Visual Search Based Indoor Localization in Low Light via RGB-D Camera

Yali Zheng, Peipei Luo, Shinan Chen, Jiasheng Hao, Hong Cheng

***Abstract*—**Most of traditional visual indoor navigation algorithms and methods only consider the localization in ordinary daytime, while we focus on the indoor re-localization in low light in the paper. As RGB images are degraded in low light, less discriminative infrared and depth image pairs are taken, as the input, by RGB-D cameras, the most similar candidates, as the output, are searched from databases which is built in the bag-of-word framework. Epipolar constraints can be used to relocalize the query infrared and depth image sequence. We evaluate our method in two datasets captured by Kinect2. The results demonstrate very promising re-localization results for indoor navigation system in low light environments.

***Keywords*—**Indoor navigation, low light, RGB-D camera, vision based.

I. Introduction

VISUAL place recognition either indoor or outdoor is a challenging problem, which is crucial to navigation systems. For a service mobile platform at home, navigation in low light is a very attractive characteristics, for example, calling a domestic robotics to take water for drink during night, or for help when an emergency occurs in the dark. Most of existing approaches for localization or place recognition relies on visual information, and the core technique – visual search technique includes images representation, image matching, and so on. They are widely used to detect loop closure in visual simultaneous localization and mapping systems (vSLAM) and re-localization in navigation system.

As all we know, camera motions and 3D structures are encoded in 2D image sequences, so RGB frames can provide rotation and translation for each position, and the rotations and translations are used to generate 3D map by applying to depth images. However, RGB frames are only captured well in visible light. In low light environments, localization methods of moving platforms from RGB images will fail, since the low quality images are captured from RGB cameras. Fortunately, most of RGB-D cameras (e.g. Kinect2, Xtion) are equipped with infrared sensors. The infrared images can be recorded in low light environments when RGB sensors does not work well. Infrared images can also be visible, however, the infrared images are quite noisy, and features are extracted at low quality, and are less distinctive compared with features from RGB frames, which degenerates the result of re-localization. And on the other hand, the depth images provide the geometry description in space but limited description to objects, which can be used to improve the mapping. Fig. 1 shows an example captured from RGB-D camera.

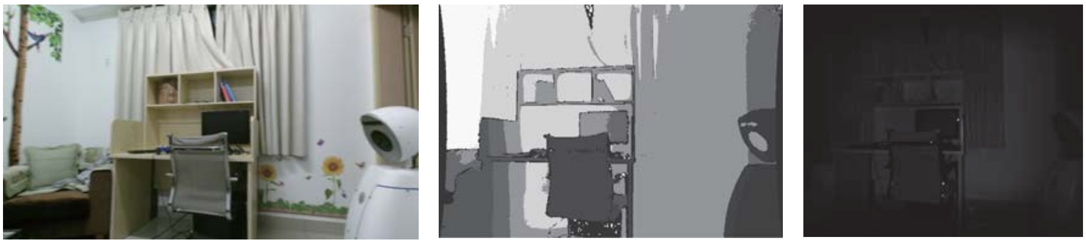
In our method, we take the infrared images and depth image pair as a query, and search the most similar images in the existing datasets to cope with indoor localization problem in low light environments. Our contribution are: 1) we try to solve the localization problem for moving platforms in low light environments; 2) we fuse the 2D infrared features and 3D depth features in the DBoW framework to improve the matching result.

II. Related Work

As the most important part of vSLAM, vision-based place recognition has been attractive to a number of researchers. Indoor localization actually is a typical place recognition in a small range of ways. A good survey of visual place recognition please refers to [1], [2]. In the survey, visual place recognition is considered from place description, mapping or retrieval, and recognition in a vast range of outdoor scenarios. Most of vSLAM systems only take 2D images as input, since 2.5D structure captured from depth sensors is hard to obtain at low-cost in outdoor environments. However, it is appropriated to take RGB-D cameras for indoor environment [3]. Generally, the valid distance of RGB-D sensors is between 0.8m and 4.0m, which is enough to indoor SLAM systems. Mur-Artal et al. created a ORB-SLAM system, and ORB features are used for all tasks like tracking, mapping, re-localization and loop closing, which achieved unprecedented performance with respect to state-of-the-art monocular SLAM systems [4]. Endres et al. presented an SLAM approach for RGB-D camera first, and was used to estimate the trajectory of a hand-held RGB-D sensors, which can generate a dense 3D map [5], and three types of features including SIFT, SURF, ORB were compared from translation and rotation in the evaluation. Engel et al. of TUM (Technical University Munich) proposed a monocular SLAM algorithm for large-scale, consistent maps of the environment [6]. Unlike most current real-time SLAM systems operating at low-level features like points, lines , or patches, Salas-Moreno et al. presented a ‘object oriented’ 3D SLAM paradigm, called SLAM++, which considered the prior knowledge that many scenes consisted of repeated structures and objects [7]. Loop closure detection is a critical procedure to obtain optimized camera trajectories and maps in SLAM systems. In [8], [9], an online loop closure detection method was proposed for large-scale and long-term operation, and a real-time solution was considered for the loop closure

Peipei Luo, Shinan Chen, Jiasheng Hao and Hong Cheng are with School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China. Yali Zheng and Peipei Luo equally contribute to the paper.

Yali Zheng is with School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China (e-mail: zhengyl@uestc.edu.cn).



a) RGB image b) Depth image c) Infrared image

Fig. 1 Example images from the RGB-D camera

detection based on a memory management [10], which kept computation time under a fixed limit.

III. Method for Indoor Navigation in Low Light

A. The Framework of Our System

Fig. 2 shows the framework of our system. First we pre-process both infrared image and depth image, the processed image pair is taken as the query image pair. Then infrared features and 3D features are extracted, and searched in a hierarchical bag-of-words tree. The output is the most similar image pair in the database. And epipolar geometry can be performed between the processed query image pair and matching image pair to localize where the query image pair are captured in the map.

B. Image Preprocessing

We observe that infrared images are only visible signals in low light from RGBD camera, however, the infrared images are very dark and low quality. And we apply feature matching methods to the raw images, they do not work well. So we consider to enhance infrared images by image processing tools. First, we perform median filtering on infrared images by using 3×3 mask to denoise. And from the histogram of infrared images, most of the pixel intensity lie between 0 and 30, and few pixels have high intensity. As all we know, histogram equalization is an algorithm to enhance images by nonlinear stretching the original intensity. In our method we take histogram equalization to transform the denoised infrared images. Fig. 5 demonstrates the comparison of before and after image preprocessing. The enhanced infrared images are used to the following steps in our method.

C. Image Representation with Infrared Feature Fused 3D Depth Feature

For each camera position, three signals (including RGB, infrared and depth frame) are captured by Kinect2. Compared with RGB images, infrared images are degenerated. 3D depth feature is not distinctive as intensity features of RGB frames, but features from depth images and features from infrared images can work together to represent one position point. We extract 1000 ORB features from each enhanced infrared image. ORB feature is proposed by Rublee et al. [11], which is binary descriptor based on BRIEF(Binary Robust Independent Elementary Features). The most important characteristic of ORB feature is at two orders of magnitude faster than SIFT, which can be implemented in real-time. A ORB feature is denoted by a 32 byte vector with 32 dims. We extract Intrinsic Shape Signatures (ISS) from each depth image [12], and use FPFH (Fast Point Feature Histogram) to describe 3D features as a vector with 33 dims.

D. Image Vocabulary and Database Generation

After we have a bunch of ORB features and depth FPFH features to represent each position from the above subsection, we quantize ORB features and ISS+FPFH depth features from infrared image and depth image into a sparse numerical vector in the framework of hierarchical bag of words [13]. We fill one zero to all 32 dims ORB features, so it can work together with 33 dims depth features to generate databases. The visual vocabulary is created offline from the training sequence by performing K-means clustering. And a hierarchical tree is built, and whose leaves are represented vocabularies in the dataset corresponding to ORB features and FPFH features. The inverse index file contains the weighted assigned for each word in the images in which they appear according to its relevance in the training set. The inverse index is easy to update when new images need to add to the dataset, and is widely used in the framework for image fast searching. Further, a direct index is used to store the features of each image, and it can be updated when a new image is added into the database. We use the DBoW2 [13] to implement and generate the image database, which is public C++ library.

E. Image Matching

Image matching is the key step to vision-based loop detection and re-localization. Given a query image pair, we need a fast matching int the database. First, we have to convert the infrared and depth image pair into a vector by the vocabulary words from the bag-of-words algorithm. We take the following formula to measure the similarity of two vector as the same in [13],

$$s(\mathbf{v_1}, \mathbf{v_2}) = 1 - \frac{1}{2}|\frac{\mathbf{v_1}}{|\mathbf{v_1}|} - \frac{\mathbf{v_2}}{|\mathbf{v_2}|}| \tag{1}$$

And the similarity score is normalized, we use the same normalization as in [13], the normalized score is defined as

$$\beta(\mathbf{v_{t_n}}, \mathbf{v_{t_j}}) = \frac{s(\mathbf{v_{t_n}}, \mathbf{v_{t_j}})}{s(\mathbf{v_{t_n}}, \mathbf{v_{t_{n-1}}})} \tag{2}$$

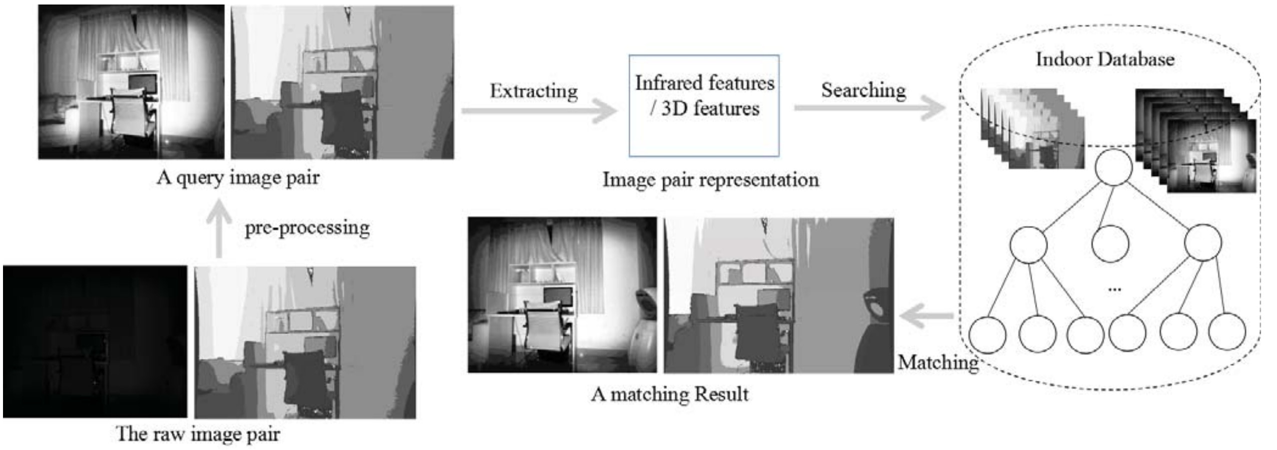


Fig. 2 The framework of our system

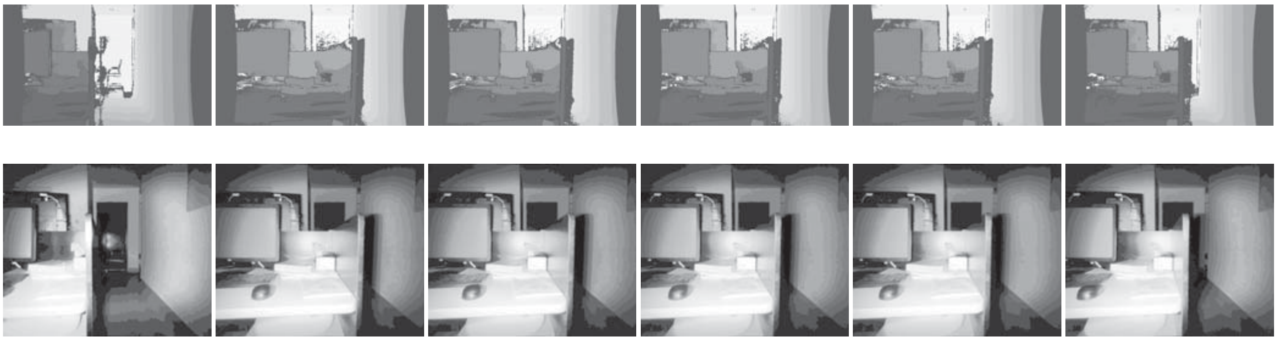


Fig. 3 Ten consecutive frames of infrared and depth images for the office dataset

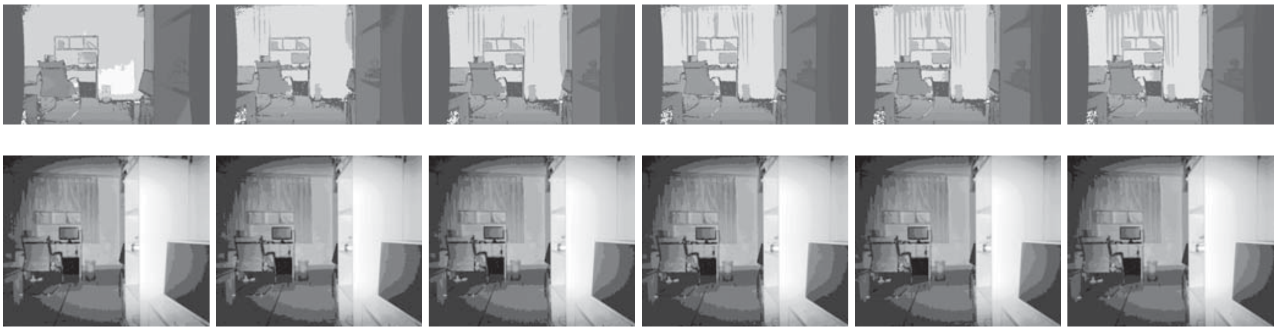
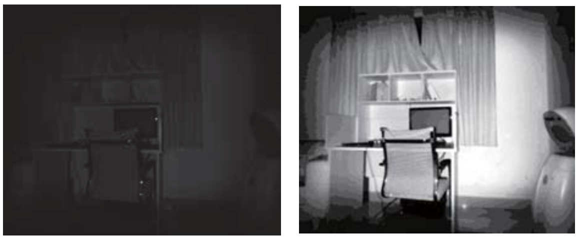


Fig. 4 Ten consecutive frames of infrared and depth images for the apartment dataset



a) The raw infrared image

b) Infrared image after image preprocessing

Fig. 5 Comparison of before and after image preprocessing

If the normalized score is greater than a threshold β, then matching is accepted, otherwise it is rejected.

IV. Experimental Results

In our experiments, we collect two databases — office and apartment in daytime by hand held microsoft Kinect2. In daytime environment, Kinect2 captures three types of images — RGB, depth, and infrared images. The RGB and depth images are used to reconstruct 3D maps to better demonstrate. We reconstruct our database in the hierarchical bag-of-words framework by features from the depth and infrared images. Then we capture the query sequences in low light, which RGB sensor of Kinect2 fails to record any useful information. Each frame of the sequences in low light is taken as the input image

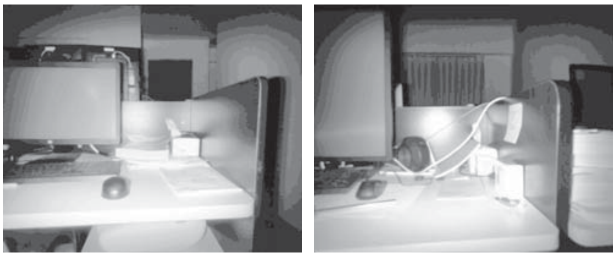


Fig. 6 A wrong match example

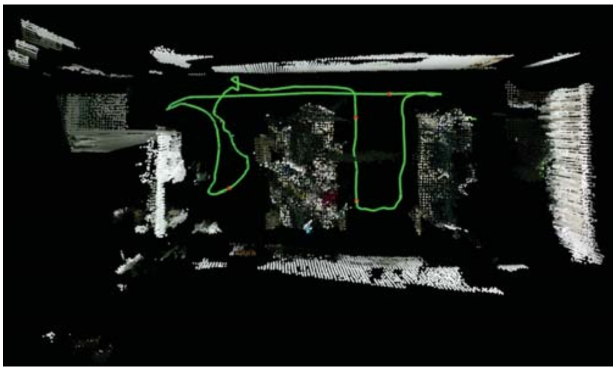


Fig. 7 Re-localization in the office dataset. Green points are the correct re-localization, while red points are the incorrect matching

pair, and is searched and matched in the databases by our proposed method. For the office database, we totally have 1340 original infrared images and depth images in the database, see Fig. 3, and generate 80679 words in the vocabulary from 2D ORB and 3D FHPH features. For the apartment sequence, we record 1045 original infrared images and depth images in the database, see Fig. 4, and generate 86283 words in the vocabulary. Figs. 7 and 8 show the re-localization results in the office and apartment datasets, respectively. The path in 3D map is the re-localization result, the green points are the correct re-localization, while the red points are the incorrect matching. We only have 6 wrong match frames for the office sequence, and 3 wrong match frames for apartment sequence (corresponding to the red points on the pathes). The wrong matches are mainly coming from the similar features, like the tables in the office are all the same. And other reason is from the narrow view field with confused corners, see Fig. 6.

V. Conclusion

Vision based indoor localization is a challenging problem due to variant indoor environments and the changing lighting, especially there is few light source during night. Many Chinese house owns the contemporary and contracted style with less decoration, which will cause vision-base methods failed easily. So different sensors will be considered in our future work to compensate for the vision based method. On the other hand, since indoor environments are well structural with lines features, especially in the corridor and so on. How to use the structural features for localization is our future work.

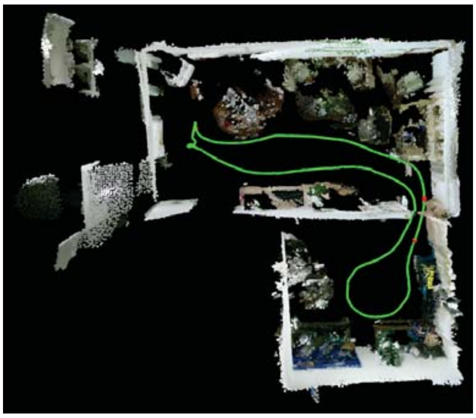


Fig. 8 Re-localization in the apartment dataset. Green points are the correct re-localization, while red points are the incorrect matching

Acknowledgment

This work is supported partially by the National Natural Science Foundation of China (No. 61603077), the Research Fund for the Doctoral Program of Higher Education of China (No.20130185120014), China Postdoctoral Science Foundation funded project (No. 2016M600732), and the Fundamental Research Funds for the Central Universities of China (No. ZYGX2016J076).

References

[1] Lowry, S., Sunderhauf, N., Newman, P., Leonard, J. J., Cox, D., Corke, P., Milford, M. J., Visual Place Recognition: A Survey, IEEE Transactions on Robotics, 2015, 31(1): 1–19.
[2] Williams B., Cummins M., Neira J., Newman P., Reid I., Tardos J. D., A comparison of loop closing techniques in monocular SLAM, Robotics and Autonomous Systems, vol. 57, no. 12, pp. 1188C1197, 2009
[3] Lee D, Kim H, Myung H., 2D image feature-based real-time RGB-D 3D SLAM, Robot Intelligence Technology and Applications, 2012: 485-492.
[4] Mur-Artal R, Montiel J. M. M., Tardos J D., ORB-SLAM: a versatile and accurate monocular SLAM system, IEEE Transactions on Robotics, 2015, 31(5): 1147-1163.
[5] Endres, F., Hess, J., Engelhard, N., Sturm, J., Cremers, D., Burgard, W., An Evaluation of the RGB-D SLAM System, ICRA, 2012.
[6] Engel J, Schops T, Cremers D., LSD-SLAM: Large-scale direct monocular SLAM, ECCV, 2014.
[7] Salas-Moreno, R,, Newcombe, R., Strasdat, H., et al., Slam++: Simultaneous localisation and mapping at the level of objects, CVPR, 2013.
[8] Labbe M, Michaud F., Appearance-based loop closure detection for online large-scale and long-term operation, IEEE Transactions on Robotics, 2013, 29(3): 734-745.
[9] Labbe M, Michaud F., Online global loop closure detection for large-scale multi-session graph-based slam, IEEE Intelligent Robots and Systems (IROS), 2014.
[10] Labb M., Michaud F., Memory management for real-time appearance-based loop closure detection, IEEE Intelligent Robots and Systems (IROS), 2011.
[11] Rublee, E., Rabaud, V., Konolige, K., et al., ORB: an efficient alternative to SIFT or SURF, IEEE International Conference on Computer Vision, 2011.
[12] Zhong, Y., Intrinsic shape signatures: A shape descriptor for 3D object recognition, IEEE International Conference on Computer Vision Workshops, 2009.
[13] Glvez-Lpez, D., Tardos, J. D., Bags of binary words for fast place recognition in image sequences(J). IEEE Transactions on Robotics, 2012, 28(5): 1188-1197.