Self-localization Using Visual Experience Across Domains

Tsukamoto Taisho

Tanaka Kanji

Abstract—In this study, we aim to solve the single-view robot self-localization problem by using visual experience across domains. Although the bag-of-words method constitutes a popular approach to single-view localization, it fails badly when its visual vocabulary is learned and tested in different domains. Further, we are interested in using a cross-domain setting, in which the visual vocabulary is learned in different seasons and routes from the input query/database scenes. Our strategy is to mine a cross-domain visual experience, a library of raw visual images collected in different domains, to discover the relevant visual patterns that effectively explain the input scene, and use them for scene retrieval. In particular, we show that the appearance and the pose of the mined visual patterns of a query scene can be efficiently and discriminatively matched against those of the database scenes by employing image-to-class distance and spatial pyramid matching. Experimental results obtained using a novel cross-domain dataset show that our system achieves promising results despite our visual vocabulary being learned and tested in different domains.

I. INTRODUCTION

In this study, we aim to solve the problem of cross-domain single-view robot self-localization. For solving SLAM and other similar problems in mobile robotics, visual localization is crucial [1]–[3], [5]. While self-localization can be done either by using prior knowledge of the problem domain [1] or without them [2], we deal with applications and scenes where a collection of visual images from different domains, termed cross-domain visual experience, is available as prior knowledge. At the same time, we require the localization algorithm to be extremely fast (to work in a fast robot navigation) and to recognize the place from a single frame [1] (i.e., without temporal tracking [7] and visual sequence measurements [8]).

One of most popular approaches to address the problem of single-view localization is bag-of-words methods [1], [5], [10], wherein a collection of local invariant visual features is extracted from an input image, and each feature is translated into a visual word by using a pre-learned library of vector-quantized features. Consequently, an input scene image is described compactly and discriminatively as an unordered collection of visual words (“bag-of-words”). However, as argued by several authors, the bag-of-words method fails badly when learned and tested in different domains; the main reasons include the following: (1) Because the bag-of-words method ignores all information about the spatial layout of the features, it limits the descriptive ability considerably [9]. (2) Essential features are lost during quantization [10].

Our work has been supported in part by JSPS KAKENHI Grant-in-Aid for Young Scientists (B) 23700229, and for Scientific Research (C) 26330297 ("The realization of next-generation, discriminative and succinct SLAM technique: PartSLAM").

T. Tsukamoto and K. Tanaka are with Graduate School of Engineering, University of Fukui, Japan. tnknknj@u-fukui.ac.jp

Fig. 1. Cross-domain scene descriptor (CD-SD). Our objective is to take an input image (top left) and the keypoints extracted from the image (top right) as input, and create a scene descriptor whose spatial layout is globally similar to that of the keypoints in the input image and locally similar to the visual patterns mined from a cross-domain vocabulary, which is a set of raw visual images collected in different domains (season + route). Shown in the lowermost image of Fig. 1 is a visualization of our scene descriptor, in which visual patterns mined from the vocabulary are placed at the locations of the corresponding keypoints.

Recently, image-to-class matching techniques have received increasing attention in cross-domain classification tasks [11]–[14]. In [11], a novel example-based non-parametric NBNN classifier was presented. This classifier combines the Naive Bayes assumption with an approximate Parzen estimate. In [12], the NBNN approach was used for domain adaptation, and it achieved state-of-the-art performance. The NBNN approach is appropriate under the following two conditions: (1) raw visual features are used without vector quantization, and (2) the image-to-class (rather than image-to-image) distances are used for scene comparison. However, in the abovementioned studies, the NBNN approach was used for image classification tasks having pre-defined scene classes. Therefore, the NBNN approach cannot be used directly for those localizations in which there is no explicit scene class; the class has to be learned by the robot itself in an unsupervised manner.

In this study, we address the above issues by mining visual experiences. Our strategy is to mine a cross-domain visual experience (i.e., a library of raw visual images collected in different domains) to find the relevant visual patterns to effectively explain the input scene. Our approach is motivated by the following facts:
A library of raw images enables a quantization-free approximation of an input scene.

The mined visual features can serve as training data for a given input database scene that characterizes a place.

It is often feasible to explain natural scenes by using visual patterns mined from such an image library [15].

Based on the above idea, we propose the approximation of an input local feature \( f = \langle p, a \rangle \), which consists of a keypoint \( p \) and an appearance descriptor \( a \), by a set of IDs \( - l_1^1, \ldots, l^K \) of \( K \) library features mined from the visual experience, in the following form: \( f' = \langle p, l_1^1, \ldots, l^K \rangle \). This is a compact representation that can be indexed and matched efficiently by an inverted file system.

As the next contribution, we show that the proposed scene descriptor successfully realizes cross-domain localization. In particular, we are interested in a specific cross-domain setting where both the seasons and routes of the visual experience are different from those of the input query/database scenes. Based on the above discussion, we view places (i.e., database images) as independent classes, and for each class, we form a class-specific set of training features, by mining the visual experience to find the relevant library features that effectively explain the input scenes. We show that both the appearance and pose of the mined visual patterns of a query scene can be matched against those of the database scenes by employing image-to-class distance [2] and spatial pyramid matching [9]. We conducted experiments using a new “cross-domain” dataset created from a publicly available image collection in our previous ICRA15 paper [17], and found that our system achieved promising results despite our visual library being learned and tested in different domains.

A. Related Work

In this study, we are interested in cross domain localization rather than scalable localization. Many of existing localization frameworks focus on scalable localization that is characterized by a large-size vocabulary. As an instance, for success of previous BoW methods [1] (including its variants for soft assignment and multiple assignment per word) in scalable localization, a library of vector quantized visual features is trained and serves as a quantizer. However, this also imposes a limitation, vector quantization errors as aforementioned. In contrast, in our approach, each library feature is directly stored without being approximated nor vector quantized.

On the other extreme, in [3], a simple solution to localization —using a “bag-of-raw-features” which matches raw SIFT-like features directly, rather than their vector-quantized representation, is presented. However, this may not be possible in practice when the database size increases because of the high dimensionality of raw SIFT-like descriptors. Our approach can be viewed as exploiting this type of raw feature matching, not for direct matching between query and database images but rather for mining an available visual experience to find discriminative visual landmarks.

The problem of cross domain localization has been attracting increasing attention in recent years [18]–[23]. [20] is based on the assumption that some of the mid- to long-term processes that cause the environment changes are (pseudo)periodic, e.g., seasonal foliage variations, daily illumination cycle and routine human activities. In [21], a calibrated, synchronized, and ground truth-aligned dataset of woodland trail navigation in semi-structured and changing outdoor environments is presented. The study described in [22] addressed the problem of change removal and presented a novel approach to learning about appearance change and generalizing the learned change to new locations. In [23], a lazy-sequences matching algorithm under substantial appearance changes was presented. Very recently, in [24], a localization approach using a map of path memories, “visual experiences”, where an experience is a single representation of the environment under particular conditions, much like a snapshot, is presented. However, none of the existing works deal with the cross domain localization from a novel perspective of fast single-view localization.

In the literature, NBNN techniques [11]–[14] have been mainly studied in the context of image categorization and classification with pre-defined scene classes (e.g., place categorization [25]). In [11], the concept of NBNN was introduced by extending NN techniques to satisfy conditions (1) and (2) stated above, and an improved image classification performance was achieved. In [13], the NBNN framework was extended to a kernelized version of NBNN. In [14], pooling strategies were introduced into the NBNN framework. Most relevant to our study is [12], in which the NBNN technique was used for domain adaptation and achieved state-of-the-art performance in addressing the cross-domain image categorization problem. In contrast, we cast our task as a robot self-localization problem which requires the localization algorithm to be extremely fast, and in which there is no explicit scene class (i.e., place); the class has to be learned by the robot itself, in an incremental manner.

Conceptually, our approach is motivated by our previous work on visual experience mining in ICRA15, IROS14, and VPRCE15 papers [17], [26], [27], and differs from all the above works on bag-of-words, direct matching, image-to-class matching and spatial pyramid matching. This study is also different from our previous work on “cross-season” localization in [17] where the vocabulary is learned and tested in mutually overlapping routes. To the best of our knowledge, these issues have not been explored in existing work.

II. APPROACH

A. Library of Visual Experience

As the first novelty of our approach, a library of raw image data for local feature vocabulary, is used as a prior without vector quantization. The library images are not required to be associated with spatial information such that the viewpoint and orientation are known. For example, they can be visual experiences obtained by the robot itself in a previous navigation, images shared by other colleague robots, or a publicly available image data resource on the web, such as Google StreetView.
Our "cross-domain" library consists of view images collected using a handheld monocular camera in a university campus over four different seasons and three different routes (Fig. 2). The routes start at three different locations inside the university campus,— some going through the main central path and others going through the pedestrian walkway along the campus wall, as shown in Fig. 2. We considered a typical scenario that deals with view images taken relatively far apart (e.g., 1m) from each other. Severe occlusion occurred in all the scenes, which were occupied by people and vehicles as dynamic entities. The datasets also have significant viewpoint change due to their handheld nature. As can be seen a single place looks quite different depending on the geometric conditions (e.g., viewpoints, fresh snow covers) and photometric conditions, making our localization task a challenging one.

B. Scene Descriptor

Similar to bag-of-words approaches the proposed scene descriptor represents an input image by unordered collections of local features; however, we do not rely on a library of vector quantized visual features (i.e., visual words) and domain-specific library learning; instead, we use a library of \( V \) raw visual features:

\[
\{z[1], \cdots, z[V]\}.
\]

(1)

Our feature library is obtained by computing SIFT local features for all images in the image library.

Our scene interpretation step begins by extracting \( N \) SIFT features from the query (or database) scene image. For each extracted feature \( x \), we search its \( K \) nearest neighbor library features \( S^{K-NN} = \{i_k | k \in [1, V] \} \) (NN features) whose distances from \( x \) are shortest over the library, and describe the search result using a \( V \)-dimensional vector \( f^{query} \) (or \( f^{database} \)) whose \( i \)-th element is a truncated similarity:

\[
f^{query}[i] = \begin{cases} 
\max(D^0 - ||x - z[i]||^2, 0) & \text{if } i \in S^{K-NN} \\
0 & \text{(otherwise)}. 
\end{cases}
\]

(2)

\[
f^{database}[i] = \begin{cases} 
1 & \text{if } i \in S^{K-NN} \\
0 & \text{(otherwise)}.
\end{cases}
\]

(3)

The resulted descriptor is a size \( N \) set of \( V \)-dim vectors \( \{f_1, \cdots, f_N\} \) which we term nearest neighbor descriptor. Since each \( V \)-dim vector is a sparse vector with only \( K \) (or \( K' \)) non-zero elements, our descriptor can be efficiently stored and compared using an inverted file system. In addition, we need to store only IDs of non-zero elements for mapped images to compact the database, and we use both IDs and similarity values for query images. We fix the parameters \( K \) and \( D_0 \) throughout the paper: \( D_0 = 200 \), \( K = 10 \) for query feature, and \( K' = 3 \) for database feature.

For scene comparison, we employ the image-to-class (rather than image-to-image) distance with the NBNN formulation, which has proven to be effective for cross domain scene recognition in [12]. For this purpose, we view places (i.e., database images) as independent classes, and for each class, we prepare a class specific set of training features. The training set is obtained by a data mining approach, in which the feature library is mined for discovering NN library features that effectively explain each feature in the database image of interest, and the training set is represented by the mined NN features for each class. A comparison between a pair of a query NN descriptor \( f^{query} \) and a database NN descriptor \( f^{database} \) is based on a similarity function in the form:

\[
I(f^{query}, f^{database}) = \max_{i=1}^V \left( f^{query}[i] \right) \left( f^{database}[i] \right). 
\]

(4)

Then, we solve the scene matching problem as a search problem in the form:

\[
e^* = \arg \max \sum_{i=1}^N \left( \max_{j=1}^N I(f^{query}_i, f^{database}_j) \right)
\]

(5)

where \( c \) is a candidate class (i.e., place) and \( x \) is the query feature.

C. Spatial Pyramid Matching

We adopt spatial pyramid matching by placing a sequence of increasingly coarser grids over the image region and taking a weighted sum of the number of matches that occur at each level of resolution. At any fixed resolution, two features are compared in terms of a given similarity measure \( I \). \( K \) is the pyramid match kernel, defined as:

\[
K(X, Y) = \frac{1}{2^L} f^0 + \sum_{l=1}^L \frac{1}{2^{L-l+1}} f^l
\]

(6)

In the original implementation of spatial pyramid matching in [9], the similarity measure \( I \) aims to measure similarity between a pair of bag-of-words vectors in terms of histogram intersection. Note that the number of matches at level \( l \) also includes all the matches found at the finer level \( l+1 \). Therefore, the number of new matches found at level \( l \) is given by \( f^l - f^{l+1} \) for \( l = 0, \cdots, L \). In addition, matches found at coarser level are penalized by using a weight \( 1/2^l \) which is inversely proportional to the cell width at level \( l \). In this study, we replace this similarity function \( I \) that measures similarity in terms of the NBNN based similarity, so that we can incorporate the robust NBNN distance within the
spatial pyramid matching framework, to obtain image level similarity. Note that the similarity function $K$ reduces to a standard NBNN similarity function when $L = 0$. The resulted scene descriptor is a sparse long $16V$-dim vector formed by concatenating the NN descriptors at the finest level.

III. EXPERIMENTS

A. Settings

We collected a new “cross-domain” dataset on three different paths #1-#3 in a university campus environment as shown in Fig. 3. We went each path for three times, collected three independent collections of images and use each for query, library and database image collection. The datasets were collected across all the four seasons over a period of one year, as shown in Fig. 2. In addition, we collected an independent set of 3,537 images in the autumn season on routes different from those above and added it to all the databases considered here, as a set of additional distracter images.

In this study, we implemented several comparing methods, FAB-MAP, TF-IDF, CPD, and an NBNN descriptor w/o spatial pyramid (NBNN-SD) and the proposed cross-domain scene-descriptor with spatial pyramid (CD-SD), for performance evaluation. FAB-MAP is appearance based localization framework based on a bag-of-words scene models and learning of co-occurrence of visual words. We used the code provided by the authors in [1]. TF-IDF is a standard method for bag-of-words image retrieval. We follow a setting in [10] and set the visual dictionary size 10,000. CPD is the technique we have proposed in previous ICRA15 paper, where a small collection of 8 visual phrases that effectively explain an input scene is discovered by a common pattern discovery between the input and the visual experience. We used the same code as in [17]. NBNN-SD is different from the proposed method only in that it does not consider spatial layout of the scene and this is realized by setting the parameter $L = 0$. Finally, CD-SD is the proposed method, described in previous section, Section II. We set the parameter $L = 2$ as a default.

Fig. 4 shows an investigation of approximation errors of the proposed approach. Recalling that our approach approximates each SIFT descriptor in a query image by an NN library feature mined from the library of visual experience, the approximation error is represented by distance between an input SIFT descriptor and the NN library descriptor. In this study, we investigate the approximation error induced by cross domain library, and compare it with that induced by non cross domain library. To this end, we compared distance from each query feature to the nearest neighbor library feature.

B. Examples of Image Retrieval

Fig. 5 shows five examples of image retrieval, from left to right, the input query image, the ground-truth database image, and images retrieved by the proposed method and by FAB-MAP. One can see that each method returns database images that are similar to query image. However, the FAB-MAP method tends to fail when there are confusing images in the database whose appearance is partially similar with the query image but with different spatial layout. In constrast, the proposed method was successful in these examples by using both appearance and spatial layout of
features as a discriminative cue.

C. Performance Results

Table I shows performance results. We evaluated the proposed CD-SD method and other comparing methods, TF-IDF, FAB-MAP, and NBNN-SD. We found that the CPD method was very slow to be tested on a database of this size, and therefore, a comparison with CPD using a relatively small database was reported in Section III-D.

Series of independent 12 × 100 retrievals are conducted for each of the 12 combinations of routes and seasons. The retrieval performance was measured in terms of the percentage (%) averaged normalized rank (ANR)—a ranking-based retrieval performance measure, for which a smaller value indicates better performance. The ranking based performance measure is more suitable for localization task than precision/recall measure that has been often used in classical image retrieval tasks which aim at finding as many relevant database images as possible for a given query image. This measure is motivated by the fact that for the sake of localization, it is sufficient to retrieve just one relevant image.

To evaluate ANR, we evaluated the rank assigned to the ground-truth relevant image for each of the 100 independent retrievals; we normalized the rank with respect to the ground-truth relevant image for each of the 100 independent retrievals. For the ground truth, an image i that is most similar to the query image is manually selected from the database images and its neighbors [i − 10; i + 10] are defined as ground truth images. Different database is prepared for different query and contains only one random ground-truth database image. Note that our database images contain spatially dense viewpoints, which makes it difficult for a localizer to distinguish between them. From Table I, one can see that our approach outperformed the BoW method in most of the retrievals considered here.

Table II shows mAP performance. We observe that mAP value tends to be a low value, as each dataset contains only one ground-truth database image in our case. One can see that the proposed CD-SD method again outperforms the other methods considered here.

D. Comparison against CPD

We conducted an independent experiment to compare the performance between the proposed CD-SD method and CPD method, due to the reason described in Section III-C. Table III shows the performance results. One can see that the proposed method clearly outperforms the CPD method. A main reason is that the CPD method uses a region-level feature, “visual phrase”, to represent the visual patterns mined from the visual experience, and it often fails to explain
Fig. 6. Histograms of visual words used for explaining individual categories (seasons+routes) of input images. Shown in each row is the histogram for each category of input scene. “AU,” “WI,” “SP,” and “SU” are seasons, while “1,” “2,” and “3” are routes.

E. Comparison against Different Vocabularies

Table II reports a result of comparing performance results between cross-domain (CD) vocabulary focused in this paper and three different vocabularies, cross-season (CS), cross-route (CR), and full (FULL) vocabularies. CS or CR is a union of vocabularies from different seasons or routes, and thus, is an easier setting than CD. FULL is the easiest setting, in which a union of all the 12 vocabularies is employed. The performance results suggest that localization by using the CR vocabulary is not as bad as localization by using the CD vocabulary. While localization by using the CR vocabulary can be considered easier than localization by using the CD vocabulary, it is challenging compared to localization by using the CS or FULL vocabulary.

F. Frequency of library images

Fig. 6 summarizes the frequency of individual vocabularies being used for explaining individual scene categories. In this study, we categorized library images into 12 disjoint classes (i.e., 4 seasons 3 routes), and investigated the frequency of library images from each class used for explaining query images. We summarize the result for each of the 12 disjoint classes and show it in the figure. We can see that the frequency of the most and the least frequently used vocabularies differ by almost three times, and that every vocabulary is used for every case considered in the current experiments.

G. Spatial Pyramid Matching

We here demonstrate the effect of spatial pyramid matching compared to conventional image-to-image matching. The key idea of spatial pyramid matching is to match sub-images at each location at each level between query and database images. To demonstrate the effect, we visualize which partitions are similar between query and database images for several examples of pairs of query and database images. To this end, we evaluated the sub-image level similarity between query and database images for each of the $\sum_{l=0}^{4} 4^l = 21$ pairs of sub-images, and selected top 5 similar sub-image pairs, i.e., 5 sub-image pairs that most contributed to the image-level similarity. Fig. 7 shows the top 5 similar sub-image pairs by overlaying bounding boxes on the original input image. We can see that the largest sub-image that corresponds to
Fig. 8. Results for scene parsing. Images in the middle row of this figure are pairs of query images and similar database images. Shown in the upper and lower rows of the figure, is the visualization of our scene parsing for the query and the database images, respectively. The essence of our scene parsing is to explain the SIFT descriptors in an input image by using its NN library descriptor from the library images. To visualize which SIFT descriptor is explained by which library image, we show a synthesized image where a small (size: 20 × 20 pixels) patch containing a SIFT descriptor in an input image is replaced by a patch of the same size mined from the library image.

the entire image region was always selected as one of the most contributed subimages for all the cases and that salient subimages of various sizes and locations were selected.

H. NN Descriptor

Fig. 8 demonstrates scene parsing for two pairs of query and database images. The essence of our scene parsing is to explain the SIFT descriptors in an input image by using its NN library descriptor from the library images. To visualize which SIFT descriptor is explained by which library image, we show a synthesized image where a small (size: 20 × 20 pixels) patch containing a SIFT descriptor in an input image is replaced by a patch of the same size mined from the library image. Fig. 8 shows a visualization result. It can be seen that structure and nature parts of a scene are well explained by visual patterns mined from very different seasons and places.

IV. Conclusions

In this study, we aimed to solve the single-view robot self-localization problem by using a cross-domain vocabulary. Further, we are interested in using a cross-domain setting, in which the visual vocabulary is learned in different seasons and routes from the input query/database scenes. Our strategy is to mine a cross-domain visual experience, a library of raw visual images collected in different domains, to discover the relevant visual patterns that effectively explain the input scene, and use them for scene retrieval. In particular, we showed that the appearance and the pose of the mined visual patterns of a query scene can be efficiently and discriminatively matched against those of the database scenes by employing image-to-class distance and spatial pyramid matching. Experimental results obtained using a novel cross-domain dataset showed that our system achieves promising results despite our visual vocabulary being learned and tested in different domains.

REFERENCES

[1] M. Cummins and P. Newman, “Appearance-only slam at large scale with fab-map 2.0,” *Int. J. Robotics Research*, vol. 30, no. 9, pp. 1100–1123, 2011.

[2] S. Se, D. Lowe, and J. Little, “Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks,” *Int. J. robotics Research*, vol. 21, no. 8, pp. 735–758, 2002.
[3] H. Zhang, “Borf: Loop-closure detection with scale invariant visual features,” in *Proc. IEEE ICRA*, 2011, pp. 3125–3130.
[4] R. Arandjelovic and A. Zisserman, “Three things everyone should know to improve object retrieval,” in *CVPR*, 2012, pp. 2911–2918.
[5] J. Knopp, J. Sivic, and T. Pajdla, “Avoiding confusing features in place recognition,” in *ECCV*, 2010, pp. 748–761.
[6] J. Sivic and A. Zisserman, “Video google: A text retrieval approach to object matching in videos,” in *ICCV*, 2003, pp. 1470–1477.
[7] D. Burschka and G. D. Hager, “V-gps (slam): Vision-based inertial system for mobile robots,” in *Robotics and Automation, 2004. Proceedings. ICRA’04. 2004 IEEE International Conference on*, vol. 1. IEEE, pp. 409–415.
[8] M. J. Milford and G. F. Wyeth, “Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights,” in *ICRA*, 2012, pp. 1643–1649.
[9] S. Lazebnik, C. Schmid, J. Ponce, et al., “Spatial pyramid matching,” *Object Categorization: Computer and Human Vision Perspectives*, vol. 3, p. 4, 2009.
[10] H. Jegou, M. Douze, and C. Schmid, “Packing bag-of-features,” in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 2357–2364.
[11] O. Boiman, E. Shechtman, and M. Irani, “In defense of nearest-neighbor based image classification,” in *CVPR*, 2008, pp. 1–8.
[12] T. Tommasi and B. Caputo, “Frustratingly easy NBNN domain adaptation,” in *ICCV*, 2013, pp. 897–904.
[13] T. Tuytelaars, M. Fritz, K. Saenko, and T. Darrell, “The nbnn kernel,” in *ICCV*, 2011, pp. 1824–1831.
[14] K. Rematas, M. Fritz, and T. Tuytelaars, “The pooled nbnn kernel: Beyond image-to-class and image-to-image,” in *ACCV*, 2013, pp. 176–189.
[15] P. Isola and C. Liu, “Scene collaging: Analysis and synthesis of natural images with semantic layers,” in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3048–3055.
[16] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, “Monoslam: Real-time single camera slam,” *IEEE Trans. PAMI*, vol. 29, no. 6, pp. 1052–1067, 2007.
[17] A. Masatoshi, C. Yuuto, T. Kanji, and Y. Kentaro, “Leveraging image-based prior in cross-season place recognition,” in *ICRA*. IEEE, 2015, pp. 5455–5461.
[18] N. Sunderhauf, M. Milford, and P. Corke, in *ICRA 2015 Workshop on Visual Place Recognition in Changing Environments (VPRiCE)*, 2015.
[19] H. Morita, M. Hild, J. Miura, and Y. Shirai, “View-based localization in outdoor environments based on support vector learning,” in *IROS*. IEEE, 2005, pp. 2965–2970.
[20] T. Krajinik, J. P. Fentanes, J. M. Santos, K. Kusumam, and T. Duckett, “Fremen: Frequency map enhancement for long-term mobile robot autonomy in changing environments,” in *ICRA15 WS VPRiCE*, 2015.
[21] J. Bruce, J. Wawerla, and R. Vaughan, “The sfu mountain dataset: Semi-structured woodland trails under changing environmental conditions,” in *ICRA15 WS VPRiCE*, 2015.
[22] S. Lowry and M. J. Milford, “Change removal: Robust online learning for changing appearance and changing viewpoint,” in *ICRA15 WS VPRiCE*, 2015.
[23] O. Vysotska and C. Stachniss, “Lazy sequences matching under substantial appearance changes,” in *ICRA15 WS VPRiCE*, 2015.
[24] C. Linegar, W. Churchill, and P. Newman, “Work smart, not hard: Recalling relevant experiences for vast-scale but time-constrained localisation,” in *Proc. IEEE International Conference on Robotics and Automation (ICRA2015)*, 2015.
[25] H. Carrillo, Y. Latif, J. Neira, and J. A. Castellanos, “Place categorization using sparse and redundant representations,” in *IROS*, 2014, pp. 4950–4957.
[26] K. Tanaka, Y. Chokushi, and M. Ando, “Mining visual phrases for long-term visual slam,” in *IROS*, 2014, pp. 136–142.
[27] Y. Kentaro, A. Masatoshi, C. Yuuto, and T. Kanji, “An experimental study of the effects of landmark discovery and retrieval on visual place recognition across seasons,” in *ICRA15 WS VPRiCE*, 2015.
[28] J. Sivic and A. Zisserman, “Video google: A text retrieval approach to object matching in videos,” in *IEEE Int. Conf. Computer Vision (ICCV)*, 2003, pp. 1470–1477.