Robotics and Autonomous Systems 77 (2016) 55–65

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

High performance loop closure detection using bag of word pairs

Nishant Kejriwal [a], Swagat Kumar [a,*], Tomohiro Shibata [b]

[a] Innovation Labs, Tata Consultancy Services, New Delhi, India

[b] Graduate School of Life Science and System Engineering, Kyushu Institute of Technology, Japan

HIGHLIGHTS

- We propose a new method for loop closure detection for topological mapping.
- It uses relative spatial co-occurrence information to improve the performance.
- We augment BoW method with a dictionary of spatially co-occurring word pairs.
- A memory map data structure is used for storing and indexing word pairs.
- We incorporate best of the existing methods to provide state-of-the-art performance.

ARTICLE INFO

Article history:
Received 29 August 2015
Accepted 9 December 2015
Available online 24 December 2015

Keywords:
Topological mapping
SLAM
BoW
BoWP
Relative spatial co-occurrence
RANSAC
Loop closure detection
Bayesian filtering

ABSTRACT

In this paper, we look into the problem of loop closure detection in topological mapping. The bag of words (BoW) is a popular approach which is fast and easy to implement, but suffers from perceptual aliasing, primarily due to vector quantization. We propose to overcome this limitation by incorporating the spatial co-occurrence information directly into the dictionary itself. This is done by creating an additional dictionary comprising of word pairs, which are formed by using a spatial neighborhood defined based on the scale size of each point feature. Since the word pairs are defined relative to the spatial location of each point feature, they exhibit a directional attribute which is a new finding made in this paper. The proposed approach, called bag of word pairs (BoWP), uses relative spatial co-occurrence of words to overcome the limitations of the conventional BoW methods. Unlike previous methods that use spatial arrangement only as a verification step, the proposed method incorporates spatial information directly into the detection level and thus, influences all stages of decision making. The proposed BoWP method is implemented in an on-line fashion by incorporating some of the popular concepts such as, K-D tree for storing and searching features, Bayesian probabilistic framework for making decisions on loop closures, incremental creation of dictionary and using RANSAC for confirming loop closure for the top candidate. Unlike previous methods, an incremental version of K-D tree implementation is used which prevents rebuilding of tree for every incoming image, thereby reducing the per image computation time considerably. Through experiments on standard datasets it is shown that the proposed methods provide better recall performance than most of the existing methods. This improvement is achieved without making use any geometric information obtained from range sensors or robot odometry. The computational requirements for the algorithm is comparable to that of BoW methods and is shown to be less than the latest state-of-the-art method in this category.

1. Introduction

Simultaneous localization and mapping (SLAM) is an important problem in mobile robotics which needs to be solved in order to achieve autonomous navigation. There exist two types of approaches to solve this problem—metric and topological. Metric SLAM aims to build a geometric map of the environment and hence requires accurate robot pose estimation. On the other hand, topological SLAM aims at building a graphical model of the environment comprising of *key locations* and their *connectivity* without explicitly making use of geometric or odometric information. Most of the topological SLAM research makes use of visual sensors which have become a common and inexpensive accessory in robotic

* Corresponding author.
E-mail addresses: nishant.kejriwal@tcs.com (N. Kejriwal), swagat.kumar@tcs.com (S. Kumar), tom@brain.kyutech.ac.jp (T. Shibata).

http://dx.doi.org/10.1016/j.robot.2015.12.003

applications. There is a third type of method called topo-metric methods [1–3] that combine metric and topological informations to obtain better performance.

One of the key aspects of a SLAM system is the loop closure detection [4,5] which requires a robot to recognize previously visited places accurately and correctly when they are revisited. The challenge lies in solving the *perceptual aliasing* problem because of which two physically distinct locations may appear similar to robot sensors.

This paper focuses on topological mapping that uses appearance (image)-based methods for loop closure detection. These methods use image similarity to identify previously visited places and hence, the success of such SLAM systems rely on having a robust place recognition algorithm. The Bag-of-Words (BoW) approach [6,4,7,8] is one of the most popular methods in this category. In this method, an image is represented as a histogram of words present in a dictionary. Usually an off-line dictionary is created by clustering similar features extracted from a large set of images. Histogram comparison is used to find the similarity between a query image acquired recently with the existing images in the map. Although BoW gives good results with very less computation time, it suffers from perceptual aliasing due to vector quantization. The problem due to quantization could be solved by using *direct feature matching* approaches [9–11] where the raw features are used directly for computing image similarity instead of their quantized representation obtained through clustering. Even though these methods are shown to provide better recall performance, they have a higher computational requirement which increases with growing map size and thus, makes it prohibitive for larger maps.

In this paper, we aim to improve the recall performance of BoW methods without sacrificing its simplicity and speed of execution. This is done by incorporating spatial co-occurrence information directly into the dictionary itself. In other words, we create a dictionary of word pairs in addition to the dictionary of individual words. A word pair is formed by observing that the spatial neighborhood of a word location may include other nearby words which, together with the former, may provide better discrimination in identifying loop closures. Since the point features (like SURF [12]) are used as words in the dictionary, the extent of spatial neighborhood for each word is defined by the scale size of the point feature. This scale size depends on the scale at which it is detected in the feature extraction algorithm and thus, is not a user-defined parameter. Since the word pairs are defined relative to the spatial neighborhood of each feature, these word pairs exhibit a directional attribute which has not been exploited so far in the literature. It is possible to extend this approach to word triplets and quadruplets with increasing cost of computation and memory storage requirement. Our consideration in this paper is limited to word pairs so as to keep the computational requirement closer to that of BoW method.

The usefulness of co-occurring words in addressing the perceptual aliasing problem is well known. For instance, Cummins and Newman [6] use a Chow–Liu tree to capture the co-occurrence information into the observation likelihood. While doing so, they do not take into account their spatial proximity to each other in the image. A pair of images will be called similar as long as they contain the same set of words in them irrespective of their spatial arrangement. Secondly, creation of Chow–Liu tree is a computationally expensive process which is usually built off-line to meet the real-time requirements. In another work, Stumm et al. [13] used co-visibility maps to incorporate the co-occurrence information into the decision making process. In this method, a graph of various landmarks is maintained by linking landmarks that are visible together and then a search is performed to find a cluster of landmarks as a clique found in the query image. This approach not only requires searching for cliques in a graph, but also requires tracking of individual landmarks over multiple frames which makes it computationally intensive compared to BoW approaches that use histogram matching for computing image similarity. In other cases, the spatial arrangement of features is used for providing better discrimination while detecting loop closures as in [4,10,7,14]. This is usually done by using RANSAC or multi-view geometry (MVG) constraints to discard outliers. These methods are known to be computationally expensive and hence, used as a second stage of verification. Some other authors have attempted to incorporate spatial information into bag of words methods as in [14,15]. In [15], spatial neighborhood is created by dividing the image into regular grids. On the other hand, the authors in [14] use fixed radial distance to decide the spatial neighborhood. Both of these methods suffer from two limitations—first, the spatial neighborhood requires user-defined parameters and second, they are not invariant to scale variations.

The proposed method differs from the above methods in two ways. First, we use relative spatial co-occurrence information that combines spatial proximity with co-occurrence information to provide better loop closure detection. This spatial occurrence has an associated directional attribute which is unique to our approach. Secondly, spatial information is used at every level of decision making unlike previous methods [4,10,7,14] that use it only as a second stage of verification. Finally, the extent of spatial neighborhood is decided automatically by using scale size of point features thereby making the algorithm scale invariant unlike methods [14,15] which use an user-defined spatial neighborhood to group features. Through experiments on standard datasets, we show that the proposed method provides significant improvement in recall performance compared to most of the existing state-of-the-art methods such as, FAB-MAP [6,16], incremental BoW [4], direct feature matching based methods [11,10] and the methods that use binary features [7,17] etc. The merit of the approach is further corroborated by making the observation that the improvement in the recall performance is more significant when the dictionary size is small (or quantization error is more). This improvement, however, is accompanied by a slight increase in the computational and memory requirement as one needs to create an additional dictionary of word pairs. We also present a completely online version of the algorithm that incorporates the best of the existing methods. This includes creating dictionary incrementally in an online fashion, using K-D tree to search for matching features, carrying out tree update at regular intervals, using Bayesian filtering to reduce transient errors in making loop closure decisions and using RANSAC as a second stage of verification.

The main contributions made in this paper are as follows: (1) We demonstrate that the recall performance of BoW approaches could be improved significantly by incorporating spatial co-occurrence information directly into the dictionary. This is done by creating an additional dictionary comprising of word pairs, which are formed by using a directional spatial neighborhood defined based on scale parameter of each point feature. This concept of relative spatial co-occurrence is a new finding which has not been exploited earlier in the literature. (2) An online version of the algorithm implementation is presented which incorporates various popular concepts like K-D tree based nearest-neighbor search for identifying potential loop closure candidates [10], tf–idf as a similarity measure [4], incrementally building dictionary [4], belief propagation based on Bayesian network to suppress transient errors in decision making [6] and RANSAC-based geometric verification stage for confirming loop closures [10,4]. This implementation which combines all of these concepts together is itself a new contribution in this field. (3) An incremental version of K-D tree implementation, available with the latest FLANN library, is used for the first time in the context of topological mapping in this paper. The features can be added to the tree incrementally with each

incoming image without having to rebuild it every time as being done in previous works [18]. This has helped us in achieving one of the lowest per image computation time for the proposed method. (4) The efficacy of the proposed method is established through an exhaustive comparison with the existing state-of-the-art methods. To our knowledge, the recall performances presented in this paper are among the best reported so far in the literature. The improvement in the recall performance is achieved without sacrificing on speed or computational requirements.

The rest of this paper is organized as follows. We provide an overview of the existing methods in the next section. The proposed method is described in Section 3. The experimental results are provided in Section 4 followed by conclusion in Section 5.

2. Related work

In this section, we provide an overview of various related works available in the literature. The Bag of Words (BoW) approach from text retrieval [19] has been used extensively for place recognition which forms an integral part of any loop closure detection algorithm. In a BoW approach, each image is represented as a histogram of word-frequency of each word present in the dictionary. An off-line dictionary is created by clustering similar visual features extracted from a large number of images. For a query image, the loop closure candidates are identified by using image similarity based on histogram matching [9]. Usually, a second stage of verification based on multi-view geometry or epipolar geometry is employed for confirming the loop closure detection [20]. Later authors, such as, Filliat and Angeli [4,21] and Nicosevici and Garcia [8] created the visual dictionary incrementally thereby making the whole process online. Many of these authors used Bayesian filtering to provide robustness against transient errors arising out of sensor noise. Cummins and Newman's FABMAP [6] incorporated the co-occurrence of words information into the observation likelihood using a Chow–Liu tree. This was shown to provide robustness against perceptual aliasing over long datasets. Many of these appearance-based methods were further extended by incorporating range information as in FAB-MAP 3D [22] or CAT-SLAM [3].

Unlike the Bag-of-Words approach (BoW) which used a quantized version of image features to represent the images, Zhang et al. [9,23] used raw features to represent images and used direct feature matching for detecting loop-closures for a given query image. The growing computational complexity with increasing map size was dealt with by carefully selecting keyframes from all the images. The growing computational complexity of direct feature matching method was tackled by Liu and Zhang [10] who used a K-D tree based search technique along with an inverted index table for finding the loop closure candidates. Similarly, Shahbazi [11] used Locality Sensitive Hashing (LSH) as an approximate nearest neighbor (ANN) search algorithm for feature matching. Even though LSH provides higher accuracy compared to other nearest neighborhood search algorithms, its computational time and memory requirement becomes prohibitive for large datasets. Quite recently, Hajebi [24] has shown that the time for feature matching process could be further reduced by using a graph-based nearest neighbor search (GNNS) algorithm by exploiting the fact that the images are acquired sequentially. This approach is also not without its own share of problems. First, graph construction is itself a computationally expensive process for a large dataset. Secondly, search time for GNNS is reduced by exploiting the assumption that the sequential images have significant overlap between them which might not be true in many cases.

Unlike the approaches that focuses on recognizing a place by a single image, there are methods that focus on recognizing a sequence of places [25] or matching trajectories [26]. These are shown to be more robust in identifying places even under extreme perceptual changes. These methods do not use Bag-of-Words (BoW) method and hence, are out of the purview of this paper as we do not focus on recognizing places under extreme perceptual changes or on finding matching trajectories through image sequences.

Growing computational and memory requirement with increasing map size is another concern which needs to be address while building maps over long distances. This is more pronounced in case of methods that use direct feature matching for loop closure detection [9,11,10]. Quite recently, Labbé and Michaud [18] proposed a novel memory architecture to address this issue of linearly increasing memory and time requirement with growing map size. While the most recently and frequently visited places were kept in a working memory (WM), the other nodes were transferred to a long term memory (LTM). This allows them to limit the number of nodes that needs to searched for detecting loop closures for every query image. They have reported one of the best recall performance for BoW based methods. We have compared the performance of our algorithm with this work to demonstrate the effectiveness of our approach.

From the feature perspective, most of the researchers have used either SIFT [27] or SURF [12] for place recognition in visual SLAM. Since the extraction time for these features are large, researchers are now increasingly using binary features such as BRIEF, ORB [8,7] or BRISK [17]. Binary features are becoming popular as they have a very compact presentation which, in turn, requires less memory and low comparison time.

Comparatively, there have been few attempts to incorporate spatial arrangement in detecting loop closures. FAB-MAP 1.0 [6] exploit the co-occurrence of words to deal with perceptual aliasing. Stumm et al. [13] used co-visibility graphs to eliminate the problem of pose selection in recognizing places. The places which appear together is connected in the graph. For place recognition, a clique of co-visible landmarks are searched in the co-visibility graph formed earlier. FAB-MAP 3D [22] also captures the arrangement of words through a random graph structure. However, they use range information to compute the pairwise distances between words in the actual 3D space.

In this paper, we aim to retain the simplicity of BoW approach while incorporating the spatial information into the algorithm. The word pair is formed based on the scale information of point features which is directional in nature. Moreover, the spatial proximity is computed in 2D image space unlike some methods which actual 3D distance to form word-pairs as in [22]. The proposed method for loop closure detection is described in the next section.

3. The proposed method

In this section, we describe our proposed approach for loop closure detection. This is an online method for creating a topological map where images are processed sequentially and does not involve any off-line pre-processing phase. We use a Bag-of-Words (BoW) method for image representation where each image is represented by a set of visual words. The standard BoW method is extended by incorporating a dictionary of word pairs which are formed by exploiting the directional spatial proximity between the words. We call our method as a Bag of word pairs (BoWP) which exploits the *relative spatial co-occurrence* of the words to overcome the perceptual aliasing problem. The details of this method is explained next in the section.

The following notations and symbols will be used for describing the proposed method. An image which is acquired at any given instant k is represented by the symbol I_k. Let the total number of nodes present in the map at this instant be M. Let us further assume that each node i, $i = 1, 2, \ldots, M$ contains l individual

words given by $\{w_1^i, w_2^i, \ldots, w_l^i\}$ and, m word-pairs given by $\{wp_1^i, wp_2^i, \ldots, wp_m^i\}$. Two separate inverted index tables are maintained which store the indices of the nodes where these words and word pairs appear.

In our method, SURF [12] descriptors are used as image features due to their robustness to photometric and geometric distortions. They also have lesser computation time as compared to SIFT [27]. Two dictionaries are created for storing words and word pair separately. While the word dictionary is created using a K-D tree, the dictionary for word pairs is creating using a multimap data structure [28]. These dictionaries are updated over time to incorporate new information obtained from the images which are acquired sequentially. For each query image, SURF descriptors are extracted and quantized to the respective words in the dictionary using an approximate nearest neighbor search in the K-D tree. The word pairs are created from these quantized words by using the extent of spatial neighborhood of each of these words. These quantized words and the corresponding word pairs are now used to compute a similarity measure between the query image and each of the existing nodes in the map based on *tf–idf* score [4]. In other words, two separate observation vectors are computed for each query image given by the following equations:

$$\mathbf{z}_k^w = \{z_i^w\}, \quad i = 1, 2, \ldots, M$$
$$\mathbf{z}_k^{wp} = \{z_i^{wp}\}, \quad i = 1, 2, \ldots, M \tag{1}$$

where z_i^w and z_i^{wp} are the tf–idf based similarity measure between the query image and the node i, $i = 1, 2, \ldots, M$ using word and word-pairs respectively. These observations are normalized to obtain an observation likelihood which is then used to compute the loop closure probability of each node using a Bayesian framework. The node with maximum probability is considered as a loop closure candidate if its probability is greater than a certain threshold θ_{lc}. A second stage of verification is carried out using RANSAC to confirm that this node is indeed a loop closure for the query image. This is done by checking if the matching coefficient P_{match} (defined later) is greater than a user-defined threshold θ_{ransac}. If this condition is not satisfied, the incoming image is considered as a new node in the map. At the end of each iteration, new word and word-pairs obtained from the query image are added to the dictionary. Inverted index tables are accordingly updated using this new information. Whenever a loop closure is detected, the information present in the new image is used to update the node. This is done by adding new words and word pairs obtained from the query image into the node description. This node update is essential to capture the minor variation that might arise when the same place is visited multiple times. Our experience shows that the overall size of node description stabilizes over time. However, it is possible to choose a fixed size for each node description and selecting features based on frequency of occurrence. This can alleviate problems of growing complexity due to quantization errors. The flowchart of the overall algorithm is provided in Fig. 1. As one can see, the algorithm consists of three major components: (1) Creation of dictionary of individual words, (2) Creating word pairs using relative spatial co-occurrence, and (3) Loop Closure Detection using Bayesian Framework. These components are described next in this section.

3.1. Creating dictionary of individual words

Our method is a completely on-line approach where the dictionaries are created incrementally as suggested in [4,8]. It is to be noted that these approaches use clustering to create visual words which leads to perceptual aliasing due to vector quantization. As opposed to this, Liu [10] uses a K-D tree and Shahbazi [11] uses LSH (locality sensitive hashing) to create off-line dictionary using the raw features as the words themselves and not their quantized

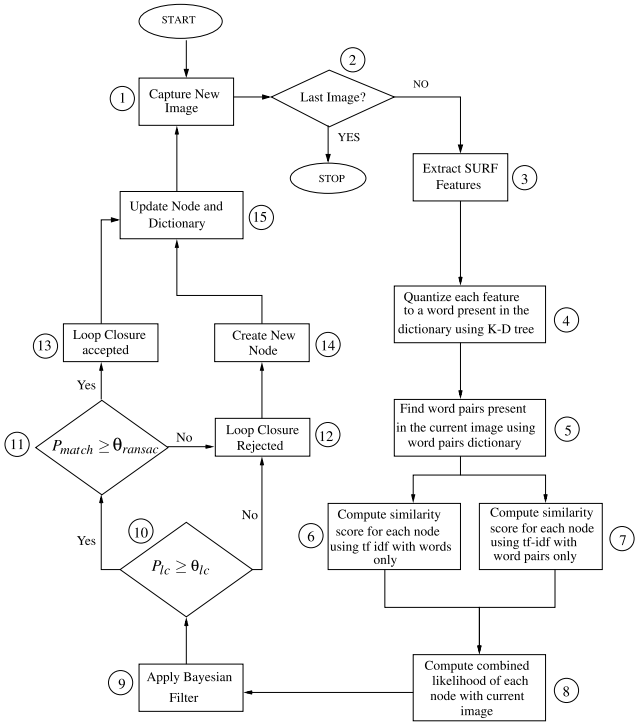

Fig. 1. Flowchart of the overall algorithm for topological mapping using Bag of Word Pairs. The key elements of the algorithm are marked with a number which are referred to when the computational complexity is discussed later in the experiment section.

representation obtained through clustering. Even though LSH provides higher recall performance and faster computation compared to K-D tree as shown by Shahbazi [11], it cannot be used for on-line update of dictionary as the re-computation of optimal LSH parameters is computationally expensive. On the other hand, the reconstruction time for K-D tree is very less as it does not involve any distance calculation unlike Hierarchical K-means [29] which uses distance computation for creating clusters.

In this work, we use a K-D tree to create a dictionary on-line using the raw features as the visual words. The new words obtained from an incoming image is added incrementally to this K-D tree using the latest version of the FLANN library (version 1.8.4) [30]. However, this may unbalance the tree over time which would, in turn, reduce the computational performance. This is avoided by reconstructing the K-D tree whenever the number of words becomes double. This is better than reconstructing the K-D tree for every query image as done by Labbé [18]. A distance ratio criterion [27] is used to avoid false matches in the K-D tree. According to this criterion, a good match should have a ratio of its distance from the closest neighbor to that from the second closest neighbor below a certain threshold θ_{DR}. All the descriptors which do not find any match in the visual word dictionary are considered as new words which are added to the dictionary.

3.2. Creating word pairs using relative spatial co-occurrence

As explained earlier, we extend the standard BoW approach by creating a dictionary of word pairs which not only takes into account the co-occurrence information of words, but also their spatial proximity to each other. This is in contrast to many of the earlier works [6,13,31,4], where co-occurrence information along with geometrical constraints is exploited to overcome the perceptual aliasing problem. This spatial proximity is defined in relation to the neighborhood of a given visual word. Hence it is termed as *relative spatial co-occurrence* which has a directional

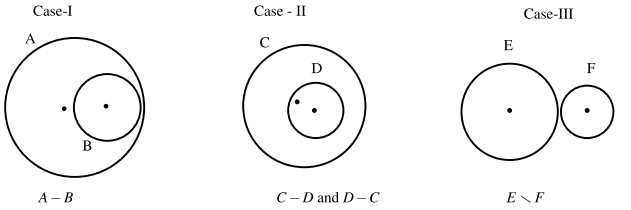

Fig. 2. Understanding relative spatial co-occurrence of words in a pair. While Case I and Case II show a valid word pair, case III does not.

attribute associated with it. This is different from other existing methods that attempt to incorporate spatial information into the Bag of Word methods as in [15,14]. Moreover, these methods use a user-defined parameter to define the spatial neighborhood unlike our method where this neighborhood is defined automatically by the scale size of the point feature. This makes our approach scale invariant and view invariant compared to the above methods.

In our case, each visual word is represented by its 128 dimensional SURF descriptor. The extent of its spatial neighborhood is defined by the scale size [12] of the descriptor. It is represented by a circle with a proportionate radius as shown in Fig. 2. A descriptor (or visual word) will be said to lie in the spatial neighborhood of another descriptor if its location represented by its center lies with in the circumference of the later descriptor. The creation of word pairs using relative spatial co-occurrence could be better understood by analyzing Fig. 2. A pair of visual words represented by $A - B$ would be considered as a valid word pair only if B lies in the spatial neighborhood of A as shown in Case-I of this figure. In this case, $A - B$ is a valid word pair while $B - A$ is not as the descriptor A does not lie in the spatial neighborhood of B. In case-II, C–D and D–C are both valid word pairs as each one lies in the spatial neighborhood of the other. Both of these word pairs are included into the dictionary without being purportedly redundant. This is because, the word pairs C–D and D–C go into different bins of the histogram when the likelihood is computed using *tf–idf* score. This is how the directional attribute comes into play in recognizing places. In case III, E–F and F–E are both invalid word pairs as they do not lie within the spatial neighborhood of each other. If only co-occurrence is considered, all the three cases would result in valid word pairs which is not the case with our approach. Fig. 3(a) shows few possible word pairs that satisfy the spatial neighborhood constraint explained above and the other Fig. 3(b) shows few common word pairs which were found in two images corresponding to the same place taken at two different times.

In a sense, the word pairs are formed by using directional spatial proximity between the words which is a new concept in BoW methods. This aspect of relative spatial co-occurrence to form word pairs has not been exploited earlier in the literature. Hence, in that respect, we consider it to be a novel contribution in the field. We call this method as a Bag of Word Pairs (BoWP) to emphasize the fact that a separate dictionary of word pairs is used in addition to the word dictionary used in the BoW methods. These word pairs are stored in a multimap data structure [28] to facilitate faster search for obtaining matching word pairs. All the new word pairs which are found in a query image are added incrementally to this data structure. This does not require re-constructing the dictionary as it happens with individual words that use K-D tree. The observation likelihood incorporates the tf–idf score obtained from individual words and word pairs using their respective inverted index tables. This information will be used for computing the loop closure probability as explained in the next section.

3.3. Loop closure detection using Bayesian framework

Since the images are acquired sequentially, we use a Bayesian framework to incorporate temporal coherence in the loop closure

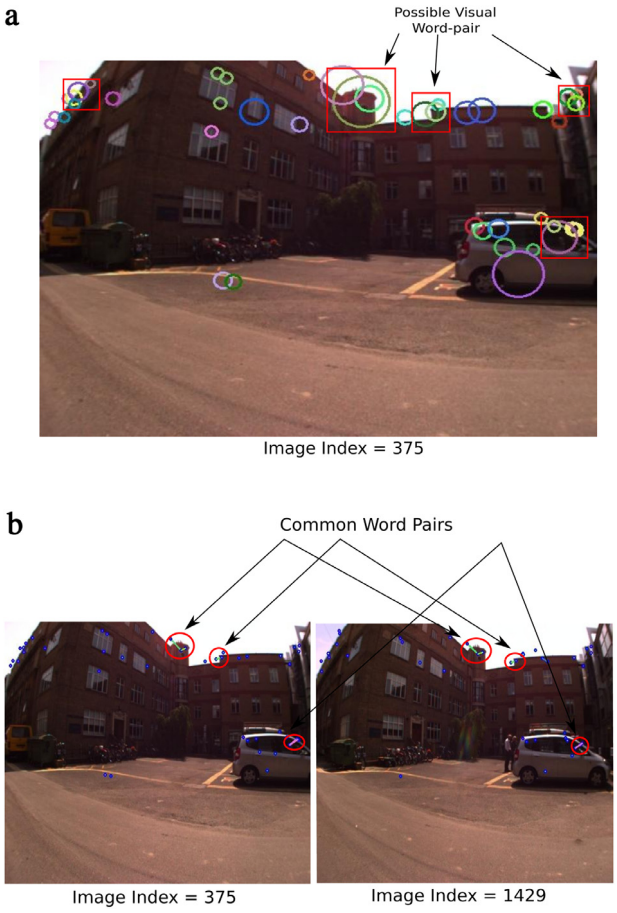

Fig. 3. Finding word pairs in a given image. (a) Shows few possible word pairs that can be found in an image. (b) Shows the common word pairs found in a pair of images taken at different time instants.

decision making process. This will provide robustness against transient errors that might occur while detecting loop closures. This Bayesian framework has been used by several authors [4,6,18] to compute the loop closure probability of each node for a given query image. In particular, we follow the approach by Angeli et al. [4] where information from multiple observations are fused to form a single observation likelihood, which is then used for computing the posterior loop closure probability. In our case, the observation likelihood has to incorporate information obtained from individual words and word-pairs.

Taking the cue from Angeli's work [4] and using Eq. (1), the individual likelihood functions are given by the following equations:

$$L_w(X = i|M) = \begin{cases} \dfrac{z_w^i - \sigma_w}{\mu_w} & \text{if } z_w^i \geq \mu_w + \sigma_w \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

and

$$L_{wp}(X = i|M) = \begin{cases} \dfrac{z_{wp}^i - \sigma_{wp}}{\mu_{wp}} & \text{if } z_{wp}^i \geq \mu_{wp} + \sigma_{wp} \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

where μ_w, σ_w and μ_{wp}, σ_{wp} are means and standard deviation for word and word-pairs observation vectors $(\mathbf{z}_w, \mathbf{z}_{wp})$ respectively and $i \in \{1, 2, \ldots, M\}$ is the index of nodes in the map. The variable X denotes the random variable that can take any node index as its value. Nodes having no similarity with current image are multiplied by 1; hence their posterior probability remains same as

the prior probability. We also compute the observation likelihood for a new location as suggested in [18]. This is given by

$$L_w(X = -1|M) = \frac{\mu_w}{\sigma_w} + 1 \tag{4}$$
$$L_{wp}(X = -1|M) = \frac{\mu_{wp}}{\sigma_{wp}} + 1$$

where the event $X = -1$ indicates the creation of a new node. The combined likelihood for each node $X = i$ is calculated as follows:

$$L(X = i|M) = L_w(X = i|M)L_{wp}(X = i|M) \tag{5}$$

where the node index i takes any value in the set $\{-1, 1, \ldots, M\}$. This likelihood is used to compute the posterior probability $P(X = i|M)$ of a node being a loop closure candidate for the query image. The node having the highest loop closure probability is declared as a possible loop closure candidate for the query image if this probability is higher than a threshold $\theta_{lc} = 0.5$, which is considered to be a constant in this work. If this condition is satisfied, a second stage of geometric verification is carried out using RANSAC to confirm the loop closure detection. The second stage essentially involves computing a similarity measure given by

$$P_{match} = \frac{2 \times N_{inliers}}{N_{image} + N_{node}} \tag{6}$$

where $N_{inliers}$ are the number of matching points obtained after applying RANSAC, N_{image} and N_{node} are the number of points in the image and node respectively. The presence of loop closure is said to be confirmed if $P_{match} \geq \theta_{ransac}$. If this condition is not satisfied, a new node is created using the query image features. It is to be noted that θ_{ransac} is the only user-defined parameter which is varied to obtain the precision–recall curve for our method.

4. Experimental results

In our approach, the dictionary of individual words and word-pairs are created incrementally by processing each image sequentially. The individual word dictionary is created using a K-D tree and the word-pair dictionary is created using a multimap data structure [28] where each word pair is stored along with their index numbers. The multimap data structure is implemented using C++ STL. We have used a nearest neighbor distance ratio $\theta_{DR} = 0.8$ to decide if a query descriptor is a new word or it matches with one of the existing words in the dictionary. Similarly, another constant $\theta_{lc} = 0.5$ is used while selecting the loop closure candidate for the query image. The final decision of loop closure is made only after making the second stage of verification that uses a variable user-defined threshold θ_{ransac}. So, this is the only user-defined parameter which is varied to obtain the precision–recall curve for our algorithm. The proposed algorithm is implemented using C/C++ and OpenCV library [32] on a GNU/Linux machine running on an Intel Core i7 processor and having 7.5 GB of RAM. Two instances of program output are shown in Figs. 4 and 5 respectively. In the first figure, BoWP selects a loop closure with higher confidence compared to BoW approach. In the second figure, BoWP is able to select the right loop closure in spite of having multiple similar candidates.

4.1. Performance comparison

In order to evaluate the performance of our approach, two different experiments are performed. In both the experiments, the recall performance of BoWP is compared with the existing state-of-the-art algorithms at 100% precision on same set of datasets. The datasets used for the experiment are the New College and the City Centre datasets originally used by FAB-MAP [6], LipIndoor

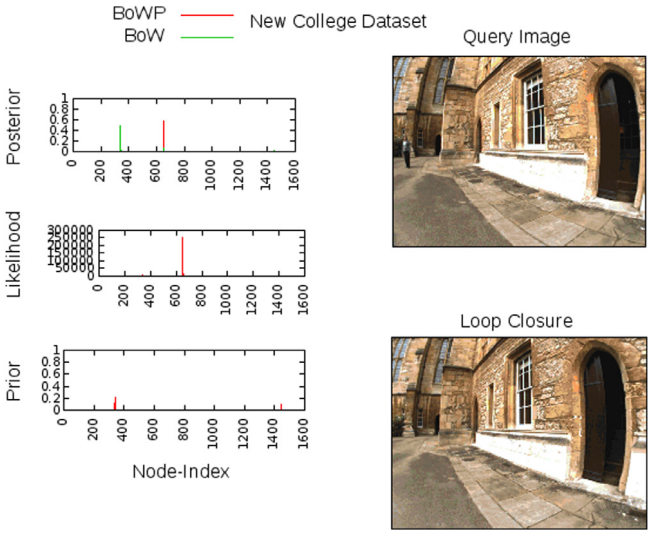

Fig. 4. An example of loop closure detection with Bayesian probabilities for a query image taken from New College dataset. As one can see, the proposed BoWP method selects a different loop closure candidate with higher belief compared to BoW method.

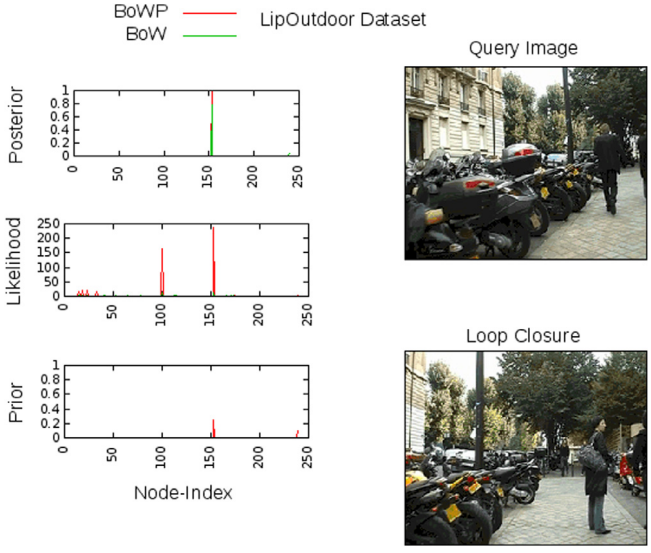

Fig. 5. An example of loop closure detection with Bayesian probabilities for a query image taken from LipOutdoor dataset. It shows how belief propagation helps in selecting the right loop closure candidate.

and LipOutdoor datasets [4], Bovisa dataset [18] and Malaga 6L dataset [35]. These two experiments differ in their operating conditions. The operating conditions for the first experiment is similar to that of the original FAB-MAP work [6] while the operating conditions in the second experiment is similar to that of RTAB-MAP work [7]. In the first experiment, a 128-dimensional SURF [12] vector is used as the image feature. The left and right side images in New College and City Centre datasets are processed separately. On the other hand, a 64-dimensional SURF vector is used for image processing in the second experiment. In addition, the stereo images of the New College and City Centre datasets are combined together to form single images for each location. The details of the experiments and their outcomes are explained below:

4.1.1. Experiment 1

In the first experiment, BoWP is compared with ten state-of-the-art algorithms such as, FAB-MAP [6,16], incremental

Table 1
Performance Comparison of BoWP with other state-of-the-art algorithms in terms of Recall (%) at 100% Precision. A '–' indicates that the corresponding value is not available for a given algorithm and a given dataset. The results for first three algorithms were obtained through our implementation. The results for next six algorithms were directly reproduced from their original papers. The BoWP implementation makes use of a 128-dimensional SURF vector.

Dataset	Image size	No. of images	Avg. no. of descriptors per image used/total number	Algorithms										
				Raw features with KD-tree [10]	Raw features with LSH [11]	Hierarchical K-Means [29]	Incremental BoW + TF–IDF [4]	Fab-Map [6]	FAST + BRIEF [7]	FAB-MAP 2.0 [16]	ORB [17]	IBuILD [33]	GMM + KD-tree [34]	BoWP
New College	640 × 480	2146	516/2169	47.5	48.8	41.6	–	47	–	–	–	–	–	**59.78**
City Centre	640 × 480	2474	437/1762	51.5	54.18	52.8	–	37	30.6	38.77	43.03	38.92	–	**66.80**
LipOutdoor	320 × 240	531	428/428	57.38	45.3	51.67	71	–	–	–	–	25.5	89	**90.9**
LipIndoor	320 × 240	388	369/369	73.4	67.3	72.5	80	23.64	–	–	–	41.9	87	**94**
Bovisa	320 × 240	2277	572/572	30.4	27	31	–	–	–	–	–	–	–	**46**
Malaga6L	1024 × 768	869	1000/2700	–	–	–	–	–	74.75	68.52	81.51	78	–	**87**

Table 2
Performance comparison of BoWP with RTAB-Map in terms of Recall at 100% Precision. The actual number of loop closures detected in each case is also reported. In this case the assumptions of RTAB-MAP are used to compute the recall performance. The implementation of RTAB-MAP and BoWP uses a 64-dimensional SURF vector.

Dataset	No. of images	RTAB-MAP [18]	PIRF 2.0 [36]	DP + MRF [37]	BoWP
New College	1073	**82**	–	57	77
City Centre	1237	84	80.04	41	**86**
LipOutdoor	531	70	–	–	**92**
LipIndoor	388	85	77.73	–	**94**
Bovisa	2277	37	–	–	**40**
Malaga6L	869	73	–	–	**84**

BoW [4], direct feature matching approaches [10,11], hierarchical k-means [29], bag of binary words [7,33], ORB based method [17] and a method based on color [34]. As mentioned above, the operating conditions in this experiment are similar to those in the original FAB-MAP work [6]. The images are processed using 128 dimensional SURF [12] vectors. The left and right side images in New College and City Centre datasets are processed separately in a sequence $\{left1 \rightarrow right1 \rightarrow left2 \rightarrow right2 \ldots\}$. The loop closure detection algorithms make use of approximately 500 descriptors from each image for recognizing places. This constitutes about 25% total number of descriptors available in each image in case of New college and City Centre datasets. Similarly, only 37% of the available descriptors were used in case of Malaga 6L dataset. In other datasets, all of the available descriptors were used by the algorithm. The images were used as they were available without making any changes to their size or resolution. The ground truth for the first two datasets are the same as those used in FAB-MAP [6]. The ground truths from LipOutdoor, LipIndoor and Bovisa datasets were obtained from RTAB-MAP implementation [18]. The ground truths for Malaga 6L dataset were created for 869 locations using GPS values. Any image that lies within 3 m of a given location is considered as a valid loop closure.

The resulting performance comparison for these datasets is shown in Table 1. We implemented the first three algorithms ourselves and reported the results obtained for these datasets. The results for the next seven algorithms were taken from the corresponding papers as these were reported. The recall performances were not available in some cases and this is indicated by a '–' in the table. As one can see, the proposed BoWP method provides better recall compared to all other methods listed in the table. It is to be noted that BoWP provides better recall performance even compared to direct feature matching methods [10,11] which are known to provide better accuracy with higher computational cost. We will show later that our computational requirements are much less compared to these methods. The table also shows that BoWP performs better than the recent methods based on binary features [17,33] which are faster to compute compared to SURF descriptors. It is, however, to be noted that our approach of word pairs with relative spatial co-occurrence is applicable to binary features as well. In that sense, the concept of spatial co-occurrence is generic in nature and is applicable all other local point features. To summarize, this table shows one of the best recall performance results reported so far in the literature and hence, demonstrates the efficacy of our algorithm.

4.1.2. Experiment 2

In the second experiment, BoWP is compared with PIRF 2.0 [36], RTAB-MAP [18] and DP + MRF [37]. The operating conditions of this experiment is similar to those used in the RTAB-MAP paper [7]. In this experiment, each pair of stereo images in New College and City Centre datasets are stitched together to form a single image representing each location. We also use 64 dimensional SURF vector instead of 128 dimensional vector used in the previous experiment. The number of descriptors that is processed for each image is limited to 400 for all datasets. The ground truths for the first five datasets are the same as those used by RTAB-MAP method [18]. The results reported for RTAB-MAP and BoWP have been obtained by running the algorithm on the same machine. The results for PIRF 2.0 and DP + MRF have been taken directly from their respective papers. The recall performance comparison among the three algorithms is shown in Table 2. As one can see, BoWP provides better recall performance compared to RTAB-MAP for all datasets except the New College dataset. The poor performance in case of New College dataset could be attributed to the fact that the ground truths selected for this dataset do not take geometric constraints into account which forms an integral part of our algorithm. This has effect in this particular dataset as the vehicle traverses the same locations from opposite directions leading to lateral inversions of the objects in the stitched image. So many of these ground truths are rejected by BoWP as the geometric constraints are violated. In other words, the ground truths selected for this dataset favor RTAB-MAP compared to BoWP.

4.2. Effect of incorporating spatial information into BoW method

The proposed bag of word pairs approach incorporates spatial and co-occurrence information directly into the creation of dictionary itself and hence, it is expected to provide better performance compared to the BoW approaches where the words are considered to be independent of each other. This fact can be verified by showing that the observation likelihood obtained with word pairs has higher information compared to that obtained using individual words alone. This is done empirically by computing the K–L divergence between the posterior and the prior probabilities in both cases as shown in Fig. 6. The higher value of K–L divergence indicates that the posterior has more information compared to its prior and this additional information is contributed only by the observation likelihood. As one can see, K–L divergence curve has a higher magnitude for word pairs (shown in red) compared to individual words (shown in blue). So, it can be said that the incorporation of spatial and co-occurrence information in BoWP provides more information which should be useful in overcoming the perceptual aliasing problem to a greater extent than BoW approaches. This is ascertained by comparing the performance of BoWP method against the BoW method for the same datasets as shown in Table 3. Note that the results of BoW reported here are obtained from our own implementation that includes several improvements over traditional methods as discussed in Section 3. As one can see, the BoWP provides higher recall compared to BoW in all the cases. The improvement is more significant in case of New college and City Centre datasets where the number of features available in each image is more. The improvement can be seen graphically in Fig. 8 where the precision–recall curves are drawn for two datasets. As one can see, dictionary of word pairs alone provides improvement over the case where only dictionary of individual words is used. This improvement in performance is further enhanced in our BoWP approach where dictionaries of both word and word-pairs are used.

Table 3
Recall performance comparison between BoWP and BoW method for 100% precision. The operating conditions are same as those in Table 1.

Dataset	# Loop closures		Recall (%)	
	BoW	BoWP	BoW	BoWP
New College	473	510	55.45	59.78
City Centre	684	740	63.36	66.80
LipOutdoor	269	271	90.2	90.9
Bovisa	134	151	41	46
LipIndoor	208	211	92	94

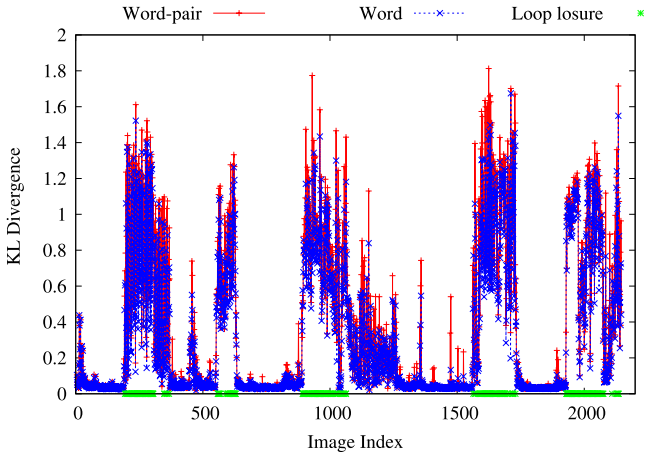

Fig. 6. The observation likelihood obtained using word pairs contains more information than that obtained using individual words. The Y-axis shows the K–L divergence between the posterior and the prior pdf. Higher value of K–L divergence indicates higher information. The figure corresponds to New College dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The benefit of BoWP over BoW is further corroborated by making the observation that the improvement in recall performance is higher when the quantization error is large. The quantization error can be controlled by varying the nearest neighbor distance ratio threshold θ_{DR}. Higher value of θ_{DR} indicates higher quantization error as more and more features now correspond to one word in the dictionary. We draw the precision–recall curve for two different values of $\theta_{DR} = \{0.8, 0.95\}$ for both the methods which is shown in Fig. 7. We can see that the improvement in recall performance with BoWP method is more significant when $\theta_{DR} = 0.95$ compared to the case when the quantization error is low ($\theta_{DR} = 0.8$). This is more clearly visible in Fig. 9 where the recall performances (at 100% precision) of BoWP and BoW are plotted against θ_{DR}. As one can see, the difference between the recall performances of BoWP and BoW is highest when the dictionary size is smallest. This is an important observation as one often needs to work with high quantization error so as to keep the computational requirement within real-time limits and BoWP ensures that the performance of algorithm is not compromised in such situations.

4.3. Computation complexity analysis

The computational complexity of our algorithm mainly depends on the dictionary size and the number of nodes in the map as in any other BoW method. The additional computation time is introduced due to the creation of a separate dictionary of word-pairs. As one can see in Fig. 10, BoWP takes at least 40–50 ms more than BoW for processing each image for City Centre dataset. This time difference is more or less remains same even with increasing map size as shown in this figure. The increase in computation is about 17%–20% for BoWP over BoW for any given dataset.

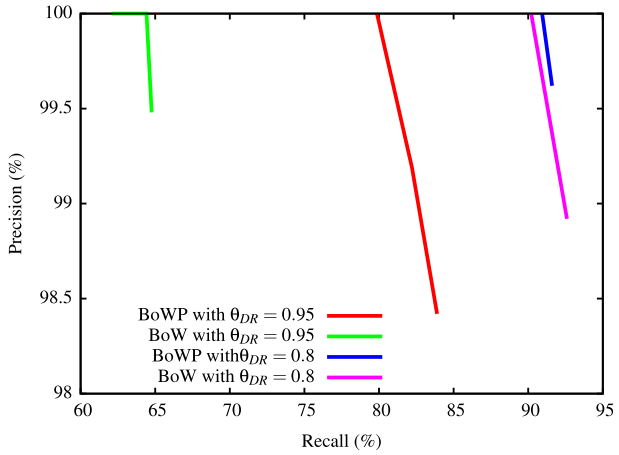

Fig. 7. Precision–recall curve for BoWP and BoW methods having different quantization error indicated by the value of θ_{DR}. The improvement in recall performance is significant when the quantization error is higher ($\theta_{DR} = 0.95$). The figure corresponds to LipOutdoor dataset.

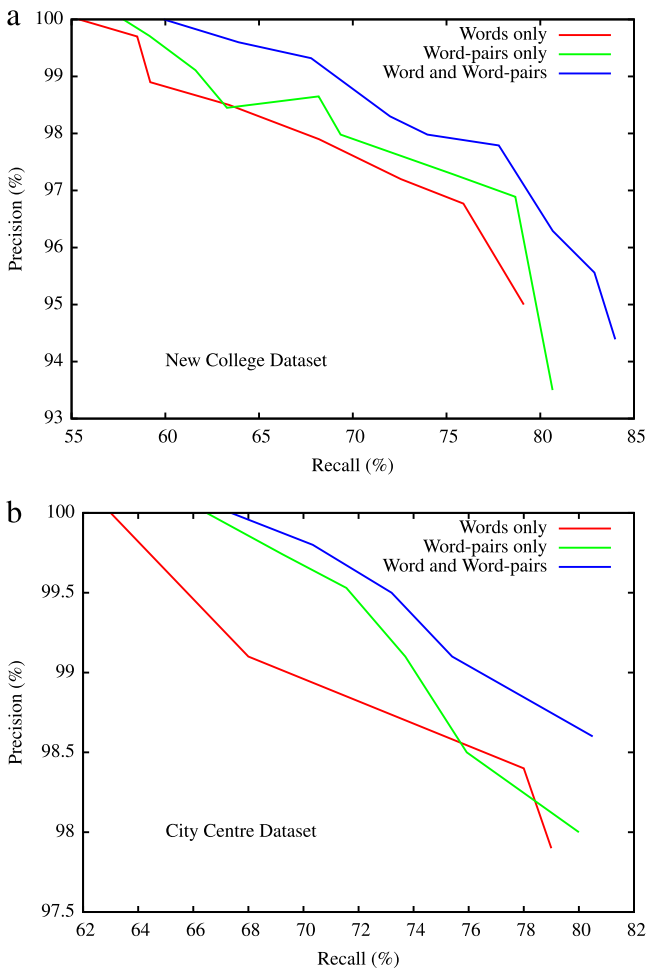

Fig. 8. Effect of Word Pairs on performance of Loop Closure Detection Algorithm. Word pairs provide better performance compared to only bag of words. This improvement is magnified in our BoWP approach where dictionary of word and word-pairs are used together. (a) P–R curve for New College Dataset (b) P–R curves for City Centre datasets.

It is worthwhile to analyze the computation time in some more detail. The time required by various aspects of the algorithm is shown in Fig. 11. As one can see, SURF extraction time constitutes the major part of the total computation time per image. This is

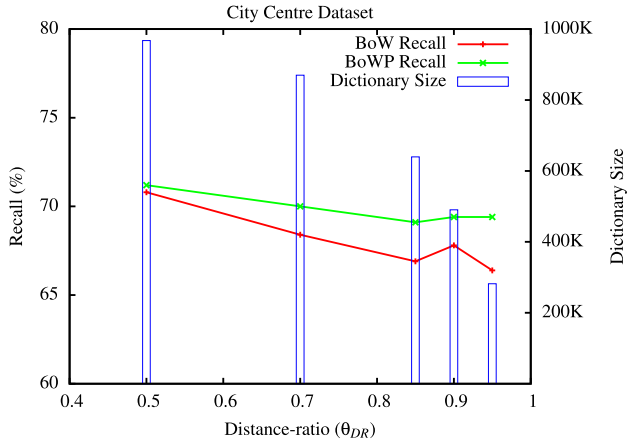

Fig. 9. Effect of BoWP with increasing quantization error. The improvement in recall performance of BoWP over BoW increases with increasing quantization error or decreasing dictionary size. The distance ratio θ_{DR} used in nearest neighbor search controls the size of dictionary.

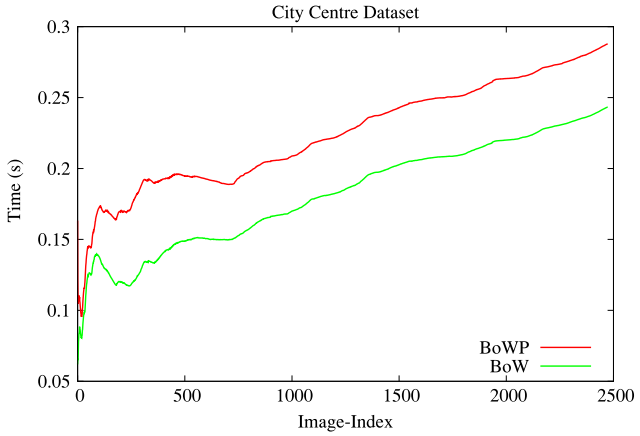

Fig. 10. Per image computation time comparison between BoWP and BoW. BoWP takes approximately 40–50 ms more time compared to BoW for City Centre Dataset. This time is fixed and does not grow with the increasing map size. Note that only 25% of available descriptors are processed in each image.

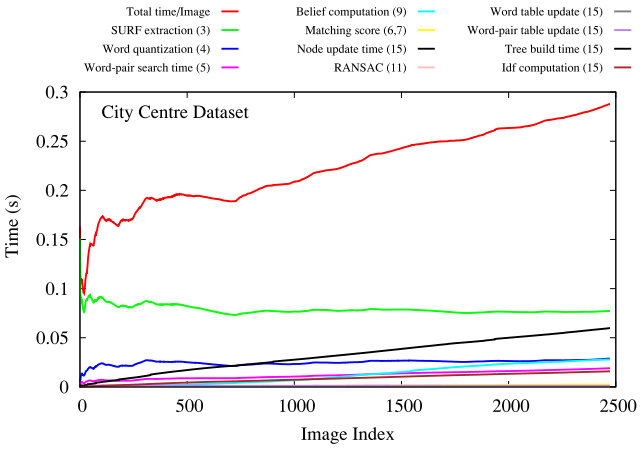

Fig. 11. Analysis of the BoWP in terms of computational complexity of each module in the algorithm and how they contribute to the total computation time per image. The numbers in the parentheses refer to the corresponding blocks in the Flow Chart provided in Fig. 1.

approximately about 200–400 ms which remains more or less constant for a given dataset. The second major component is contributed by the K-D tree building time which increases with the increased map size. In our method, new features are added to the K-D tree incrementally whenever a loop closure is found. The K-D tree is rebuilt only when the dictionary size doubles itself. This is unlike RTAB-MAP [18] where the tree is rebuilt at every iteration. Remaining components take very less time, however, increase very slowly with the increasing map size. The problem of linearly increasing computation time could be overcome by using the memory architecture of RTAB-MAP [18] where the less frequently encountered nodes are transferred to a long term memory (LTM) and the most recently and frequently visited nodes are kept in working memory (WM). This helps in achieving a bounded limit for per image computation time even with a growing map size.

We tried to compare the computational requirements of BoWP with the existing methods. The outcome is shown in Table 4. The computation time for FAB-MAP [6] and Incremental BoW [4] were taken from those reported by the authors of the PIRF-2.0 method [36]. The computation time for BoWP and RTAB-MAP is reported by considering only 400 descriptors per image. The computation time for RTAB-MAP [18] is obtained from our own implementation on the same system. As one can see, BoWP takes minimum computation time per image compared to all these methods. The computation time can be further reduced by using binary features such as ORB [17] or FAST/BRIEF [7] with slight decrease in the recall performance.

To summarize, it can be said that BoWP provides better recall performance compared to most of the existing methods. The incremental method for creating dictionary and building KD-tree helps in reducing the computation time per image. As per our knowledge the results reported in this paper are the best so far in this category. This does not include the methods that use geometric informations obtained from robot odometry or range sensors for verifying or confirming loop closures.

5. Conclusion

Reliable mapping requires having a robust loop closure detection algorithm. The challenge lies in getting higher recall performance at 100% precision with reasonable computational requirements. The Bag of words (BoW) approach is known to be fast and easy to implement but provides a low recall performance. We propose to improve the recall performance of BoW approaches by incorporating spatial neighborhood information into the dictionary itself. The extent of this spatial neighborhood is defined by the scale size of the point features. Since the word pairs are defined relative to the spatial neighborhood of each point feature, it exhibits a directional attribute. The resulting approach is called a bag of word pairs which is shown to provide higher recall performance compared to most of the existing methods. This improvement is achieved without making use of any geometric information obtained from the use of range sensors or robot odometry. There are a couple of extensions for this work which would be taken up in the future. One direction would be to deal with the growing computational complexity with increasing map size. The growing computational complexity can be dealt with by having a fixed size dictionary which is updated to retain only most frequent words in it. A forgetting factor may be included to emulate human memory model where older features would be gradually forgotten. However, it is yet to be seen how it would affect the recall performance of the proposed approach. Secondly, we plan to include this mapping method into a SLAM algorithm by incorporating camera ego motion into it. The resulting algorithm will be made available free for others to try and improve.

Acknowledgment

This work was partly supported by MEXT KAKEN (23120005).

Table 4
Average computation time (in seconds) per image for different datasets. BoWP takes lowest time for processing each image. The values for RTAB-MAP and BoWP are obtained through our own implementation. The values for other algorithms were taken from those reported in [36].

Dataset	Algorithms				
	FAB-MAP [6]	PIRF 2.0 [36]	Incremental BoW [4]	RTAB-MAP [18]	BoWP
New College	–	–	–	0.603	0.441
City Centre	0.466	0.852	5.82	1.33	0.393
LipIndoor	0.588	0.084	0.255	0.141	0.069
LipOutdoor	–	–	–	0.404	0.120
Bovisa	–	–	–	0.523	0.209

References

[1] W. Maddern, M. Milford, G. Wyeth, Towards persistent localization and mapping with a continuous appearance-based topology. in: Proceedings of Robotics: Science and Systems. Sydney, Australia, July 2012.
[2] Y. Latif, C.C. Lerma, J. Neira, Robust loop closing over time, in: Proceedings of Robotics: Science and Systems. Sydney, Australia, July 2012.
[3] W. Maddern, M. Milford, G. Wyeth, CAT-SLAM: probabilistic localisation and mapping using a continuous appearance-based trajectory, Int. J. Robot. Res. 31 (4) (2012) 429–451.
[4] A. Angeli, D. Filliat, S. Doncieux, J.-A. Meyer, Fast and incremental method for loop-closure detection using bags of visual words, IEEE Trans. Robot. 24 (5) (2008) 1027–1037.
[5] B. Williams, M. Cummins, J. Neira, P. Newman, I. Reid, J. Tardos, A comparison of loop closing techniques in monocular slam, Robot. Auton. Syst. 57 (12) (2009) 1188–1197.
[6] M. Cummins, P. Newman, FAB-MAP: Probabilistic localization and mapping in the space of appearance, Int. J. Robot. Res. 27 (6) (2008) 647–665.
[7] D. Galvez-Lopez, J.D. Tardos, Bags of binary words for fast place recognition in image sequences, IEEE Trans. Robot. 28 (5) (2012) 1188–1197.
[8] T. Nicosevici, R. Garcia, Automatic visual bag-of-words for online robot navigation and mapping, IEEE Trans. Robot. 28 (4) (2012) 886–898.
[9] H. Zhang, B. Li, D. Yang, Keyframe detection for appearance-based visual slam, in: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2010, pp. 2071–2076.
[10] Y. Liu, H. Zhang, Indexing visual features: Real-time loop closure detection using a tree structure, in: 2012 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2012, pp. 3613–3618.
[11] H. Shahbazi, H. Zhang, Application of locality sensitive hashing to realtime loop closure detection, in: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2011, pp. 1228–1233.
[12] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF), Comput. Vis. Image Underst. 110 (3) (2008) 346–359.
[13] E. Stumm, C. Mei, S. Lacroix, Probabilistic place recognition with covisibility maps, in: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2013, pp. 4158–4163.
[14] N. Morioka, S. Satoh, Building compact local pairwise codebook with joint feature space clustering, in: Computer Vision–ECCV 2010, Springer, 2010, pp. 692–705.
[15] E. Johns, G.-Z. Yang, Feature co-occurrence maps: Appearance-based localisation throughout the day, in: 2013 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2013, pp. 3212–3218.
[16] M. Cummins, P. Newman, Appearance-only slam at large scale with FAB-MAP 2.0, Int. J. Robot. Res. 30 (9) (2011) 1100–1123.
[17] R. Mur-Artal, J.D. Tardós, Fast relocalisation and loop closing in keyframe-based slam, in: 2014 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2014, pp. 846–853.
[18] M. Labbé, F. Michaud, Appearance-based loop closure detection for online large-scale and long-term operation, IEEE Trans. Robot. 29 (3) (2013) 734–745.
[19] J. Sivic, A. Zisserman, Video google: A text retrieval approach to object matching in videos, in: Ninth IEEE International Conference on Computer Vision, 2003. Proceedings, IEEE, 2003, pp. 1470–1477.
[20] K. Konolige, J. Bowman, J. Chen, P. Mihelich, M. Calonder, V. Lepetit, P. Fua, View-based maps, Int. J. Robot. Res. (2010).
[21] A. Angeli, S. Doncieux, J.-A. Meyer, D. Filliat, Real-time visual loop-closure detection, in: IEEE International Conference on Robotics and Automation, 2008. ICRA 2008, IEEE, 2008, pp. 1842–1847.
[22] R. Paul, P. Newman, FAB-MAP 3D: Topological mapping with spatial and visual appearance, in: 2010 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2010, pp. 2649–2656.
[23] H. Zhang, Borf: Loop-closure detection with scale invariant visual features, in: 2011 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2011, pp. 3125–3130.
[24] K. Hajebi, H. Zhang, An efficient index for visual search in appearance-based SLAM, 2013. arXiv Preprint arXiv:1309.7170.
[25] M.J. Milford, G.F. Wyeth, SeqSLAM: visual route-based navigation for sunny summer days and stormy winter nights, in: 2012 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2012, pp. 1643–1649.
[26] S. Lynen, M. Bosse, P. Furgale, R. Siegwart, Placeless place-recognition, in: 3D Vision (3DV) 2014 2nd International Conference on. Vol. 1, IEEE, 2014, pp. 303–310.
[27] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[28] J. Bergin, Sets, maps, multisets, and multimaps, in: Data Structure Programming, Springer, 1998, pp. 239–266.
[29] D. Nister, H. Stewenius, Scalable recognition with a vocabulary tree, in: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Vol. 2, IEEE, 2006, pp. 2161–2168.
[30] M. Muja, D.G. Lowe, Fast approximate nearest neighbors with automatic algorithm configuration. In: VISAPP (1), 2009, pp. 331–340.
[31] Z. Wu, Q. Ke, M. Isard, J. Sun, Bundling features for large scale partial-duplicate web image search, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009, IEEE, 2009, pp. 25–32.
[32] G. Bradski, OpenCV, 2008.
[33] S. Khan, D. Wollherr, IBuILD: Incremental bag of binary words for appearance based loop closure detection, in: 2015 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2015, pp. 5441–5447.
[34] M. Boulekchour, N. Aouf, Efficient real-time loop closure detection using GMM and tree structure, in: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2014), IEEE, 2014, pp. 4944–4949.
[35] J.L. Blanco, Malaga dataset-parking 6l. 2009. URL: http://www.mrpt.org/a-repository-of-robotics-datasets/dataset-malaga-dataset-2009-parking-6l/.
[36] N. Tongprasit, A. Kawewong, O. Hasegawa, PIRF-Nav 2: Speeded-up online and incremental appearance-based slam in an indoor environment, in: 2011 IEEE Workshop on Applications of Computer Vision (WACV), IEEE, 2011, pp. 145–152.
[37] R. Anati, K. Daniilidis, Constructing topological maps using Markov random fields and loop-closure detection, in: Advances in Neural Information Processing Systems, 2009, pp. 37–45.

Nishant Kejriwal obtained his Bachelor's degree in Computer Science from Indian Institute of Technology Jodhpur in 2012. Since then, he is working as a researcher at Innovation Labs in Tata Consultancy Services. His research interests include Machine Learning, Robotics and Computer Vision.

Swagat Kumar obtained his Master's and Ph.D. in Electrical Engineering from Indian Institute of Technology Kanpur in the year 2004 and 2009 respectively. He was a post-doctoral researcher at Kyushu University, Fukuoka, Japan during 2009–10. He worked as an assistant professor at Indian Institute of Technology Jodhpur for about two years before joining Tata Consultancy Services as a scientist in 2012. His research interest includes robotics, computer vision, control systems and machine learning. He is a member of IEEE Robotics and Automation Society.

Tomohiro Shibata received his B.E., M.E. and Ph.D. in 1991, 1993, and 1996 from the University of Tokyo. He is currently a professor at the Graduate School of Life Science and Systems Engineering of Kyushu Institute of Science and Technology. His main research interest is on understanding and assisting motor control and decision making by humans by using interdisciplinary approaches. He received a young investigator award from the Robotics Society of Japan (1992), the best paper award from the Japanese Neural Network Society (2002, 2015), the Neuroscience Research Excellent Paper Award from the Japan Neuroscience Society (2007), and the Best Application Paper Award of IROS (2015).

He is a visiting professor at the Nara Institute of Science and Technology and Chubu University, an Editorial Board Member of Neural Networks, an executive board member of the NPO Agora Music Club.