A Two-Layered Pipeline for Topological Place Recognition Based on Topic Model

Xingliang Dong[1], Jing Yuan[1,2(✉)], Fengchi Sun[3], and Yalou Huang[3]

[1] College of Computer and Control Engineering, Nankai University, Tianjin, China
dongxingliang_1@163.com, yuanj@nankai.edu.cn

[2] Information Technology Research Base of Civil Aviation Administration of China,
Civil Aviation University of China, Tianjin, China

[3] College of Software, Nankai University, Tianjin, China
{fengchisun,huangyl}@nankai.edu.cn

Abstract. In this paper, a two-layered pipeline for topological place recognition is proposed. Based on the Bag-of-Words (BoW) algorithm, Latent Dirichlet Allocation (LDA) is applied to extract the topic distribution of single images. Then, the environment is partitioned into different topological places by clustering the images in the topic space. By selecting the relevant topological place for the newly acquired image according to its topic distribution in the first level, we can reduce the number of images to perform the image-to-image matching in the second step while preserving the accuracy. We present evaluation results for City Centre data set using Precision-Recall curves, which reveal that our algorithm outperforms other state-of-the-art algorithms.

Keywords: Place recognition · Topological place discovery
Latent Dirichlet Allocation

1 Introduction

Visual place recognition [1] with mobile robots is an important problem for vision-based navigation and simultaneous localization and mapping (SLAM) [2, 3]. It is a technique that the mobile robot uses the images collected by the visual sensor to determine whether the current place is a new place or the one that has been visited. During the last decade, many novel methodologies have been presented in the literature aiming to address the place recognition task. A well known method is the Bag-of-Words (BoW) algorithm [4–7], which is borrowed from the realm of information retrieval. The framework of BoW is described below. Firstly, all of the local descriptors are collected, and then a clustering

J. Yuan—This work was sponsored in part by Natural Science Foundation of China under grant 61573196, in part by the Natural Science Foundation of Tianjin under grant 15JCYBJC18800 and 16JCYBJC18300 and in part by the Open Project Foundation of Information Technology Research Base of Civil Aviation Administration of China under grant CAAC-ITPB-201503.

J. Yang et al. (Eds.): CCCV 2017, Part III, CCIS 773, pp. 430–441, 2017.
https://doi.org/10.1007/978-981-10-7305-2_37

algorithm is run on this data set. After clustering, each cluster center is selected as a visual word. All cluster centers are selected to form a visual dictionary. A histogram of the term frequency is extracted as image features. This approach is initially inspired for solving image retrieval problems, which represents an image as a vector by defining the term frequency of a visual word.

However, although an image is represented as a collection of visual words, there are still some noisy visual words, which are useless for place recognition. Aiming at this problem, we introduces LDA [8] into place recognition task. The procedure of LDA applied in our work is shown in Fig. 1. Each visual topic is viewed as a mixture of various visual words. Via LDA, the visual topic distribution of each image is assigned according to the visual words extracted from it. The effect of noisy visual words on image features can be reduced. Afterwards, Affinity Propagation (AP) [9] algorithm is performed in visual topic space to partition the environment into different topological places. Then, each topological place is represented by the cluster center, which is the most representative image in this cluster. When a new image is acquired by the mobile robot, Nearest Neighbor (NN) algorithm is applied to find the best matching topological place. Finally, an image-to-image matching is performed to find the final loop closure. The contributions of this work are as follows.

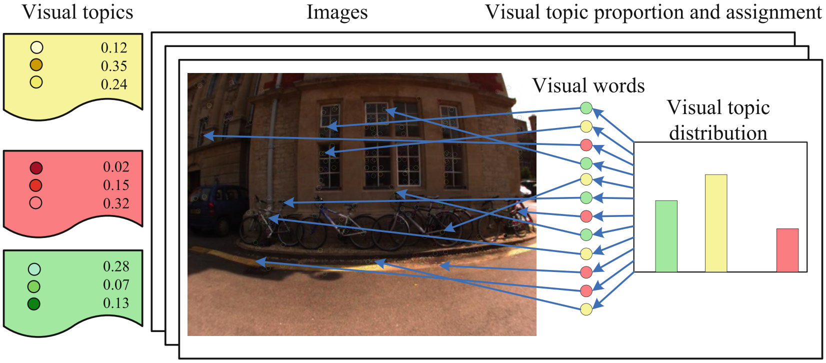


Fig. 1. The procedure of LDA in our work.

(1) The novelty of the presented method lies upon the partition of the matching procedure into two stages: initially between the image and the topological places and later between single images. In the first stage, the images are automatically divided into different topological places by clustering. Rather than matching all of the images in the image database, the time cost of image matching in the first stage is greatly reduced by matching all of the topological places. And then, the images belong to the relevant topological place are chosen as matching candidates. In the second stage, geometrical validation is performed to find the final loop closure between the newly acquired image and matching candidates.

(2) To the best of our knowledge, this is the first attempt at applying LDA in the place recognition task. The visual topic distribution of the image in visual topic space is not only used to discover topological places in the environment, but also to recognize topological places at the first step. Using the visual topic distribution of the image to represent image features, rather than the term frequency, the effect of noisy visual words can be reduced.

The rest of the paper is organized with the following structure: Sect. 2 briefly discusses the related work in the area of appearance-based place recognition. In Sect. 3 the proposed algorithm is described in detail, while Sect. 4 provides experimental results on City Centre data set. Finally, Sect. 5 serves as an epilogue for this paper where our final conclusions are drawn.

2 Related Work

BoW, as a well established technique, has been widely used in the past decade to address the place recognition task [3–7,10–15]. The conventional BoW algorithms are composed of three major steps, selection of local descriptors, generation and organization of the visual dictionary and the measure of similarity between two different images.

There are many kinds of local descriptors, such as SIFT [16], SURF [17], BRIEF [18], ORB [19], which have been successfully applied to the place recognition task. The real-valued descriptors, such as SIFT [16], SURF [17], are more robust against illumination and rotation, while the binary descriptors, such as BRIEF [18], ORB [19], are very fast to compute. Recently, the notion of the use of local descriptors by matching consecutive images, instead of individual images, has been reported in the literatures. Kawewong et al. [5,11] aims to find some descriptors which appear together in a couple of successive images, called PIRF, which are invariant to change of the environment to a large extent. Zhang et al. [14] catches the change of two matched local descriptors from motion dynamics to generate the visual dictionary.

The visual dictionary generation process can be divided into two categories: one is offline and the other is online and incremental. The work presented in [3,4,6,10,12,13,15] can be placed in the first category. This kind of algorithms is easy to generate a visual dictionary by using a clustering algorithm, such as K-Means [4,10], hierarchical K-Median [3,6,12,13,15]. However, it is sensitive to the appearance of the environments. When facing an new environment, the visual dictionary constructed by a known data set may be invalid. The online and incremental algorithms [5,7,11,14] can solve this problem. Accordingly, additional computational cost of update and maintenance of the visual dictionary has to be introduced.

The measure of similarity between two different images is usually calculated with L1-score [3,5,6,11,12,15]. As well, probabilistic models can be applied to calculate the similarity between two different images, such as FAB-MAP [4,10]. Both kinds of similarity can be utilized to find matching candidates.

However, because of the exchangeability of visual words, the geometrical consistency among the local descriptors is ignored, which produces a lot of false matches. Therefore, geometrical validation [2,3,6,12,13,15] is always implemented between the image and matching candidates to find the final loop closure.

Recently, some of two-layered place recognition algorithms [13,15,20], have been proposed. The first level is to find correlations between the image and different image sets, and the images belonging to the relevant image set are chosen as matching candidates. The second level performs geometrical validation between the image and these matching candidates to find the final loop closure. Mohan et al. [13] decomposed the set of images into independent environments and selected the best matching environment before the place recognition. Bampis et al. [15] segmented the image data set into different image sets based on their spatiotemporal proximity. Korrapati et al. [20] adopted image sequence partitioning techniques to group panoramic images together as different image sets.

Despite their success achieved in term of place recognition, there are still lots of noisy visual words in a single image, which are affect the performances of place recognition. Therefore, in this paper, in order to further reveal the nature of the image, LDA [8] is introduced into the place recognition task to mining the visual topic distribution of the image in the visual topic space. Then, through clustering the topic distribution of the image in the topic space, we can find all topological places in the environment. At last, a two-layered pipeline for topological place recognition is proposed, one of which is to find correlations between the image and topological places and the other is geometrical validation on these individual images.

3 Approach

The preparation before our proposed algorithm includes the offline creation of a visual dictionary. In this paper, PIRF [5,11] descriptors are extracted to yield a generic training set. Then, the open source implementation called OpenFABMAP [21] is utilized to create a visual dictionary with default parameters. Finally, we create 3982 words in the final dictionary. In the following, we first provide a brief description of LDA and its usage in extracting visual topics from images. Subsequently, the process of discovering topological places in the environment is discussed in detail. Finally, we elaborate the two levels of matching: the first level is to find correlations between the image and topological places and the second one is to match these individual images.

3.1 Latent Dirichlet Allocation

In the following, we briefly introduce the application of LDA in our work. Suppose we are given a collection D of images $\{I_1, I_2, \cdots, I_M\}$. Each image I is a collection of PIRF descriptors $I = \{\omega_1, \omega_2, \cdots, \omega_N\}$, where ω_i is the visual word index representing the i-th PIRF descriptor. A visual word is the basic item from a visual dictionary indexed by $\{1, 2, \cdots, V\}$.

The LDA model assumes there are K underlying latent visual topics. Each visual topic is represented by a multinomial distribution over the V visual words. An image is generated by sampling a mixture of these visual topics, then sampling visual words conditioning on a particular visual topic. The generative process of LDA for an image I in the collection can be formalized as follows:

I. Choose $\theta \sim \mathrm{Dir}(\alpha)$;
II. For each of the N visual words ω_n:
  (a) Choose a visual topic $z_n \sim \mathrm{Mult}(\theta)$;
  (b) Choose a visual word ω_n from $\omega_n \sim \mathrm{p}(\omega_n|z_n, \beta)$, a multinomial probability conditioned on z_n.

$\mathrm{Dir}(\alpha)$ represents a Dirichlet distribution with the parameter α, and $\mathrm{Mult}(\theta)$ represents a Multinomial distribution with the parameter θ. The parameter θ indicates the mixing proportion of different visual topics in a particular image. α is the parameter of a Dirichlet distribution that controls how the mixing proportions θ vary among different images. β is the parameter of a set of Dirichlet distributions, each of them indicates the distribution of visual words within a particular visual topic. Learning a LDA model from a collection of images $D = \{I_1, I_2, \cdots, I_M\}$ involves finding α and β that maximize the log likelihood of the data $l(\alpha, \beta) = \sum_{d=1}^{M} \log P(I_d|\alpha, \beta)$. This parameter estimation problem can be solved by Gibbs sampling [22]. Then, each image I is represented by a distribution of different visual topics $T_i = \{t_1, t_2, \cdots, t_K\}$, where t_j is the probability that the i-th image contains the j-th visual topic.

3.2 Clustering and Topological Place Discovery

In practical situations, a topological place can be defined as an image set, which contains a group of images which are similar to each other in appearance and continuous in actual positions. Thus, by clustering images in the visual topic space, the environment can be partitioned into different image sets, each of which represents a topological place. However, the number of topological places of the environment is usually hard to specify in advance. Therefore, we choose Affinity Propagation (AP) [9] as the clustering algorithm, which do not need to specify the cluster number in advance and can work for any meaningful measure of similarity between data samples, to discover different topological places in the environment.

In this paper, the similarity between two images contains two parts: one is cosine similarity and the other is the spatial potential function. The definition is shown in (1).

$$sim(i,j) = \frac{T_i \cdot T_j}{\|T_i\| \cdot \left\|T_j\right\|} + \frac{1}{e^{|i-j|}} \tag{1}$$

where i, j are indices of images, T_i is the visual topic distribution of image I_i, $T_i \cdot T_j / \|T_i\| \cdot \left\|T_j\right\|$ is the cosine similarity between I_i and I_j and $e^{-|i-j|}$ is the spatial potential function between I_i and I_j. The cosine similarity part represents that images with similar visual topic distributions can be clustered together. The space potential function guarantees that images that are contiguous in the environment will be clustered together.

After clustering, each cluster represents a topological place. Each cluster center, which is defined as an *exemplar* in AP, is chosen as the representative image of the topological place, which is used in the first step of place recognition.

3.3 Two-Layered Topological Place Recognition

After discovering all topological places in the environment, each topological place is represented by the *exemplar* of each cluster, which is the most representative image in each cluster. When a new image I_t is acquired, we can use the well-trained LDA model to infer its visual topic distribution. Then, we match I_t with all topological places to determine which topological place it belongs to as our first level of recognition. The similarity between the image and the topological place is calculated by cosine similarity. The best matching topological place $P_q = \{I_i | i = k, \cdots, k+m\}$ is found by Nearest Neighbor (NN) algorithm. All images belonging to P_q are chosen as matching candidates.

In the final step, geometrical validation is implemented between the related images using PIRF descriptors. For each $I_i \in P_q$, we adopt RANSAC to compute the fundamental matrix between I_t and I_i. If it fails to provide a fundamental matrix, we reject the loop closure between I_t and I_i. At the same time, if the number of inliers of PIRF descriptors of I_i is less than a constant f, the loop closure between I_t and I_i is also rejected. If more than one I_i meets the above criteria, the one with the most inliers will be chosen as the final loop closure.

4 Experimental Results

In this section, we present the experimental results of the proposed method and compare them with other state-of-the-art methods, including FAB-MAP 2.0 and DBoW2.

4.1 Experimental Setup

We implement the proposed algorithm on City Centre data set collected by Cummins and Newman [4]. The robot traveled twice around a loop with total path length of 2 km along public roads near the city center. The data set includes 1237 image pairs, which were collected by two cameras facing the left and right of the robot, respectively. The data set contains many dynamic objects, such as traffic and pedestrians. The ground truth of the trajectory is shown in Fig. 2. The yellow curve is the robot travelling trajectory, which is obtained by GPS. The loop part of this data set is from the 677^{th} image pairs to the 1237^{th} image pairs.

In this paper, we put SIFT [16] descriptors into the PIRF 2.0 [11] framework. The size of the sliding window in the PIRF 2.0 framework is set to 2.

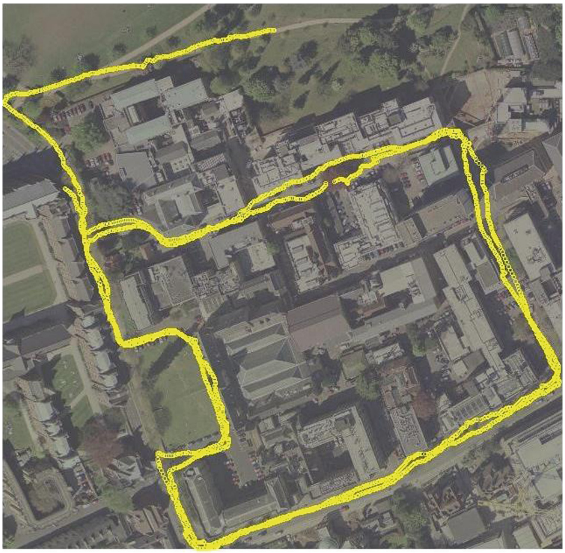


Fig. 2. Motion trajectory of the robot in the City Centre data set. (Color figure online)

As mentioned above, the open source implementation called OpenFABMAP [21] is utilized to create the visual dictionary with default parameters. Finally, we are able to create 3982 words in the final visual dictionary. The number of visual topics is set to 300 in this paper. α is chosen small enough to ensure that each image contains one visual topic as far as possible. β is chosen small enough to make sure that each visual word belongs to only one visual topic. α and β are both set to 0.01.

4.2 Validity of Topological Places

Experimental results of topological place discovery for the first loop of the City Centre data set are shown in Fig. 3, which is drawn by GPS. In the first loop, 143 different topological places are discovered in total. Continuous marks with a same color represent a topological place. We use different colors to make a distinction between two consecutive topological places. It can be seen that images belonging to the same topological place are continuous in actual positions.

Figure 4 shows the left side images of the 71^{st}–73^{rd} topological places of the City Centre data set. It is easy to find the differences among the three topological places. The *exemplar* of each topological place is marked by a red box. It can be observed that the appearance of each *exemplar* is significantly different with others. Thus, the experimental results demonstrate that the proposed topological place discovery algorithm can automatically discover different topological places in the environment and ensure that the images, which belong to a same topological place, are similar with each other in appearance and continuous in terms of physical positions.

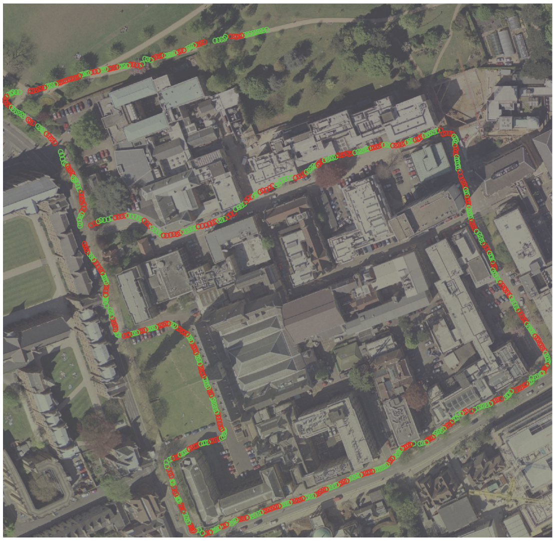


Fig. 3. Topological places discovered from the first loop of the City Centre data set.

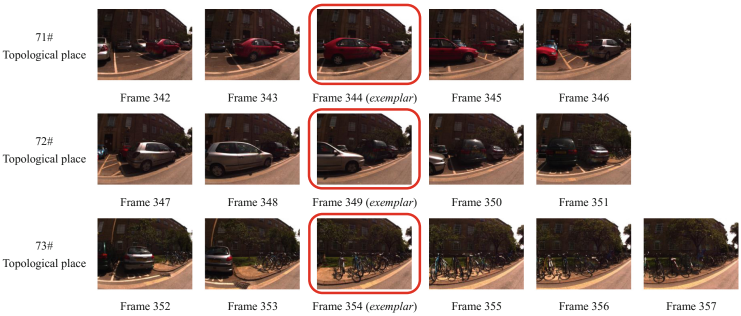


Fig. 4. Three obtained topological places on City Centre data set (left side of image pairs only). (Color figure online)

4.3 Overall Performance

With a view to measuring the performance of our system, we have chosen to use Precision-Recall curves. Precision is defined as the ratio between the correct and the total number of detected loop closures. On the other hand, recall is equal to the ratio of the correctly detected loop closures, to the total number of loops the data set contains.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

where the TP is true positive, i.e., the correctly detected loop closures, FP is false positive, i.e., the incorrectly detected loop closures, and FN is false negative, i.e., the loop closures that the system erroneously discards. Figure 5 shows the Precision-Recall curve for the City Centre data set. We obtained this curve by varying the constant f. It can be seen that our proposed algorithm retains 71.83% Recall with 100% Precision.

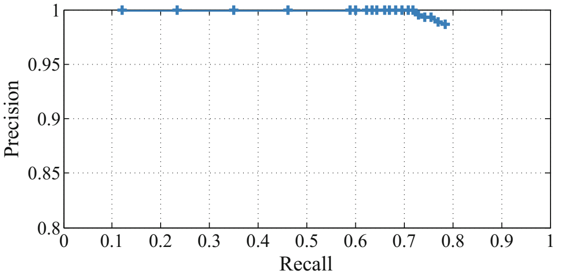


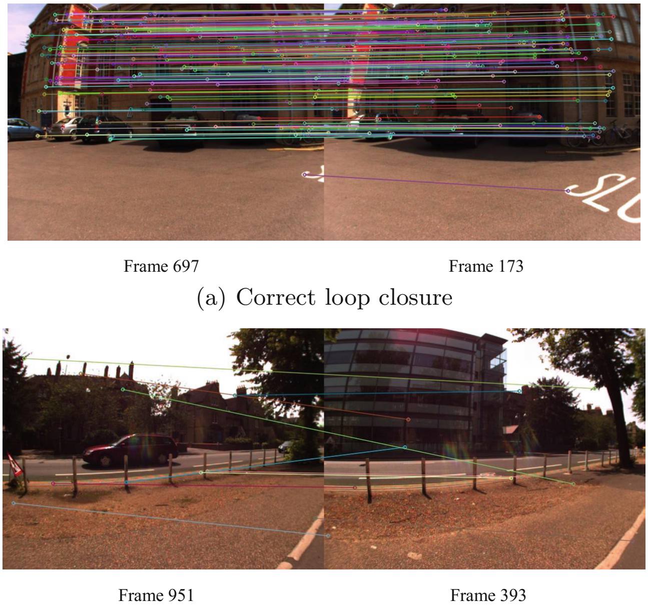
Fig. 5. Precision-Recall curve, for City Centre data set, illustrating the recognition performance of the proposed method.

Comparative results are presented in Table 1 with some of the most well established techniques available in the literatures for the City Centre data set, viz. [6,10,12]. The achieved performance of the aforementioned algorithms was obtained straightforwardly from their respective papers. It can be observed that our algorithm achieves better results than the other three state-of-the-art algorithms.

Table 1. Comparison among four algorithms.

Data set	Approaches	Precision (%)	Recall (%)
City Centre	DBoW2 [6]	100	30.61
	FAB-MAP 2.0 [10]	100	38.77
	Mur-Artal [12]	100	43.03
	Proposed method	**100**	**71.83**

Finally in Fig. 6, example frames on City Centre data set are presented and lines depict final corresponding PIRF descriptors. Figure 6(a) shows an correct loop closure that our algorithm is able to accept. Although the visual angle of the robot has changed, our algorithm is still able to detect the loop closure between the 697^{th} frame and the 173^{rd} frame. An incorrect loop closure that our algorithm is able to reject is shown in Fig. 6(b). It can be observed that the 951^{st} frame and the 353^{rd} frame are quite similar with each other in some local



(a) Correct loop closure

(b) Incorrect loop closure

Fig. 6. Example frames on City Centre data set.

regions. That means there are a lot of common visual words between the 951^{st} frame and the 353^{rd} frame. Because of the exchangeability of visual words, the visual topic distributions of these two frames assigned by LDA are similar with each other. Nevertheless, our algorithm is still able to reject the loop closure between the 951^{st} frame and the 353^{rd} frame according to the number of inliers in PIRF descriptors. Therefore, the experimental results demonstrate that the proposed algorithm can achieve a high performance in place recognition task, even facing dynamic environments with large variability.

5 Conclusions

In this paper, LDA is introduced to assign the visual topic distributions of each image. Image features are represented by the visual topic distributions in the visual topic space. Then, by clustering images in the visual topic space, the environment can be automatically partitioned into different topological places. When a new image is acquired, a two-layered pipeline for topological place recognition is proposed to improve the performance of place appearance. Compared with the state-of-the-art place recognition algorithms, our algorithm is more robust against variation of the place appearance. Furthermore, a complete representation for the visual topics can help the robot to better understand and interpret the environments, even to extract the semantic cues, which can bridge the gaps of understanding and expression for the environmental information between the human and the robot.

References

1. Lowry, S.M., Sünderhauf, N., Newman, P., Leonard, J.J., Cox, D.D., Corke, P.I., Milford, M.J.: Visual place recognition: a survey. IEEE Trans. Robot. **32**(1), 1–19 (2016)
2. Konolige, K., Agrawal, M.: Frameslam: from bundle adjustment to real-time visual mapping. IEEE Trans. Robot. **24**(5), 1066–1077 (2008)
3. Mur-Artal, R., Montiel, J.M.M., Tardós, J.D.: ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE Trans. Robot. **31**(5), 1147–1163 (2015)
4. Cummins, M., Newman, P.M.: FAB-MAP: probabilistic localization and mapping in the space of appearance. Int. J. Robot. Res. **27**(6), 647–665 (2008)
5. Kawewong, A., Tongprasit, N., Tangruamsub, S., Hasegawa, O.: Online and incremental appearance-based SLAM in highly dynamic environments. Int. J. Robot. Res. **30**(1), 33–55 (2011)
6. Gálvez-López, D., Tardós, J.D.: Bags of binary words for fast place recognition in image sequences. IEEE Trans. Robot. **28**(5), 1188–1197 (2012)
7. Khan, S., Wollherr, D.: IBuiLD: incremental bag of binary words for appearance based loop closure detection. In: IEEE International Conference on Robotics and Automation, ICRA 2015, Seattle, WA, USA, 26–30 May 2015, pp. 5441–5447 (2015)
8. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
9. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science **315**(5814), 972–986 (2007)
10. Cummins, M., Newman, P.M.: Appearance-only SLAM at large scale with FAB-MAP 2.0. Int. J. Robot. Res. **30**(9), 1100–1123 (2011)
11. Kawewong, A., Tongprasit, N., Hasegawa, O.: Pirf-Nav 2.0: fast and online incremental appearance-based loop-closure detection in an indoor environment. Robot. Auton. Syst. **59**(10), 727–739 (2011)
12. Mur-Artal, R., Tardós, J.D.: Fast relocalisation and loop closing in keyframe-based SLAM. In: 2014 IEEE International Conference on Robotics and Automation, ICRA 2014, Hong Kong, China, 31 May–7 June 2014, pp. 846–853 (2014)
13. Mohan, M., Gálvez-López, D., Monteleoni, C., Sibley, G.: Environment selection and hierarchical place recognition. In: IEEE International Conference on Robotics and Automation, ICRA 2015, Seattle, WA, USA, 26–30 May 2015, pp. 5487–5494 (2015)
14. Zhang, G., Lilly, M.J., Vela, P.A.: Learning binary features online from motion dynamics for incremental loop-closure detection and place recognition. In: 2016 IEEE International Conference on Robotics and Automation, ICRA 2016, Stockholm, Sweden, 16–21 May 2016, pp. 765–772 (2016)
15. Bampis, L., Amanatiadis, A., Gasteratos, A.: Encoding the description of image sequences: a two-layered pipeline for loop closure detection. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2016, Daejeon, South Korea, 9–14 October 2016, pp. 4530–4536 (2016)
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)

17. Bay, H., Ess, A., Tuytelaars, T., Gool, L.J.V.: Speeded-up robust features (SURF). Comput. Vis. Image Underst. **110**(3), 346–359 (2008)
18. Calonder, M., Lepetit, V., Strecha, C., Fua, P.: BRIEF: binary robust independent elementary features. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 778–792. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15561-1_56
19. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.R.: ORB: an efficient alternative to SIFT or SURF. In: IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, 6–13 November 2011, pp. 2564–2571 (2011)
20. Korrapati, H., Courbon, J., Mezouar, Y., Martinet, P.: Image sequence partitioning for outdoor mapping. In: IEEE International Conference on Robotics and Automation, ICRA 2012, 14–18 May 2012, St. Paul, Minnesota, pp. 1650–1655 (2012)
21. Glover, A.J., Maddern, W.P., Warren, M., Reid, S., Milford, M., Wyeth, G.: OpenFABMAP: an open source toolbox for appearance-based loop closure detection. In: IEEE International Conference on Robotics and Automation, ICRA 2012, 14–18 May 2012, St. Paul, Minnesota, pp. 4730–4735 (2012)
22. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Nat. Acad. Sci. U.S.A. **101**(Suppl 1), 5228–5235 (2004)