MVA2015 IAPR International Conference on Machine Vision Applications, May 18-22, 2015, Tokyo, JAPAN

Leveraging Image based Prior for Visual Place Recognition

Tsukamoto Taisho Tanaka Kanji
University of Fukui
3-9-1, Bunkyo, Fukui, Fukui, JAPAN
tnkknj@u-fukui.ac.jp

Abstract

In this study, we propose a novel scene descriptor for visual place recognition. Unlike popular bag-of-words scene descriptors which rely on a library of vector quantized visual features, our proposed descriptor is based on a library of raw image data, such as publicly available photo collections from Google StreetView and Flickr. The library images need not to be associated with spatial information regarding the viewpoint and orientation of the scene. As a result, these images are cheaper than the database images; in addition, they are readily available. Our proposed descriptor directly mines the image library to discover landmarks (i.e., image patches) that suitably match an input query/database image. The discovered landmarks are then compactly described by their pose and shape (i.e., library image ID, bounding boxes) and used as a compact discriminative scene descriptor for the input image. We evaluate the effectiveness of our scene description framework by comparing its performance to that of previous approaches.

1 Introduction

Scene description is an important first stage in visual place recognition (VPR), which allows one to search through a pre-built image database to find visually similar views. The most popular scene description method is to translate each image into a bag of vector-quantized visual features, termed as visual words, and then apply document retrieval techniques that are based on the bag-of-words document model (BoW) [1]. Many recent VPR systems are based on the BoW scene description scheme. Despite its computational efficiency and robustness, these BoW scene descriptor -based VPR systems suffer from vector quantization errors, and often fail to handle the appearance changes across views that appear in practice [2].

In this study, we address this issue by leveraging image based prior. Unlike popular BoW scene descriptors which rely on a library of vector quantized visual features, our proposed method is based on a library of raw image data, such as publicly available photo collections from Google StreetView and Flickr. These library images need not be associated with spatial information such as the viewpoint and orientation of the scene, and are thus cheaper than the database images; furthermore, these library images are readily available, which is an added advantage. In our approach, the descriptor directly mines the image library to identify landmarks (i.e., image patches) that suitably match an input query/database image. The discovered landmarks are then compactly described by their pose and shape, i.e., library image ID, bounding boxes (BB), and used as a compact discriminative scene descriptor for the input image. We evaluate the effectiveness of our scene description framework by comparing its performance to that of previous approaches.

The problem associated with conventional scene descriptors for VPR have been studied extensively. Local feature approaches such as BoW scene descriptors have been widely studied considering various aspects, including self-similarity of images [3], quantization errors [4], query expansion [5], database augmentation [6], vocabulary tree [7], global spatial geometric verification as post-processing [8], and pyramid matching to capture spatial context [9]. Previous researches on VPR have shown that the BoW scene model is not sufficiently discriminative and is often unsuccessful at capturing the appearance changes across views [2]. Global feature approaches such as GIST feature descriptor [10] (in which a scene is represented by a single global feature vector) focus on the compactness of scene description and have high matching speeds. Other possible representations include those that describe a scene as a collection of meaningful parts, such as object models [11] and part models [12]. Although these approaches may potentially provide rich information about a scene, existing techniques rely on a large amount of training examples to learn about the models under supervision. Note that our use of a publicly available photo collection (e.g., Flickr) is different from that of large-scale geo-localization [13] where the collection is directly utilized as the database rather than a library.

This study is motivated by the authors' previuos works on a novel data mining approach to scene description [14, 15, 16]. [14] built a prototype method called "common landmark discovery", in which landmark objects are mined through common pattern discovery (CPD) between an input image and known reference images. This framework has been further extended for large-scale visual place recognition by introducing efficient CPD techniques in [15]. The data mining approach has been utilized for single-view cross-season place recognition in [16], where objects whose appearance remain the same across seasons are utilized as valid landmarks. The effectiveness of the scene description framework was evaluated by comparing its performance to that of previous BoW approaches, and by adapting the Naive Bayes Nearest neighbor (NBNN) distance metric [17] to our scene description framework, ("NBNN scene descriptor"). In contrast, the current study further investigates the effectiveness of the proposed approach from a novel perspective of landmark mining.

2 VPR Framework

The VPR framework consists of three main steps, including scene parsing, scene description, and scene retrieval. First, during scene parsing, an input scene is analyzed, and landmarks are discovered that effectively explain the input image. Second, the framework describes the input scene using IDs of the library images and BBs that crop landmark objects within each library image. Scene descriptors are also computed for all images in the image database. Finally, the third step involves the retrieval of database images using the computed scene descriptors as the query.

For the above mentioned method, we assume a dictionary or library of random L_o view images to be given. The library images need not required associated with spatial information such as the viewpoint and orientation. A small subset of $L (L \leq L_o)$ appropriate library images that are most similar to a given input image are selected and used to interpret the image. Our experimental results suggest that high recognition performance tends to be associated with the coverage of the database images provided by these library images.

2.1 Scene Parsing

We consider scene parsing as data mining over the image library. Our scheme begins by over-segmenting the input scene image into a set of R superpixels and clustering them into a set of K scene parts, which will serve as landmark candidates. Then, it evaluates the usefulness of each landmark region in terms of the saliency of the region. It selects K landmark regions with the highest usefulness score, and translates each of these into a compact VLAD code. VLAD codes are also computed for each landmark for all images in the image library. Then, the image library is searched using the K VLAD codes as query, and a score is assigned to each library image in terms of the sum of the reverse rank $\sum_{i=1}^{K} 1/r_i$ of the individual VLAD-based ranking results $r_i (i = 1, \cdots, K)$. For image segmentation, $R = 72$ superpixels are produced by SLIC superpixel [18], and clustered into $2R - 1$ landmark regions using hierarchical region clustering method provided in [19]. For saliency evaluation, the PCA-based distinctiveness score that has been described in [20] is evaluated for all the SIFTs belonging to the region and these are summed up to obtain the region's saliency. To calculate the VLAD codes, method used in [21] is employed. The number of landmarks K per image controls the reliability-efficiency tradeoff of our data mining and currently was set to a relatively high value $K = 40$ (i.e., put weights on reliability) during our study.

2.2 Scene Description

We describe a scene using L landmarks and each landmark is described as a pairing of a landmark image ID and a BB of landmark region with respect to the landmark image. The procedure for discovering landmark images was as discussed in the previous subsection. However, the problem of determining the BB has not been addressed yet. In the proposed method, we extract sets of SIFT features from the input and the library images, F_Q and F_L, in addition, the nearest point to each $f \in F_Q$ among the F_L points in the 128-dim SIFT descriptor space, and then use keypoints $\{(x, y)\}$ of the nearest point to compute the BB. For noise reduction, only the middle 80% x (or y) values are

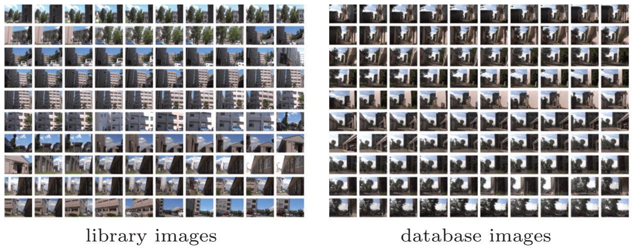

Figure 1. Snapshot of our image collection captured at a University Campus.

used for the computation after all the x (or y) values are sorted numerically. As a consequence, our scene descriptor is of the form:

$$\{\langle I_i, B_i \rangle\}_{i=1}^{L}, \tag{1}$$

where I_i is the ID of landmark image, B_i is the BB consisting of the top left and the bottom right node, (x_i^{min}, y_i^{min}) and (x_i^{max}, y_i^{max}), of BB.

2.3 Scene Retrieval

In this final step, we search the image database and score each database image using the scene descriptor. To build the database, the image ID I_i with the BB B_i for each database image is stored in an inverted file using the element I_i as index. This structure is an array of L_o inverted lists, one per library image ID. For database retrieval, each I_i of a given query image is used as the index and all the database images assigned to the inverted list associated with this I_i are returned. To evaluate the similarity between the query and each of the returned database images, we use the number of common I_i between the image pair as the primary similarity measure, and the area of overlap between the BB pair as the secondary similarity measure.

3 Experimental Results

To evaluate our proposed method, we used an image dataset consisting of view images captured at a university campus, using a handheld camera as the vision sensor. Occlusion is severe in the scenes, and people and vehicles are dynamic entities occupying the scene. We took nine different paths three times each, to collect three independent collections of images of each path, and used each of them for query, library and database image collections. The size of each query and library imageset was 100. The sizes of the database imagesets were 338, 406, 474, 529, 371, 340, 354, 397 and 328. Fig.1 shows examples of library and database images. It can be seen that the database consists of near duplicate images, which makes our scene retrieval a challenging task.

Fig.2 shows some examples of scene parsing. The first column in Fig.2 shows the input image and the following $L = 20$ columns show the L landmark images and their BBs that describe the input image. Further, it is evident that not all the selected landmark images look similar to the input query image they describe. Despite this fact, many of the landmark images actually contribute to obtaining discriminative scene descriptors as we report in the following results.

Figure 2. Scene parsing. The first column shows the input view image that is to be described. The columns numbered 1-20 show the $L = 20$ library images used for describing the input view image.

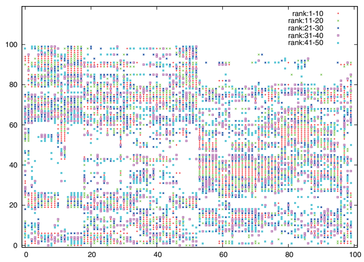

Figure 3. Relationship between input and library images. The x-axis represents the ID of input image that needs to be classified, whereas the y-axis indicates the ID of library images used.

Table 1. Performance results.

dataset	BoW	VLAD	IP	
			w/o BB	w/ BB
0	31.7	26.9	24.2	22.0
1	38.7	27.8	21.9	21.2
2	34.4	14.0	15.3	14.8
3	27.5	20.8	21.6	19.6
4	28.9	17.5	16.2	14.8
5	21.6	17.6	16.9	15.4
6	21.7	27.1	26.4	24.1
7	28.9	28.2	23.1	21.1
8	26.4	23.7	25.2	22.1

Fig.3 shows the relationship between input and library images. In the figure, “rank” means that the ranking assigned by our library image selection at the image description stage. For instance, when we set $L = 20$, only “rank:1-10” and “rank:11-20” images are used for description. We observe that only a small subset of library images tend to contribute to the retrieval performance.

Table 1 lists performance results. We evaluated the proposed image based prior method (“IP”) in terms of the retrieval accuracy and compare it with the BoW method (“BoW”) [1], and VLAD [21]. For the BoW method, we employed a visual feature descriptor and a vocabulary provided in [1]. For VLAD, we employed the code used in [21]. A series of independent 100×9 retrievals were conducted for each of the 100 random query images of all the 9 different paths. The retrieval performance was measured in terms of the averaged normalized rank (ANR) as percentage; the ANR is a ranking-based retrieval performance measure wherein a smaller value indicates a better retrieval performance. To evaluate ANR, the rank assigned to the ground-truth relevant image was evaluated for each of the 100 independent retrievals, and then the rank was normalized on the basis of the database size and these ranks were averaged over the 100 retrievals. From Table 1, one can observe that our approach outperformed both BoW and VLAD in most of the retrievals considered in this study.

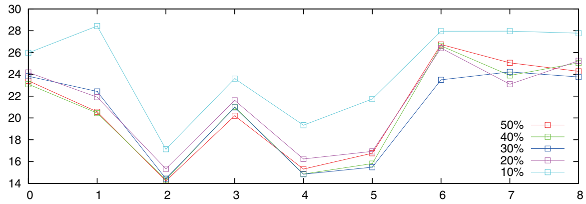

Figure 4. Graph showing the effect of the number of landmarks used per image during the description process.

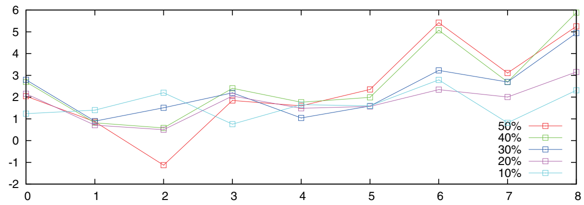

Figure 5. Comparison between cases with and without BBs.

We also investigate the influence of the parameter L, i.e., the number of landmarks used for scene modeling. Fig.4 shows the ANR performance for different settings of the parameter L, including $L = 10$, 20, 30, 40 and 50. As can be seen, the results are comparable to each other. An exception is the case where $L = 10$, where the number of landmarks are too small to make our bag of landmarks based representation less discriminative.

We also investigated the effect of using BBs on the retrieval performance. In this study, we conducted another set of experiments using the proposed scene descriptor without using the BBs, as a proof-of-concept, and compared the recognition performance against that of the proposed descriptor. Fig.5 shows the comparison of results of the proposed descriptor with and

Figure 6. Snapshot that shows the cases in which our method fails.

without the BBs. The vertical axis in this figure is the ANR performance of the case using BBs subtracted from that of the case without using BBs. It can be seen that the ANR performance shows an improvement when the BBs are used for most of the cases considered in this study. A notable exception is the case where L is set to a relatively large value, e.g., 50. This is due to a large number of landmark images that naturally include dissimilar scenes as we already showed in Fig.3, and BBs of landmarks with respect to such dissimilar landmark images provide less meaningful and less reliable information. However, it should be noted that even such dissimilar landmark images do actually improve the scene retrieval performance as we can see in Fig.4.

Fig.6 reports some examples of failure cases. For each row, the first column shows the query images, the 2nd, 3rd, and 4th columns show the images that received higher similarity score than the ground-truth images when the proposed method was used, and the last column shows the ground-truth images. As can be seen, the proposed approach can be confused if some database images with locally similar but globally dissimilar structures that cannot be captured by "bag-of-X" scene model are included. However, the issue of the globally dissimilar structure can be mitigated by introducing some extension to the BoX model such as spatial pyramid matching; this will form part of our future work.

4 Conclusions

The primary contribution of this paper is the proposal of a simple and effective approach to VPR. Unlike popular BoW scene descriptors which rely on a library of vector quantized visual features, our descriptor is based on a library of raw image data, such as publicly available photo collections from Google StreetView and Flickr; our method directly mines the library to discover landmarks (i.e., image patches) that effectively explain an input query/database image. The discovered landmarks are then compactly described by their pose and shape (i.e., library image ID, BBs) and used as a compact discriminative scene descriptor for the input image. Experiments using a challenging dataset validate the effectiveness of the proposed approach.

References

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

[2] B. Yao, G. R. Bradski, and F. Li, "A codebook-free and annotation-free approach for fine-grained image categorization," in *CVPR*, 2012, pp. 3466–3473.

[3] J. Knopp, J. Sivic, and T. Pajdla, "Avoiding confusing features in place recognition," in *ECCV*. Springer, 2010, pp. 748–761.

[4] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *CVPR*, 2012, pp. 2911–2918.

[5] O. Chum, A. Mikulik, M. Perdoch, and J. Matas, "Total recall ii: Query expansion revisited," in *CVPR*, 2011, pp. 889–896.

[6] P. Turcot and D. G. Lowe, "Better matching with fewer features: The selection of useful features in large database recognition problems," in *ICCV Workshops*, 2009, pp. 2109–2116.

[7] G. Schindler, M. Brown, and R. Szeliski, "City-scale location recognition," in *CVPR*, 2007, pp. 1–7.

[8] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *CVPR*, 2007, pp. 1–8.

[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] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid, "Evaluation of gist descriptors for web-scale image search," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2009, pp. 19:1–19:8.

[11] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Advances in neural information processing systems*, 2010, pp. 1378–1386.

[12] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman, "Blocks that shout: Distinctive parts for scene classification," in *CVPR*, 2013, pp. 923–930.

[13] A. R. Zamir and M. Shah, "Accurate image localization based on google maps street view," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 255–268.

[14] A. Masatoshi, T. Kanji, I. Yousuke, C. Yuuto, and H. Shogo, "Common landmark discovery for object-level view image retrieval: Modeling and matching of scenes via bag-of-bounding-boxes," in *ACPR*, 2013.

[15] T. Kanji, C. Yuuto, and A. Masatoshi, "Mining visual phrases for long-term visual slam," in *IROS*, 2014, pp. 136–142.

[16] L. I. based Prior in Cross-Season Place Recognition, "Ando masatoshi and chokushi yuuto and tanaka kanji and yanagihara kentaro," in *ICRA*, 2015.

[17] T. Tommasi and B. Caputo, "Frustratingly easy nbnn domain adaptation," in *ICCV*, 2013, pp. 897–904.

[18] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. PAMI*, vol. 34, no. 11, pp. 2274–2282, 2012.

[19] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders, "Segmentation as selective search for object recognition," in *ICCV*, 2011, pp. 1879–1886.

[20] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?" in *CVPR*, 2013, pp. 1139–1146.

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