# An Improved Bag of Words Method for Appearance Based Visual Loop Closure Detection

Zhu Huishen[1], Xie Ling[1], Yu Huan[1], Wang Liujun[1]

1. School of Automation, Beijing Institute of Technology, Beijing 100081
E-mail: 564872185@qq.com

**Abstract:** Loop closure detection is an essential component of many robotics applications such as SLAM (Simultaneous Localization and Mapping) and place recognition. This paper presents an appearance based loop closure detection algorithm based on bag of words method and inverse depth of feature words. Our proposed approach represent the feature points and descriptors in the images as the visual words according to the off-line generated visual directory to simplify the similarity comparison between the images. For the evaluation criteria, on the basis of tf-idf scores of visual words, the inverse depth of words are also introduced into the process of loop closure detection. Considering the temporal consistency of image sequences, the co-visibility graphs are also applied to verify the real loop from all candidates. At last, some experiment results are showed and analyzed to illustrate the feasibility and performance of our algorithm in different situations and environments.

**Key Words:** SLAM, loop closure detection, bag of words, inverse depth

## 1 INTRODUCTION

Simultaneous localization and mapping (SLAM) has been a prevalent research topic in autonomous robotics and computer version aspects for some decades. SLAM aims to build the model of environment around and estimate motion and pose trajectory at the same time without any prior environment information that is important to autonomous robot in autonomous navigation. Nowadays, quite a few SLAM studies focus on visual sensor systems which equip mono, stereo and RGB-D camera as localization sensor for its inexpensive price and rich information compared with other sensors like laser and LIDAR. Loop closure detection is a key module of SLAM system whose function is to discern the previously reached scenes to add extra time interval constraints besides the adjacent frame constraint generated by feature point matching in the front end. Without loop closure detection, the SLAM system is unable to eliminate the accumulated error produced by the front end and build the global consistent trajectory and map.

As an independent part of the SLAM system, loop closure detection does not use the geometry information previously calculated to identify preceding visited places. Currently, most of the related research is appearance based method which just utilize the previously extracted feature points and descriptors to estimate whether current location has been reached. Bag of Words (BoW) is the main current method used in many mainstream projects, which classifies the features and descriptors as different visual words. In this way, the task turns from judging the similarity of images to verdict the difference of the vectors representing the words each image processes. However, current loop closure detection methods especially those who are based on appearance of images often faced with perceptual aliasing and perceptual variability problem [3]. Perceptual aliasing means the detection algorithm mistakes two images taken from different places for they are from the same place. This problem happens when two places look similar in appearance. Perceptual variability takes place when the algorithm failed to recognize the used visited places which is related to dynamic scenes, illumination changes and photographic effects. Meanwhile, the system's real-time request is also a challenging factor in designing the algorithm.

In this paper, a novel loop closure detection method based on bag of visual words is proposed to improve the precision and recall performance. At the same time, considering the high real-time request of the algorithm, we not only make use of the features and descriptors extracted before also use the depth information to help to verify whether the loop is correct. In addition to this, to avoid the perceptual aliasing problem, we take advantage of the co-visibility graph to combine those feature words to formation word blocks which helps to avoid mismatching.

## 2 RELATED WORKS

Since appearance based loop closure detection has been established as main stream approach, many researchers pay a high attention to it and they have achieved quite a few achievements and results.

### 2.1 Features and Descriptors

The first matter needs to be considered is the types of features. Scale-invariant Feature Transform (SIFT) [4] has been widely used in loop closure research since it was present. As a high performance descriptor which is robust to the change of illumination, scale and rotation, the most serious disadvantage of SIFT descriptor is the time consuming problem. In this situation, Speeded-Up Robust Features (SURF) [5] which keeps the merit of SIFT and

978-1-5386-1243-9/18/$31.00 ©2018 IEEE

reduce the complexity of extracting process is an ideal descriptor to be used in loop closure research. Furthermore, some binary descriptors are also proved to be capable of solving the problem such as BRIEF and ORB [6].

### 2.2 Loop Closure Methods

Before starting to explain our algorithm, we firstly introduce some primary contribution to let readers have an overview of loop closure detection methods. Bag of words method [7], as mentioned above, extract features and descriptors and then take different value descriptors as different visual words recorded in a dictionary. After that an image is expressed as a histogram vector composed of all of the visual words extracted from it. Angeli [8] proposed two visual vocabularies including appearance and color. They have a combined action adding to a Bayesian filter to estimate the probability of loop closure. One of the most pioneering work is the Fast Appearance-Based Mapping (FAB-MAP) [9] published by Cummins and Newman. They use an off-line pre-trained vocabulary to have an on-line test which have a high performance on 70km and 1000km experiment. Lopez [10] simplify the descriptors, using binary words to meet the real time request and improve the running performance. The use of direct index and inverse index effectively reduce the execution time by nearly 90%. Moreover, the K-means cluster methods are applied in the establishment of vocabulary. The choice of vocabulary's data structure is set in [13], considering the convenience of adding new visual words.

In recent years, Labbe [11] propose a new method called Real-Time Appearance-Based Mapping (RTAB-MAP) using working memory (WM) and long-term memory (LTM) to discriminate locations need to be compared. This approach is proved to be useful in long-term working and can keep processing time fast enough. IBuILD [12] is Incremental bag of binary words for appearance based Loop Closure detection. This approach build directory online and establish a likelihood function to calculate the candidate loop. Another innovate is the binary words is created through tracking the feature points among continuous frames to guarantee the words robustness.

## 3 IMAGE PROCESSING

In this section, we start to describe and explain our algorithm flow and the innovative places. To begin with, we need to introduce the process of getting all data and information we need including visual words, dictionary establishment and depth index from the raw images.

### 3.1 Extraction of Visual Words

As mentioned in section2.1, considering the performance and running time requirement of the system, we choose the binary features Oriented FAST and Rotated BRIEF ORB [6] which can be calculated quickly meanwhile have the high quality on scale and orientation. The FAST feature points are extracted according to the value of gray scale just like corner points. The length of descriptor we choose is 256 bits, each bit of descriptor can be calculated as:

$$\mathrm{v}_{\mathrm{i}}(p)=\begin{cases}0 & p+x_i<p+y_i\\ 1 & otherwise\end{cases}\qquad \forall i\in[1,256] \tag{1}$$

where $\mathrm{v}_\mathrm{i}$ is the value of the i-th bit descriptor, $\mathrm{x}_\mathrm{i}$, $\mathrm{y}_\mathrm{i}$ is the gray scale value of random chosen location which established through Gaussian distribution in advance around the point. p represents the point being computed now. Such binary descriptors are not only fast to calculate but also easy to be compared. We use the Hamming distance to calculate the distance between the descriptors:

$$\mathrm{d}_{\mathrm{Ham}}[v(p),v(q)]=\sum_{i}^{256}XOR[v(p),v(q)] \tag{2}$$

where v(p), v(q) is the descriptor. XOR is the exclusive-OR bit operation. After getting features and descriptors, we need to turn them into visual words and convert images to sparse vectors.

### 3.2 Establishment of Visual Directory

Like most of bag of words works, tree data structure [7] is used to build the dictionary, every layer use kmeans++ [17] clustering to classify the descriptors. The sketch map of dictionary structure is shown in figure1, in which only the leaf nodes store visual words and the medial nodes are just used to find the words. The whole storage capacity of the directory is $k^d$. When searching for the given word, we just need to compare with the clustering center d times to locate the word, which promising the searching time efficiency of O(logN).

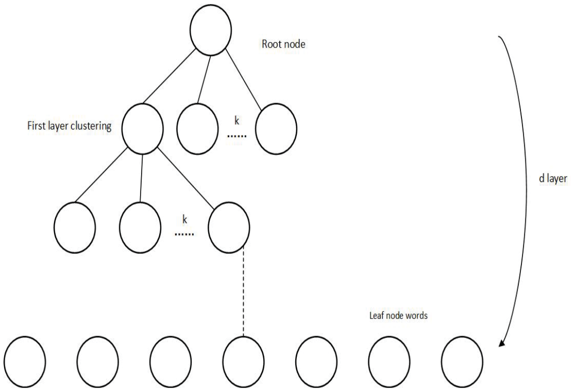

Fig 1. Sketch map of tree structure visual directory.

Meanwhile, to distinguish the importance of each word and assist to operate follow-up loop closing, direct index[18] are applied which is used to store the father nodes of the words in the directories that can speed up the comparison.

### 3.3 Depth Solver

In this section, we propose to recommend the usage of depth information in the loop closure detection which is rarely applied before. As mentioned in part2.2, quite a few works prefer to add a structure consistency or geometrical checking to avoid mistake detection, the method they use always have to use extra feature points to proceed matching and computing fundamental matrix in epipolar geometry. However, such kinds of method is time consuming and repetitive with front end solution. In this situation, we introduce a novel method using the relative inverse depth information to improve the precision-recall performance and reduce the complexity of the whole algorithm.

The depth of the feature points is available whatever the kind of camera sensor is. For the RGB-D camera, the depth can be directly got through using time of flight (TOF). For the monocular and stereo camera, the depth of features can be solved by triangulation which just use the coordinates of coupled matched points in left and right frames (while for the monocular camera is two continuous frames) through least square method. The specific principle are given as following: set $x_1$ and $x_2$ as the normalization coordinates of matched feature points, they satisfy:

$$d_1 x_1 = d_2 R x_2 + t \tag{3}$$

where the R and t is the external reference of two frames, represent rotation and translation respectively. d represents the depth of the point which is need to be calculated. Left multiply by $x_1^{\wedge}$ ($x_2^{\wedge}$) which represents the anti-symmetry matrix on both side of the formula, $d_2$ ($d_1$) can be calculated out as followed:

$$d_1 x_1^{\wedge} x_1 = 0 = d_2 x_1^{\wedge} R x_2 + x_1^{\wedge} t \tag{4}$$

where $d_1$,$d_2$ we get here is the depth value in the camera coordinate system, to obtain the solution of the depth value in the world coordinate system we can just use the pose R and t at present to resolve the answer, the formula is similar to (3). For the feature points recorded in the directory as visual words in the image, we calculate their depth and record it into the vocabulary refer to tf-idf (detail in section4.1) score. In fact, the practical depth value is not suitable to be used directly for the huge diversity between them, so we use inverse depth [17] to reach a higher quantitative value stability. As shown in figure2, O1, O2 are the focus points of the camera, p1, p2 are the matched features in images, P is the point in the real word whose depth can be calculated through p1 and p2 in triangulation.

## 4 LOOP CLOSURE ALGORITHM

After the image processing introduced in last section, the next step is to make use of these data to finish the detection of loop closure. The procedure in detail is proposed in this section.

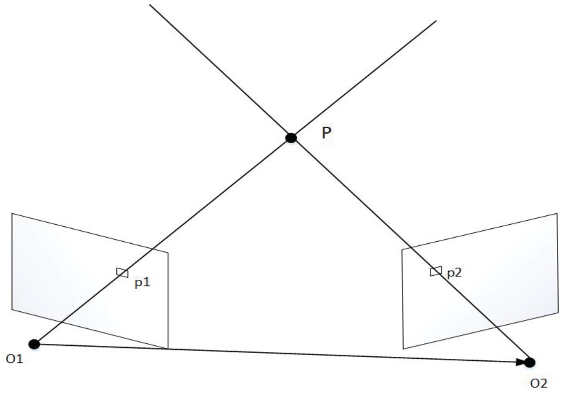

Fig 2. Schematic map of getting the depth of feature points through triangulation.

### 4.1 Similarity Judgement

Although the feature descriptors have been converted into visual words based on the vocabulary, not all of the visual words have the same discrimination and differentiation degrees for place recognition. For example, some words are very common that can be find in many frames, in this situation, such kind of visual words just have a little utility in loop detection. We adopt the TF-IDF (Term Frequency - Inverse Document Frequency) method which is usually used in text retrieval to distinguish the importance of different words just like [10] [19].

In the bag of words model, IDF can be calculated before processing the images, in other word, the IDF property is determined when building the dictionary. IDF can be defined as the proportion of the number of features in the leaf node $w_i$ to all the features in the directory. If we define n as the total number of feature words, $n_i$ is the number of words under the $w_i$, the expression of inverse document frequency is:

$$idf(i) = \log \frac{n}{n_i} \tag{5}$$

on the other side, Term frequency aim at the frequency of certain feature words in the single frame:

$$TF_i = \frac{n_i}{n} \tag{6}$$

where $n_i$ is the number of times the word $w_i$, n is the total number of words appear in this frame.

Hence, the weight of the word $w_i$ equals to:

$$\eta_i = TF_i \times IDF_i \tag{7}$$

After considering the weight of words, for the certain image A, the feature points extracted from A can corresponding to many visual words. In addition to this, as illustrated in 3.2, the average inverse depth of each visual words is also added in the bag of words, becoming bag of depth words:

$$\begin{aligned} A &= \{(w_1,d_1,\eta_1),(w_2,d_2,\eta_2)\cdots(w_n,d_n,\eta_n)\} \overset{\Delta}{=} v_A \\ s_A &= \{(w_1,\eta_1),(w_2,\eta_2)\cdots(w_n,\eta_n)\} \\ d_A &= \{(w_1,d_1),(w_2,d_2)\cdots(w_n,d_n)\} \end{aligned} \tag{8}$$

In fact, $v_A$ is very likely to be a sparse vector that contains lots of zero, for the similar features are classified into the same node. The nonzero parts indicate what words the image A contain, and the values are the corresponding TF-IDF score.

### 4.2 Loop Candidate

In view of the binary descriptors and sparse vectors, L-1 norm is appropriate to the comparison.

$$\begin{aligned} s(A,B) &= 1 - \frac{1}{2}\left| \frac{s_A}{|s_A|} - \frac{s_B}{|s_B|} \right| \\ &= -\frac{1}{2}\left[ \left|(w_A,\eta_A)-(w_B,\eta_B)\right| - \left|(w_A,\eta_A)\right| - \left|(w_B,\eta_B)\right| \right] \end{aligned} \tag{9}$$

It’s worth noting that the $v_A$ and $v_B$ here only include the word grade, and the depth of words are calculated otherwise in the part 4.4, for the weight and grade method are not the same.

Then, the threshold to determine whether the image is accepted needs to be confirmed. Considering the appearance of different environment situations have otherness. We introduce a prior similarity which is normalized through dividing s(A,B) by s(A,A-1) to response the relevance between the image and its last image, to improve the adaptability in different environment. After that, all similarity score is normalized according to this value. The formula is written as followed:

$$S'(A,B)=\frac{s(v_A,v_B)}{s(v_A,v_{A-1})} \tag{10}$$

in this way, we make use of the previous images similarity grade to avoid absolutely threshold because the value we get here overcome the otherness of different environment. The threshold γ can be set according to the demand of precision and recall performance. If we want to reduce perceptual aliasing, the threshold can be set higher, and vice versa.

### 4.3 Temporal Consistency

In the previous process, we can get several candidates of loop closure, to affirm which image is the most accurate one, we need to do verification to decide whether the images refer to the used visited place. The first approach is to verify the temporal consistency of the image. The continuous images of the query image are also probably be similar to the loop closure image. Compared with other works such as [18], directly group the close images, we use the co-visibility graph to improve the precision and strengthen the verification as a novelty.

The co-visibility graph is composed of the frame pose as vertex and the mutual landmarks projected by corresponding feature points as edges. Via the co-visibility graph, we can get all adjacent spatial relationship of the query frame and the loop closure frame. Based on the information of co-visibility graph, we can merge the candidate frames that represent the same scene, and choose the highest performance frame as the best candidate of this scene.

### 4.4 Depth and Geometry Verification

After the temporal consistency check, the excess loop candidates from the same frames have been rejected, the next step is to affirm the correct loop for the query image.

As mentioned in section 3.2, the depths of all points have been translated from the camera coordinate system into the world coordinate system, therefore if two features represent the same landmark, their depth should be nearly equivalent, if in consideration of the error from the front end visual odometry, a bias threshold α is set for the depth verification. We also use the tf-idf score in the query image vector as the weight of depth words, and compute the L-2 norm:

$$d(A,B)=\sqrt{\sum_{i=1}^{n}\eta_i^A(d_i^A-d_i^B)^2} \tag{11}$$

where the symbols have all been illustrated in formula (8) in 4.1 and the threshold α is 0.5×d(A,A-1). As a novelty comparing with other works, we no longer use the geometry verification that not only need to compare and match features, but also computing the fundamental matrix between query and candidate images. Our method links the spatial depth information and visual words to reach the maximum utilization of existing data, which proved to be feasible.

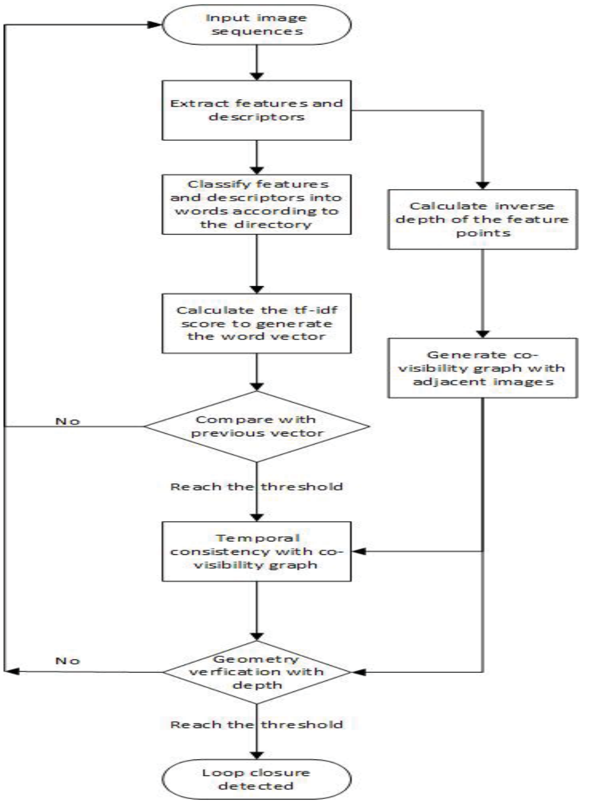
Input image sequences
Extract features and descriptors
Classify features and descriptors into words according to the directory
Calculate inverse depth of the feature points
Calculate the tf-idf score to generate the word vector
Generate co-visibility graph with adjacent images
No
Compare with previous vector
Reach the threshold
Temporal consistency with co-visibility graph
No
Geometry verfication with depth
Reach the threshold
Loop closure detected

Fig 3. Float diagram of our proposed algorithm.

## 5 EXPIREMENT EVALUATION

In this section, we show some experiment results of our proposed algorithm compared with some main current algorithm on public dataset and image sequences collected by ourselves. The evaluation criteria of loop detection is the precision and recall ratio, the precision ratio means the proportion of real loop closure in all loop detected. The recall ratio is the number of real loop detected divided by all loop closure in real.

### 5.1 Verification on Self Collected Images

Firstly，to prove the effectiveness and robustness of our algorithm, we design some experiments including indoor and outdoor environments and the image data is collected by ourselves. The camera sensor we use is the loitor stereo camera and the specific experiment conditions are recorded in Table1.

The scenes of indoor experiments are taken in the meeting room and library. The trajectory is approximate to a rectangle. To test the property of our algorithm, we take images in different scales, disparate directions from the same scene. As shown in figure 4 and 5, when facing with the change of scales and orientations, our algorithm is still robust enough to detect the loop closure relationship between the related images and reject the incorrect queries, at the same time, the usage co-visibility graph make the algorithm choose the most suitable images.

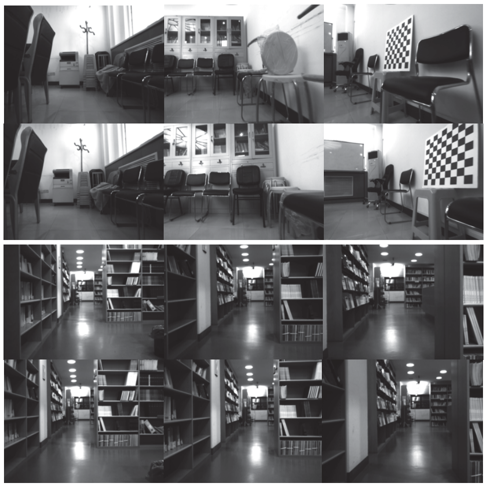

Fig 4. Some loop images detected in indoor environment.

The outdoor data is gathered in the road of school which is in a dynamic environment as figure6 showed. The running cars, bicycles and pedestrians make the appearance of the same places have a change, meanwhile, the illumination condition is more complex than indoor environments. All these factors increase the possibility of perceptual aliasing, our algorithm add the depth information of those high discrimination feature points that reduce the effect of dynamic things.

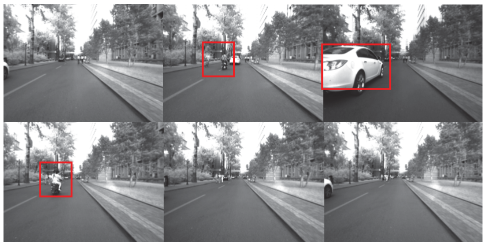

Fig 5. Examples of loop images detected on the road of school.

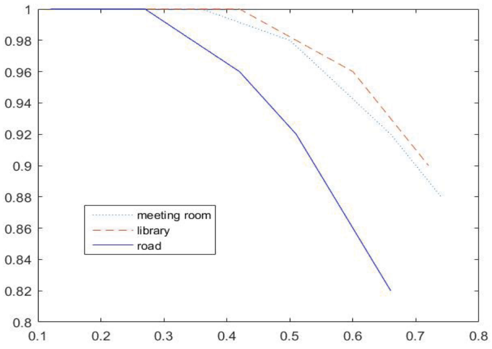

Fig 6. The P-R curve in different environments

To make a conclusion, through three tests under different situation, we find that our algorithm as a method including several novel improvement has the ability to detect the loop in different environment and proved to have a high performance in various external conditions.

Table1. Experiment conditions

| Environment | meeting room | library | road |
|---|---|---|---|
| FPS | 15 | 15 | 30 |
| illumination | artificial&sun | artificial | sun light |
| dynamic | half static | static | dynamic |
| Camera carrier | dolly | handhold | car |
| Feature extraction time | 7.1ms/img | 6.3ms/img | 7.0ms/img |
| Loop detection time | 16ms/img | 15ms/img | 16ms/img |
| Running distance | 10m | 6m | 1km |
| Image resolution | 752×480 | 640×480 | 752×480 |

### 5.2 Experiments on Public Datasets

In the second experiment, we choose the benchmark dataset Malaga2013 which is usually used in loop closure research as the image data, meanwhile to directly reflect the property of the algorithm we proposed, we compared it with traditional bag of words and FAB-MAP2 [10].

As shown in figure 7, the path has many loop that is very suitable to evaluate the performance of loop closure algorithm, the environment around is dynamic.

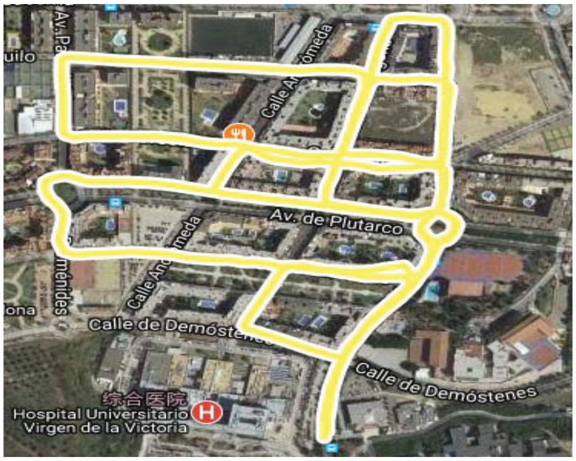

Fig 7. The ground truth trajectory of the dataset in Malaga2013

From the figure 8, we can find that when the threshold of similarity score is reduce, the FAB-MAP and traditional bag of binary words method's recall performance is not satisfactory for the verification of detection cannot exclude all wrong loop closure. In comparison, our method can keep relatively high precision performance when the threshold of similarity score is decrease for the reason of the introduction of depth information

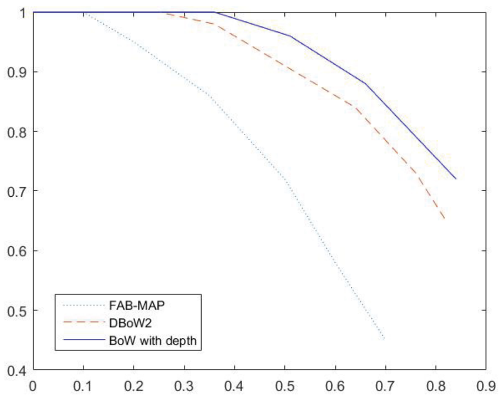

Fig 8. P-R curve of three loop closure methods on the dataset.

## 6 CONCLUSION AND FUTURE WORKS

In this paper, we present a novel loop closure detection method based on the bag of words methods with binary features. Meanwhile, considering the lack of spatial information and geometry constraint, which is the inherent weakness of appearance based bag of words method, our algorithm introduce the depth information of visual words to enhance the spatial qualification. Furthermore, the utilization of the co-visibility graph in the temporal consistency check is also a novelty of our work. The experiment we did also test and verify the feasibility of our algorithm.

In the future, we want to implement our method on incremental on-line directory to apply the depth information into the course of establishing the visual vocabulary. Then, more detailed and impeccable experiments need to be done especially in many challenging environments and long-term trajectory.

## REFERENCES

[1]Zhang G, Lilly M J, Vela P A. Learning binary features online from motion dynamics for incremental loop-closure detection and place recognition[C] IEEE International Conference on Robotics and Automation. IEEE, 2016:765-772

[2]Lowry S, Sunderhauf N, Newman P, et al. Visual Place Recognition: A Survey[J]. IEEE Transactions on Robotics, 2016, 32(1):1-19.

[3]Gutmann J S, Konolige K. Incremental mapping of large cyclic environments[C] IEEE International Symposium on Computational Intelligence in Robotics and Automation, 1999. Cira '99. Proceedings. IEEE, 2002:318 - 325.

[4]Lowe DG. Distinctive Image Features from Scale-Invariant Keypoints[J]. International Journal of Computer Vision, 2004, 60(2):91-110.

[5]Bay H, Ess A, Tuytelaars T, et al. Speeded-Up Robust Features (SURF)[J]. Computer Vision & Image Understanding, 2008, 110(3):346-359.

[6]Rublee E, Rabaud V, Konolige K, et al. ORB: An efficient alternative to SIFT or SURF[C] IEEE International Conference on Computer Vision. IEEE, 2011:2564-2571.

[7]Csurka G, Dance C R, Fan L, et al. Visual categorization with bags of keypoints[J]. Workshop on Statistical Learning in Computer Vision Eccv, 2004, 44(247):1--22.

[8] A. Angeli, D. Filliat, S. Doncieux, et al. Fast and Incremental Method for Loop-Closure Detection Using Bags of Visual Words[J]. IEEE Transactions on Robotics, 2008, 24(5):1027-1037.

[9]Cummins M, Newman P. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance[M]. Sage Publications, Inc. 2008.

[10] Galvez-Lopez D, Tardos J D. Real-time loop detection with bags of binary words[C] Ieee/rsj International Conference on Intelligent Robots and Systems. IEEE, 2011:51-58.

[11] Labbe M, Michaud F. Appearance-Based Loop Closure Detection for Online Large-Scale and Long-Term Operation[J]. IEEE Transactions on Robotics, 2013, 29(3):734-745.

[12] Khan S, Wollherr D. IBuILD: Incremental Bag of Binary Words for Appearance Based Loop Closure Detection[J]. 2015.

[13] Girdhar Y, Dudek G. Online Visual Vocabularies[C]// Computer and Robot Vision. IEEE, 2011:191-196.

[14] Kejriwal N, Kumar S, Shibata T. High performance loop closure detection using bag of word pairs[M]. North-Holland Publishing Co. 2016.

[15] Li Y, Hu Z, Huang G, et al. Image Sequence Matching Using both Holistic and Local Features for Loop Closure Detection[J]. IEEE Access, 2017, PP(99):1-1.

[16] Erhan C, Sariyanidi E, Sencan O, et al. Patterns of approximated localised moments for visual loop closure detection[J]. Iet Computer Vision, 2017, 11(3):237-245.

[17] Civera, Javier, Davison, Andrew J, Montiel, J. M Martínez. Inverse Depth Parametrization for Monocular SLAM[J]. IEEE Transactions on Robotics, 2008, 24(5):932-945.

[18] Galvez-Lopez 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.