Loop-Closure Detection based on Image Retrieval with Location Information

Sang-Eun Lee*, Jae Seok Jang*, Eun-Ju Yang*, Dae-Hwa Seo*† and Soon Ki Jung*‡
*Center for Embedded Software Technology,
Kyungpook National University
http://www.cest.re.kr
†Graduate School of Electronics Engineering,
College of IT Engineering, Kyungpook National University
‡School of Computer Science and Engineering,
College of IT Engineering, Kyungpook National University
Email: skjung@knu.ac.kr

***Abstract*—In visual simultaneous localization and mapping, there are some efforts to reduce drift, which can be produced from sensor and processing errors. One of the drift reduction techniques is closed loop detection, which can be defined as the process to determine which one is the most similar to the current keyframe from previous visited keyframes. Popular approaches to detect closed loop take a vocabulary tree based image retrieval framework using local feature descriptors. However, these local feature descriptors do not have the location information. Therefore, in the image retrieval step, incorrect images can be carried out. In this paper, the location embedded loop-closure method is proposed. The proposed approach starts from the vocabulary tree based image retrieval technique. To detect loop candidate correctly, we use database that composed of a hierarchical bag-of-words and direct and inverse indexes. We additionally store feature location vector that the visual words of the images and their associated nodes at a certain level of the vocabulary tree to the inverse index table, and indicates in which keyframe the feature was in and the location of the visual words in vocabulary tree. According to the feature location vector, the visual words are generated and employ to search keyframes. Experimental results demonstrate that our approach detects the reliable loop candidate for loop-closure step when the image has the similar appearance at the different location.**

***Index Terms*—loop-closure detection; visual SLAM; vocabulary tree**

I. Introduction

Visual simultaneous localization and mapping (vSLAM) can be defined as the process to keep a trajectory of a robot (agent) and reconstruct 3D world map using images, which are captured from the visual sensors equipped with a robot, in unknown environments. In past decades, the vSLAM comes one of the most active research areas in computer vision for robotics, because it can be customized to enormous applications in real life such as an intelligent driving system, visual navigations[1], undersea or planet exploration[2][3] and entertainments in an indoor or outdoor environment[4].

In order to improve the accuracy of tracking robot poses and reconstructed 3D environment map, the drifts due to potential errors from associated sensors or processing noise have to be minimized. The drifts produced through processing noises can be considered as unstable depth of 3D map points, incorrect matches and propagated errors. The depth ambiguity can yield incorrect robot pose in the tracking process. However, this error can be resolved through multiple view triangulation[5] or optimization of a parameterized 3D map point[6][7]. Another one is incorrect corresponding points, which can be produced from feature descriptor matching. In many localization applications, the incorrect matches in feature matching step are rejected as robust estimation for outlier removal[8]. The other is the propagated error throughout tracking and mapping processes. In spite of the efforts to estimate an accurate robot pose and build a high precision global map using optimization, a little mount of processing error cannot be avoided. These little mount of errors can produce the drifts over time. In case of structure from motion (SFM), this propagated error can be solved the global bundle adjustment[7][9]. However, the global bundle adjustment requires huge mount of resources and high computational complexity. Therefore, the local bundle adjustment or loop-closure detection algorithms are more suitable in real-time applications like vSLAM or visual odometry.

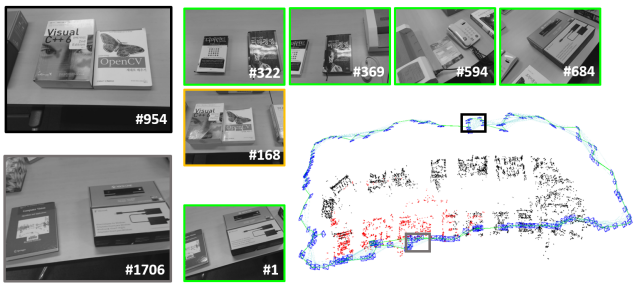

Fig. 1. **The comparison of image retrieval between a conventional and proposed method.** The green images indicate retrieved images by the proposed method. A loop candidate image for loop-closure is selected successfully. Whereas the traditional method demonstrated as the yellow box, loop-closure performs incorrectly because it considers only the matching image.

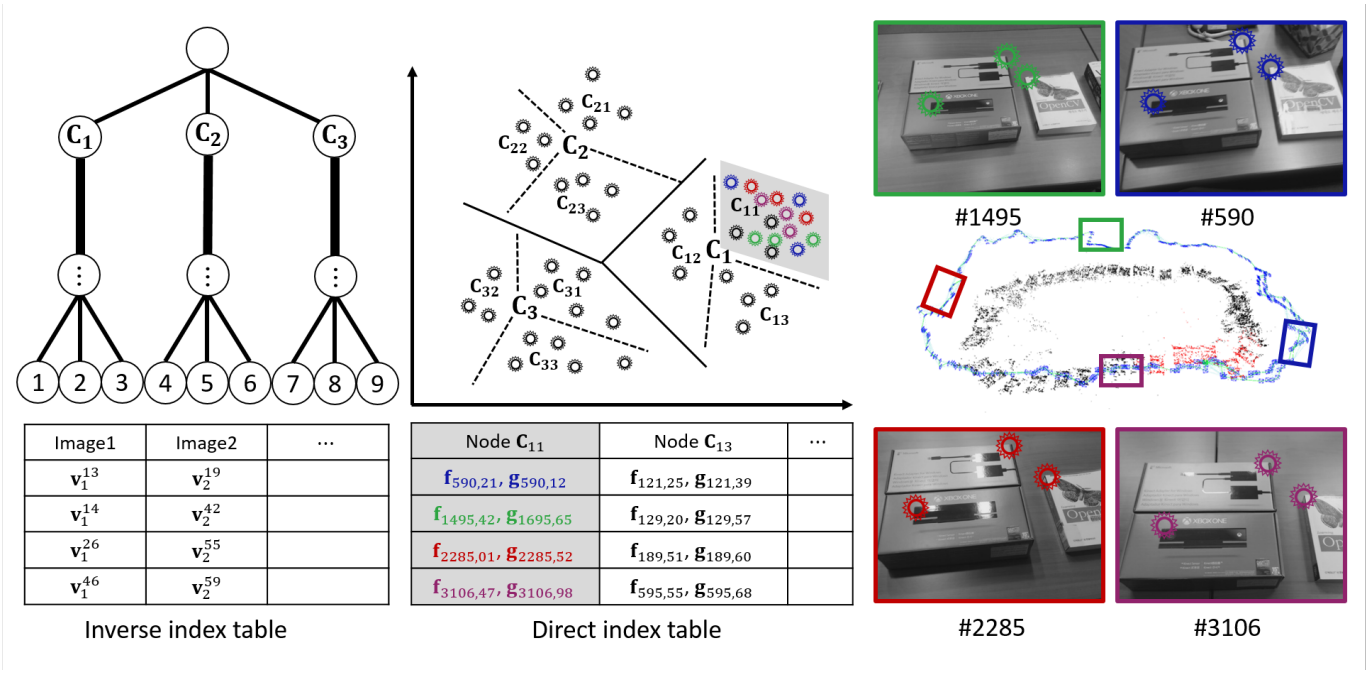

Fig. 2. **Scheme of proposed method.** Visual vocabulary tree with direct and inverse indices. The inverse index is modified by adding location information to classify correctly as loop candidates in loop-closure step.

Although there are many strategies to reduce the propagated error, one of key processes is to recognize the previous visited place. This place recognition process in vSLAM can be defined as loop-closure detection. Therefore, in this paper, we focus on the vocabulary tree based loop-closure detection. To be specific, our proposed approach is able to cope with the incorrect loop detection by image retrieval as shown in Fig. 1. The rest of this paper is organized as follows. In Section II, the related studies to detect loop-closure are described. In Section III, the proposed loop-closure detection is explained. In Section IV, the performance of the proposed methods with qualitative analysis is represented. Finally, we conclude the proposed method in Section V.

II. Related Works

In this section, we summarize the loop-closure detection techniques from the previous visited keyframes.

In order to detect a closed loop in a global environment, bag-of-words (BoW) based approaches using visually distinctive local feature descriptors, for example SIFT[10], SURF[11], BRIEF[12], ORB[13] and so on, appeared on captured images are proposed in many researches in vSLAM[14][15]. Traditionally, the features for loop closing works use SIFT or SURT because they have robustness to lighting, scale and rotation changes. In addition binary features that have the advantages in compact, less memory and fast such as BRIEF and ORB for loop-closure detection are proposed[17]. BoW approach includes that transforming each keyframes to the quantized vector (visual words) using the local feature descriptors, finding the corresponding keyframes to the current keyframes using the image retrieval method, and fusing the closed loop from most similar keyframe. As described, each keyframes are represented by visual words from a tree, this method is helpful for rapid comparison between each keyframes in Database and the current keyframe rather than feature descriptor matching. The commonly used the quantization methods are KD-tree[10], approximated K-Means[18] and vocabulary tree[19].

In order to recognize the previous visited place, there are two main approaches, which are direct matching[20][21] and image retrieval based approach[15][16][22][23]. In the direct matching, the tree for global localization is built using the descriptors associated with reconstructed 3D map points. When a query image is given, the system finds the corresponding 3D map points to the extracted 2D feature points on the query image. From the given 2D-3D corresponding points, the camera pose is estimated. Therefore, the direct matching is more suitable to global localization. The image retrieval based approach takes an off-line learning process to build a tree using huge mount of images. A Database can be constructed by the visited keyframes during on-line vSLAM. When a query image is given, the generating DB can carry out the top-k ranked images. From these carried out images, the system can recognize that whether the query image is the revisited place or not. For recognizing step, the feature matching between the query image and the retrieved images needs the additional computation time. For this reason, the direct and inverse indexing, which allow direct accessing from a tree to keyframe and vice verse, methods are proposed[23].

In spite of the high speed keyframe matching performance, there is ambiguity in retrieving keyframes because the local

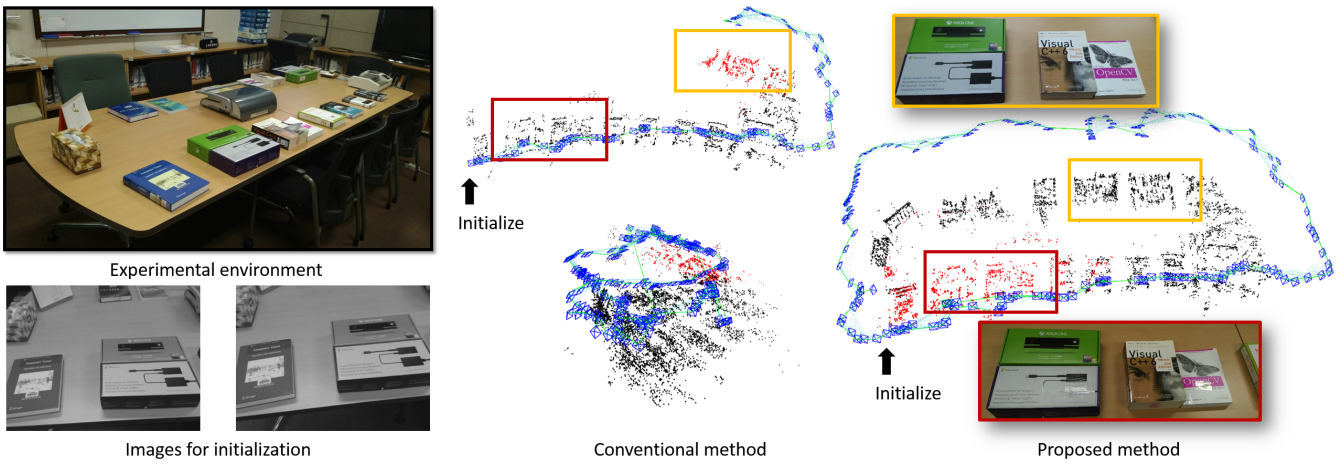

Fig. 3. **Experimental setup and qualitative result.** The left images show experimental environment and the initialization step for tracking. The right images demonstrate the result of a qualitative comparison the between conventional and proposed method in the image sequences where the similar appearance occurs at one time.

feature descriptors, which are not included the location information, are only used in the building a tree step. Therefore, we propose the visual loop-closure detection method based on image and feature position in image retrieval step.

III. Image Retrieval-Based Loop-Closure Detection

For loop-closure detection, the bag-of-words, one of appearance-based scene matching technique based on a visual vocabulary tree is employed. The visual vocabulary module with bag-of-words is based on DBoW2[17]. The vocabulary tree is created with ORB descriptor in offline by numerous indoor and outdoor images. And we use a database to retrieve images close to any given one as shown in Fig. 2. The vocabulary words are the leaf node of the tree. The inverse index stores the weight of the words in the image which they appear. The bag-of-words vector v_i^j denotes the value of visual word j for image i. In direct index table, the feature vector $f_{i,k}$ denotes the feature k for image i and stores the features of the images and their associated nodes at certain level of the tree. The right image in Fig. 2 indicates that similar images are sparsely located in image sequences. Consequently, the visual words that they have a similar appearance are an identical source of certain level in the tree. In this regard, The traditional visual vocabulary tree for image retrieval is likely to occur a significant problem that process carries loop-closure out since the camera location is not globally considered. This has the limitation of not including keyframes viewing the same scene at the other location in a different time. To deal with this issue, each word of bag-of-words in the database is embedded with a location information where it is located on a global map. In detail, the location of a camera and bag-of words in a keyframe are estimated by tracking process and build their relationship based at a particular node level of the tree. We use feature position vector $g_{i,j}$ that denotes the words and location information and their associated nodes at certain level of the tree. It is stored with feature vector in direct index table. In other words, we store the feature position vector that expresses the location of the visual words in vocabulary tree and in which keyframe the feature was. Firstly, we compute the Euclidean distance between query keyframe and all keyframes, and sort out keyframes according to the distance value. The visual words in bag-of-words are totally gathered from particular nodes that query keyframe and sorted keyframes have in common. According to the location information and particular node level, the bag-of-words employ to search keyframes that sharing the enough visual words. Finally, even though visual words that they have similar appearance are close in the node, we are able to correctly classify them at exploring step in vocabulary tree. Particularly, the loop-closure candidates base on image retrieval approach are exactly identified by the result in consideration of neighboring camera poses.

IV. Experimental Results

We have experimented in a laptop with Intel i7-4270HQ processor and 16GB RAM. Our algorithm is implemented in C++ Robot Operating System (ROS). The proposed method is qualitatively evaluated by a monocular sequence as shown in Fig. 3. There are things on a desk for tracking easily such as books, a calendar and boxes. The monocular sequence presents a loop-closure around the desk in an indoor environment. It could obtain that we could compare to handle the loop-closure test between the traditional and the proposed method. Our algorithm is based on DBow2[17], place recognition module and is qualitatively compared with it. The conventional method in Fig. 3, image retrieval for loop-closure detection is failed when the same objects appeared at the different location. As shown in Fig. 1 the retrieved image at the different location has the most similar image which has been visited before. On the contrary, the retrieved image sharing enough the visual words,

having the most similar appearance and located in neighboring the query image by the proposed method is obviously detected as a loop candidate image. The loop candidate provide to perform loop-closure successfully.

V. Conclusion and Discussion

The object of our study is to improve the performance of vSLAM. In order that, loop-closure detection is crucial for enhancing the robustness of the algorithm because the more poses are incrementally increased, the more the drift becomes larger. To deal with this, we have loop detection based on image retrieval by the vocabulary tree approach included the location information. As shown in Fig. 3, the result shows that keyframes viewing the same scene at the other location in different time is not detected as the loop candidate to perform loop-closure process. This is because, the visual words to search keyframes are already removed by feature location vector at their certain node levels before exploring the vocabulary tree. As a result, our algorithm is able to detect reliable loop candidate for loop-closure step even if the image has the similar appearance at the different location. In the future, we could explore that our algorithm working with vSLAM and is applied to the mobile platform in the real world.

Acknowledgment

Following are results of a study on the "Smart Hospital Management Service Based on IoT and RTLS" Project, supported by technical commercialization of Daegu Innopolis.

References

[1] Mercedes Benz, Mercedes S Class AUTONOMOUS DRIVING DEMO Intelligent Drive, *"https://www.youtube.com/watch?v=ROhT00vSvHk"*, 2014.

[2] H. Moravec, Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover, Ph.D. Dissertation, *Stanford University*, Stanford, CA, 1980.

[3] S. Lacroix, A. Mallet, R. Chatila and L. Gallo, Rover Self Localization in Planetary-like Environments, *Proceedings of International Symposium Artificial Intelligence, Robotics, and Automation for Space*, 1999.

[4] Quill Pen Studio, AR Invaders, "$https : //www.youtube.com/watch?v = Vl_L0ilxqzs$", 2012.

[5] M. Rumpler, A. Irschara and H. Bischof, Multi-View Stereo: Redundancy Benefits for 3D Reconstruction, *Photogrammetric Engineering and Remote Sensing*, vol. 76, no. 10, pp. 1123-1134, 2010.

[6] L. Matthies and S.A. Shafer, Error Modeling in Stereo Navigation, *IEEE Journal of Robotics and Automation*, vol. 3, no. 3, pp. 239-248, 1987.

[7] N. Snavely, S.M. Seitz and R. Szeliski, Modeling the World from Internet Photo Collections, *International Journal of Computer Vision*, vol. 80, no. 2, pp. 189-210, 2008.

[8] M.A. Fischler and R.C. Bolles, Random Sample Consensus: a Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography, *Communications of the ACM*, vol. 24, no. 6, 1981.

[9] S. Agarwal, N. Snavely, S.M. Seitz and R. Szeliski, Bundle Adjustment in the Large, *European Conference on Computer Vision*, 2010.

[10] D.G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[11] H. Bay, A. Ess, T. Tuytelaars and L.V. Gool, Speeded-Up Robust Features, *Journal of Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346-359, 2008.

[12] M. Calonder, V. Lepetit, C. Strecha and P. Fua, BRIEF: Binary Robust Independent Elementary Features, *Proceedings of the 11th European conference on Computer vision: Part IV*, 2010.

[13] E. Rublee, V. Rabaud, K. Konolige and G. Bradski, ORB: An Efficient Alternative to SIFT or SURF, *Proceedings of International Conference on Computer Vision*, 2011.

[14] M. Labbé and F. Michaud, Appearance-Based Loop Closure Detection for Online Large-Scale and Long-Term Operation, *IEEE Transactions on Robotics*, vol. 29, no. 3, pp. 734-745, 2013.

[15] A. Angeli, D. Filliant, S. Doncieux and J.A. Meyer, Fast and Incremental Method for Loop-Closure Detection Using Bags of Visual Words, *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1027-1037, 2008.

[16] A. Angeli, S. Doncieux and D. Filliant, Real-Time Visual Loop-Closure Detection, *IEEE International Conference on Robotics and Automation*, pp. 1842-1847, 2008.

[17] Galvez-Lopez, Dorian and Tardos, J. D., Bags of Binary Words for Fast Place Recognition in Image Sequences, *IEEE Transactions on Robotics*, 2012.

[18] J. Philbin, O. Chum, M. Isard, J. Sivic and A. Zisserman, Object Retrieval with Large Vocabularies and Fast Spatial Matching, *IEEE International Conference on Computer Vision and Pattern Recognition*, 2007.

[19] D. Nister and H. Stewénius, Scalable Recognition with a Vocabulary Tree, *IEEE International Conference on Computer Vision and Pattern Recognition*, 2006.

[20] Y. Li, N. Snavely and D.P. Huttenlocher, Location Recognition using Prioritized Feature Matching, *European Conference on Computer Vision*, 2010.

[21] T. Sattler, B. Leibe and L. Kobbelt, Fast Image-Based Localization using Direct 2D-to-3D Matching, *IEEE International Conference on Computer Vision*, 2011.

[22] R.M. Artal, J.M.M. Monitel and J.D. Tardós, ORB-SLAM: A Versatile and Accurate Monocular SLAM System, *IEEE Transactions on Robotics*, vol. 31, no.5, pp. 1147-1162, 2015.

[23] R.M. Artal, and J.D. Tardós, Fast Relocalization and Loop Closing in Keyframe-Based SLAM, *IEEE International Conference on Robotics and Automation*, 2014.