Scene Retrieval by Unsupervised Salient Part Discovery

Sugegaya Naotoshi Tanaka Kanji Yanagihara Kentaro
University of Fukui
3-9-1, Bunkyo, Fukui, Fukui, JAPAN
tnkknj@u-fukui.ac.jp

Abstract

While bag-of-words (BoW) scene descriptor has been widely used for scene retrieval applications, the BoW descriptor alone often fails to capture local details of a scene and produces poor results. In this paper, we address this issue by a simple effective approach, "unsupervised salient part discovery", in which a set of salient parts are discovered via scene parsing and used as additional queries for the scene retrieval. Further, we also address the issue of discovering salient parts in a scene, and present a solution that provides similar parts for similar scenes. Multiple ranking results from the individual part queries are then integrated into a final ranking result by adopting an unsupervised rank fusion technique. Experimental results using challenging scene dataset validate the effectiveness of our approach.

1 Introduction

Visual search, object matching and many other scene retrieval applications rely on representing images with compact discriminative scene descriptors. The most popular descriptor is bag-of-words (BoW), a histogram of frequencies of visual words obtained by quantizing local image descriptors to a visual vocabulary. This study is built on a recently developed BoW scene description scheme, vector of locally aggregated descriptors (VLAD) [1], which achieves state-of-the-art discriminativity while maintaining extreme compactness.

This study is motivated by the observation that BoW scene descriptor alone fails to capture local details of a scene. Typically, BoW descriptor is sensitive to view changes, and often produces poor results in scene retrieval [2]. A simple solution is to use a bag of raw local SIFT-like descriptors (e.g., bag-of-raw-features [3]). However, it may not be possible in practice as it requires to memorize and match a large number of high dimensional local SIFT-like descriptors. Instead, we introduce a global-local hybrid descriptor, called part descriptor, which describes a relatively large region covering 30%-80% of the entire image. By integrating a set of local BoW part descriptors and the global BoW scene descriptor, we achieve a practical compactness-discriminativity tradeoff.

Further, we also address the issue of discovering useful parts in a scene. This is different from the problem of object segmentation, i.e., segmenting an image into meaningful parts such as objects, which is a core problem in the field of machine vision [4, 5, 6]. Our goal is to realize consistent segmentation for similar view images, allowing one to obtain similar part descriptors for similar scenes. Our solution is composed of three distinct steps. (1) For stable part segmentation, we borrow techniques from the hierarchical region clustering algorithms [7], which provide a pool of homogeneous scene parts with different levels of inter-region appearance similarity. (2) To select a small number of effective scene parts, we adopt the PCA-based saliency evaluation measure recently developed in [8] for the application of part-based scene description. (3) To integrate multiple ranking results from individual part retrievals, we develop a novel ranking scheme based on an adaptive ranking strategy.

Our contributions are summarized as follows: (1) We provide a simple effective approach, "unsupervised salient part discovery", in which a set of useful parts are discovered via scene parsing and used as additional queries for scene retrieval. (2) We provide an effective strategy to integrate both the color and the SIFT cues by combining the part segmentation and the part retrieval. (3) We present a practical scene retrieval system that achieves practical discriminativity-compactness tradeoff. (4) Experimental results using challenging scene dataset validate the effectiveness of the proposed scheme.

Bag-of-words (BoW) scene descriptors have been extensively studied in the context of scene recognition [9, 10, 11]. There are also several techniques to extend BoW from various aspects, including self-similarity of images [12], quantization errors [13], query expansion [14], database augmentation [11], vocabulary tree [15], global spatial geometric verification as post-processing [16], spatial pyramid matching for capturing spatial context [17], and various strategies for feature pooling [18]. Most of the above techniques extending BoW can also be adopted in extending our approach. The scene recognition task addressed in this paper is different from scene categorization where the goal is to classify a scene into pre-learned scene categories. In scene categorization literature, BoW has been combined with discriminative learning techniques such as SVM and achieved high recognition performance. For instance, [19] developed a discriminative scene learning framework in which the structure of a database image is represented by a graph and used for graph based ranking and re-ranking that improve BoW. Some methods describe a scene as a collection of meaningful parts, such as object models [20] and part models [21]. Although these approaches may potentially provide rich information of a scene, existing techniques rely on a large amount of training examples to learn the models under supervision.

2 Scene Retrieval Framework

Our part-based scene description approach combines two different types of cues, color cue and SIFT cue, to achieve better scene retrieval performance (Fig.1). We observe that color cue provides inter-region similarity information that is useful for image segmentation,

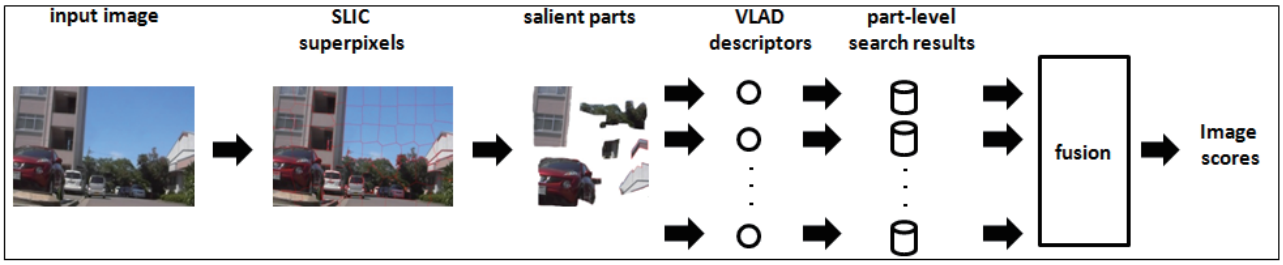


Figure 1. Algorithm pipeline.

while SIFT cue provides inter-image invariant information that is useful for image matching. We also observe that many part hypotheses are typically produced by color-based part segmentation, but they can be effectively verified by SIFT-based part matching in a hypothesize-and-verification fashion [22].

Based on the observations, our scene retrieval framework consists of four distinct steps:

1. Segment the scene into a pool of part regions,
2. Evaluate saliency of each part and select salient parts with highest saliency,
3. Compute scene/part descriptors,
4. Retrieve the database using each descriptor as query and integrate the multiple retrieval results.

Each of the above steps 1)-4) will be explained in detail in the following subsections.

2.1 Segmentation

Our strategy for part segmentation is to segment an image into homogeneous color regions. We adopt the hierarchical region clustering strategy [7] as it consistently produces similar region segments for similar views. Our framework begins by extracting superpixels using the SLIC algorithm in [23] as initial candidates of homogeneous color regions. It then iteratively merges a pair of similar regions at a time to produce an additional region candidate. For each iteration, it selects a pair of neighboring regions with highest similarity among all neighboring region pairs in the pool of candidate regions, and merges the selected region pair to produce a larger and less homogeneous region. Dissimilarity between a region pair is measured in terms of Euclidean distance in RGB color space. Given N initial superpixels, we finally obtain $(2N-1)$ region candidates in total.

2.2 Saliency

We adopt PCA-based saliency measure in [8]. The basic idea of PCA-based saliency is to use PCA to capture dominant variations among patterns, and to evaluate the distinctiveness of a pattern by measuring the PCA distance. Our framework begins by extracting SIFT keypoints from the entire image, and computing a SIFT descriptor for each keypoint. The PCA is then applied to those computed SIFT descriptors to extract the principal modes of variation. The L_1 norm of a SIFT descriptor is computed in $D' = 10$-dim PCA coordinates, and is used as the SIFT's distinctiveness. We compute the saliency score for every SIFT feature in the part region and use the total score as the region's saliency score.

2.3 Descriptors

Our part selection strategy is motivated by the fact that all the $(2N-1)$ part candidates output by the previous part segmentation stage often are not equally important. Our selection framework evaluates saliency of each candidate region, and then selects K part regions with highest saliency score. Then, it translates those SIFT descriptors that belong to each scene part into a discriminative compact VLAD descriptor in [1]. In our part-based scene descriptor, every query/database image is represented by a pairing of a VLAD scene descriptor and a set of $(K-1)$ VLAD part descriptors.

2.4 Retrieval

The scene retrieval aims to rank all the database images according to similarity to query. We do a series of K independent scene retrievals using each of the K VLAD scene/part descriptors as query, and obtain multiple ranking results from individual retrievals. Our basic idea is to consider multiple search engines for the multiple part queries and fuse individual search results by using rank fusion, similar to Reciprocal Rank Fusion introduced in [24]. Rank fusion techniques are low cost and unsupervised; i.e., they do not require individual engines to return similarity scores nor supervised training data. Currently, multiple search engines share a single common database to save the total spatial cost. More formally, to integrate the multiple ranking results from K queries, we score a database image by integrating reverse ranks from individual scene/part retrievals in the form:

$$S = w\ r_0^{-1} + (1-w) \sum_{i=1}^{K-1} r_i^{-1} \qquad (1)$$

Here, r_0 and r_i $(i \in [1, K-1])$ indicate the ranking results of scene and each i-th part retrieval. In default, $K = 40$ and $w = 0.5$.

3 Experimental Evaluation

For evaluation, we use image dataset consisting of view images taken around a university campus, using a handheld camera as the vision sensor. We went along nine different paths, some of them going through the main central path and others going along the pedestrian walkway along the campus wall, as can be seen

Table 1. Performance results.

data ID	BoW	VLAD	PSD			
			K:10	K:20	K:30	K:40
1	31.7	26.9	25.1	25.5	21.6	22.1
2	38.7	27.8	24.4	23.8	21.8	17.7
3	34.4	14.0	14.7	13.4	11.5	10.7
4	27.5	20.8	19.1	17.1	16.8	16.0
5	28.9	17.5	16.5	14.7	13.6	13.7
6	21.6	17.6	17.2	15.1	14.5	12.4
7	21.7	27.1	27.7	25.9	20.9	16.3
8	28.9	28.2	29.8	26.0	24.8	21.3
9	26.4	23.7	24.8	24.8	22.7	18.3

in Fig.2. Occlusion is severe in all the scenes, and people and vehicles are dynamic entities occupying the scene. We traversed each path twice and obtained a pair of collections of view images for database building and scene retrieval for each path. All images are 1000 x 667 RGB color images. For each path, we collected sets of 338, 406, 474, 529, 371, 340, 354, 397, 328 database images respectively for each dataset, and we also collected sets of 100 query images for each dataset. The dataset consists of many near duplicate images, which makes our scene retrieval a challenging task.

By performing several experiments, we compare the performance of the proposed part-based scene description method, denoted as "PSD", with other two baseline methods, bag-of-words ("BoW") [25] and VLAD ("VLAD") [1]. For VLAD, we interpret a view image into one VLAD scene descriptor. For PSD, we interpret an image into K VLAD descriptors, consisting of one scene descriptor and $(K-1)$ part descriptors, using the scene description scheme described in section 2.3. We conducted a series of 100 independent retrievals for each of the 100 query images and for each of the 9 different datasets. Retrieval performance is measured in terms of averaged normalized rank (ANR) in percent (%). ANR is a ranking-based retrieval performance measure, where the smaller value is better. To compute 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.

Table 1 shows the ANR performance for individual methods for each dataset. It can be seen that the proposed PSD method clearly outperforms both the BoW and VLAD methods. We are also interested in and investigated the relationship between the number K of scene/part descriptors per image and the scene retrieval performance. Table 1 reports the results. It can be seen that the proposed system still almost outperforms other methods even when the number of parts is reduced to K=20.

Fig.2 shows example results of five independent scene retrievals. The first and second columns of Fig.2 are the query input and the ground truth database images. The third and the fourth columns compare the results of database images top-ranked by the BoW and the proposed methods. As can be seen, our approach provides an accurate image retrieval by combining both local part descriptors and global scene descriptor that capture local details while recognizing global layout of the scene.

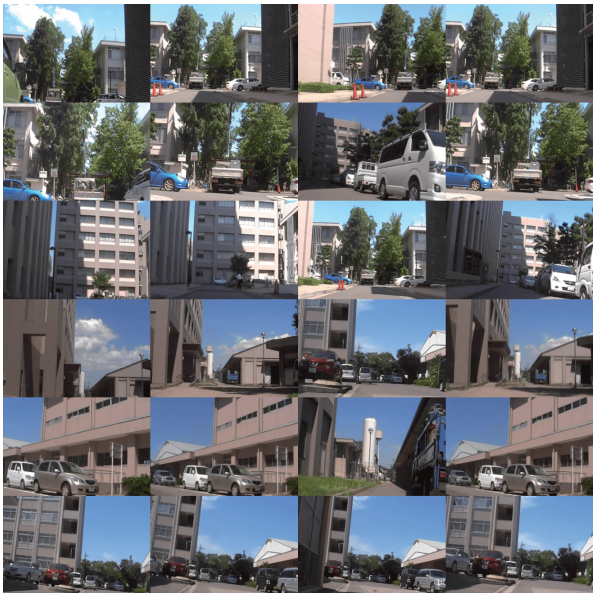


Figure 2. Scene retrieval results.

Fig.3 illustrates results of part matching for an example pair of relevant query and database images. In this figure, four pairs of parts that received highest part-level similarity in terms of the reverse rank r_i^{-1} are selected and shown in each row. In Fig.3a, the largest scene parts that correspond to the entire image regions are selected. In Fig.3b, relatively small parts that consist of buildings are selected as matched parts.

4 Conclusions

This paper proposed two distinct contributions with regard to BoW scene descriptor in scene retrieval applications. First, we provided a simple effective approach, "unsupervised salient part discovery", in which a set of useful parts are discovered via scene parsing and used as additional queries for scene retrieval. Our framework combines color and SIFT cues in an effective manner in order for part segmentation, part-based scene description and retrieval. As a next contribution, we presented a practical scene retrieval system that achieves practical discriminativity-compactness tradeoff. Experimental results using challenging scene dataset showed our approach improves over baseline approaches.

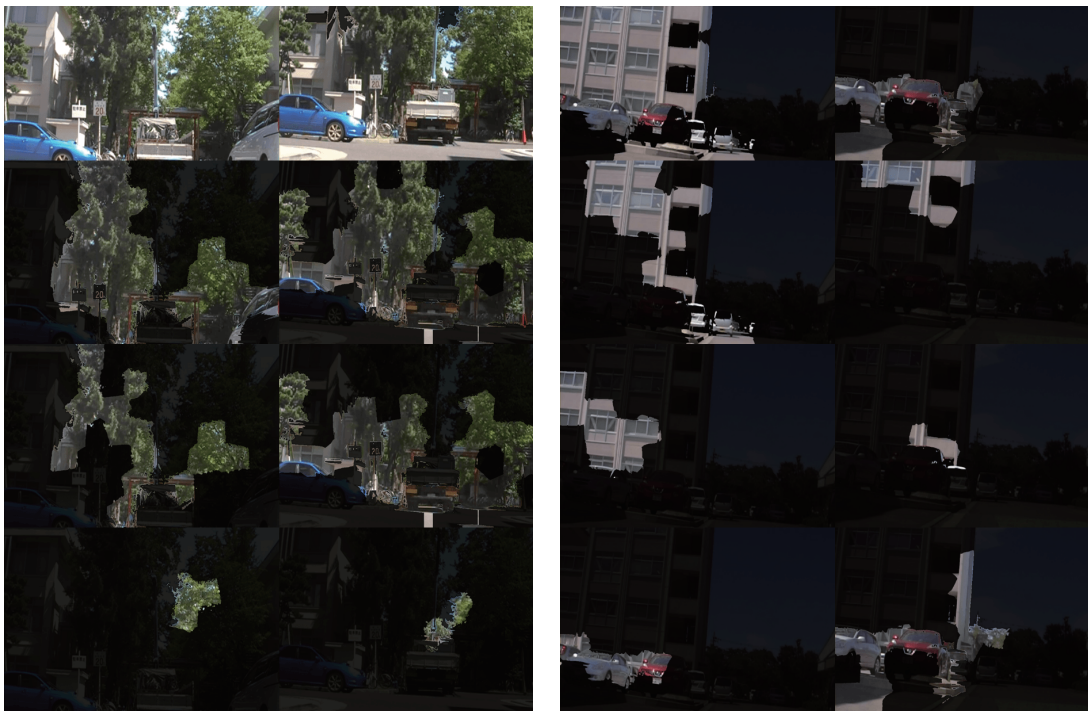
References

[1] 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.

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

[4] A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: A benchmark," in *Proc. ECCV*, 2012, pp. 414–429.



a b

Figure 3. Examples of part matching beween a query and a database images.

[5] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. IEEE CVPR*, 2013, pp. 2083–2090.

[6] S. Manen, M. Guillaumin, and L. V. Gool, "Prime object proposals with randomized prim's algorithm," in *ICCV*, 2013, pp. 2536–2543.

[7] 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.

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

[9] J. Knopp, J. Sivic, and T. Pajdla, "Avoiding confusing features in place recognition," in *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I*, 2010, pp. 748–761.

[10] A. Kim and R. M. Eustice, "Combined visually and geometrically informative link hypothesis for pose-graph visual slam using bag-of-words," in *IROS*, 2011, pp. 1647–1654.

[11] 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.

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

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

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

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

[16] 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.

[17] S. Lazebnik, C. Schmid, J. Ponce, *et al.*, "Spatial pyramid matching," *Object Categorization: Computer and Human Vision Perspectives*, vol. 3, p. 4, 2009.

[18] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *CVPR*, 2010, pp. 2559–2566.

[19] S. Cao and N. Snavely, "Graph-based discriminative learning for location recognition," in *CVPR*, 2013, pp. 700–707.

[20] R. Anati, D. Scaramuzza, K. G. Derpanis, and K. Daniilidis, "Robot localization using soft object detection," in *Proc. IEEE ICRA*, 2012, pp. 4992–4999.

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

[22] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[23] 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.

[24] G. V. Cormack, C. L. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *SIGIR*, 2009, pp. 758–759.

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