Mining Visual Phrases for Long-Term Visual SLAM

Tanaka Kanji Chokushi Yuuto Ando Masatoshi

***Abstract*— We propose a discriminative and compact scene descriptor for single-view place recognition that facilitates long-term visual SLAM in familiar, semi-dynamic and partially changing environments. In contrast to popular bag-of-words scene descriptors, which rely on a library of vector quantized visual features, our proposed scene descriptor is based on a library of raw image data (such as an available visual experience, images shared by other colleague robots, and publicly available image data on the web) and directly mine it to find visual phrases (VPs) that discriminatively and compactly explain an input query / database image. Our mining approach is motivated by recent success in the field of common pattern discovery—specifically mining of common visual patterns among scenes—and requires only a single library of raw images that can be acquired at different time or day. Experimental results show that even though our scene descriptor is significantly more compact than conventional descriptors it has a relatively higher recognition performance.**

I. INTRODUCTION

Long-term visual SLAM, in familiar, semi-dynamic, and partially changing environments is an important area of research in robotics [1]–[8]. The SLAM task can be viewed as a combination of two subtasks: visual map building and robot self-localization. A robot incrementally builds an environment map using images viewed during visual robot navigation, while simultaneously using the map to localize itself with respect to the environment. These two subtasks respectively involve incremental construction and retrieval of a database of view images. The view image retrieval process is the primary focus of this paper (Fig.1).

In this paper, we focus on a compact discriminative scene descriptor for single-view place recognition. Unlike typical long-term SLAM scenarios that rely on the assumption of view sequence measurements (e.g., SeqSLAM [1]), we tackle the challenging task of single-view place recognition with important applications, in which the robot's views only sparsely overlap with pre-mapped views. The main problem we faced is the question of how to describe a scene discriminatively and compactly—both of which are necessary in order to cope with changes in appearance and a large amount of visual information.

We address this issue by mining visual phrases. In the field of computer vision, visual phrase is a method used to enhance the discriminative power of visual features, where co-located features in a visual image are grouped together to form a visual phrase [9]–[11]. In contrast to popular bag-of-words scene descriptors, which rely on a library of vector quantized visual features (e.g., FAB-MAP [12]), our proposed scene descriptor is based on a library of raw image data (such as an available visual experience, images shared by other colleague robots, and publicly available image data on the web) and directly mines it to find visual phrases (VPs) that discriminatively and compactly explain an input query/database image. Our mining approach is motivated by recent success in the field of common pattern discovery (CPD) [13]–[15], specifically, mining of common visual patterns among scenes, and requires only a single library of raw images that can be acquired at different time or day. In contrast to typical supervised VP learning frameworks [13], our mining approach is unsupervised, and thus enables a robot to learn a compact and discriminative scene model without human intervention. Our implementation of CPD is inspired by the robust high-speed CPD algorithm and randomized visual phrase (RVP) in [13]. The results of evaluations of our method conducted in challenging visual image retrieval experiments and performance comparisons with the conventional Bag-of-Visual-Features and FAB-MAP techniques show that even though our scene descriptor is significantly more compact, it has a higher recognition performance than these techniques.

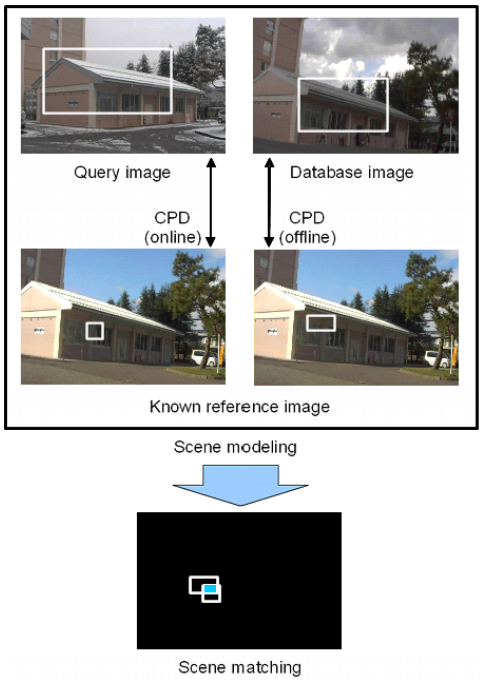

Fig. 1. Modeling and matching a pair of scene images ("query", "database") using our scene descriptor. A raw image matching process ("CPD") mines an available visual experience ("known reference image") to find discriminative visual phrases that effectively explain an input query / database image. The scene matching problem then becomes one of comparing reference image ID and bounding boxes between query and database scenes.

Our work has been supported in part by JSPS KAKENHI Grant-in-Aid for Young Scientists (B) 23700229, and for Scientific Research (C) 26330297.

K. Tanaka, Y. Chokushi, and M. Ando are with Faculty of Engineering, University of Fukui, Japan. tnkknj@u-fukui.ac.jp

II. RELATED WORKS

Existing approaches to long-term SLAM are broadly divided into those that describe a variety of visual appearances of scenes in a single map and those in which multiple independent maps are employed to describe different visual experiences. [1] developed a robust state-of-the-art SLAM framework, called SeqSLAM, for cross-season navigation tasks separated by months or years and opposite seasons. However, the SeqSLAM algorithm explicitly assumes that image sequence measurements are available for robot localization and relies on an image sequence-based scene descriptor. [2] proposed a robust approach that can capture the typical time-varying appearance of an environment in multiple different maps, with the number of experiences required tending to a constant. [3] showed that by quantizing local features in both feature and image space, discriminative statistics can be gained on the co-occurrences of features at different times throughout the day. However, it is not clear whether these approaches can create compact map representations because they directly memorize multiple varieties of visual experiences. A notable exception is [4], in which the issue of compactness is addressed with a question: "How little and what quality of visual information is needed to localize along a familiar route?" Although impressive results have been demonstrated, it also relies on the assumption of image sequence measurements.

Scene descriptors for SLAM problems have been studied extensively. Global feature approaches such as SeqSLAM [1] (in which a scene is represented by a single global feature vector) are compact and have high matching speeds; however, they are not robust, and often require image sequence measurements. Local feature approaches have long outperformed global feature approaches. For example, FAB-MAP [12], in which a scene is represented and matched as a bag-of-visual-features (BoVF), is one of the most successful algorithms. However, BoVF-based methods suffer from high spatial costs because they represent a scene using numerous small local features [16], and each feature consumes several bits, even when a succinct inverted file system is being used.

Other related work include scene models (such as spatial context [13]–[15], part model [17]–[19], and object model [20]–[22]) designed for specific computer vision applications. While these approaches achieve accurate recognition with compact representation of view images, they all assume some level of human supervision in assembling training datasets and learning VP/part/object detectors. Our proposed method differs from these scene models in the formulation of the problem. In our formulation, visual phrases are discovered via a CPD method, and represented in the form of VPs, which enables an autonomous robot to learn a compact scene model without human intervention.

Long-term SLAM for changing environments is a rapidly growing area of research [23]–[26]. [23] presented a framework that uses the stored distinct visual appearances of a workspace, i.e., visual experiences, to improve localization on future visits, and introduced a novel introspective process, executed between sorties. [24] presented a visual mapping system that uses only the input from a stereo camera, and which continually updates an optimized metric map in large indoor spaces with movable objects, e.g., people, furniture, partitions. [25] investigated the persistent navigation and mapping problem over a two-week period in the context of an autonomous robot that performs mock deliveries in a working office environment, and presented a solution based on the biologically inspired visual SLAM system, RatSLAM. [26] investigated the robustness of the place recognition of the SeqSLAM algorithms in changing environments across all four seasons on a 3000 km journey. In contrast, we focus on the issue of visual experience mining for unsupervised descriptor learning.

Our earlier works also focused on scene descriptors, (e.g., global GIST feature descriptor [27], local shape context descriptor [28], and part-based scene descriptor [29]) and SLAM in changing environments (e.g., map updating [30] and change detection [31]). In this paper however, we focus on the discriminativity and compactness of scene descriptors and the problem of long-term SLAM.

III. APPROACH

SLAM's two main subtasks, visual map building and robot self-localization, involve incremental construction and retrieval of a database of view images, respectively:

1) Either subtask interprets each view image in the query or the database as a scene descriptor, and then
2) the localization subtask searches the database to find similar descriptors to the query descriptor.

Then, the image with the highest similarity score is viewed as the localized image. These subtasks are respectively outlined in **Alg.1** and **Alg.2**, and discussed in detail in the subsections below.

Algorithm 1 Mapping for Our Model

Input: Input image $\mathscr{I}$, and reference images $\{\mathscr{R}_j\}_{j=1}^{L}$.
1: Compute bag-of-visual-features W of the input image $\mathscr{I}$.
2: Retrieve most similar reference images $\{\mathscr{R}_j\}_{j=1}^{J}$ to $\mathscr{I}$.
3: **for** $j = 1$ to J **do**
4: Sample subimages $\{\mathscr{I}_i\}_{i=1}^{I}$ from the input image $\mathscr{I}$.
5: **for** $i = 1$ to I **do**
6: Perform CPD between $\mathscr{I}_i$ and $\mathscr{R}_j$.
7: Crop bounding box $\mathscr{B}_{i,j}$.
8: **end for**
9: **end for**

For the above interpretation, we assume that a dictionary or a library of random L reference view images is given. The reference images are not required to be associated with spatial information such that the viewpoint and orientation are known. Such images are cheaper than the mapped images with spatial information required by the map database, and are more readily available. For example, they can be a visual experience obtained by the robot itself in a previous navigation, or shared by other colleague robots, e.g., via information sharing networks [5]. They could also be publicly available resource image data on the web, such as Google

Algorithm 2 Localization for Our Model

Input: Input image $\mathscr{I}$, and reference images $\{\mathscr{R}_j\}_{j=1}^{L}$.
1: Compute bag-of-visual-features W of the input image $\mathscr{I}$.
2: Retrieve most similar reference images $\{\mathscr{R}_j\}_{j=1}^{J}$ to $\mathscr{I}$.
3: **for** $j = 1$ to J **do**
4: Sample subimages $\{\mathscr{I}_i\}_{i=1}^{I}$ from the input image $\mathscr{I}$.
5: **for** $i = 1$ to I **do**
6: Perform CPD between $\mathscr{I}_i$ and $\mathscr{R}_j$.
7: Crop bounding box $\mathscr{B}_{i,j}$.
8: **end for**
9: **end for**
10: **for all** database image $\mathscr{I}'$ **do**
11: Look up bounding box representation $\{\mathscr{B}'_{i,j}\}$ of $\mathscr{I}'$.
12: Compute image level similarity $f(\mathscr{I}, \mathscr{I}')$.
13: **end for**

StreetView. A small subset of J appropriate reference images most similar to a given input image are selected and used to interpret the image. Our experimental results suggest that high localization performance tends to be associated with coverage of the robot's route by these library images. We discuss the view image library issue in Section III-A.

Next, we perform CPD between an input and the reference images to mine a set of VPs that effectively explain the input image. Any CPD algorithm can be adopted, but for our purposes, we utilize the RVP algorithm [13] because it provides fast and stable detection of common visual patterns and can generally handle scale variations among objects without relying on any image segmentation or region detection. We describe the CPD algorithm used in Section III-B.

Next, we obtain a bag-of-bounding-boxes (BoBB) representation, which consists of J pairings of

- a reference image ID (an integer),
- I visual phrases (BBs on the reference image),

as a scene descriptor. Because a BB is a much lower-dimensional representation than many existing feature descriptors such as 128 dimensional SIFT vectors, the search for similar BBs to a query BB can be done quite quickly. We discuss scene descriptors in Section III-C.

A. Mining Visual Experience

To select the most similar J reference images $\{\mathscr{R}_j\}_{j=1}^{J}$ to a given input image, we perform similarity search over the library of L reference images. The similarity search algorithm is designed on the bag-of-raw-features image model [32], in which every image is modeled as an unordered collection of raw feature vectors (such as SIFT visual features). The pairwise similarity between the input and a reference images is evaluated as the number of similar SIFT matches between the image pair. Approximate near neighbor search (ANN) [33] can be used to efficiently search for similar SIFTs to an input query SIFT. The similarity search is iterated for every SIFT feature in the input image, and those J similar images that are supported by the highest number of SIFT matches are considered as the relevant reference images (**Alg.1 Line 2**, **Alg.2 Line 2**).

Unlike the popular bag-of-words image model, our similarity search does not rely on quantized SIFT vectors. Instead, our image model is based on the precise, raw SIFT features. Although it is computationally more demanding, the search process is fast because our model only requires a small library.

B. Mining Visual Phrases

We adopt the RVP algorithm (**Alg.3**) to mine a set of VPs that effectively explain an input image. The RVP algorithm addresses the problem of common object search over reference images given an image of the target object [13]. In its implementation, a set of I subimages $\{\mathscr{I}_i\}_{i=1}^{I}$ are randomly sampled from the input image $\mathscr{I}$. Then, each subimage $\mathscr{I}_i$ is viewed as a hypothesized target image for the problem of object search over a reference image $\mathscr{R}_j$. Although initial localization of such a randomly sampled target sub-image within the input image may be inaccurate, the target subimage is recropped and expected to be sufficiently localized by means of the RVP algorithm, as shown below.

Algorithm 3 Common Pattern Discovery

Input: Input image $\mathscr{I}$, and reference image $\mathscr{R}$.
1: **for all** pixel p on the voting image V **do**
2: Initialize the pixel value $V(p)$ to 0.
3: **end for**
4: Look up bag-of-words histogram H of the input image $\mathscr{I}$.
5: **for** $k = 1$ to K **do**
6: Partition reference image $\mathscr{R}$ into $M \times N$ patches $\{\mathscr{R}_{m,n,k}\}$.
7: **for** $m = 1$ to M **do**
8: **for** $n = 1$ to N **do**
9: Look up bag-of-words histogram $H_{m,n,k}$ of $\mathscr{R}_{m,n,k}$.
10: Compute similarity $S_{m,n,k}$ between H and $H_{m,n,k}$.
11: **for all** pixel that belongs to the patch $p \in \mathscr{R}_{m,n,k}$ **do**
12: $V(p) \leftarrow V(p) + S_{m,n,k}$.
13: **end for**
14: **end for**
15: **end for**
16: **end for**

The RVP algorithm employs a bag-of-visual-words image representation style, and thus requires a visual word vocabulary. In our approach, we use the set of V visual features in the library images as the vocabulary. To translate a given visual feature, we run the ANN to find similar visual features in the library, followed by a verification step to ensure that the normalized L1-distance between the SIFT descriptor pair is smaller than 0.4. We then assign their feature IDs as the visual words, i.e., multiple visual word per feature.

We then independently and randomly partition reference image $\mathscr{R}$ a total of K times into $M \times N$ non-overlapping rectangular patches (**Alg.3 Line 6**). The result is a pool of $M \times N \times K$ image patches (in our experiments, $32 \times 16 \times 200$ patches) or visual phrases $\{\mathscr{R}_{m,n,k} |\ m \in [1,M],\ n \in [1,N],\ k \in [1,K]\}$ each of which is characterized as a V-dimensional histogram, $H_{m,n,k}$, recording the visual word frequency of $\mathscr{R}_{m,n,k}$ (**Alg.3 Line 9**). Since in the k-th partition, each pixel, t, falls into a single patch, $\{P_{m,n,k} |\ t \in P_{m,n,k},\ k \in [1,K]\}$, there are a total of K patches containing t after K rounds

of partitioning. Each patch is viewed as a VP, and provides more contextual information than a typical visual word. For each pixel, t, the confidence that the pixel is part of the target object is measured by the average similarity between the input, $\mathscr{I}_i$, and the visual phrase, $\mathscr{R}_{m,n,k}$, over all the K patches that contain t (**Alg.3 Lines 10-13**).

Following the assignment of a confidence score to each pixel, we obtain a voting map for each image pair, $\mathscr{I}_i, \mathscr{R}_j$. Object localization then becomes the task of segmenting the dominant region in the form of bounding box $\mathscr{B}^*_{i,j}$ from the reference image of interest, $\mathscr{R}_j$. Intuitively, the optimal bounding box should be the one whose sum of confidence scores over all the pixels inside the bounding box are higher than all other potential bounding boxes. Further, the integral image [34] can be used to efficiently compute the sum of the values in the rectangular regions defined by these bounding boxes. The size of a bounding box should be sufficiently small that it can be localized well, and should not exceed 10% of the area of the reference image.

C. Scene Descriptor

The BoBB scene descriptor consists of J pairings of a reference image ID (i.e., an integer) and a set of I visual phrases (BBs on the reference image). A BB carries appearance information of a VP as it indicates the VP region within the reference image. Suppose that a function $Overlap(\mathscr{B}_{i,j}, \mathscr{B}'_{i',j'})$ returns the area of overlap between a given BB pair $\mathscr{B}_{i,j}$, $\mathscr{B}'_{i',j'}$ when they belong to the same reference image or zero otherwise. Note that our current implementation ensures that each bounding box is well localized, i.e., smaller than 10% of the image area, and we have already found that there is no need to penalize the size of the bounding boxes. A large value for $Overlap(\mathscr{B}_{i,j}, \mathscr{B}'_{i',j'})$ indicates that the VPs cropped by the BBs are similar between the image pair, and vice versa. By aggregating the VP-level similarity, we obtain the image-level similarity (**Alg.2 Line 12**):

$$f(\mathscr{I}, \mathscr{I}') = \frac{1}{IJ} \sum_{j=1}^{J} \sum_{i=1}^{I} \max_{i',j'} Overlap(\mathscr{B}_{i,j}, \mathscr{B}'_{i',j'}). \quad (1)$$

Since a BB can be compactly represented by a 4D parameter (a much lower-dimensional representation than other local feature descriptors such as 128D SIFT vectors), the search for BBs similar to a query BB can be conducted very rapidly.

IV. EXPERIMENTAL EVALUATION

In this section we present and discuss the results of experiments conducted using the proposed framework.

The dataset used in the experiments consisted of sequences of view images taken around a university campus, using a handheld camera as a monocular vision sensor. [1] Fig.2 is a bird's eye view of our experimental environments and viewpoint paths. We considered a typical scenario that deals with view images that are taken relatively far apart (1 m-5 m) from each other, significantly reducing the memory required to describe a given path. The sequences start at 10 different locations inside the university campus—some going through the main central path, and others going along the pedestrian walkway along the campus wall, as shown in the figure. Occlusion is severe in all the scenes, and people and vehicles are dynamic entities occupying the scene. In order to evaluate the performance of the framework, we traversed each path twice, and acquired a pair of view image sequences for mapping and localization for each path. A random collection of 100 view images over various days and times were acquired along each path, and used as the size $L = 100$ library of reference images. In addition, a challenging "cross season" dataset, in which the input and the reference images were acquired in different seasons, was also created and utilized. Fig.3 shows samples of the view images used in the evaluation.

For CPD, our method selected a set of $J = 4$ reference images from the size $L = 100$ library and learned $I = 4$ VPs for each reference image, based on the SIFT similarity search and the CPD. Fig.4 shows examples of the results obtained. It can be seen that common VPs were successfully discovered for the relevant image pairs. Even when there was no identical object between the input and the reference images, our

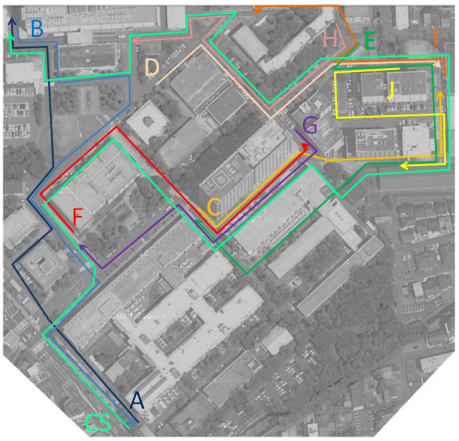

Fig. 2. Experimental environments and viewpoint paths (Bird's eye view). a,b,c,d,e,f,g,h,i,j: 10 paths used for quantitative evaluation. CS: path used for the "cross season" case.

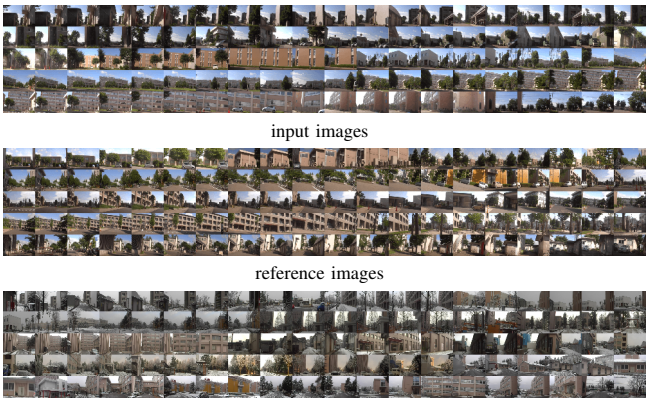

Fig. 3. Samples of view images used for evaluation.

[1] Note that there is nothing in our algorithm that requires sequential image sequences—in contrast to typical SLAM algorithms such as SeqSLAM. We chose to use the image sequence only for the convenient presentation.

RVP-based method robustly and efficiently detected similar objects by aggregating the patch-level similarity between the image pair.

We evaluated the proposed VP method ("VP") in terms of the retrieval accuracy and compared the results with that obtained by bag-of-visual-features ("BoVF") and FAB-MAP 2.0 ("FAB-MAP") [12]. For BoVF, we weighted the original BoVF vectors using a standard TF-IDF weighting scheme, and a vocabulary with 16K words. For FAB-MAP, we used the same code used by the authors in [12]. We conducted a series of 100 independent retrievals for each of 100 random query images and for each of the 10 different paths. Retrieval performance is measured in terms of averaged normalized rank (ANR)—a ranking-based retrieval performance measure, where the smaller value is better—in percent (%). To evaluate ANR, we evaluated the rank assigned to the ground-truth relevant image for each of the 100 independent retrievals, and then normalized the rank with respect to the database size and computed the average over the 100 retrievals.

In Fig.5, it can be seen that our approach outperforms BoVF and FAB-MAP in most of the retrievals considered here, even though our BoBB descriptor is significantly more compact. The result for BoVF is not as impressive as we had expected. Matched features often occupy only a small portion of an image, and as a result, are difficult for the BoVF method to be identify. In contrast, our matched VP-based framework achieves much better retrieval performance

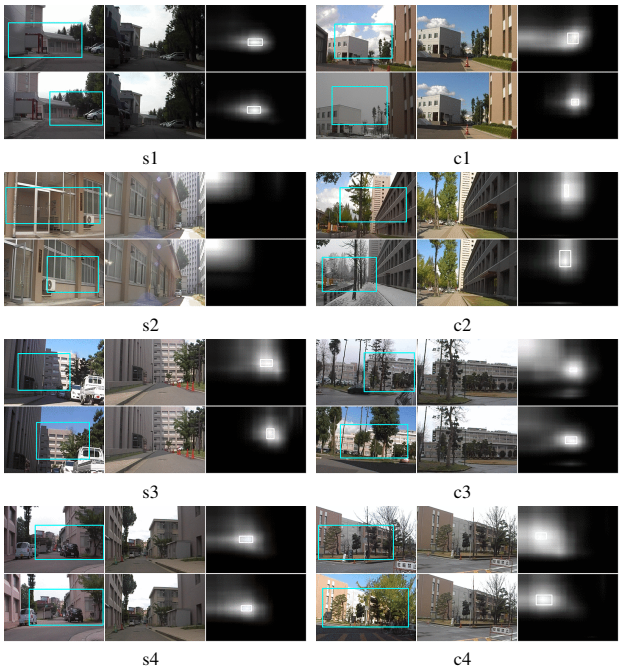

Fig. 4. Examples of common pattern discovery (CPD). s1-4: The single season case. c1-4: The "cross season" case. For each panel, the top row shows CPD for a query image and the bottom row shows CPD for the ground truth database image. For each panel, the left column shows the input image, the middle column shows the reference image selected for CPD, and the right column shows the CPD results, i.e., voting map and BB.

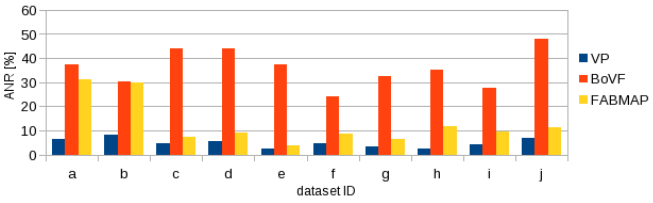

Fig. 5. Quantitative performance.

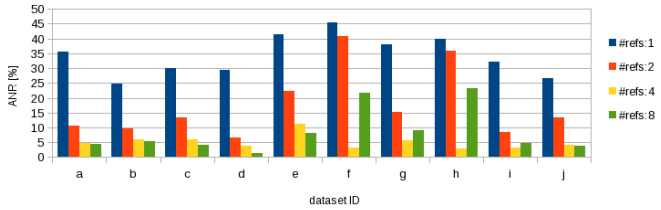

Fig. 6. Results for various #reference images.

while requiring only $IJ = 16$ indexes per image.

Fig. 6 shows the ANR performance for various setting of the #reference images parameters. It is clear that good and stable results are obtained when the number of reference images is sufficiently large. The larger number of reference images tend to have better common VPs.

Fig. 7 illustrates the benefits of utilizing both the reference image IDs and the pose and the shape of the bounding boxes (BBs) for scene description. In the figure, the results of CPD, i.e., the reference image IDs and the pose and the shape of BBs, are plotted on a single 2D plane for a short sequence of view images. It can be seen that although the reference image IDs are not sufficiently discriminative, the pose and shape of the BBs provide additional discriminative information, that can be captured by our bounding box -based scene descriptor.

Fig. 8 shows how the sensitivity of retrieval performance correlates with the choice of library. In this experiment, we are particularly interested in understanding the impact of the choice of the library on the performance. It is clear that the use of visual experiences from different viewpoint paths is

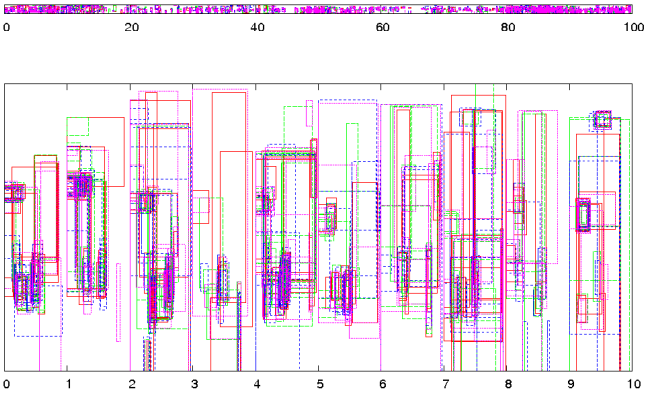

Fig. 7. Selected reference images and their bounding boxes. Top: BBs for reference images. For visualization, the BB for each n-th reference image is normalized to fit within an area $[n-1,n]\times[0,1]$. Bottom: x10 close-up for $x \in [0,10]$.

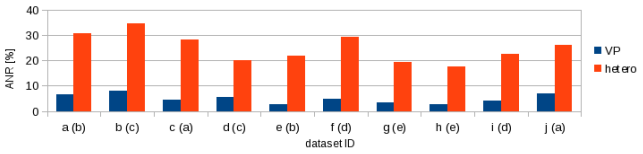

Fig. 8. Sensitivity of retrieval performance to the choice of library. The five libraries (a)-(e) are used to explain two different sets of query and database images. "hetero": performance when using the library from different viewpoint path. $(\cdot)$ indicates the path ID of the library.

not very effective. Because our method is designed to explain an input image using a pool of cropped reference images, it is not suitable for general cases in which entire regions of the input image are not similar to any reference image. We plan to derive a means of automatically and adaptively choosing the library for a given set of database images in future work.

Fig. 9 visualizes the frequency of each reference image being selected, and gives examples of the most and least frequently selected reference images overlaid with all the bounding boxes. As can be seen from the figure, the reference images where non-common objects or non-discriminative objects such as trees occupy the entire image region tend to be less frequently selected than other reference images. Further, it shows that the frequency of the most frequent reference image is 10 times higher than that of the less frequent 50% of images in the library.

We also evaluated the methods on a challenging "cross season" scenario. The database and reference image sequences were acquired at different time of day in the autumn, while the query image sequence was acquired in the winter. Further the sequences were acquired in very different illumination conditions and changes in appearance due to fresh snow cover (see Fig.3). Fig. 10 shows the results for the proposed framework compared with FAB-MAP. It can be seen that the proposed method is more robust despite the

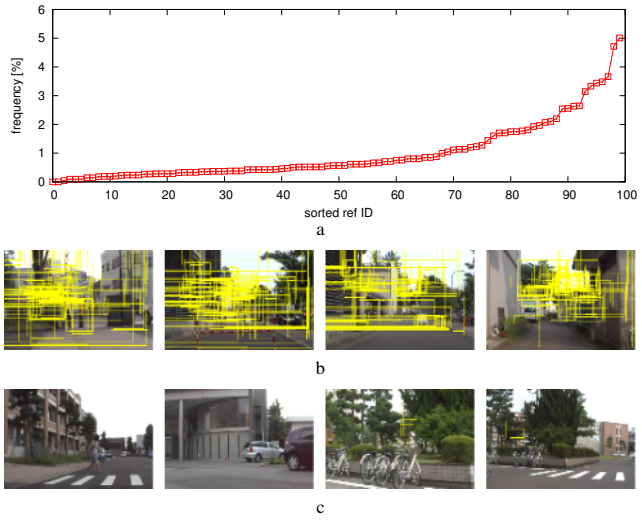

Fig. 9. Example results of selecting reference images. (a) Frequency of each reference image being selected. (b/c) The four most / least frequent reference images overlaid with all the bounding boxes.

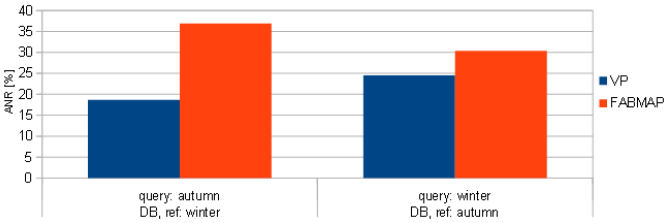

Fig. 10. Performance on "cross season" case.

difficult conditions. Even though entire regions were often dissimilar between images, parts of them often similar and were captured by our VP-based approach.

The advantage of our approach is even more obvious from a spatial cost point of view. For example, when $I, J = 4$, $L = 100$, our scene descriptor consists of four reference image IDs, each of which is represented by 22 bits (i.e., $\log_2 {}_L C_J$), and $IJ = 16$ bounding boxes. This is extremely compact, even when compared with other compact local feature approaches such as BoVF and FAB-MAP, where an image is typically represented by numerous words or entries to the inverted file system, each of which consumes a few bits. Even global feature approaches such as the GIST feature descriptor still consume thousands of bits per image. A notable exception is those that rely on advanced vector quantization techniques, such as VLAD feature descriptor and the compressed GIST global feature employed in our earlier work. Discriminativity preserving vector quantization for further compacting the proposed BoBB scene descriptor will be addressed in future work.

V. DISCUSSIONS

Our experimental results suggest that high localization performance tends to be associated with coverage of the robot's route by these library images. In future, we plan to investigate how comprehensively do these library images need to cover the path. For example if any building is missing in the library set, it is more difficult to produce meaningful results. We are also interested in investigating how different library sizes (e.g., $L = 10, 20, 50$) affect the localization performance. Combining the proposed visual phrase framework with traditional visual word framework, i.e., bag-of-words, can effectively address the coverage area issue. When no sufficient match is found by the proposed CPD technique, it can be viewed that the scene cannot be explained by the available visual experience. Such a scene might be better explained by the more primitive, visual words, at the cost of increased space for storing the visual words.

The current implementation of the framework is primarily aimed at demonstrating the effectiveness of the CPD techniques for compact and discriminative scene modeling. The bottleneck for a real-time application would be the speed of the CPD processing. The time complexity of CPD is linear to the number of candidate library images that are output by the visual experience mining. To address the time cost for CPD, we plan to leverage more advanced mining algorithms, instead of the simple ANN algorithm

used in our current study, to achieve better scalability for increasing numbers of library images. A key observation is that we can reformulate the visual experience mining as a self-localization problem, i.e., localizing the viewpoint with respect to a map of library images [29], and thus leverage the large body of self-localization algorithms in the literature. In future, we plan to expand in this direction.

VI. CONCLUSIONS

In this paper, we proposed a visual phrase approach to the problem of view image retrieval in partially changing environments. The main novelty of this approach lies in the fact that common visual phrases are mined in an unsupervised manner via CPD, and can be used for compact characterization and efficient retrieval of view images. This contrasts with the existing supervised frameworks with prelearned visual phrases commonly found in the field of computer vision. Our novel approach enables a robot vision system to learn a compact and semantic scene model without requiring human intervention. Another novelty lies in the use of a bounding box -based phrase annotation scheme for a compact and discriminative scene descriptor. Experimental results and performance comparisons with existing BoVF and FAB-MAP frameworks show that it is possible to have high retrieval performance despite the fact that our scene descriptor is significantly more compact.

REFERENCES

[1] M. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *ICRA*, 2012, pp. 1643–1649.

[2] W. Churchill and P. Newman, "Practice makes perfect? managing and leveraging visual experiences for lifelong navigation," in *ICRA*, 2012, pp. 4525–4532.

[3] E. Johns and G.-Z. Yang, "Feature co-occurrence maps: Appearance-based localisation throughout the day," in *ICRA*, 2013, pp. 3212–3218.

[4] M. Milford, "Vision-based place recognition: how low can you go?" *I. J. Robotic Res.*, vol. 32, no. 7, pp. 766–789, 2013.

[5] A. Cunningham, K. M. Wurm, W. Burgard, and F. Dellaert, "Fully distributed scalable smoothing and mapping with robust multi-robot data association," in *ICRA*, 2012, pp. 1093–1100.

[6] A. S. Huang, M. E. Antone, E. Olson, L. Fletcher, D. Moore, S. J. Teller, and J. J. Leonard, "A high-rate, heterogeneous data set from the darpa urban challenge," *I. J. Robotic Res.*, vol. 29, no. 13, pp. 1595–1601, 2010.

[7] N. Carlevaris-Bianco and R. M. Eustice, "Long-term simultaneous localization and mapping with generic linear constraint node removal," in *IROS*, 2013, pp. 1034–1041.

[8] J. McDonald, M. Kaess, C. D. C. Lerma, J. Neira, and J. J. Leonard, "Real-time 6-dof multi-session visual slam over large-scale environments," *Robot Auton Systems*, vol. 61, no. 10, pp. 1144–1158, 2013.

[9] Q.-F. Zheng, W.-Q. Wang, and W. Gao, "Effective and efficient object-based image retrieval using visual phrases," in *ACM Int. Conf. Multimedia*, 2006, pp. 77–80.

[10] J. Yuan, Y. Wu, and M. Yang, "Discovery of collocation patterns: from visual words to visual phrases," in *IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[11] M. A. Sadeghi and A. Farhadi, "Recognition using visual phrases," in *IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1745–1752.

[12] M. Cummins and P. Newman, "Highly scalable appearance-only slam - fab-map 2.0," in *Robotics: Science and Systems*, 2009.

[13] Y. Jiang, J. Meng, and J. Yuan, "Randomized visual phrases for object search," in *IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3100–3107.

[14] H.-K. Tan and C.-W. Ngo, "Common pattern discovery using earth mover s distance and local flow maximization," in *IEEE Int. Conf. Computer Vision (ICCV)*, 2005, pp. 1222–1229.

[15] M. Cho, Y. M. Shin, and K. M. Lee, "Unsupervised detection and segmentation of identical objects," in *IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 1617–1624.

[16] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704–1716, 2012.

[17] M. Fischler and R. Elschlager, "The representation and matching of pictorial structures," *IEEE Trans. Computers*, vol. C-22, no. 1, pp. 67 – 92, 1973.

[18] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in *ECCV workshop on statistical learning in computer vision*, 2004, pp. 17–32.

[19] P. Arbelaez, B. Hariharan, C. Gu, S. Gupta, L. D. Bourdev, and J. Malik, "Semantic segmentation using regions and parts," in *IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3378–3385.

[20] L.-J. Li, H. Su, E. P. Xing, and F.-F. Li, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Conf. Neural Information Processing Systems (NIPS)*, 2010, pp. 1378–1386.

[21] L. Bo, X. Ren, and D. Fox, "Unsupervised Feature Learning for RGB-D Based Object Recognition," in *ISER*, June 2012.

[22] S. Parizi, J. Oberlin, and P. Felzenszwalb, "Reconfigurable models for scene recognition," in *IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2775–2782.

[23] W. Churchill and P. Newman, "Continually improving large scale long term visual navigation of a vehicle in dynamic urban environments," in *Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on*, Sept 2012, pp. 1371–1376.

[24] K. Konolige and J. Bowman, "Towards lifelong visual maps," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, 2009, pp. 1156–1163.

[25] M. Milford and G. Wyeth, "Persistent navigation and mapping using a biologically inspired slam system," *I. J. Robotic Res.*, vol. 29, no. 9, pp. 1131–1153, 2010.

[26] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? challenging seqslam on a 3000 km journey across all four seasons," *Workshop on Long-Term Autonomy held in conjunction with the International Conference on Robotics and Automation (ICRA)*, 2013.

[27] K. Ikeda and K. Tanaka, "Visual robot localization using compact binary landmarks," in *ICRA*, 2010, pp. 4397–4403.

[28] K. Saeki, K. Tanaka, and T. Ueda, "Lsh-ransac: An incremental scheme for scalable localization," in *ICRA*, 2009, pp. 3523–3530.

[29] S. Hanada and K. Tanaka, "Part-slam: Unsupervised part-based scene modeling for fast succinct map matching," in *IROS*, 2013, http://rc.his.u-fukui.ac.jp/PARTSLAM.pdf.

[30] H. Zha, K. Tanaka, and T. Hasegawa, "Detecting changes in a dynamic environment for updating its maps by using a mobile robot," in *IROS*, 1997, pp. 1729–1734.

[31] K. Tanaka, Y. Kimuro, N. Okada, and E. Kondo, "Global localization with detection of changes in non-stationary environments," in *ICRA*, 2004, pp. 1487–1492.

[32] H. Zhang, "Borf: Loop-closure detection with scale invariant visual features," in *ICRA*, 2011, pp. 3125–3130.

[33] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *Int. Conf. Computer Vision Theory and Application*. INSTICC Press, 2009, pp. 331–340.

[34] P. Viola and M. Jones, "Robust real-time object detection," in *Int. J. Computer Vision*, 2001.