

An incremental loop closure detection method based on depth information for indoor dynamic scenes

Zhongliang Deng, Qi Wu, Jichao Jiao, Yuan Sun, Yaokai Mo, Cheng Zhang

Beijing University of Posts and Telecommunications

Abstract—We present a RGB-D based place recognition algorithm methods for large-scale dynamic indoor environment. In contrast with other methods, our methods focus on the scene that dynamic objects are remarkable and avoid the unnecessary calculate expenses on the whole image sequences. In order to prohibit the influence of dynamic object, the depth information was applied to discriminate the image foreground information and filter it out. We conducted our methods on the public dataset and tested on the real environment. The results demonstrate that the proposed methods can effectively improve the loop-closure system in indoor dynamic environments.

Index Terms—Dynamic Scene, Loop Closure, SLAM, Place Recognition.

I. INTRODUCTION

The Loop Closure detection is an essential component of SLAM (Simultaneous localization and mapping) system, which makes it possible to associate a previously visited location with another one recently visited.

Visual place recognition plays a core part in loop closure system [1]. RGB-D cameras, such as kinect[2], have become a standard sensor for robots today. The RGB-D camera helps robots to identify the visited places and eliminates the cumulative error in the pose calculation[3]. However, as time elapsed, the appearance of a place is not immutable all the time. The challenges of lighting and viewpoint changes are all likely to be the factor of failure identification. Moreover, similar appearance in multiple places may make the robot misunderstand where its original position is. To simplify the problem formulation, there are some excellent pre-work studied on static environment[4], [5], [6], which will be demonstrated in Section II. In this paper, we will focus on the indoor dynamic scenes.

The vision robots record the scenes that relay on the captured image. The moving objects, such as human, animals and vehicles, are all outliers to the recognition system[7]. It is natural that people are attracted by the characteristics of the scene rather than concentrating on these noises. However, robots cannot distinguish the outliers, which are also called foreground information. Our aim is to make robots handle the indoor dynamic scene more intelligently. In other words, the dynamic objects, which is a disturbing portion of image, can be handled by viewing them as noise. There are some solutions to eliminate the noise information when objects are remarkable.

Inspired from this idea, we intend to identify place more intelligently. We use the depth information to help us figure out the relationship between moving objects distance and the image portion. Different strategies are adopted to eliminate the

impact when it comes to different distance. Furthermore, we incorporate our dynamic objects elimination algorithm into a loop-closure detection system. The elimination approach plays pre-processing stage, which filters out data was related to moving objects. We tested our approach by using a public RGB-D dataset and demonstrated it in various indoor dynamic environments. The experimental result demonstrates that our approach is able to improve the detection accuracy.

The remainder of this paper is organized as follows. The Section II reviews related work about loop closure detection. Section III explains an overview of our place recognition system and the combination of loop closure detection. Section IV discusses the experimental results. The conclusion and future work will be discussed in the last section.

II. RELATED WORK

In this section, we review the algorithms which are designed for the dynamic environment. Particularly, we draw our attention to the feature selection, background modeling and the loop closure detection.

A. Describe places in changing environment

There are two different approaches to depicting the scene[8], [9], [10], [11]: The first is to use the global descriptors to present the whole images such as the Gist[8], [9]. Considering that the global features do not contain about the image contents, it is obviously not suitable for building dynamic environment signatures. The whole image descriptors used in the SeqSLAM SLAM system demonstrate robustness against the environmental change[12]. However, it also suffers from other problems, such as the sensitivity to viewpoint change.

Local feature descriptors are designed to extract interesting or notable parts of the images. In order to find the condition-invariant description of the dynamic environment, the local feature descriptors are designed to be scale, rotation and illumination-invariant[1]. None of tested features is robust under all conditions. However, the SIFT features are more robust to the SURF features in the lighting changing condition[10]. We adopt the SIFT features to make our algorithm more robust in the indoor dynamic environment.

B. Deep learning features in the dynamic scene

The deep learning methods recently are used for extracting robust features for place recognition in changing environments[13], [14]. The utility of CNNs takes advantage of the network learning ability to extract the generic features

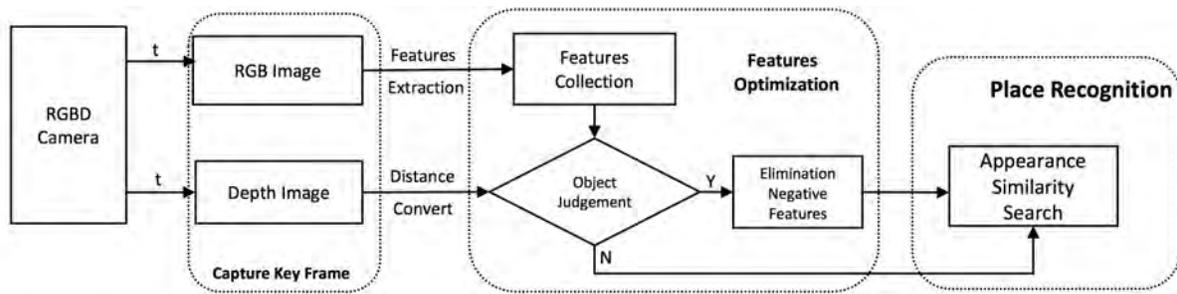


Fig. 1. The overview of our system. The algorithm extracts features from the color image and uses the depth information to eliminate the outliers in the dynamic environment, which helps the algorithm be more robust in the dynamic environment.

for place recognition. Different network layers can optimize features to make them more robust to against the visual appearance changes. However, the deep feature can not be adopted due to its poor real-time and generalization ability. In order to use our algorithm in real life scene, we will adopt the traditional SIFT feature which is more robust to the viewpoint changes.

C. Background model in dynamic scene

The previous section demonstrates the way how to find out the invariant points in the dynamic scene, while there are other methods trying to form the static background model. It seems that the background initialization is an adaptive program for this task. According to Pierre-Marc Jodin[15], the definition of initializing a background model may be defined as follows: the static camera captures the environment scene with the moving objects on top of it. The purpose of this algorithm is to try to recover the background without foreground objects. Deok-Hwa Kim[16] devises the background model-based SLAM according to this theory. However, applications designed for indoor environment, the combination of camera-induced motion and object motion will cause the algorithm to fail.

D. Loop closure detection

In SLAM approach, the loop closure detection plays a role of data association to remind the robot whether it has get back to a previously visited area or not. Independent of pose estimation and location process, the outcome of detection system affects the robot to correct the estimated pose from the cumulative error. The appearance-based methods for this task have achieved great results, a representation of this is the FAB-MAP system[4]. The system with an omnidirectional camera shows a perfect result in trajectories 70KM and 1000KM in length with no false positive. However, when the system faced with similar structures in different environments, the robustness of the system will decrease. Angeli's work employs an incremental fashion of bag-of-words method to build a coherent structure between corresponding images[5]. Differs from the probability methods, with the idea of the temporal consistency, Dorian[6] uses the bag of binary words to build the DBoW2 which exploited by the famous ORBSLAM system. Although using the binary form of pretrained vocabulary

to satisfy the time limits, real-time constraints also limit the performance in online large-scale environment. Therefore, the RTABMAP system was emerged[17]. There are also some process[18][19][20] made in the bag-of-words algorithm to make the loop detection system have better efficiency. However, the dynamic scene is a disaster environment for these algorithms.

Although there are some work[12][14] for these annoying scenes, the SeqSLAM[12] is developed based on the similarity of image sequences and the work of Silvia[14], trying to use the high level features extracted by CNN, and the study make the detection system more robust to seasonal changes. While there is another challenge in the dynamic scene, the negative features brought by the moving object in the indoor scene. The BaMVO system[16] tries to solve these problems by recovering the background model. However, as we mentioned above, the ego-motion of camera drive decreases the system robustness. There are some highlights in Yuxiang Sun's work[21], which eliminates motion by calculating the camera's ego-motion through depth images. To our regret, these algorithm has nothing to do with the scene when moving objects are remarkable and real-time ability limits the system works online. In our work, we use the depth information to determine whether there are moving objects in the scene and eliminate the negative features to improve the detection system accuracy which will be discussed in the next section.

III. FRAMEWORK OVERVIEW

Enlightened by the observation of outdoor scene, we build our own place recognition system for dynamic scene with the wandering moving objects.

In the study of dynamic scenes, the traditional thinking patterns of object tracking and background information construction should be abandoned. We analysed the relationship of distance between objects and robots with the recognition system failure, we put more focus on the infected scene where the impact of moving objects can not be overlooked. Therefore, with the help of depth information, we discard all the negative features in a certain location to improve the detection accuracy.

Furthermore, in order to test the practical performance of our algorithm, we integrate our algorithm in a Bayesian-based loop closure detection system to operate in a real environment. Figure 1 shows an overview of our system.

A. Preliminaries

It is known to all that the region of object in the image get larger and larger when the object advances towards the camera. Figure 2 reveals this phenomenon. From the experiment, we can figure out the relationship between the proportion of object in the image and the distance between the robot and unexpected people in the environment.

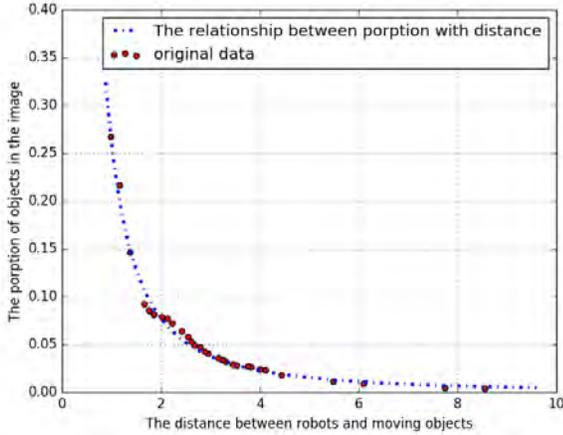


Fig. 2. The blue line demonstrates that the porption of pedestrian in the image has an inverse proportion relationship with the distance between robots and moving objects.

As figure 3 shows, we use the pinhole camera model to analyze how the 3D points in the environment are converted to the image point through the RGBD camera.

Where the 3D point p in homongeneous representation is $p = (x, y, z, 1)^T$, f_x, f_y are the camera focal lengths and o_x, o_y are the camera center coordinates. To simplify the cumbersome calculations, we roughly thought that the shape of objects in the image is rectangular. Set the top left corner point of region as P_{LU} and the right down corner point as

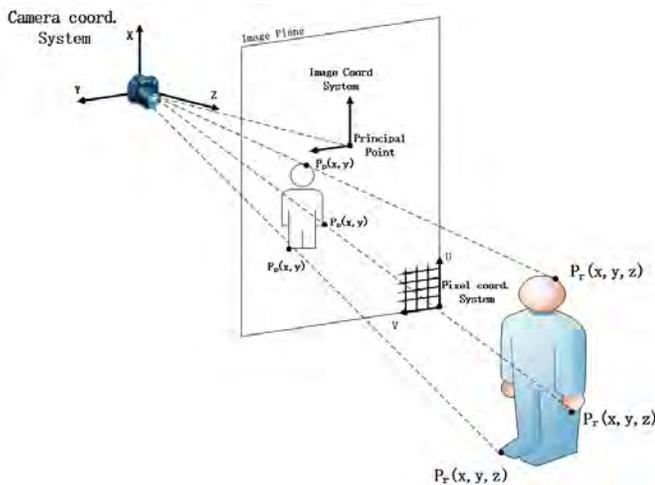


Fig. 3. This figure shows that how to project the 3D objects into the image pixel

P_{RD} . So the point represented as:

$$x = \pi(p) = [\frac{xf_x}{Z} + o_x, \frac{yf_y}{Z} + o_y]$$

and the area of the object is

$$A = \frac{X_R - X_L}{Z} \times \frac{y_R - y_L}{Z} = k \frac{1}{Z^2}$$

The formula shows that the area of object has the inverse ratio to the square of the distance Z square and the scale factor k is related to object real size.

B. The Features

According to the previous section, we would adopt the traditional SIFT feature, which is more robust to the viewpoint and rotation changes. According to Krystian Mikolajczyk[22], the characteristic is irrelevant with the image resolution. However, the feature points extracted from finer scales are more than the coarser scale. For the place recognition task, the pedestrian wandering in the indoor environment is disturbing noise. The features located on the pedestrian are called negative feature.

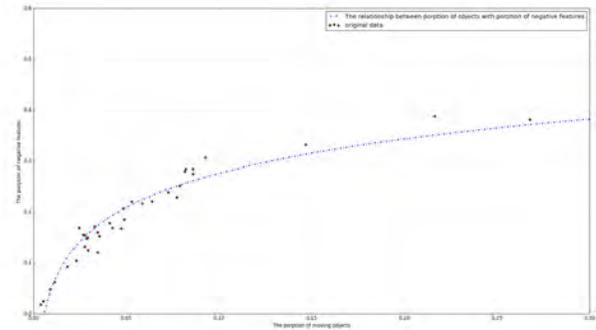


Fig. 4. The porportion of negative features with the distance between robots and pedestrian

Figure 4 demonstrates the relationship between the proportion of negative features and the porportion of object in the image. We discover that when proportion of objects is about 8%, which means the object is 1.8 meters away from the camera, the explosion of negative features will cause the place recognition system to failure. Therefore, we should eliminate the impact of negative features, which will be detailed in the next section.

C. The elimination of Negative Features

From figure 4, we know that feature has a close relationship with distance. The negative features are always gathered within a certain distance. Therefore, we use the depth information to help us determine if there are moving objects in the scene. Just like figure 5. There are two implicit assumptions. The first is that the robots should advance along a straight line steadily. If this assumption is not true, the distance between the robots and various objects will change dramatically and the features' location will be drifted away. Another implicit assumption is

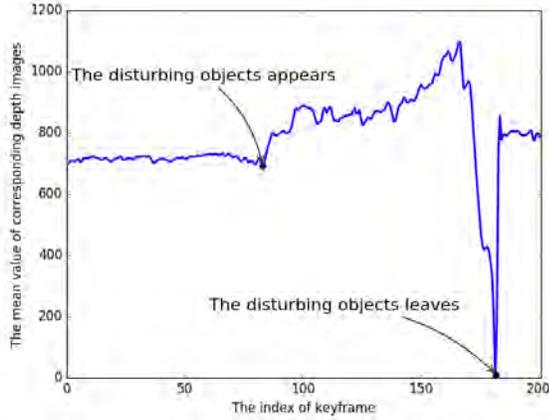


Fig. 5. The variety of the mean depth value when moving objects appear in our environment.

that the environment will have enough features. If the scenes are featureless, our robots will be confused in front of the white wall and do not know what to do next. Fortunately, these issue are not critical in most indoor environments. Our technique extracts the mean values of a sliding window over the collected depth information stream used to differentiate the negative features and the postive features.

The dramatic changes of the value rate shows that there are some changes in the environment. One change is that the robots moving from one area to another and another is the moving objects appear in the indoor environment. For the first cases, the change will not last long when robots move from one room to another. We calculate the current mean depth value of the current scene and compare it with the adjacent keyframes. If fluctuations occur within only a few frames and return to stability quickly, then it cannot be proved that there are dynamic targets in the current scene. For the second case, the appeared disturbing objects will interfere the depth value and last a long time until the objects disappear from the sight.

The moving objects appear in front of the robots or come from behind the robots. Like figure 6 shows, when pedestrians come from a distant place, the difference between the neighboring average depth values is much higher than the normal one. From the last section, the distant objects will not bring an impact to the detection process. When pedestrian approach, the number of negative features will increase rapidly. The impact should not be underestimated when pedestrian is 2 meters away from the robots, meanwhile the magnitude of changing rate is amazing and the $V_{current}$ is also higher than normal. The features we extracted from 1.5m to 2m distances will be seen as negative features, which means it will be discarded. As the objects get closer and closer, the mean depth value will also get decrease. When the $V_{current}$ is less than normal value and at a high changing rate, we realize that the perspective of robots is completely blocked by the moving objects and the negative features of 0m to 0.8m will also be abandoned.

On the contrary, when objects come behind the robots, like figure 7, the $V_{current}$ decrease rapidly to a smaller value at a large changing rate. As in the above case, we can easily

remove the negative features and improve the recognition rate with the change rate of mean depth value. The steps of methods are summarized in Algorithm1.

Algorithm 1 The Elimination of Negative Features.

Input: The captured RGB Image; The captured Depth Image;
The global variable V_{last} ; The global variable $COUNT$;
The global variable $Sign_{last}$

Output: The optimized SIFT Features

- 1: Extracting the SIFT Features from the RGB Image;
 - 2: Calculating the average pixel value of depth image $V_{current}$
 - 3: $V_{difference} \leftarrow V_{current} - V_{static}$
 - 4: $Sign_{current} \leftarrow sign(V_{difference})$
 - 5: **if** $V_{difference} > threshold_{depthdifference}$ **then**
 - 6: **if** $COUNT > 5$ and $Sign_{current} == Sign_{last}$ **then**
 - 7: **if** $V_{current} > meandepth_{normal}$ **then**
 - 8: Eliminate the features between 1.8m to 4.5m;
 - 9: **else**
 - 10: Eliminate the features between 0m to 0.8m;
 - 11: **end if**
 - 12: **else**
 - 13: $COUNT ++$;
 - 14: **end if**
 - 15: **end if**
 - 16: $V_{last} = V_{current}$;
-

D. Loop Closure detection

Dedicated to make our algorithm implemented under real-time constraints and operated in large-scale indoor environment, we integrated our method into a probability based loop-closure detection system. The system is composed of two modules. The more easily revisited locations are stored into the working memory(WM) and others left in the Long-term memory(LTM). Inspired by the Bayesian filter frame work, our algorithm keeps track of loop-cloure hypotheses by evaluating the similarity between the current Location L_t and the previously visited location L_i stored in the WM. The S_t represents the loop-closure hypotheses at time t . That event $S_t = i$ implies that the current image I_t is similar to past image I_i . The event $S_t = -1$ represents there are no viewpoints X_i and X_t are close, which means there is no loop-closure detected at time t. In order to find the corresponding past image when a loop closure occurred,

we need to estimate the full posterior, $p(S_t|I_t)$ for all states $i = -1, \dots, t - p$. According to the Bayes' rule and Markov assumption, the posterior can be decomposed into

$$(S_t|L_t) = \underbrace{\eta p(S_t|L_t)}_{Observation} p(S_t|L_{t-1}) \quad (1)$$

The observation model $p(L_t|S_t)$ is derived from a likelihood function $\zeta(S_t|L_t)$, current location L_t is compared with the

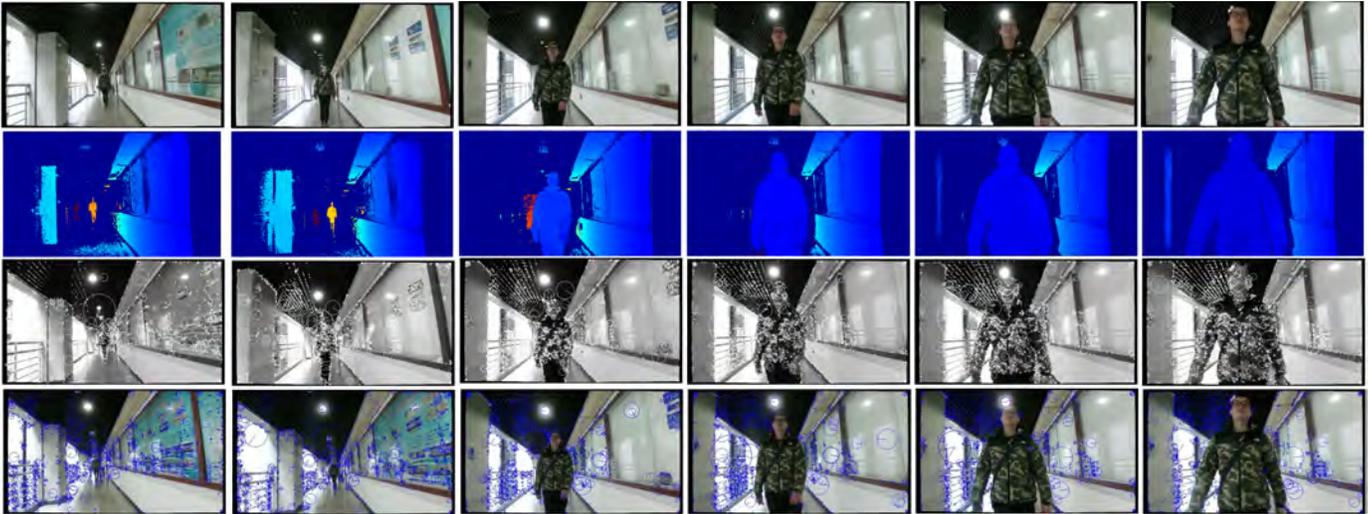


Fig. 6. This figure shows that how the negative features change when pedestrain advance to the robots. The row one is the RGB picture captured by the camera, and the row two is the changing of depth value. The row 3 shows the original features and the row 4 shows the negative features which is bothering us and the row 5 shows the optimization of sift features throughing the depth values



Fig. 7. This figure shows that how the negative features change when pedestrain advance to the robots. The row one is the RGB picture captured by the camera, and the row two is the changing of depth value. The row 3 shows the original features and the row 4 shows the negative features which is bothering us and the row 5 shows the optimization of sift features throughing the depth values

matched words to evaluate the score with each loop closure state $S_t = j$ where $j = 0, \dots, t_n$.

The $p(S_t|L_{t-1})$ indicates the belief at time t, system can get the actual evaluation for the posterior. Decomposed as follows:

$$\sum_{i=-1}^{t_n} \underbrace{p(S_t|S_{t-1})}_{Transition} = i p(S_{t-1} = i|L_{t-1}) \quad (2)$$

The belief of next loop closure can be separated by the recursive part of the filter and the transition model that used to estimate the robot's motion between t and t-1.

IV. EXPERIMENTS

In this section, we demonstrate the experimental results and discussions. In the first part, we use our self-generated

dataset to qualitatively illustrate our negative signature elimination method in a typical dynamic environment. In the second part, we integrate the proposed elimination approach into the RTABMAP system, which is an appearance-based Loop closure detection system. We also show results from a hand-held Kinect. Note that although the tremble affects the experiment results, the slight hand shaking will keep the system work correctly. A PC with an Intel i7 CPU and 8GB memory is our experiments platform. The hand-held camera we selected is Kinect V2, which provides the 960×540 RGB-D resolution.

A. The negative features elimination

In this section, we generate different RGB-D datasets with dynamic objects. The first case is a long hallway with a walking person as the moving object. We simulate the movement

of pedestrians in the indoor environment. One is the pedestrian approaches from a distance and another is the pedestrian comes from the back of the RGB-D sensor. The two image sequences are recorded in two different hallways respectively.

Figure 6 and figure 7 qualitatively show the experimental results using these two sequences. As mentioned above, our algorithm is dedicated to eliminate the negative features gathered within a certain range. Although it may misclassify some positive features as negative, the optimization significantly helps the loop closure system to avoid the interference brought by the moving pedestrians. The number of negative features that we can eliminate is presented in table 1:

TABLE I

THE PROPORTION OF NEGATIVE FEATURES WHEN ROBOTS GET CLOSER TO PEDESTRIAN

distance(m)	number of negative features	proportion of negative features
8.5	97	2.9%
6.1	144	5.0%
3.7	564	18.3%
1.6	975	27.1%
0.9	1406	39.3%

We also test our algorithm in our dormitory and our laboratory. Like figure 8 and figure 9. Because they are too small to form a loop. We can only see the result of feature extraction. Features are almost located on the background, which means it can neglect the disturbing noise in the dynamic scene and improve the loop closure detection accuracy.

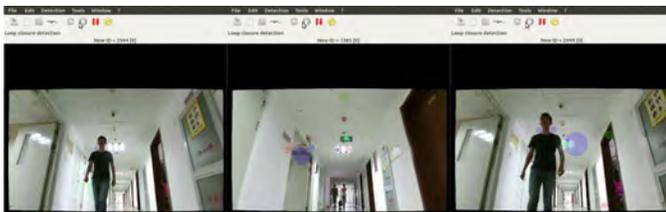


Fig. 8. We use the kinect2 to capture the image in the our dormitory. The student come back to the dorm is the disturbing noise to us. Through our algorithm, features almost locate on the background.

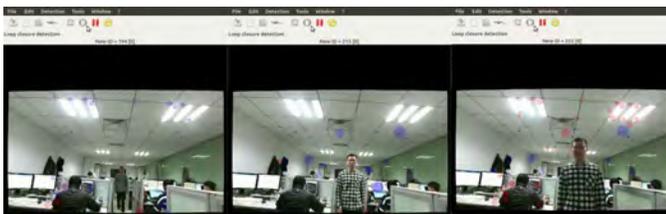


Fig. 9. We also use the kinect2 to capture the image in the our laboratory. Classmates go in and out of the laboratory are all dynamic objects need to be neglect. Through our algorithm, features almost locate on the background.

B. Loop Closure Test

In this section, we integrate our algorithm into the RTABMAP Loop closure detection system and test its detection ability.

1) *Community Datasets:* Although there is no community loop closure detection dataset in RGBD format, we build a series of depth images with depth of 0 to verify our algorithm ability.

We test our algorithm on the Lip6Indoor dataset. The result shows that our algorithm is not worse than the original RTABMAP algorithm, just like figure 10 shows. The parameters used for the experiment is: $T_{rehearsal}$ to 20%, STM size to 30, T_{loop} to 12%. Figure 10 shows that our algorithm is as good as the RTABMAP system and the execution time of our algorithm only costs 144ms on each frames, which means it can also maintain the real-time performance.

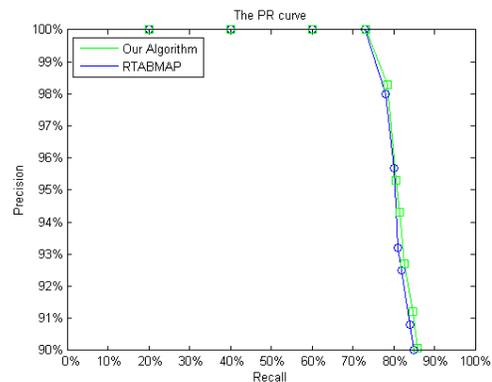


Fig. 10. The trajectory of the TUM dataset and the red lines represent the correct loops.

2) *Our Indoor Environment:* We test our algorithm in our dataset. The image sequences we captured by KinectV2 is in the second floor of the main teaching building. It is a large-scale circular indoor environment with two long similar corridors. There are different posters on the wall to help us to distinguish the two corridors. We regard the classmates who shuttle between the two corridors as moving objects. We compare our algorithm with RTABMAP on this dataset, the result of RTABMAP is represented as follow :

Figure 11 shows the result of using the RTABMAP detection system. The system mistakenly regards that the moving objects as the part of indoor scene. The emergence of a dynamic target make the system mistakenly thought it had come to a place it has been to. In our algorithm, with the help of depth information, it can make use of the static scene information to make the right loop without the inference of moving objects. The result of our algorithm is represented as follow:

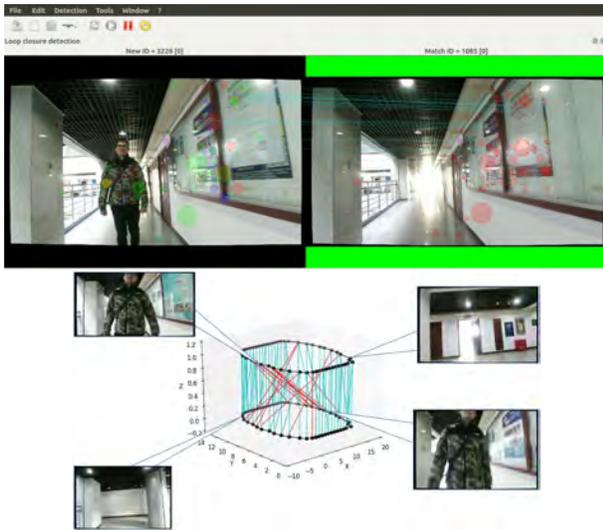
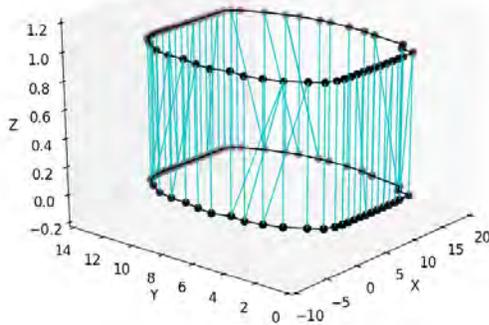


Fig. 11. The result of our algorithm test in the loop closure detection system. The left image shows that after optimization, the system can differentiate the foreground and background information and make the features almost located on the background. The right image is the schematic diagram of RTABMAP system result. The red lines represent the incorrect detection which links different places while the blue lines represent the correct links



The table 2 shows that our algorithm is more accurate in indoor environment. Furthermore, when moving objects appear, the probability of RTABMAP system error is 68.1%, while in our program, the probability is 11.3%. Regarding to processing time, the mean time of every frames is 0.826s for the whole experiment, which means that our algorithm meets the real-time constraint of 1HZ.

TABLE II
THE RESULT OF OUR ALGORITHM COMPARE WITH THE RTABMAP

Algorithm	false postive
RTABMAP	3.61%
Our Algorithm	1.62%

V. CONCLUSION

We propose a novel real-time RGB-D loop closure detection method based on the changes in depth information captured by the RGB-D sensor, which can handle the dynamic indoor environment when the moving objects appear. We use the variety of depth information to help the detection system to eliminate the negative features brought by the moving objects and incorporate this method into the RTABMAP system. We tested our method in the common dataset and real-life environment. The experiment results show that our method can effectively remove the negative features located on the moving objects which means reducing the impact caused by the moving objects and enhance the detection accuracy for the loop closure system. In our future work, we wish to use the deep learning methods to help us get the depth information from the RGB images which means that our method can be applied to the monocular system.

VI. ACKNOWLEDGEMENTS

The project sponsored by the National Key Research and Development Program (no. 2016YFB0502002), and the Fundamental Research Funds for the Central Universities(2018PTB-00-10).

REFERENCES

- [1] Lowry S, Sünderhauf N, Newman P, et al. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 2016, 32(1):1–19.
- [2] Lachat E, Macher H, Landes T, et al. Assessment and calibration of a RGB-D camera (kinect v2 sensor) towards a potential use for close-range 3D modeling. *Remote Sensing*, 2015, 7(10):13070–13097.
- [3] Newman P, Ho K. SLAM-loop closing with visually salient features. *Proceedings of Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on. IEEE, 2005. 635–642.*
- [4] Cummins M, Newman P. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research*, 2008, 27(6):647–665.
- [5] Angeli A, Filliat D, Doncieux S, et al. Fast and incremental method for loop-closure detection using bags of visual words. *IEEE Transactions on Robotics*, 2008, 24(5):1027–1037.
- [6] Gálvez-López D, Tardos J D. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 2012, 28(5):1188–1197.
- [7] Spinello L, Arras K O. People detection in RGB-D data. *Proceedings of Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on. IEEE, 2011. 3838–3843.*
- [8] Oliva A, Torralba A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 2001, 42(3):145–175.
- [9] Oliva A, Torralba A. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 2006, 155:23–36.
- [10] Lowe D G. Object recognition from local scale-invariant features. *Proceedings of Computer vision, 1999. The proceedings of the seventh IEEE international conference on, volume 2. Ieee, 1999. 1150–1157.*
- [11] Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features. *Proceedings of European conference on computer vision. Springer, 2006. 404–417.*
- [12] Milford M J, Wyeth G F. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. *Proceedings of Robotics and Automation (ICRA), 2012 IEEE International Conference on. IEEE, 2012. 1643–1649.*
- [13] Gao X, Zhang T. Loop closure detection for visual slam systems using deep neural networks. *Proceedings of Control Conference (CCC), 2015 34th Chinese. IEEE, 2015. 5851–5856.*
- [14] Cascianelli S, Costante G, Bellocchio E, et al. Robust visual semi-semantic loop closure detection by a covisibility graph and CNN features. *Robotics and Autonomous Systems*, 2017, 92:53–65.

- [15] Jodoin P M, Maddalena L, Petrosino A, et al. Extensive benchmark and survey of modeling methods for scene background initialization. *IEEE Transactions on Image Processing*, 2017, 26(11):5244–5256.
- [16] Kim D H, Kim J H. Effective background model-based RGB-D dense visual odometry in a dynamic environment. *IEEE Transactions on Robotics*, 2016, 32(6):1565–1573.
- [17] 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.
- [18] Huishen Z, Ling X, Huan Y, et al. An improved bag of words method for appearance based visual loop closure detection. *Proceedings of 2018 Chinese Control And Decision Conference (CCDC)*. IEEE, 2018. 5682–5687.
- [19] Garcia-Fidalgo E, Ortiz A. iBoW-LCD: An Appearance-based Loop Closure Detection Approach using Incremental Bags of Binary Words. *arXiv preprint arXiv:1802.05909*, 2018.
- [20] Garcia-Fidalgo E, Ortiz A. Loop Closure Detection Using Local Invariant Features and Randomized KD-Trees. *Proceedings of Methods for Appearance-based Loop Closure Detection*. Springer, 2018: 69–98.
- [21] Sun Y, Liu M, Meng M Q H. Improving RGB-D SLAM in dynamic environments: A motion removal approach. *Robotics and Autonomous Systems*, 2017, 89:110–122.
- [22] Mikolajczyk K, Tuytelaars T, Schmid C, et al. A comparison of affine region detectors. *International journal of computer vision*, 2005, 65(1-2):43–72.