Object-Level View Image Retrieval via Bag-of-Bounding-Boxes

Ando Masatoshi Chokushi Yuuto Inagaki Yousuke Hanada Shogo Tanaka Kanji

***Abstract*— We propose a novel bag-of-words (BoW) framework to build and retrieve a compact database of view images, toward robotic localization, mapping and SLAM applications. Our method does not explain an image by many small local features (e.g. bag-of-SIFT-features) as most previous methods do. Instead, the proposed bag-of-bounding-boxes (BoBB) approach attempts to explain an image by fewer larger object patterns, which leads to a semantic and compact image descriptor. To make the view retrieval system more practical and autonomous, the object patterns are discovered in an unsupervised manner, via common pattern discovery (CPD) between the input and a known reference images, which does not require pre-trained object detector. Moreover, our CPD task does not rely on good image segmentation and can handle scale variations, exploiting the recently developed CPD technique, spatial random partition. By exploiting traditional bounding box -based object annotation and knowledge transfer, we compactly describe an image in a form of bag-of-bounding-boxes (BoBB). With a slightly modified inverted file system, we efficiently index/search the BoBB descriptors. Experiments with publicly available "RobotCar" dataset show that the proposed method achieves accurate object-level view image retrieval with significantly compact image descriptors, e.g. 20 words per image.**

I. Introduction

View image retrieval on *compact* database of view images is a fundamental building block for robotic localization, mapping and SLAM systems [1]–[3]. Applications include large scale maps and information sharing, where the spatial cost for storage [1], [2] and information transfer [3] of view database becomes critical issue. One of best known ways to address this problem is the popular bag-of-visual-features (BoVF) [4]–[7], which was originally inspired by the traditional bag-of-words (BoW) model from text information retrieval, and where the indexing (or retrieval) process proceeds as follows:

1) extract local visual features from an input database (or query) view image;
2) translate the features into visual words using a feature dictionary;
3) index (or exact search) the inverted file system using the visual words.

Our approach proposed in this paper also follows a similar pipeline consisting of three steps 1)-2)-3), but it does not explain an image by *many small local features* (e.g. bag-of-SIFT-features) as most BoVF frameworks do. Instead, we attempt to explain an image by *fewer larger object patterns*, which leads to a semantic and compact image descriptor.

This work was partially supported by MECSST Grant (23700229, 30325899), by KURATA grants and by TATEISI Science And Technology Foundation.

The authors are with Graduate School of Engineering, University of Fukui, Japan. tnkknj@u-fukui.ac.jp

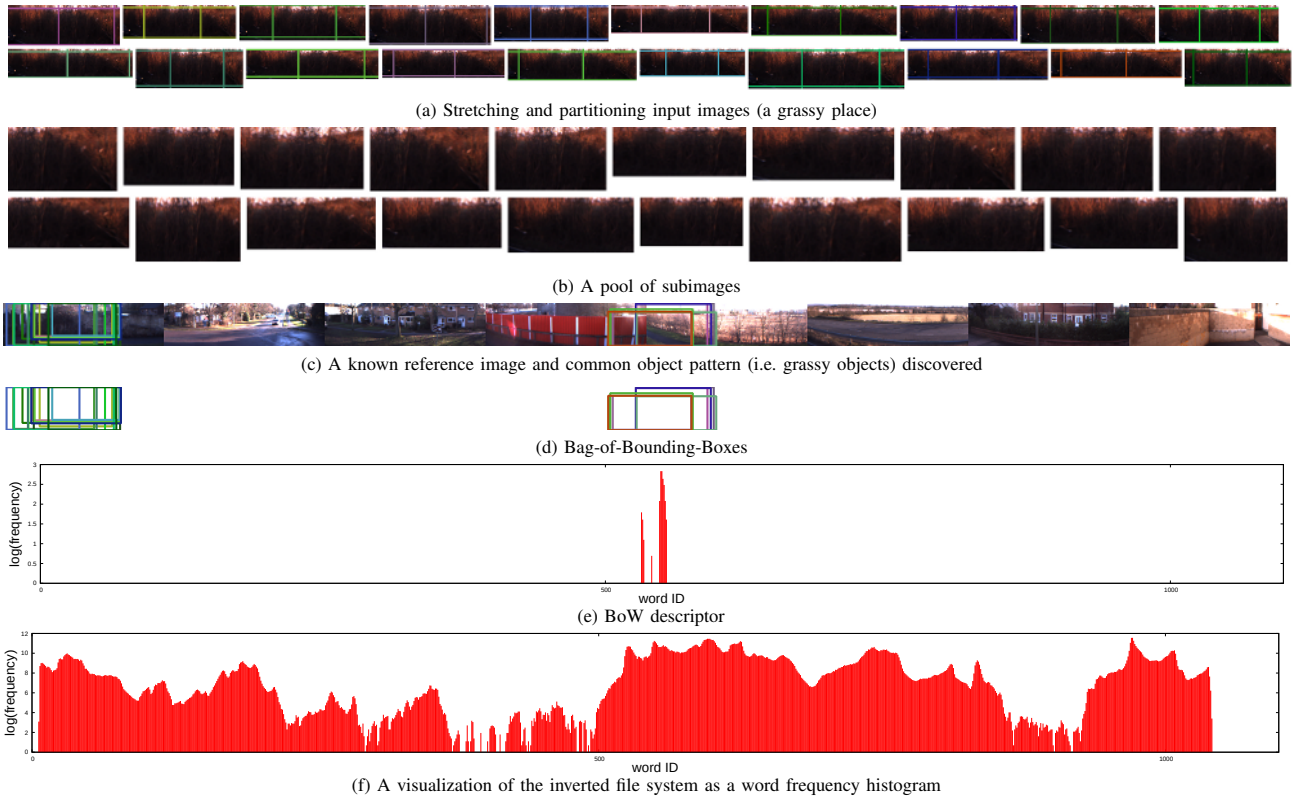
This study is motivated by recent success in object-level correspondence techniques (e.g. co-segmentation) for common pattern discovery, i.e. mining common object patterns across images [8]–[11]. A known limitation of feature-level correspondence techniques (e.g. BoVF) is that they are largely influenced by the extracted features, and cannot exploit further information beyond the detected features, whose size and shape are typically small and must be defined prior to the feature extraction (i.e. 1st) stage. To counter this, different lines of researches on object-level correspondence, including common pattern discovery [8], co-segmentation [9], subimage search [10], and visual phrase [11] have been developed. By simultaneously looking at a pair of images, those techniques attempt to find larger object-level correspondence based on the fact that true correspondences are supported by larger object region than false ones.

We are particularly inspired by the spatial random partition (SRP), a common pattern discovery (CPD) technique originally proposed in [12] and recently developed in [11], where an input image is characterized by a pool of overlapping subimages randomly sampled from it. For CPD, each subimage is queried and matched against the subimage pool, based on the fact that a common pattern is likely to be present in a good number of subimages across different images. From our viewpoint of object-level view retrieval, SRP has several desirable properties: 1) It does not rely on good image segmentation techniques; 2) It does not require a priori knowledge on how many common object patterns exist in the input views; 3) It does not rely on quantization of visual features; and 4) It is able to handle scale variations of the object. Our proposed approach is designed to leverage those desirable properties of SRP.

In this paper, we focus on use of the object-level correspondence techniques within the general BoW framework. Accordingly, our indexing (or retrieval) process is slightly different from that of the BoVF framework, and proceeds as follows (Fig.1):

1) extract *object patterns that well explain an input image from a known reference image;*
2) translate the *object patterns discovered* to visual words;
3) index (or *similarity search*) the inverted file system using the visual words.

Following the BoW literature, the 1st and 2nd stages for database images are done in offline and ready for parallelization and large-scale view retrieval. At the 1st stage, a known reference image is simply used as a view dictionary, in contrast to the pre-learned feature dictionary used by traditional BoVF frameworks.



(a) Stretching and partitioning input images (a grassy place)

(b) A pool of subimages

(c) A known reference image and common object pattern (i.e. grassy objects) discovered

(d) Bag-of-Bounding-Boxes

(e) BoW descriptor

(f) A visualization of the inverted file system as a word frequency histogram

Fig. 1. Our BoW pipeline. An input image is stretched vertically and horizontally as shown in (a), and randomly partitioned into a pool of subimages (b). The subimages are matched between the input image and a known reference image, based on the fact that a common pattern is likely to be present in a good number of subimages across different images (c). The resulted bag-of-bounding-boxes (d) is used as a compact BoW descriptor (e) for indexing and retrieving the inverted file system (f).

There are five key properties about the proposed approach:

- An image is *semantically* characterized by object-level information, in contrast to feature-level image characterization in existing BoW frameworks;
- The object pattern discovery process is *unsupervised*, without requiring manually labeled examples and/or pre-trained object detector;
- An image is *compactly* described in a form of bag-of-bounding-boxes (BoBB), employing traditional BB-based object annotation and knowledge transfer [13];
- The BoBB framework inherits the *efficiency* in indexing and retrieval from the general BoW framework by using a slightly modified inverted file system;
- The BoBB framework leverages the state-of-the-art CPD technique, spatial random partition (SRP) [11], which has desirable properties as aforementioned.

Experiments with publicly available "RobotCar" dataset [1] show that the proposed approach achieves accurate object-level view image retrieval using significantly compact description of view images, e.g. 20 words per image.

II. Relation to Other Work

Image retrieval in a large number of images has recently received increasing attention [1], [7], [14]–[22]. Previous studies have dealt with various aspect of the BoW framework, including the quantization method and its speed [1], [7], [14], the post processing based on a global spatial geometric verification [15], the matching distance of descriptors [16], and with various types of visual features including local feature (e.g. SIFT, SURF), global feature (e.g. GIST), filter bank (e.g. color, texture, object), and other feature modeling techniques. While most of the above systems work on large image databases, several efforts also focused on *compactness* of the image database [19]–[22]. [21] has improved the memory usage per image introducing a method for projecting the BoW vectors onto a set of pre-defined sparse projection functions. In [22], we also employed the BoW projection technique and used it within a multi-cue BoW framework for scalable scene retrieval applications. However, almost all of those efforts to compact view database focused on feature-level correspondence, and little study has attempted on the object-level correspondence, as we propose to do in this study.

The problem of object retrieval, whose goal is to accurately locate the target object in image collections [11], is clearly different from our view retrieval problem. Object retrieval is a challenging task due to the fact that the target object usually occupy only a small portion of an image with cluttered background, and can differ significantly from the query in scale, orientation, viewpoint and in color. One of most effective ways to address this problem is the use of spatial context [11], where the input images are partitioned into small subimages and then matched against one another based on the fact that a common object pattern is likely to co-exist in a good number of subimages across different images. However, existing works focus on object retrieval

tasks, and often concerned with setting where feature-based inference is possible, e.g. demanding rich features for geometric verification. From our view retrieval standpoint, the object retrieval approaches would waste a large amount of memory resource to index those individual objects, and not suited for our objective, i.e. compact description of view images.

Although object-level scene representation is a central importance in robotic mapping, localization and SLAM [23], existing efforts to compact the view image database focus on feature-level approaches relying on dimension reduction techniques. [24] developed a self-localization system by combining the SIFT feature descriptor with principal component analysis (PCA) dimension reduction techniques, and achieved accurate track of the position of a robot in a real environment. Many efforts have also been made on various types of feature descriptors and advanced dimension reduction techniques [1], [2], [25]–[28]. In our previous papers, we also have developed localization systems exploiting dimension reduction techniques, including locality sensitive hashing (LSH) [26], semantic hashing (SH) [27], and compact projection (CP) [28]. In contrast, our current paper focuses on an object-based scene characterization.

The problem of common pattern discovery (CPD), multiple objects co-segmentation, or co-recognition, which aims at automatic discovery of common object patterns across images is an active and open research issue [8], [29]–[31]. Because no prior knowledge is available on the common object patterns, this task is very challenging, and much more difficult than traditional tasks such as detection and retrieval of object patterns, since the search space (e.g. appearance, size, shape, number of objects) is enormous. The existing solutions include earth mover's distance (EMD) [8] and other model learning techniques, co-segmentation [29], and correspondence growing [30]. In [31], we also have developed a CPD technique in a form of correspondence growing algorithm by employing a probabilistic MCMC framework. However, improving existing common pattern discovery techniques is not the objective of our current paper. In this study, we focus on *use* of common pattern discovery as a method for the object-level image characterization within the general BoW framework.

III. Bag-of-Bounding-Boxes (BoBB) Framework

The proposed BoBB framework is slightly different from the BoVF framework in three important aspects: 1) definition of visual word; 2) representation of dictionary; and 3) search criteria. We describe the basic idea behind each of them in the following, and then explain the BoBB framework step-by-step in subsections III-A, III-B, III-C, III-D, and III-E.

First, we define a visual word as a common object pattern that well explains an input query/database image, discovered from a known reference image via common pattern discovery (CPD). Accordingly, our visual word extraction process becomes an iteration of the CPD between an input and the reference images, which consists of hypothesization and verification of common object patterns: 1) Each iteration begins by randomly stretching and shrinking each image to deal with variations of scale, viewpoint and occlusions (Fig.1a); 2) For the hypothesization, inspired by the spatial random partition technique [11], a pool of subimages are randomly sampled from both images and each pair of subimages is used as a hypothesis of common object pattern (Fig.1b); 3) For the verification, correspondence between the subimage pair is verified (Fig.1c) by using any type of correspondence measure (e.g. EMD [8], multiple objects co-segmentation [29], correspondence growing [30]); In this paper's experiments, the normalized image correlation will be used as the correspondence measure; 4) Common object patterns discovered are compactly described in a form of bag-of-bounding-boxes (Fig.1d), employing traditional bounding box -based object annotation and knowledge transfer [13].

Second, we use a known reference image as a view dictionary. This is in contrast to the feature dictionary used by the BoVF framework for dimension reduction or quantization of visual features. To make the view retrieval more practical and autonomous, we do not assume any special indexing architecture for the dictionary, such as ImageNet. Instead, our view dictionary consists of raw images (e.g. JPEG images) being acquired by the robot-self or shared via distributed robot networks, without supervised categorization. Although a dictionary for the BoW framework in general should be designed to contain visual words that are frequently used [4], it is beyond the scope of this paper to discuss such an optimal design or adaptive learning of the dictionary image. In this paper's experiments, we will simply use dictionary images consisting of 8-64 raw images, as shown in Fig.3.

Third, our search criteria is based on similarity search, in contrast to the exact search (either single- or multi- probe strategy [21]) commonly used by the BoVF framework. Because our visual word is defined as a bounding box with its pose and shape attributes, the similarity used for search criteria is designed to evaluate similarity of those attributes. In the current paper, the area of overlap between bounding boxes will be simply used as the similarity measure.

A. Problem: View Image Retrieval

The goal of view image retrieval is to retrieve images similar to a given query image I^Q by comparing the query image I^Q and each image I^D in the image database $D = \{I^D\}$, given a reference image R and an object-level correspondence measure S.

B. Common Pattern Discovery

Unlike the BoVF framework, the database building process consists of an iteration of the common pattern discovery between an input I and the reference R images, which proceeds as follows:

1) randomly partition the input and the reference images I and R for multiple times, and obtain a pool of overlapping subimages $\{I_k\}$ and $\{R_k\}$ (Fig.1a,b);
2) evaluate the likelihood of each subimage pair (I_k, R_k) being a match pair by using the correspondence measure S;

3) rank all the subimage pairs in descending order of the likelihood score;
4) select a set $\{(I_k, R_k)\}_{k=1}^{T}$ of T top ranked subimage pairs as common object patterns (Fig.1c).

Currently, the likelihood at the step 2 is evaluated by comparing geometry and appearance between the subimage pair. More formally, height and width of bounding box is compared between the subimage pair, and if width or height of taller bounding box does not exceed a pre-defined ratio $(1+r)$ than shorter box, the subimage pair is viewed as a potential match, and then, the likelihood for such a potential match is evaluated by the given correspondence measure: $S(I_k, R_k)$.

C. Visual Word Extraction

We compute a bounding box for each of the regions of the T common object patterns output by the above process, and represent it by the coordinates $x_{\min}$, $x_{\max}$, $y_{\min}$, $y_{\max}$ (Fig.1d). The pose $(x_{\min}, y_{\min})$ and the shape (w,h) of each bounding box, where $2w$ and $2h$ respectively represent the width and height of the box, is computed and the 4D parameter $(x_{\min}, y_{\min}, w, h)$ is mapped to the visual word.

D. Indexing

The procedure for indexing the inverted file system given bag-of-bounding-boxes (BoBB) descriptors is straightforward. The BoBB descriptor is represented by a 4D parameter and transformed to a 1D visual word. Since we have to store one entry for each bounding box existing in the pool [21], each input image requires space linear to the number of bounding boxes (i.e. visual words) per image.

E. Similarity Search

Given a query image I^Q, the similarity search process aims to retrieve and score images in the database D, and proceeds in the following steps:

1) extract a bag-of-bounding-boxes $\{I_i^Q\}$ from the query image I^Q in the same manner as in III-B, III-C (Fig.1d);
2) For each query BB $I_i^Q = (x, y, w, h)$,
   a) retrieve images $D_i (\subset D)$ whose BB can overlap with I_i^Q or belongs to the following area

$$Z = [x-(2+r)w, x+(2+r)w] \times [y-(2+r)h, y+(2+r)h] \times [w/(1+r), w(1+r)] \times [h/(1+r), h(1+r)],$$

   in the BB parameter space;
   b) evaluate similarity between every pair of BBs from I^Q and each $I_j^{D_i} (\in D_i)$, according to the area of overlap between BBs;
3) compute the aggregate score v^j for each of the retrieved database images D_i, while set score $v^j = 0$ for those database images that are not retrieved;
4) Rank all the database images in descending order of the aggregate score v^j.

Fig. 2. Input images and retrieval results for 5 different retrieval tasks. In each of them, the retrieval was successful and the relevant images are assigned high ANR rankings [%], 3, 3, 9, 4 and 2, respectively.

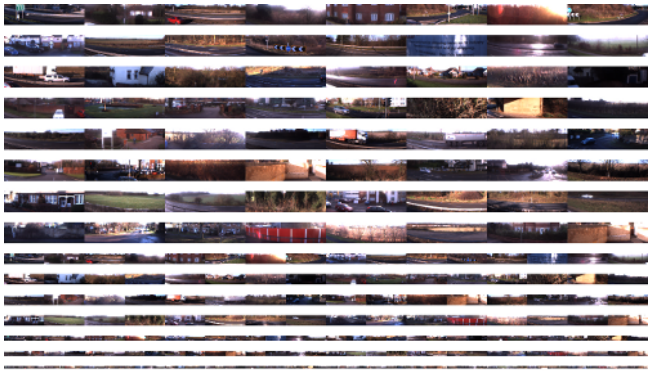


Fig. 3. Reference images. From top to bottom, images named "1", "2", "3", "4", "5", "6", "7", "8", "12", "34", "56", "78", "1234", "5678", "12345678" are shown (zoom in for detail).

IV. Experiments

We conducted view retrieval experiments by utilizing the "RobotCar" dataset provided by the authors of [1] ("FAB-MAP 2.0"). The original dataset consists of GPS, stereo, and omni-directional image data acquired by a car robot during its driving 1,000km in outdoor environments. An omni-directional image consists of 5 images from each of the five side-facing cameras #1-#5 of Ladybug cameras mounted on the robot car. For view retrieval experiments, we chose images from the camera#1 which is directed to the right and use image data within a rectangular region, $y \in [100, 250]$, and GPS data for ground truth. We expect the input images to be contaminated by variations in viewpoints, illumination and partial occlusions. To counter this, each view is represented by a small set of 10 frames sampled from a short frame sequence, and similarity between a given view pair is defined as the average of the 10×10 frame pairs from the view pair. Each retrieval experiment uses independent dataset, which consists of a query view and a size $N = 100$ view database. Each database consists of one relevant view and a set of random $(N-1)$ distracter views which do not overlap with either query or relevant view. We utilize the information of loop closing provided as a part of the "RobotCar" dataset, and use each of the beginning and the ending of a loop respectively as a query and a relevant views. The resulted datasets consist of images

TABLE I

ANR PERFORMANCE OF DIFFERENT BOW FRAMEWORKS

Dataset name	ANR(%)
"RobotCar" 10K words BoVF vocabulary	42.59
1K words BoVF vocabulary (20 words per image)	41.85
BoBB vocabulary (20 words per image)	**35.38**

TABLE II

INFLUENCE OF SUBIMAGE PROPERTIES.

			crop			
			w: 0.5		w: 0.9	
			h: 0.5	h: 0.9	h: 0.5	h: 0.9
scale	w: 1.0	h: 1.0	38.73	36.19	41.22	44.41
		h: 1.25	39.14	36.66	38.44	44.60
	w: 1.25	h: 1.0	38.72	**35.38**	42.01	45.35
		h: 1.25	40.34	36.89	40.56	44.13

taken of the same scene from different viewpoints, and the appearance variations are attributed to various factors including viewpoint change, illumination change, occlusion, which makes the view retrieval tasks challenging.

For performance evaluation, we use the averaged normalized rank (ANR) [4] as performance measure. The normalized correlation is used as the correspondence measure S. The number of common patterns per image is set $T = 20$ as default. The size of bounding box for a size $w \times h$ input image is set $(w \cdot crop.w) \times (h \cdot crop.h)$, where $crop.w = 0.5$ and $crop.h = 0.9$ are used as default. The scaling factor for shrinking/stretching bounding boxes along the horizontal and vertical directions are respectively set $scale.w = 1.25$ and $scale.h = 1.0$ in default. The default reference image is constructed by appending 8 images sampled from the image set, each of which does not overlap with any query or database image, and shown as "1" in Fig.3.

Although our BoBB (i.e. object-level) framework is complementary to existing BoVF (i.e. feature-level) frameworks, for the sake of evaluation, the BoVF framework was also implemented and compared with the proposed BoBB framework. In this study, two types of BoVF view retrieval systems were developed. One is based on the BoVF data provided as a part of the "RobotCar" dataset. We weighted the original BoVF vectors with standard TF-IDF weighting scheme, then index and retrieve the view database using an inverted file system, and evaluate the performance using the ANR measure. From our standpoint, a major inconvenience of the above publicly available BoVF data is that its vocabulary is learned from images acquired by the whole omni-directional camera, i.e. not the camera#1 we use. We hence constructed another independent BoVF data which is learned from the images acquired by the camera#1. A set of training images that are independent from the query and the database images are randomly sampled from the entire image set, and for each training image, a bag of SIFT feature vectors are extracted at keypoints extracted by a grid sampling technique, and then quantized into a bag of visual words using the approximated k-means (AKM) quantization technique in [15]. In this study, we set the vocabulary size to 1K words. Table I reports the ANR performance comparing the proposed BoBB framework with the other two BoVF frameworks. The BoVF frameworks on our view retrieval problem is not as impressive as we expected. This is because of that in the current dataset, matched objects often occupy only a small portion of an image, which are very difficult to be identified (Fig.2). Although the object retrieval techniques (e.g. geometric verification) have been used to counter this problem in literature, they require many words per image, and not suited for compact view database as discussed in II. In contrast, our BoBB framework based on compact description of object patterns achieves much better retrieval performance with requiring only 20 words per image.

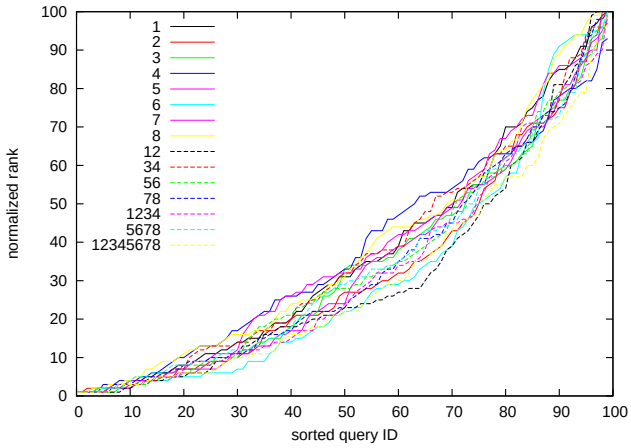


Fig. 4. ANR performance for different reference images.

So far, the experiments have focused on the case where an input image is characterized by a pool of *small* subimages. To evaluate effectiveness of this strategy, we also implemented an alternative strategy where the pool of subimages consists of *big* subimages almost the same size (e.g. 90%) as the input image, and compare it with the proposed strategy. In this study, we evaluate 16 different cases (crop.w, crop.h, scale.w, scale.h)= $\{0.5, 0.9\} \times \{0.5, 0.9\} \times \{0.5, 0.9\} \times \{0.5, 0.9\}$. Tab II reports the comparison results. It can be seen that the strategies with crop.w=0.5 clearly outperform the ones with crop.w=0.9. This is due to the fact in the current car robot applications, there are large variations in the viewpoint particularly in the horizontal direction, and setting the parameter to a small value crop.w=0.5 allows the robot to adaptively learn the size and pose of the subimages according to those variations.

One of key properties of the proposed BoBB framework is that the BoBB image descriptor is strongly dependent on the choice of the reference image that is used for common pattern discovery. We are particularly interested in understanding the impact of the choice of the reference image on the retrieval performance. Thus, we further conducted series of independent retrieval experiments using 15 different reference images, which is created from 8 reference images with the same size shown as "1"-"8" in Fig.3 by appending horizontally a pair of images (e.g. "1234" is an append of a pair of "12" and "34"), as shown in Fig.3. The graph in Fig.4 reports the ANR performance for each of the 15 reference images, where the vertical axis is the normalized rank [%] and the horizontal axis is the sorted query ID [%]. It can be seen that the proposed BoBB framework is

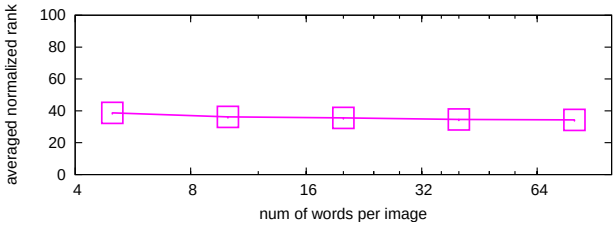


Fig. 5. ANR performance vs. num of words per image.

stable and successful for almost all the reference images used in this study. However, since our algorithm is designed to represent an input image by a pool of cropped reference images, our algorithm would not be suitable for general cases where whole regions of the input image is dissimilar from the reference image, e.g. obviously not suited for the case where the reference image is indoor. In the future we shall study a way for automatically choosing the reference images adaptively for a given set of database images.

To investigate the relationship between the number of words per image and the retrieval performance, we conducted additional retrieval experiments, using different number of words per image, 5, 10, 20, 40, and 80. For each case, we got the ANR performance 38.76, 36.19, 35.58, 34.61, 34.33, as summarized in Fig.5. The large number of words per image was used, the better was the ANR performance. However, increasing the number of words per image requires larger number of entries per image and decrease the compactness of the image database, thus there is a tradeoff between compactness and retrieval performance. In future, we would like to explore methods to improve this tradeoff.

V. Conclusions

We proposed a novel BoW approach, bag-of-bounding-boxes (BoBB), to build and retrieve a compact view image database, which is characterized by (1) *semantic* object-level image characterization, (2) *unsupervised* scene modeling, (3) *compact* view image descriptor, (4) *efficient* indexing and retrieval, and (5) the state-of-the-art CPD techniques. Experiments on challenging outdoor datasets show that our framework is insensitive to system parameters and robust to variations in the viewpoint, contaminations of images by noise, color and partial occlusions. Future work will explore the optimization and adaptive learning of dictionary image for unseen environments and compact the description of view images which are enabled by the proposed bag-of-bounding-boxes framework.

References

[1] Mark Cummins and Paul Newman. Highly scalable appearance-only slam - fab-map 2.0. In *Robotics: Science and Systems*, 2009.

[2] William P. Maddern, Michael Milford, and Gordon Wyeth. Capping computation time and storage requirements for appearance-based localization with cat-slam. In *ICRA*, pages 822–827, 2012.

[3] Andreas Wendel, Michael Maurer, Gottfried Graber, Thomas Pock, and Horst Bischof. Dense reconstruction on-the-fly. In *CVPR*, pages 1450–1457, 2012.

[4] Josef Sivic and Andrew Zisserman. Video google: Efficient visual search of videos. In *Toward Category-Level Object Recognition*, pages 127–144, 2006.

[5] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[6] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. In *CVPR*, pages 257–263, 2003.

[7] David Nistér and Henrik Stewénius. Scalable recognition with a vocabulary tree. In *CVPR*, pages 2161–2168, 2006.

[8] Hung-Khoon Tan and Chong-Wah Ngo. Common pattern discovery using earth mover's distance and local flow maximization. In *ICCV*, pages 1222–1229, 2005.

[9] Dorit S. Hochbaum and Vikas Singh. An efficient algorithm for co-segmentation. In *ICCV*, pages 269–276, 2009.

[10] Christoph H. Lampert, Matthew B. Blaschko, and Thomas Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *CVPR*, 2008.

[11] Yuning Jiang, Jingjing Meng, and Junsong Yuan. Randomized visual phrases for object search. In *CVPR*, pages 3100–3107, 2012.

[12] Junsong Yuan and Ying Wu. Spatial random partition for common visual pattern discovery. In *ICCV*, pages 1–8, 2007.

[13] Matthieu Guillaumin and Vittorio Ferrari. Large-scale knowledge transfer for object localization in imagenet. In *CVPR*, pages 3202–3209, 2012.

[14] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, 2008.

[15] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[16] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, pages 304–317, 2008.

[17] Li-Jia Li, Hao Su, Eric P. Xing, and Fei-Fei Li. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, pages 1378–1386, 2010.

[18] Herve Jegou, Matthijs Douze, and Cordelia Schmid. On the burstiness of visual elements. In *CVPR*, pages 1169–1176, 2009.

[19] Ondrej Chum, James Philbin, Michael Isard, and Andrew Zisserman. Scalable near identical image and shot detection. In *CIVR*, pages 549–556, 2007.

[20] Ondrej Chum, James Philbin, and Andrew Zisserman. Near duplicate image detection: min-hash and tf-idf weighting. In *BMVC*, 2008.

[21] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Packing bag-of-features. In *ICCV*, pages 2357–2364, 2009.

[22] Kanji Tanaka and Kensuke Kondo. Multi-scale bag-of-features for scalable map retrieval. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, pages 793–799, 2013.

[23] Javier Civera, Dorian Gálvez-López, Luis Riazuelo, Juan D. Tardós, and J. M. M. Montiel. Towards semantic slam using a monocular camera. In *IROS*, pages 1277–1284, 2011.

[24] Maren Bennewitz, Cyrill Stachniss, Wolfram Burgard, and Sven Behnke. Metric localization with scale-invariant visual features using a single perspective camera. In *EUROS*, pages 195–209, 2006.

[25] Maurice F. Fallon, Hordur Johannsson, and John J. Leonard. Efficient scene simulation for robust monte carlo localization using an rgb-d camera. In *ICRA*, pages 1663–1670, 2012.

[26] Kanji Tanaka and Eiji Kondo. A scalable algorithm for monte carlo localization using an incremental e^2lsh-database of high dimensional features. In *ICRA*, pages 2784–2791, 2008.

[27] Kouichirou Ikeda and Kanji Tanaka. Visual robot localization using compact binary landmarks. In *ICRA*, pages 4397–4403, 2010.

[28] Tomomi Nagasaka and Kanji Tanaka. An incremental scheme for dictionary-based compressive slam. In *IROS*, pages 872–879, 2011.

[29] Gunhee Kim, Eric P. Xing, Fei-Fei Li, and Takeo Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *ICCV*, pages 169–176, 2011.

[30] Jan Cech, Jiri Matas, and Michal Perdoch. Efficient sequential correspondence selection by cosegmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1568–1581, 2010.

[31] Yuuto Chokushi, Kanji Tanaka, and Masatoshi Ando. Common landmark discovery in urban scenes. *IAPR Int. Conf. Machine Vision Applications*, 2013.