermediate and the last layer of the backbone are fed into branch *Dense Desc* and branch *NetAVLAD* respectively. The fully convolutional structure of the backbone makes it possible that the input images are with an arbitrary size or aspect ratio.

4.1. *Dense Desc*: Local Features

As we have presented in the previous section, the infrared modality is leveraged to extract local features owing to its stable image details. Meanwhile, we generate sparse point cloud from the depth image.

The input image of *Dual Desc* is with the size of $W_i \times H_i \times 3$. The branch *Dense Desc* receives the intermediate feature maps F_l from the backbone network, and generates dense local descriptors with the size of $W_l \times H_l \times C_l$. Subsequently, the layer

Figure 3: The schematics of the proposed deep network *Dual Desc*. The branch *NetAVLAD* generates global descriptors for place retrieval, meanwhile, the branch *Dense Desc* generates 2D local features and 3D point cloud for geometric verification.

F_l is L2-normalized in the channel dimension, so as to generate local descriptors

$$\{\mathbf{d}_{ij} \in \mathbb{R}^C | i = 1, 2, \dots, W_l, j = 1, 2, \dots, H_l\}. \quad (1)$$

The local descriptors are associated with mesh key points, which are represented as

$$\mathbf{k}_{ij} = (u, v)_{ij} = (ri - \frac{r}{2}, rj - \frac{r}{2}). \quad (2)$$

In the equation, the stride $r = W_i/W_l = H_i/H_l$. As for a 3D point $\mathbf{p}_{ij}$, the depth value z is the average of the depth values within the image patch

$$ri \leq u \leq r(i+1), \quad rj \leq v \leq r(j+1). \quad (3)$$

In view that IR images are with the same field of view as depth images on the RealSense camera, a 2D local feature $(\mathbf{k}, \mathbf{d})$ extracted from the IR image is exactly in coincidence with a 3D point $\mathbf{p}$ extracted from the depth image. The 2D point $\mathbf{k}$ is in the image coordinate, meanwhile the 3D point $\mathbf{p}$ is in the camera coordinate. The transformation relation between the two coordinate is shown as

$$x = \frac{u - c_x}{f_x}z, \quad y = \frac{v - c_y}{f_y}z. \quad (4)$$

In the equations, (c_x, c_y) represents the principal point of camera, and f_x and f_y denotes the focal length (in pixel) of the horizontal and vertical direction.

4.2. NetAVLAD: Global Descriptor

Inspired by the state-of-the-art NetVLAD (Arandjelović et al., 2018) and DELF (Noh et al., 2017), we propose the branch *NetAVLAD* for the place retrieval task. It extracts attentive deep features and is constituted of an attention module (A) and a pooling module (VLAD), which are added on the last layer of the backbone network (Net).

The last feature maps of the backbone is denoted as F_g , which are with the size of $W_g \times H_g \times D_g$. The feature maps

can be seen as $W_g \times H_g$ column-like local descriptors with the dimension of D_g , and can be denoted as

$$\{F_g^i \in \mathbb{R}^{D_g} | i = 1, 2, \dots, W_g \times H_g\}. \quad (5)$$

Different from NetVLAD where all the local features contribute equally, the attention module of NetAVLAD re-weights local features, which impels the discriminate features to be assigned with large weights during feature pooling. The score function $s(F_g)$ transforms the input feature maps F_g into an attention map with the size of $W_g \times H_g \times 1$. Specifically, a point-wise convolution (with 1×1 kernel) with a ReLU activation reduces the channels of input feature maps by a half. Then, another point-wise convolution with a Softplus activation reduces the channels of input feature maps to one. The score function generates the attention information of the input feature maps, which is leveraged in the subsequent feature pooling phase.

The pooling module is subsequent to the attention module. Firstly, F_g is L2-normalized as F_n in the channel dimension, which is represented as

$$F_n^i = \frac{F_g^i}{\max(\|F_g^i\|_2, \epsilon)}. \quad (6)$$

In the equation, ϵ is a small value preventing the denominator from zero. In the D_g -dimensional space, we set K_c cluster centers and denote them as $\{C_k \in \mathbb{R}^{D_g} | k = 1, 2, \dots, K_c\}$, where $K_c = 64$ in this paper. Then each feature F_n^i is assigned to those cluster centers proportional to their proximity, and this procedure is called soft assignment. The assignment coefficient $a_k(F_n^i)$ of the feature F_n^i on the cluster center C_k is represented as

$$a_k(F_n^i) = \frac{\exp(-\alpha\|F_n^i - C_k\|^2)}{\sum_{k'=1}^{K_c} \exp(-\alpha\|F_n^i - C_{k'}\|^2)}. \quad (7)$$

It is worthwhile to note that the soft assignment coefficient will be larger if the feature is closer to the cluster center. This assignment operation could be implemented conveniently by a point-wise convolutional layer and a Softmax layer. In this paper, the convolutional layer has a kernel size of 1×1 , a stride of 1, and an output channels of K_c . After reshaping the feature maps to the size of $A_g \times K_c$, where $A_g = W_g \times H_g$, we apply Softmax layer and yield the assignment coefficients $\{a_k(F_n^i)\}$. Then, the attentive feature residuals for the k -th cluster center are denoted as

$$V_k = \sum_{i=1}^{A_g} s(F_g^i) a_k(F_n^i) (F_n^i - C_k). \quad (8)$$

Having L2-normalized each D_g -dimensional V_k vector, we concatenate those vectors into a global descriptor. After global L2-normalization, the descriptor $\mathbf{G} \in \mathbb{R}^{D_g \times K_c}$ is obtained finally and leveraged for image retrieval.

4.3. Network Training

Training the proposed network *Dual Desc* is to enable both global and local descriptors the robust performance on the issues of visual variations.

Table 3: The Image Numbers of the Pittsburgh Dataset

Dataset	Query Set	Database Set
Training Set	7,320	10,000
Validation Set	7,608	10,000
Testing Set	6,818	10,000

Dataset. The discriminate ability of *NetAVLAD* descriptors is obtained 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. Pittsburgh Dataset (Arandjelović et al., 2018) is such a public dataset, whose images were captured in the urban streets. Each image in the dataset is associated with a GNSS label. The dataset is composed of three subsets, and the numbers of images are shown in Table 3.

Data augmentation. To cope with the motion blur problem caused by the wearable camera in assistive navigation, the training images are augmented to make them more close to the practical images. The data augmentation methods used in this paper include image jitter, image blur, and image rotation. Before training, each image is processed randomly by one of the augmentation operations or not processed.

Random image jitter refers to change the brightness, the contrast, and the saturation of an image randomly, which simulate the illumination variations in the practical scenarios. Random image blur leverages the motion blur kernels with random sizes and orientations to convolve an image. Random image rotation is rotating an image by a random angle around image center. Those two augmentation methods simulate the motion effects of wearable cameras.

Loss function. The triplet ranking loss function (Arandjelović et al., 2018) impels the global descriptor of the query image $\mathbf{G}_q$ to close to that of positive database images $\mathbf{G}_p$ rather than that of negative images $\mathbf{G}_n$, which robustifies the adaptability of the descriptor under visual variations. In detail, the loss function is defined as

$$Loss = \frac{1}{N_n} \sum_j \max[\min_i (\|\mathbf{G}_q - \mathbf{G}_p^i\|_2) + m - \|\mathbf{G}_q - \mathbf{G}_n^j\|_2, 0]. \quad (9)$$

In the equation, N_n is the number of negative descriptors in a training tuple. The positive descriptors $\{\mathbf{G}_p^i\}$ are filtered with the geographical distance from the query image. In this paper, the distance threshold is set as 10 meters. Meanwhile, the negative descriptors $\{\mathbf{G}_n^j\}$ in the training tuple are selected using the hard negative mining method from all of the negative database descriptors. In this paper, the negatives are defined as results whose geographical distance from the query image farther than 25 meters. Herein, the number of negatives is at most 10 in a training tuple. The parameter m in the loss function denotes the margin of negative distance beyond the positive distance, which is set to 0.3 in this paper. In this work, we do not set extra optimizer for the branch *Dense Desc*.

Backbone. The model *Dual Desc* is based on a backbone network, which could be one of popular convolutional networks. In this paper, after fully connected layers removed, both

Table 4: The Layer Structure and Output Size of Backbone Networks

Index	VGG-16	Wide ResNet-18
0	Input ($H \times W \times 3$)	
1	Conv3-ReLU $H \times W \times 64$	Conv7-BN-ReLU $H/2 \times W/2 \times 64$
2	Conv3-ReLU $H \times W \times 64$	Conv3-BN-ReLU $H/2 \times W/2 \times 64$
3	Max Pooling $H/2 \times W/2 \times 64$	Conv3-BN+Res(1)-ReLU $H/2 \times W/2 \times 64$
4	Conv3-ReLU $H/2 \times W/2 \times 128$	Conv3-BN-ReLU $H/2 \times W/2 \times 64$
5	Conv3-ReLU $H/2 \times W/2 \times 128$	Conv3-BN+Res(3)-ReLU $H/2 \times W/2 \times 64$
6	Max Pooling $H/4 \times W/4 \times 128$	Conv3-BN-ReLU $H/4 \times W/4 \times 128$
7	Conv3-ReLU $H/4 \times W/4 \times 256$	Conv3-BN+Res(5-Conv1-BN)-ReLU $H/4 \times W/4 \times 128$
8	Conv3-ReLU $H/4 \times W/4 \times 256$	Conv3-BN-ReLU $H/4 \times W/4 \times 128$
9	Conv3-ReLU $H/4 \times W/4 \times 256$	Conv3-BN+Res(7)-ReLU $H/4 \times W/4 \times 128$
10	Max Pooling $H/8 \times W/8 \times 256$	Conv3-BN-ReLU $H/8 \times W/8 \times 256$
11	Conv3-ReLU $H/8 \times W/8 \times 512$	Conv3-BN+Res(9-Conv1-BN)-ReLU $H/8 \times W/8 \times 256$
12	Conv3-ReLU $H/8 \times W/8 \times 512$	Conv3-BN-ReLU $H/8 \times W/8 \times 256$
13	Conv3-ReLU $H/8 \times W/8 \times 512$	Conv3-BN+Res(11)-ReLU $H/8 \times W/8 \times 256$
14	Max Pooling $H/16 \times W/16 \times 512$	Conv3-BN-ReLU $H/16 \times W/16 \times 512$
15	Conv3-ReLU $H/16 \times W/16 \times 512$	Conv3-BN+Res(13-Conv1-BN)-ReLU $H/16 \times W/16 \times 512$
16	Conv3-ReLU $H/16 \times W/16 \times 512$	Conv3-BN-ReLU $H/16 \times W/16 \times 512$
17	Conv3 $H/16 \times W/16 \times 512$	Conv3-BN+Res(15)-ReLU $H/16 \times W/16 \times 512$

ConvN denotes the convolutional layer with the kernel of N by N, and BN denotes batch normalization layer. Res(*) means the residual branch added to the main network sequence, and * represents the structure of residual branch.

VGG-16 (Simonyan and Zisserman, 2015) and Wide ResNet-18 (Zagoruyko and Komodakis, 2016) are utilized as backbone networks. The convolutional structures of the two networks are shown as Table 4. There are 17 layers in the VGG-based backbone network, including 13 convolutional layers and 4 max-pooling layers. Wide ResNet-18 used in this paper also involves 17 layers, which are organized in a convolution layer and 4 basic blocks, each of which contains 4 convolutional layers. In the basic block, the first convolutional layer down-samples feature maps in the dimensions of height and width, and the second and fourth layers receive the shortcut from one of the lower layers.

Training configurations. In order to achieve better training results, the training procedure starts from the pre-trained

parameters on the large-scale image datasets, e.g. ImageNet (Deng et al., 2009) or Places (Zhou et al., 2018). This is because that the pre-trained networks feature some degree of description capability, training *Dual Desc* based on those backbone networks is easier compared with training the network from scratch.

The stochastic gradient descent (SGD) optimizer is used to training the parameters, which includes parameters of higher convolutional layers in the backbone network, parameters of attention section and parameters of VLAD section. The learning rate of the backbone network and the VLAD section is 10^{-4} for Wide ResNet-18 or 10^{-5} for VGG-16. Meanwhile, the learning rate of the attention section is ten times of other sections. The weight decay is set to 0.01, the momentum is set to 0.9, and the learning rate is halved for every 10 epochs.

5. Pipeline of Hierarchical Visual Localization

The pipeline of the proposed assistive visual localization is shown as Figure 4. A multimodal image tuple is composed of a color image, a depth image, and an infrared image. Having extracted global and local descriptors from multimodal images by *Dual Desc* network (presented in section 4), we retrieve the top- K candidates using global descriptor, filter the candidates with geometric verification and finally obtain the localization result by sequence-based matching.

5.1. Place Retrieval: Coarse Localization

As a holistic vector representing the input image, the N -dimensional global descriptor $\mathbf{G} \in \mathbb{R}^N$ ($N = D_g \times K_c$) serves in the phase of image retrieval for coarse localization. The top- K nearest neighbors of a query global descriptor $\mathbf{G}_q$ are retrieved from the database descriptor set $\{\mathbf{G}_d\}_S$, where S denotes the cardinality of the set. The GNSS data are leveraged in image retrieval, if they are available. For each query image, only the database images within a GNSS distance of d_{GNSS} are taken as the candidates of image retrieval. The place retrieval is shown as Algorithm 1.

Algorithm 1 Place Recognition

Input: $(\mathbf{G}_q, P_q)$: query global descriptor and GNSS position; $\{(\mathbf{G}_d, P_d)\}_S$: database global descriptors and GNSS positions

Output: Top- K nearest neighbors

- 1: Filter database with GNSS information, i.e. $\{(\mathbf{G}_d, P_d)\}_{S'} \leftarrow \{(\mathbf{G}_d, P_d) | \text{distance}(P_q, P_d) < d_{GNSS}\}$
 - 2: Retrieve top- K nearest neighbors $\{\mathbf{G}_d\}_K$ from $\{\mathbf{G}_d\}_{S'}$
-

In view of the high dimension of global descriptors, the randomized k-d forest is utilized to retrieve the approximate nearest neighbors of the query image from database images. During constructing the searching index, we use Euclidean distance as the distance function of global features. The number of k-d trees is set as 5 in this paper. The implementation of descriptor retrieval relies on the fast library for approximate nearest neighbors (FLANN) (Muja and Lowe, 2009).

5.2. Geometric Verification: Fine Localization

Utilized in the phase of geometric verification for fine localization, the local features of an image are denoted as

$$L = \{(\mathbf{L}_{2D}, \mathbf{L}_{3D})\}_l, \mathbf{L}_{2D} = (\mathbf{k}, \mathbf{d}), \mathbf{L}_{3D} = \mathbf{p}. \quad (10)$$

The number of local features in an image is $l = W_l \times H_l$. Each local feature is constituted of a 2D local feature $\mathbf{L}_{2D}$ and a 3D point $\mathbf{L}_{3D}$. The 2D feature is composed of a key point $\mathbf{k} = (u, v)$ and a M -dimensional local descriptor $\mathbf{d} \in \mathbb{R}^M$. Referring to the previous section, $M = C_l$. The 3D point $\mathbf{p}$ is denoted as (x, y, z) , where z value of point is derived from the depth image. We use both 2D features and 3D sparse point cloud to execute geometric verification, so as to select the optimal database image out of the retrieved top- K candidates. The pipeline of geometric verification is shown as Algorithm 2.

Algorithm 2 Geometric Verification

Input: $\{(\mathbf{k}_q, \mathbf{d}_q)\}$: the 2D local features of a query image; $\{\mathbf{p}_q\}$: the 3D point cloud of a query image; $\{(\mathbf{k}_d, \mathbf{d}_d)\}_K$: the 2D local features of K database images; $\{\mathbf{p}_d\}_K$: the 3D point clouds of K database images

Output: At most two verified database image

- 1: **for all** $\{(\mathbf{k}_d, \mathbf{d}_d)\}$ in $\{(\mathbf{k}_d, \mathbf{d}_d)\}_K$ **do**
 - 2: Match each $\mathbf{d}_q$ with $\mathbf{d}_d$ in the Euclidean space
 - 3: Form the matches into the correspondences $\{(\mathbf{k}_q : \mathbf{k}_d)\}$
 - 4: Estimate H matrix with RANSAC
 - 5: Get the number of inlier correspondences
 - 6: **end for**
 - 7: Rank K candidates by inlier number, return the database image with the most inliers as the 2D verification result
 - 8: **for all** $\{\mathbf{p}_d\}$ in $\{\mathbf{p}_d\}_K$ **do**
 - 9: Estimate R and t between $\{\mathbf{p}_q\}$ and $\{\mathbf{p}_d\}$ with ICP
 - 10: **end for**
 - 11: Rank K candidates by transition t , return the database image with the smallest t as the 3D verification result
 - 12: Get rid of low-confidence candidates
-

2D geometric verification. Local descriptors $\{\mathbf{d}\}$ are leveraged to find the correspondences between the query image and each candidate database image. We search the first and the second nearest database descriptor $\mathbf{d}_1$ and $\mathbf{d}_2$ of a query descriptor $\mathbf{d}_0$ in the M -dimensional Euclidean space. The first nearest neighbor $\mathbf{d}_1$ is the matched correspondence of $\mathbf{d}_0$ only if the Euclidean distance between the query descriptor and the nearest neighbors satisfies the ratio test

$$\|\mathbf{d}_0 - \mathbf{d}_1\|_2 < th \|\mathbf{d}_0 - \mathbf{d}_2\|_2. \quad (11)$$

The ratio threshold th is related to the visual variations of images (e.g. illumination, rotation, and scale) and local descriptors themselves. If the threshold is smaller, the qualified correspondences will be fewer, and vice versa. In this paper, the ratio threshold is determined as 0.7 by trial and error. In this way, we obtain the robust correspondences between query local features and the database local features. Therefore, a correspondence can also be denoted as the key point relationship,

Figure 4: The hierarchical visual localization is composed of three stages of place retrieval, geometric verification, and sequence matching.

i.e. $\{(\mathbf{k}_q = (u, v) : \mathbf{k}_d = (u', v'))\}$. The local correspondences in the homogeneous coordinate are associated by the perspective transformation H (Homography matrix) (Hartley and Zisserman, 2003), which is represented as

$$\begin{bmatrix} u'_i \\ v'_i \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{bmatrix} \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = H \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix}. \quad (12)$$

In view that the number of correspondences succeeds the minimal requirement of getting an H matrix solution, RANSAC algorithm (Fischler and Bolles, 1981) is utilized to get the best solution. The optimization criterion is to minimize re-projection error, which is defined as

$$error = \|\mathbf{k}_d - H\mathbf{k}_q\|_2. \quad (13)$$

According to whether fitting the final H matrix, the correspondences fall into two groups: inliers and outliers. The re-projection error threshold is set as 4 pixels in this paper, the correspondence is defined as an outlier if it exceeds the threshold. Finally, the correspondence inliers of Homography estimation are used to rank the K database images, and the candidate with the most inliers is the optimal 2D verified candidate.

3D geometric verification. Thanks to the three-dimensional sparse points derived from depth images, we use the iterative closest point (ICP) to register two groups of point cloud $\{\mathbf{p}_q\}$ and $\{\mathbf{p}_d\}$. Given two sets of 3D points, the ICP algorithm finds the optimal rotation R and transition t , when the registration error $E(R, t)$ of the two groups of point cloud gets the minimal value.

$$E(R, t) = \frac{1}{n} \sum_{i=1}^n d_i, \quad d_i = \|\mathbf{p}_{di} - (R\mathbf{p}_{qi} + t)\|_2 \quad (14)$$

Following Birdal's ICP implementation (Birdal, 2015), we make modifications to the vanilla ICP pipeline in this paper. Firstly, point sampling is omitted and all of the 3D points are leveraged in the algorithm, in that the point cloud is sparse. Moreover, the correspondences obtained by descriptor matching are taken as the initial point correspondence, which aims to lead to a more correct pose. The stopping condition of iterative computation is 500 iterations or registration error descent of less than 0.5%. In every iteration (except for the first iteration), the correspondences with distances $d > 3\sigma$ are refused

before pose estimation, where σ is the standard deviation of correspondence distance distribution.

After pose estimation, the transition distance of the query and the candidate is taken as the re-ranking criterion. The candidate with the least distance is selected as the optimal 3D candidate.

Through 2D and 3D geometric verification, two database images are selected out of K candidates as the potential results of visual localization. Subsequently, we get rid of the low-confidence candidates by the number of Homography inliers and the point cloud transition distance. The 2D candidate with more than 300 inliers and the 3D candidate with transition distance of fewer than 15 meters are processed in the following sequence matching.

5.3. Sequence Matching: Final Decision

In order to robustify the visual localization further, sequence matching, which utilizes the geometric verification results of successive frames, determines the final result of visual localization for each query image. Inspired by SeqSLAM (Milford and Wyeth, 2012), the online sequence matching proposed by our previous work OpenMPR (Cheng et al., 2019) is modified to adapt to this work, which is presented in Algorithm 3.

Algorithm 3 Sequence Matching

Input: $\{(q, d_{2D}, d_{3D})\}$: verified database candidates of query images

Output: Matching results of all query images

- 1: **for all** query-database pairs **do**
 - 2: Set a cone-like region around the query-database pair
 - 3: Count verified candidates number within the region
 - 4: Calculate the matching score s of the pair
 - 5: **end for**
 - 6: For each query, let the database with highest score as the potential result
 - 7: Get rid of low-confidence results by thresholding
-

In the preceding phase, we got at most two potential database images for one query image. As shown in Figure 5, we construct a matching matrix with the size of $N_d \times N_q$, where N_d and N_q denote the number of database images and the number of query images respectively. For the candidates in the matching matrix, the element where a verified result locates is set to 1, which is represented by the blue square in Figure 5. Otherwise,

Figure 5: The online sequence matching strategy used in this paper.

the element is set to 0. That is to say, at most two elements are non-zero in a row of the matching matrix.

As shown in Figure 5, taking a query-database pair (i, j) as the vertex, a cone-like searching region is defined by sequential length n_q , maximal velocity v_{max} and minimal velocity v_{min} . The velocity refers to the speed ratio of query sequence to database sequence, i.e. $v = \delta d / \delta q$. From the cone region, it can be seen that the online sequence matching only makes use of the *past* query images, regardless of the *future* query images.

Within the cone region, the sum of element values is defined as s_{match} . Then, the score $s_{i,j}$ of a query-database pair (i, j) is defined as

$$s_{i,j} = \frac{s_{match}}{n_q}. \quad (15)$$

In this way, we get a matching score for each query-database pair. For each query image, the latent best-matching result is the database image with the highest score.

Finally, the matching score of the best query-database pair is evaluated by windowed uniqueness thresholding (Milford and Wyeth, 2012) to remove low-confidence matching results. For any query image (e.g. the i -th image), assuming that the j -th database image features the highest score ($s_{i,j}$) in the database sequence, we set a window in the database score sequence and take $s_{i,j}$ as the center. Among the scores out of the window, the highest score is searched and denoted as $s'_{i,j}$. Representing the robustness level of the latent best-matching result, the ratio of windowed uniqueness is defined as

$$w = \frac{s_{i,j}}{s'_{i,j}}. \quad (16)$$

The latent sequence matching result whose uniqueness ratio is larger than the threshold is a qualified result. Otherwise, the latent result is abandoned. This paper follows the parameter settings used in OpenSeqSLAM2.0 (Talbot et al., 2018) that window size is set to 15 and the ratio threshold is set to 1.1.

6. Experiments

In this section, we evaluate the performance of descriptor network *Dual Desc* on the public dataset. Then, the real-world experiments of the proposed multimodal visual localization system are illustrated and discussed.

6.1. Evaluation of Dual Desc

In order to validate the effectiveness of the proposed descriptor network, *Dual Desc* is evaluated by the image retrieval task on Pittsburgh Testing Dataset. Using VGG-16 as the backbone network, we follow the conclusion in (Arandjelović et al., 2018) that the optimal retrieval performance is obtained when only the layers 15-17 (listed in Table 4) are trained. As for the backbone of Wide ResNet-18, we train the last 8 layers (i.e. the layers 10-17). The backbones VGG-16 and Wide ResNet-18 are pre-trained on ImageNet and on Places365 (Zhou et al., 2018) respectively. The query image is localized correctly if one of images in the top- K retrieved results is within the range of 25 meters from the query image. Recall, which is the percentage of correctly recognized queries, is chosen as the performance indicator.

As shown in Table 5, both the proposed global descriptor *NetAVLAD* (denoted as w/ Att.) and its counterpart *NetVLAD* (denoted as w/o Att.) are tested for image retrieval on Pittsburgh Testing dataset. For both backbone networks, the *NetAVLAD* with the attention module achieves better retrieval recall compared with its counterpart. For the VGG-based network, *NetAVLAD* attains the recall gain by 0.20%, 0.28%, and 0.18% for top-5, top-10, and top-20 retrievals respectively. For the ResNet-based network, *NetAVLAD* attains the recall gain by 0.03%, 0.08%, and 0.14% for top-5, top-10, and top-20 retrievals respectively. The performance improvement is not so significant considering the recall is already high for *NetVLAD*, but the improvement is rather stable for the two backbones under different top- K retrievals. Although the top-1 recall of VGG-based *NetAVLAD* is slightly lower than that of *NetVLAD*, it does not affect the performance of the whole visual localization system, because the proposed pipeline conveys multiple retrieval candidates to the next phase.

Table 5 also lists the retrieval performance of the original VGG-based *NetVLAD* (Arandjelović et al., 2018) and the VGG-based SPE-VLAD (Yu et al., 2020), which is a state-of-the-art variant of *NetVLAD*. The original *NetVLAD* we used in this paper is implemented by van Noord (2019). The recall improvement of *NetAVLAD* is derived from the proposed attention mechanism and data augmentation methods.

6.2. Real-world Test

The real-world dataset under the scenario of assistive navigation is leveraged to validate the proposed hierarchical visual localization pipeline.

6.2.1. Experimental Configurations and Indicators

In this paper, the top-5 candidates are obtained from the phase of image retrieval. The local features derived from *Dense Desc* are obtained from the intermediate feature maps of backbone networks. Specifically, the ninth layer of Wide ResNet-18 and the ninth layer of VGG-16 are chosen as the input of *Dense Desc*. Therefore, the ratio of image size to feature maps size is $r = 4$. The threshold of GNSS distance d_{GNSS} is set as 15 meters, which is determined by the positioning accuracy of GNSS device. As for the parameters in the phase of sequence

Table 5: The Recall of NetAVLAD Descriptor with Different Training Configurations on Pittsburgh Testing Dataset

	Top 1	Top 5	Top 10	Top 20
VGG-16 w/o Att.	79.86%	91.01%	93.84%	95.94%
VGG-16 w/ Att.	79.84%	91.21%	94.12%	96.13%
WRRes-18 w/o Att.	81.78%	90.87%	93.79%	95.91%
WRRes-18 w/ Att.	81.95%	90.90%	93.87%	96.05%
NetVLAD (Arandjelović et al., 2018)	81.88%	91.23%	93.66%	95.75%
SPE-VLAD (Yu et al., 2020)	-	89.23%	92.50%	94.62%

WRRes-18 denotes Wide ResNet-18, and Att. denotes attention module.

matching, the sequence length $n_q = 20$, the maximal velocity $v_{max} = 2.5$, and the minimal velocity $v_{min} = 0.4$, which are the optimized parameters obtained by our previous work OpenMPR (Cheng et al., 2019).

Similar to OpenMPR, we use F_1 score, precision, recall and mean localization error (MLE) as the performance indicators of visual localization. If no database image is matched in a localization result, the result is defined as a negative result. Otherwise, it is defined as a positive result. If a negative result should have matched with a database image according to the ground truth, it is defined as a false negative (FN) result. If a positive result is correct (i.e., within the tolerance), it is defined as a true positive (TP) result. Otherwise, the positive result is defined as a false positive (FP) result. Precision is the proportion of true positives out of all predicted positives, and recall is the proportion of true positives to all of actual positives.

$$Precision = \frac{TP}{TP + FP}, \quad (17)$$

$$Recall = \frac{TP}{TP + FN}. \quad (18)$$

Moreover, F_1 score is a performance indicator that synthesizes both precision and recall by

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \quad (19)$$

In order to evaluate the localization accuracy, we use mean localization error to represent the averaged index difference between positive results and their ground truths.

6.2.2. Ablation Study

In order to validate the necessity of RGB-D-IR multimodal images, we carried out the ablation study of image modalities, where only RGB images or RGB-IR images are used in the hierarchical visual localization. For the cases without depth images, the 3D geometric verification is omitted in the pipeline. When only RGB images are available, both global descriptors and 2D features are extracted from RGB images. The visual localization results using different modalities on the Yuquan multimodal dataset are shown in Table 6.

From the experimental results, *Dual Desc* with RGB-D-IR images features a higher F_1 score than *Dual Desc* with RGB-IR or RGB images. It is concluded that both depth modality and infrared modality contribute to the localization robustness, and indeed promote the visual localization performance in the

Table 6: The Ablation Study on the Yuquan Multimodal Dataset

Subset	Method	F_1	Precision	Recall	MLE
Subset-1	Res+RGB	0.56	61.3%	52.1%	5.40
	Res+RGB-IR	0.72	73.0%	70.1%	5.10
	Res+RGB-D-IR	0.82	80.7%	82.7%	2.33
	VGG+RGB	0.81	86.8%	75.9%	0.91
	VGG+RGB-IR	0.83	87.3%	79.3%	0.85
	VGG+RGB-D-IR	0.84	87.5%	80.5%	0.86
Subset-2	Res+RGB	0.59	58.8%	58.8%	5.55
	Res+RGB-IR	0.71	72.1%	70.5%	2.51
	Res+RGB-D-IR	0.77	77.0%	76.6%	2.75
	VGG+RGB	0.65	63.4%	66.1%	3.66
	VGG+RGB-IR	0.73	74.3%	73.1%	2.90
	VGG+RGB-D-IR	0.75	71.0%	80.0%	3.92

scenario of assistive navigation. Compared with the results of RGB modality, the F_1 score of RGB-IR increases by more than 0.1 on Subset-2 and Subset-1 with the VGG backbone. The significant performance improvement illustrates that the local features extracted from the infrared modality benefit geometric verification than those from RGB images. The reason is that the infrared images captured by our device feature less motion blur and thereby contain more detailed texture information. Moreover, the 3D sparse point cloud derived from depth images improves the pipeline performance, which is natural because depth images provide the scene shape for the stage of geometric verification. What deserves to mention is that the mean localization error of RGB-D-IR is slightly lower than that of RGB-IR, which is caused by the recall improvement, but it is still within the acceptable range.

In the experiment, both VGG-16 and Wide ResNet-18 are utilized for the backbone of *Dual Desc* separately. The two kinds of backbone networks achieve similar performance. In terms of mean localization error and F_1 score, the VGG-based *Dual Desc* is better on Subset-1, but the ResNet-based *Dual Desc* is better on Subset-2. Considering Subset-2 is with more visual variations, the ResNet-based model seems to have stronger description capability on real-world visual localization, whose reason may be that the more complicated network structures like shortcuts strengthen the generalization of the descriptor network.

6.2.3. Performance Comparison

Different baselines are compared with the proposed hierarchical visual localization pipeline, and the experimental results are summarized in Table 7. The GNSS-based localization, the state-of-the-art place recognition algorithm OpenSe-

Table 7: The Visual Localization Results on the Testing Datasets

Subset	Method	F_1	Precision	Recall	MLE
Subset-1	GNSS	-	0	-	37.79
	OpenSeqSLAM2.0	0.54	65.2%	46.4%	-
	OpenMPR	0.94	88.7%	100.0%	1.33
	Ours (Res)	0.82	80.7%	82.7%	2.33
	Ours (VGG)	0.84	87.5%	80.5%	0.86
Subset-2	GNSS	-	0	-	41.09
	OpenSeqSLAM2.0	0.35	40.6%	31.3%	-
	OpenMPR	0.73	57.8%	99.3%	7.59
	Ours (Res)	0.77	77.0%	76.6%	2.75
	Ours (VGG)	0.75	71.0%	80.0%	3.92

qSLAM2.0 (Talbot et al., 2018), and our preliminary work OpenMPR are chosen as the baselines. Moreover, OpenABLE (Arroyo et al., 2018) is also tested on the Yuquan dataset, but it fails to achieve visual localization under such challenging scenarios. In view of the large mean localization errors, the performance of merely using GNSS data is not acceptable for assistive navigation. This is caused by the limited positioning accuracy of the used portable GNSS device, which also illustrates the necessity of visual localization.

Compared with OpenSeqSLAM2.0, the proposed pipeline promotes the localization F_1 score by a large margin (more than 0.3), especially on the more challenging Subset-2. OpenSeqSLAM2.0 leverages two cone regions in the sequence matching scheme for place recognition. Apart from that, it shares the same parameters with our method in this paper. However, due to the naive image description method, OpenSeqSLAM2.0 cannot cope with the place recognition issues of visual variations. Compared with our previous work OpenMPR that multiple descriptors based on multimodal images are integrated into sequence matching, the proposed unified descriptor network *Dual Desc* and the pipeline based on it achieves superior localization performance in terms of the localization error. Even on the challenging Subset-2, the mean localization error of the proposed pipeline is superior to that of OpenMPR, and is less than the error tolerance of 5.

Some visualization results on the Yuquan multimodal dataset are shown in Figure 6, where ResNet-based and VGG-based *Dual Desc* networks are leveraged separately. Those localization results are uniformly sampled from the query image sequence. For Subset-1, the image motion blur and the apparent viewpoint variation between query and database images do not impede the localization performance. From the results of Subset-2, we can find that the illuminance variation is the overwhelming visual changes between query and database images. Moreover, the dynamic objects (pedestrians) are also different in query and database images, e.g. query 20, query 100, and query 220. In this condition, the visual localization results are still satisfactory.

6.2.4. Computational Efficiency

Real-time performance is of vital importance for assistive navigation. Therefore, the computational efficiency of the proposed visual localization pipeline is investigated on different platforms the assistive device Intoer, a desktop with Intel Core i5-6500 CPU @ 3.20GHz $\times$ 4 and 16G RAM, and the NVIDIA

Table 8: The Latency of the Proposed Pipeline

	DE (ms)		VL (ms)		Sum (s)	
	VGG	Res	VGG	Res	VGG	Res
TX2	181	125	2,663	2,300	2.8	2.4
Desktop	419	200	2,238	2,639	2.7	2.8
Intoer	7322	3981	10,052	9,172	17.4	13.2

DE denotes Descriptor Extraction, and VL denotes Visual Localization.

Jetson TX2 (NVIDIA, 2020). It is worthwhile to note that GPU is not deployed on the desktop and the Intoer, but is available on TX2. The challenging Subset-2 of the Yuquan multimodal dataset is utilized to analyze the time consumed on *Dual Desc* descriptor extraction and subsequent visual localization phases. The latency results of VGG-based and ResNet-based *Dual Desc* are shown separately in Table 8, where the phase of visual localization includes image retrieval, 2D-3D geometric verification, and sequence matching.

Descriptor extraction is implemented in Python, while the subsequent visual localization is implemented in C++. It can be concluded from the experimental statistics that the proposed pipeline processes one query frame for around 2 seconds on the TX2 and the desktop, which basically satisfies the real-time requirement of assistive navigation. However, due to the limited computational capability, the assistive device Intoer yields a much longer computational latency, which is not so ideal in the practical use.

Luckily, there are still multiple solutions to promote real-time performance in the future. Multi-threading programming is bound to improve the frame rate of the proposed pipeline by a large margin, considering there are multiple parallel procedures in the proposed pipeline, like 2D-3D geometric verification. Moreover, implementing the descriptor extraction phase of the proposed algorithm in C++ may accelerate the network inference speed. Furthermore, compact deep networks can be chosen as the backbone of *Dual Desc*, as we came up with in Cheng et al. (2020).

7. Conclusions

Aiming at visual localization in assistive navigation, we propose a novel visual localization pipeline that utilizes images in different modalities to generate robust local and attentive global descriptors and then obtain precise results using geometric verification and sequence matching for assistive navigation. The codes of the proposed visual localization pipeline are available at <https://github.com/chengricky/DualDesc>.

The assistive visual localization confronts appearance variations between query and database images. The appearance variations include illuminance changes, viewpoint changes, dynamic objects and motion blur. Facing to those issues, we propose a unified descriptor network *Dual Desc* to extract both local and global descriptors from multimodal images. The proposed unified descriptor network *Dual Desc* surpasses the state-of-the-art NetVLAD and its variants. Validated on the real-world dataset captured by the wearable assistive device, the proposed multimodal visual localization algorithm achieves satisfactory performance.

(a)

(b)

Figure 6: Some results of visual localization by the proposed pipeline on the Yuquan multimodal (a) Subset-1 and (b) Subset-2. The image indices of the query (left), the ResNet-based result (middle), and the VGG-based result (right) are labeled above the images, and the correct/incorrect localization results are denoted with the green/red color.

In the future, we plan to improve the robustness of visual localization further, and develop visual localization system based on more visual modalities, such as panoramic images.

Acknowledgments

This work has financial support from the ZJU-Sunny Photonics Innovation Center (No. 2020-03).

References

- Al-Khalifa, S. and Al-Razgan, M. (2016). Ebsar: Indoor guidance for the visually impaired. *Computers & Electrical Engineering*, 54:26–39.
- Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J. (2018). NetVLAD: CNN architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1437–1451.
- Arroyo, R., Alcantarilla, P. F., Bergasa, L. M., and Romera, E. (2018). Are you able to perform a life-long visual topological localization? *Autonomous Robots*, 42(3):665–685.
- Bai, D., Wang, C., Zhang, B., Yi, X., and Yang, X. (2018). Sequence searching with CNN features for robust and fast visual place recognition. *Computers & Graphics*, 70:270–280.
- Birdal, T. (2015). ICP registration using efficient variants and multi-resolution scheme. <https://www.mathworks.com/matlabcentral/fileexchange/47152-icp-registration-using-efficient-variants-and-multi-resolution-scheme>.
- Bourne, R. R. A., Flaxman, S. R., Braithwaite, T., Cicinelli, M. V., Das, A., Jonas, J. B., Keeffe, J., Kempen, J. H., Leasher, J., Limburg, H., Naidoo, K., Pesudovs, K., Resnikoff, S., Silvester, A., Stevens, G. A., Tahhan, N., Wong, T. Y., Taylor, H. R., Bourne, R., Ackland, P., Ardit, A., Barkana, Y., Bozkurt, B., BRAITHWAITE, T., Bron, A., Budenz, D., Cai, F., Casson, R., Chakravarthy, U., Choi, J., Cicinelli, M. V., Congdon, N., Dana, R., Dandona, R., Dandona, L., Das, A., Dekaris, I., Monte, M. D., Deva, J., Dreer, L., Ellwein, L., Frazier, M., Frick, K., Friedman, D., Furtado, J., Gao, H., Gazzard, G., George, R., Gichuhi, S., Gonzalez, V., Hammond, B., Hartnett, M. E., He, M., Hejtmancik, J., Hirai, F., Huang, J., Ingram, A., Javitt,

- J., Jonas, J., Joslin, C., Keeffe, J., Kempen, J., Khairallah, M., Khanna, R., Kim, J., Lambrou, G., Lansingh, V. C., Lanzetta, P., Leasher, J., Lim, J., LIMBURG, H., Mansouri, K., Mathew, A., Morse, A., Munoz, B., Musch, D., Naidoo, K., Nangia, V., PALAIOU, M., Parodi, M. B., Pena, F. Y., Pesudovs, K., Peto, T., Quigley, H., Raju, M., Ramulu, P., Resnikoff, S., Robin, A., Rossetti, L., Saaddine, J., SANDAR, M., Serle, J., Shen, T., Shetty, R., Sieving, P., Silva, J. C., Silvester, A., Sitorus, R. S., Stambolian, D., Stevens, G., Taylor, H., Tejedor, J., Tielsch, J., Tsilimbaris, M., van Meurs, J., Varma, R., Virgili, G., Volmink, J., Wang, Y. X., Wang, N.-L., West, S., Wiedemann, P., Wong, T., Wormald, R., and Zheng, Y. (2017). Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. *The Lancet Global Health*, 5(9):e888 – e897.
- Brilhault, A., Kammoun, S., Gutierrez, O., Truillet, P., and Jouffrais, C. (2011). Fusion of artificial vision and gps to improve blind pedestrian positioning. In *2011 4th IFIP International Conference on New Technologies, Mobility and Security*, pages 1–5. IEEE.
- Cheng, R. (2019). OpenMPR. <https://github.com/chengricky/OpenMultiPR>.
- Cheng, R., Wang, K., Bai, J., and Xu, Z. (2019). OpenMPR: Recognize places using multimodal data for people with visual impairments. *Measurement Science and Technology*, 30(12):124004.
- Cheng, R., Wang, K., Bai, J., and Xu, Z. (2020). Unifying visual localization and scene recognition for people with visual impairment. *IEEE Access*, 8:64284–64296.
- Cheng, R., Wang, K., Lin, L., and Yang, K. (2018). Visual localization of key positions for visually impaired people. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2893–2898.
- Deng, J., Dong, W., Socher, R., Li, L., and and (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.
- Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395.
- Galvez-López, D. and Tardos, J. D. (2012). Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197.
- Garg, S., Sunderhauf, N., and Milford, M. (2019). Semantic-geometric visual place recognition: a new perspective for reconciling opposing views. *The International Journal of Robotics Research*, 0(0):1–26.
- Gleason, C., Ahmetovic, D., Savage, S., Toxtli, C., Posthuma, C., Asakawa, C., Kitani, K. M., and Bigham, J. P. (2018). Crowdsourcing the installation and maintenance of indoor localization infrastructure to support blind navigation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–25.
- Gomes, J. P., Sousa, J. P., Cunha, C. R., and Morais, E. P. (2018). An indoor navigation architecture using variable data sources for blind and visually impaired persons. In *2018 13th Iberian Conference on Information Systems and Technologies (CISTI)*, volume 2018-June, pages 1–5. IEEE.
- Hartley, R. and Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge University Press.
- Heikkilä, M., Pietikäinen, M., and Schmid, C. (2006). Description of interest regions with center-symmetric local binary patterns. In Kalra, P. K. and Peleg, S., editors, *Computer Vision, Graphics and Image Processing*, pages 58–69, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Jégou, H., Douze, M., Schmid, C., and Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3304–3311.
- Keselman, L., Woodfill, J. I., Grunnet-Jepsen, A., and Bhowmik, A. (2017). Intel(R) RealSense(TM) stereoscopic depth cameras. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1267–1276.
- Kim, J.-E., Bessho, M., Kobayashi, S., Koshizuka, N., and Sakamura, K. (2016). Navigating visually impaired travelers in a large train station using smartphone and bluetooth low energy. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing - SAC '16*, pages 604–611, New York, New York, USA. ACM Press.
- KrVision (2019). Into: auxiliary glasses for people with visual impairments (in Chinese). <http://www.krvision.cn/cpjs/>.
- Lin, S., Cheng, R., Wang, K., and Yang, K. (2018). Visual Localizer: Outdoor localization based on ConvNet descriptor and global optimization for visually impaired pedestrians. *Sensors*, 18(8):2476.
- Lowry, S., Sünderhauf, N., Newman, P., Leonard, J. J., Cox, D., Corke, P., and Milford, M. J. (2016). Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1):1–19.
- Milford, M. J. and Wyeth, G. F. (2012). SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In *2012 IEEE International Conference on Robotics and Automation*, pages 1643–1649.
- Muja, M. and Lowe, D. G. (2009). Fast approximate nearest neighbors with automatic algorithm configuration. In *International Conference on Computer Vision Theory and Applications (VISAPP'09)*.
- Mur-Artal, R. and Tardós, J. D. (2017). ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262.
- Naseer, T., Burgard, W., and Stachniss, C. (2018). Robust visual localization across seasons. *IEEE Transactions on Robotics*, 34(2):289–302.
- Noh, H., Araujo, A., Sim, J., Weyand, T., and Han, B. (2017). Large-scale image retrieval with attentive deep local features. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3476–3485.
- NVIDIA (2020). Jetson tx2 developer kit. <https://developer.nvidia.com/embedded/jetson-tx2-developer-kit>.
- Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175.
- Pepperell, E., Corke, P., and Milford, M. (2016). Routed roads: Probabilistic vision-based place recognition for changing conditions, split streets and varied viewpoints. *The International Journal of Robotics Research*, 35(9):1057–1179.
- Piasco, N., Sidibé, D., Demonceaux, C., and Gouet-Brunet, V. (2018). A survey on visual-based localization: On the benefit of heterogeneous data. *Pattern Recognition*, 74:90 – 109.
- Qiao, Y., Cappelle, C., and Ruichek, Y. (2017). Visual localization across seasons using sequence matching based on multi-feature combination. *Sensors*, 17(11):2442.
- Radenović, F., Tolias, G., and Chum, O. (2016). CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 3–20, Cham. Springer International Publishing.
- Revaud, J., Almazán, J., de Rezende, R. S., and de Souza, C. R. (2019). Learning with average precision: Training image retrieval with a listwise loss. *CoRR*, abs/1906.07589.
- Sarlin, P.-E., Cadena, C., Siegwart, R., and Dymczyk, M. (2019). From coarse to fine: Robust hierarchical localization at large scale. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Sünderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B., and Milford, M. (2015). On the performance of ConvNet features for place recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4297–4304.
- Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Sivic, J., Pajdla, T., and Torii, A. (2018). Inloc: Indoor visual localization with dense matching and view synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Talbot, B., Garg, S., and Milford, M. (2018). 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)*, pages 7758–7765.
- van Noord, N. (2019). PyTorch implementation of NetVLAD. <https://github.com/Nanne/pytorch-NetVlad>.
- Yang, K., Wang, K., Bergasa, L., Romera, E., Hu, W., Sun, D., Sun, J., Cheng, R., Chen, T., and López, E. (2018). Unifying terrain awareness for the visually impaired through real-time semantic segmentation. *Sensors*, 18(5):1506.
- Yang, K., Wang, K., Cheng, R., Hu, W., Huang, X., and Bai, J. (2017). Detecting traversable area and water hazards for the visually impaired with a pRGB-D sensor. *Sensors*, 17(8):1890.
- Yang, X. and Cheng, K. T. (2014). Local difference binary for ultrafast and distinctive feature description. *IEEE Transactions on Pattern Analysis and*

Machine Intelligence, 36(1):188–194.

- Yu, J., Zhu, C., Zhang, J., Huang, Q., and Tao, D. (2020). Spatial pyramid-enhanced NetVLAD with weighted triplet loss for place recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 31(2):661–674.
- Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks. In Richard C. Wilson, E. R. H. and Smith, W. A. P., editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press.
- Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2018). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464.
- Zhu, Y., Wang, J., Xie, L., and Zheng, L. (2018). Attention-based pyramid aggregation network for visual place recognition. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, pages 99–107, New York, NY, USA. ACM.