CGI2017 manuscript No.
(will be inserted by the editor)

Rank-based Voting with Inclusion Relationship for Accurate Image Search

Jaehyeong Cho · Jae-Pil Heo · Taeyoung Kim · Bohyung Han · Sung-Eui Yoon

Abstract We present a rank-based voting technique utilizing inclusion relationship for high quality image search. Since images can have multiple regions of interest, we extract representative object regions using a state-of-the-art region proposal method tailored for our search problem. We then extract CNN features locally from those representative regions and identify inclusion relationship between those regions. To identify similar images given a query, we propose a novel similarity measure based on representative regions and their inclusion relationship. Our similarity measure gives a high score to a pair of images that contain similar object regions with similar spatial arrangement. To verify benefits of our method, we test our method in three standard benchmarks and compare it against the state-of-the-art image search methods using CNN features. Our experiment results demonstrate effectiveness and robustness of the proposed algorithm.

Jaehyeong Cho
KAIST, Republic of Korea
E-mail: dil122001@gmail.com

Jae-Pil Heo
Sungkyunkwan University, Republic of Korea
E-mail: jaepilheo@gmail.com

Taeyoung Kim
KAIST, Republic of Korea
E-mail: retupmoc@kaist.ac.kr

Bohyung Han
POSTECH, Republic of Korea
E-mail: bhhan@postech.ac.kr

Sung-Eui Yoon
KAIST, Republic of Korea
E-mail: sungeui@kaist.edu

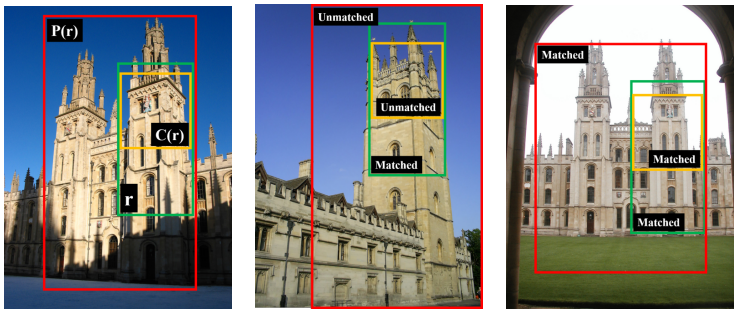
1 Introduction

Image search is a task to identify visually similar images from an image database given a query image, by identifying matching features. This is one of the most fundamental tools in many image processing and graphics applications, since the performance of many high-level problems often depends on the quality of image search.

Many interesting and novel data-driven applications have been proposed thanks to the advance of image search and other related techniques. Some of well-known applications include photo tourism [28] and scene completion [7]. Furthermore, the data-driven approach keeps to widen its scope into many other directions [24, 17]. The final quality of these data-driven approaches are affected significantly by the search accuracy, since they commonly assume that identified images share similar patches and objects. As a result, we are interested mainly in improving the accuracy of image search in this paper.

Image search starts from representing images with feature vectors, which have been investigated extensively. Recently, deep convolutional neural networks (CNNs) [19, 26, 29, 8] demonstrate outstanding classification performance, and the CNN features obtained from a few hidden layers present great generalizability to many other domains or tasks [2, 22, 5].

A naïve integration of global CNN features to recent image search techniques demonstrated better performance [2] than the algorithms based on existing handcrafted image descriptors. such as VLAD [14]. Recent approaches [1, 22] improve accuracy further by incorporating spatial information of images. While these recent CNN-based image search methods utilize spatial information, they still rely only on weaker information compared to techniques developed for object detection



(a) A query image (b) A similar, but incorrect image (c) A correct tested image

Fig. 1 Our method extracts representative regions from images and identify similar images by matching those regions. We also improve matching accuracy by considering the inclusion relationship between regions. (a) shows related regions given a query region r shown in the green box. Yellow and red boxes represent child $C(r)$ and parent $P(r)$ regions of the query region, respectively; i.e., $C(r) \subseteq r \subseteq P(r)$. (b) The region r of the query image is matched with the green box of this test image. Nonetheless, their related regions do not match. As a result, we can conclude this image to be different from the query image. (c) Most of related regions of the query image match with those of this tested image, resulting in matching with the query image.

and localization. This paper discusses how to utilize such candidate regions for image search. One of technical challenges is that most of those candidate regions may not contain any meaningful objects, since many region proposal techniques have been designed to achieve a high recall. These false-positive candidate regions deteriorate accuracy of image search.

Main contributions. We propose a novel, rank-based voting technique, which identifies representative object regions and retrieves similar images based on those regions. Our method identifies such representative regions by utilizing a state-of-the-art region proposal method, while adjusting the objectness measure to be invariant to object sizes (Sec. 3.1). To achieve a high matching quality between regions of images, we introduce a novel similarity measurement method based on a voting scheme, which considers spatial relationship between regions, especially, inclusion relationship (Sec. 3.3). Using the spatial relationship, we can cull out incorrect matching results, resulting in substantial accuracy improvement (Fig. 1).

To verify the effectiveness of our approach, we evaluate our method over three standard benchmark datasets [11,21,20] and compare it against state-of-the-art image search methods also utilizing CNNs. Overall, our method achieves higher search accuracy across all the three benchmarks compared to the tested methods with the same CNN network. Such successful results are attributed to the accurate identification of representative regions and reliable similarity measurement between regions. Source codes of our work are available as an open-source project.

2 Related Work

In this section, we discuss prior methods that are directly related to our work.

2.1 Convolutional Neural Networks and Image Search

Deep convolutional neural networks (CNNs) have achieved remarkable performance improvement in various recognition tasks. The main reason to make CNNs successful is their representation power; CNNs capture high-level semantic information in images much more accurately [2,22,5] than traditional hand-crafted feature descriptors. Moreover, the CNNs pre-trained on a large-scale image datasets are often directly applicable to other domains and tasks, and their performance is even boosted by fine-tuning existing models with additional data.

There are several existing works in image retrieval based on CNN representations. Babenko *et al.* [2] used activations of top three fully connected layers as global descriptors of the input image. Also, they showed that the features can become more discriminative for particular datasets by retraining the network with specific purpose. Some recent papers proposed a feature aggregation approach to generate improved global descriptors [1,16,30,6]. These ideas are similar to methods designed for dense-SIFTs, *i.e.,* VLAD [14]. In an aggregation step, SPoC [1] gives larger weights to features near the center as a simple weighting heuristic, and CroW [16] gives different weights according to the location and channel of each feature. Similarly, R-MAC [30] aggregates region feature vectors, which are generated by collecting the maximum activation of each region.

(a) Query w/ mask (b) Completed image (c) Retrieved image

Fig. 2 This shows examples of the scene completion using our method. (a) show queries with red masks for the completion, while (b) are completed images using the retrieved images (c) by our method in *Oxford 5k*.

Gordo *et al.* [6] computes CNN features by leveraging a three-stream Siamese network with a triplet ranking loss. This method also suggests an aggregated feature among multiples features by taking the maximum activations.

Departing from these aggregation approaches, our method extracts features from multiple regions and utilizes them for achieving high search accuracy.

2.2 Region Detection

It has been widely accepted to extract and use discriminative regions from images for achieving a high search accuracy. One of famous object proposal detection methods is selective search [25], which constructs object proposals from the hierarchical bottom-up image segmentation. Krähenbühl *et al.* [18] introduced geodesic object proposals, where seeds for objects are selected by minimizing the geodesic distance with all the objects. Cheng *et al.* [4] observed that various sizes of objects have the correlation with norms of gradients, when their patches are resized to the 8 by 8 fixed resolution.

Recently, neural networks considering regions have been proposed. Faster R-CNN applies the regression to predict box coordinates, while considering multiple anchors [23]. Also, inside-outside net [3] considered context information for regions. Our proposed method can be combined with these recent region network for achieving higher search accuracy. Our proposed method uses a region proposal method and improve precision by considering the spatial relationship among regions.

2.3 Similarity Measure

It is also important to have a similarity measure that gives a small distance to a pair of similar images. For this purpose, it is common to use simple vector-to-vector distances such as *L1*, *L2*, and *cosine* metrics.

Some techniques extracted multiple features like our approach and thus used similarity measures for these features. Razavian *et al.* [22] extracted features from uniformly generated sub-patches, and measured similarity by averaging the minimum *L2* distance between sub-patches of each image. Recently, Xie *et al.* [31] extracted multiple features from their manually defined objects, and applied Naive-Bayes Nearest Neighbor (NBNN) search in order to utilize semantic category information of images. This method, however, requires categorized image datasets.

Our method utilizes rank-based voting scheme, which directly considers the spatial relationship between regions, and thus improves the accuracy of search method.

3 Our Approach

In this section, we describe our method that extracts multiple regions from an image and utilize the relationship among those regions. At a high level, our approach is divided into four parts: representative region selection, feature extraction, similarity measurement, and compression/indexing parts. We first explain how to extract regions from an image.

3.1 Representative region selection

The state-of-the-art object localization methods show impressive recalls for detecting objects by employing region proposal techniques [4, 25, 18]. Although these approaches are useful to improve recalls in object detection, it is inappropriate to directly use such region proposals to our problem of image search. Especially, we have found that most candidate regions, e.g., more than 90%, generated by recent region proposal methods, i.e., selective search [25] and BING [4], do not match with their ground truth bounding boxes; see the table in the supp. report. It is thus unlikely to achieve high search accuracy based on representations involving such unimportant regions.

To address the aforementioned problems, we decide to leverage only top-k confident regions. Many effective confidence measuring methods make it possible to determine the ranking of region proposals Thankfully, among the state-of-the-art object localization methods, BING [4] proposes candidate regions by estimating their

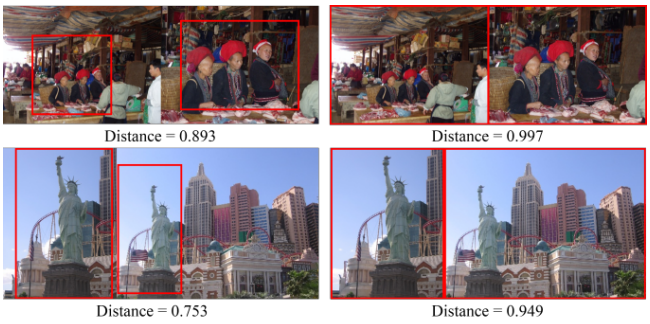


Fig. 3 Comparison of the Euclidean distances between features extracted only from chosen regions (red boxes in left) and extracted from the whole images (right). Features extracted from chosen regions represent the contents of an image clearly, and enable similarity matching to be more accurate.

objectness in an efficient manner. We therefore decide to utilize this objectness metric for selecting top-k representative regions among many candidates. We tune the objectness score to be more suitable for image search, and select top-150 regions based on the score. Details of the tuning is described in our supplementary material, and the analysis for k, the number of selected regions, is available in Sec. 4.6.

3.2 Feature extraction

After selecting k representative regions from an image, we extract a feature from an image box of each region. We can use any CNN feature as our image descriptor for each region. Additionally, we store inclusion relations of regions together, which are used to improve the confidence of matching (Sec. 3.3).

The features extracted from representative regions is more appropriate than features extracted from the whole image, especially for the image search task. As an example, Fig. 3 shows that while two images contain the same object, their Euclidean distance between features extracted from the whole images is far. On the other hand, by extracting features only from representative regions, we achieve a shorter distance, resulting in better image search. In Sec. 4.2, we show that our method improves recall of image retrieval, compared with using features extracted from the whole images.

During the feature extraction, we also store inclusion relations of regions. There exist many kinds of regions such as regions representing some parts of objects, indicating whole objects, and including even backgrounds. We utilize such inclusion relationships between two regions for accurately identifying similar images as shown in Fig. 1(a).

3.3 Similarity measurement

Most existing image search techniques adopted a type of aggregation that computes a single feature from many local features [27, 14, 1]. One of well-known aggregation techniques includes the bag-of-feature model [27]. These techniques then employ a distance metric, *e.g.,* Euclidean or cosine distances, between those aggregated features to compute a similarity between two images. We can also use such a similarity metric by aggregating features extracted from regions into a single feature vector.

Unfortunately, we found that the traditional distance metrics produce suboptimal results to our method, because the aggregated features dilute the content information extracted from multiple image regions. Instead, we propose a voting-based similarity measure, which is well suited for our features extracted from multiple image regions.

Let $q_i \in \mathcal{R}(I_q)$ be an i-th selected region from the query image I_q, where $\mathcal{R}(I_q)$ denotes a set of regions proposals in the image I_q. Since the order of regions does not matter in our method, the feature set of a query image is then given by $\mathcal{F}(I_q) = \{\mathbf{f}(q_i) \mid i = 1, ..., k\}$, where $\mathbf{f}(q_i)$ denotes the feature descriptor for q_i and k is the number of representative regions extracted from the image. For each region q_i, we retrieve images from the database that contain a region close to q_i based on the Euclidean distance in the feature space. Suppose that an ordered set of retrieved regions for the query region q_i is denoted by $\mathcal{D}_i = \{d_{ij} \mid j = 1, ..., v\}$, where j is the rank index in terms of the Euclidean distance and v is the number of retrieved regions.

We identify similar images using a new similarity measure, which is based on a voting scheme defined on region proposals. The proposed voting algorithm provides a score for a region of an image in the database, where the score is composed of two factors:

$$\text{VotingS}(q_i, d_{ij}) = \text{RankS}(q_i, d_{ij}) \cdot \text{RelS}(q_i, d_{ij}), \quad (1)$$

where $\text{RankS}(\cdot,\cdot)$ and $\text{RelS}(\cdot,\cdot)$ denote ranking and relation scores between two regions, respectively. Note that $\text{RankS}(\cdot,\cdot)$ computes rank-based region similarity and $\text{RelS}(\cdot,\cdot)$ considers the inclusion relationship with related regions of input proposals. The final similarity measure considered with all regions between the query image I_q and an arbitrary image I in the database is given by:

$$\text{Sim}(I_q, I) = \sum_{q_i \in \mathcal{R}(I_q)} \max_{d_{ij} \in \mathcal{D}_i \cap \mathcal{R}(I)} \text{VotingS}(q_i, d_{ij}), \quad (2)$$

where $\mathcal{D}_i \cap \mathcal{R}(I)$ indicates a set of retrieved regions for q_i among regions from the image I. This similarity metric

simply sums voting scores, each of which is computed by the best match from a region q_i of the query to another region d_{ij} in the test image I.

We first define our rank-based region similarity, $\mathrm{RankS}(\cdot,\cdot)$, which utilizes a rank of a retrieved region from a query region. Simply, the ranking score of d_{ij} for q_i is determined as:

$$\mathrm{RankS}(q_i, d_{ij}) = 1/j. \tag{3}$$

Intuitively, as we get a bigger rank order, we give a lower similarity score to the region d_{ij}.

We now define our relationship score $\mathrm{RelS}(\cdot,\cdot)$. Our main idea of using the inclusion relationship is that while regions of a database image can match with query regions, their related regions may or may not be matched (Fig. 1(b)). Depending on whether related regions match well or not, our confidence level of the matching between the query image and the tested database image can vary significantly. To realize this intuition, we introduce the relationship score $\mathrm{RelS}(q_i, d_{ij})$.

For an arbitrary region r, we let $P(r)$ be a set of parent regions including the box of r, and $C(r)$ be another set of child regions included in the box of r. We first define a set of matched parent regions between q_i and d_{ij} as follows:

$$\mathrm{MatP}(q_i, d_{ij}) = \left\{ d_p \mid d_p \in \mathcal{D}_x \cap P(d_{ij}) \; for \; q_x \in P(q_i) \right\}, \tag{4}$$

where $\mathcal{D}_x$ is a retrieved region set of q_x, a parent region of q_i. Intutively speaking, $\mathrm{MatP}(q_i, d_{ij})$ contains the parent regions of d_{ij} that are also retrieved for parent regions of q_i. A set of matched child regions $\mathrm{MatC}(\cdot,\cdot)$ is defined in the same manner. We then define a set of matched related regions between q_i and d_{ij} as follows:

$$\mathrm{MatRels}(q_i, d_{ij}) = \mathrm{MatP}(q_i, d_{ij}) \bigcup \mathrm{MatC}(q_i, d_{ij}). \tag{5}$$

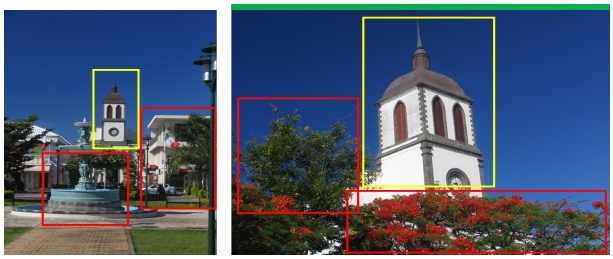
We assume that better matched images are likely to have bigger sets of $\mathrm{MatRels}(\cdot,\cdot)$ for the query region q_i. The relationship score $\mathrm{RelS}(q_i, d_{ij})$ is finally defined based on the matched related regions:

$$\mathrm{RelS}(q_i, d_{ij}) = 1 + |\mathrm{MatRels}(q_i, d_{ij})|, \tag{6}$$

where 1 represents the matching of q_i and d_{ij} themselves, and the cardinality of $\mathrm{MatRels}(q_i, d_{ij})$ indicates the number of matching within their related regions.

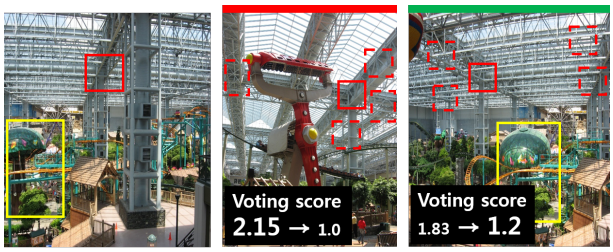
Fig. 4 shows visually how our rank-based voting method achieves high matching quality. Fig. 4(a) shows matched regions in yellow boxes and unmatched regions in red boxes between a query and its ground-truth test image. In case of using cross-matching [22] or NBNN [31] for the similarity measurement between multiple regions, all the regions in two images contribute to their similarity. In this approach, dissimilar regions are also matched and lower down the quality of measuring the similarity between two images. On the other hand, our method measures the similarity only with matched regions by utilizing the ranking set of $\mathcal{D}_i$, which is top-v close regions to the query region q_i.

Fig. 4(b) points out a problem caused by frequently appearing objects. Objects in the red boxes are visually very similar, so they can have a much higher similarity score than other matches, overwhelming important matches between the object shown in the yellow box. As a simple, yet effective way to normalize the impact of each region in the query, we use a ranking score instead of the direct *L2* similarity for the final search



(a) Ignore dissimilar regions (red boxes) with voting

(b) Deemphasize frequently appearing objects, resulting in short L2 distances (red boxes) by the rank



(c) Handle burstiness of local regions by having a maximum single vote; dotted boxes are not considered, so the impact of other matches relatively increases (yellow boxes). As a result, the voting score, 1.2, for the positive true is higher than that, 1.0, of the positive false.

Fig. 4 Left images are queries, and right images are test images. Green and red bars on top of test images indicate correct and incorrect matched results, respectively. Our rank-based voting method handles these issues commonly raised by using local features. Yellow boxes represent similar regions, while red ones indicate less important matches.

result as shown in Eq. 3. For frequently appearing objects, there are many images containing such objects and thus a ranking of a test image can be very low, as demonstrated in Fig. 4(b); note that this is a similar concept to the term of inverse document frequency (IDF), a common technique deemphasizing frequently appearing words in the text search.

Fig. 4(c) demonstrates the burstiness problem, which can occur for local image descriptors [12] including our approach. Our method, however, naturally prevents the burstiness, since each region in our method votes only once for a test image, by utilizing the maximum score as shown in Eq. 2. Finally, the RelS$(\cdot,\cdot)$ term considered spatial relationship among regions and improve the matching confidence (Fig. 1). Effects of this term is elaborated in Sec. 4.4.

3.4 Scalable encoding and indexing

Our method aims to achieve high search accuracy by representing an image with multiple regions. As a result, it may require a large amount of memory for encoding all the images for large image datasets. In order to improve both memory and computational efficiencies, we adopt Principle Component Analysis (PCA) with whitening [10] and Product Quantization (PQ) [13] for our features.

PCA is well-known method for dimensionality reduction, and Jégou *et al.* [10] showed that it could even improve the quality of features by applying whitening. We also apply PCA and whitening to the features first, and convert them into compact codes using PQ. In the search time, our method first retrieves a shortlist with a size M using the compact codes, and then performs our voting method on the shortlist using the PCA features. Details of our implementation can be found in the supp. report.

We also utilize IVFADC [13] to retrieve a shortlist for our voting method. IVFADC is built upon the inverted index defined with a coarse quantizer such as K-means clustering. Each feature is indexed based on the inverted index and stored as a compact code compressed by PQ. At the searching stage, we first collect nearest-neighbor candidates by accessing the inverted index, and re-rank them according to the estimated distance between a query feature and a compact code. Since we perform the search by using the inverted index and accessing compact codes before computing the shortlist, it drastically improves the memory footprint required and access time. More details on the IVFADC and shortlist computation can be found in [13,9].

Table 1 Comparison of search accuracy w/ and w/o using $RelS(\cdot,\cdot)$. $RelS(\cdot,\cdot)$ improves the accuracy significantly for difficult queries that produce frequent mismatches w/o using the term in *Oxford 5k*. Considering the term increases 0.137 AP on average over these four queries.

queries	w/ $RelS(\cdot,\cdot)$	w/o $RelS(\cdot,\cdot)$
christ church 5	**0.461**	0.304
cornmarket 2	**0.782**	0.641
balliol 1	**0.758**	0.625
magdalen 3	**0.506**	0.387

Table 2 Comparison of the average recalls of our approach w/ local features extracted from regions against the single global feature extracted from a whole image.

Feature types	*Holidays* recall@20	*Oxford 5k* recall@200	*UKB* recall@10
Single global	0.831	0.618	0.931
Multiple regional	**0.977**	**0.865**	**0.990**

4 Experiment

We now present various experimental results on three standard benchmarks. We also compare the search accuracy of our image search algorithms against the state-of-the-art techniques using CNN-based representations.

4.1 Datasets

Our experiments are performed on the following three standard benchmark datasets:

Holidays dataset [11]. This dataset consists of 500 groups of photographs, and each group represents the same scene or object taken in vacation. The total number of photographs is 1,491, and the number of queries is 500. The performance is measured by the mean average precision (mAP) over the queries. For a fair comparison, we also use the manually rotated version of the dataset as adopted by previous works [2,1,16,6].

UKB dataset [20]. This dataset is composed of 10,200 photographs for 2,550 indoor objects, so there are four images taken from different viewpoints for each object. Each image is used as a query, and the performance is measured by the average number of retrieved images representing the same objects within the top-4 over all the queries.

Oxford 5k dataset [21]. The dataset consists of 5,062 Oxford landmark images, collected from Flickr. The dataset has 11 query landmarks, each of which has 5 query images. The performance is evaluated with mAP over the queries. For each query, the dataset provides a single bounding box containing the exact landmark for each query; note that the bounding box is provided only for queries, not for test images. Since the bounding box provides accurate region-of-interest, it is desirable

Table 3 Comparison of the accuracy between a general image search scheme and our proposed method w/ different image descriptors. Across all different cases and benchmarks, our method improves the accuracy over a general image search using the single global feature and $L2$ distance metric.

Holidays (mAP)

search scheme	Caffe	CaffePCA	VGG	VGGPCA
Single-$L2$	0.691	0.743	0.642	0.697
Ours	0.879	0.907	0.884	**0.917**

Oxford 5k (mAP)

search scheme	Caffe	CaffePCA	VGG	VGGPCA
Single-$L2$	0.302	0.356	0.319	0.415
Ours	0.678	0.735	0.661	**0.768**

UKB (recall@4)

search scheme	Caffe	CaffePCA	VGG	VGGPCA
Single-$L2$	0.832	0.877	0.806	0.866
Ours	0.943	0.957	0.952	**0.970**

to use this additional information, and thus we utilize the bounding box as one of our region proposals. We also use the well-known *Oxford 105k*, which add additional 100k distractor images to *Oxford 5k*. We use the benchmark for scalability tests.

Holidays consists of outdoor scene images, *Oxford 5k* contains building images, and *UKB* consists of indoor object images. Thanks to the distinct characteristics of the benchmarks, we can validate not only effectiveness, but also robustness of our approach.

We perform various memory reduction and quantization methods for our method, resulting in two orders of magnitude of memory reduction compared to uncompressed features. On average, our method takes 0.7 s for searching images given a query image on *Oxford 105k*. Details on these methods are available at the supp. report.

4.2 Benefit of using multiple regional descriptors

Before showing the effect of our similarity measurement, we first demonstrate the benefit achieved mainly from our image representation. For this, we compare recalls of two different types of features: the global feature extracted from a whole image and our local feature obtained from regions, under a simple similarity measure, the $L2$ distance.

As shown in Table 2, our method using local features from the regions provides higher recalls than the method using a single global feature across all the benchmark datasets. Our multiple features capture the details of an image effectively that can be missed in the global feature, and this is the main reason for improved recalls.

4.3 Effects of the proposed method

We now demonstrate the effectiveness of our proposed method using the multiple features and the rank-based voting method as the similarity measurement. We compare its accuracy against that of a general image search scheme that uses the single global feature and L2 distance metric. To test robustness of our method, we perform experiments with four different image descriptors, which are features from CaffeNet and VGG-19 without any post-processing and ones with PCA-whitening.

Table 3 shows the accuracy of two different methods combined with each descriptor. In spite of the diverse conditions, our approach consistently improves the accuracy for all the cases over the general image search method. Based on the high recall from our image representation, our rank-based voting effectively matches and scores regional features and returns accurate search results. On average, our method improves the accuracy relatively by 29% on *Holidays*, 13% on *UKB*, and surprisingly, 105% on *Oxford 5k*. Thanks to the small visual variances of buildings and frequent viewpoint changes in *Oxford 5k*, multiple local features work much better than a single global feature on this dataset.

4.4 Comparison with the state-of-the-arts

So far, we discussed benefits of our image representation and similarity measure methods. We now compare the search accuracy of our approach with the state-of-the-art image search methods that also utilize CNNs. Before comparing the accuracy with other methods, we categorize the methods according to their main objectives. The first category, the improved descriptors category, includes approaches that propose compact descriptors with improved accuracies [2, 1, 16, 30, 6]. The second category, improved similarity measurements category, consists of approaches proposing novel similarity measurements for utilizing multiple regional features [22, 31].

The main objective of the first category is improving discriminative power of a descriptor, while maintaining the memory efficiency. On the other hand, the second category focuses on improving search results by effectively utilizing larger amount of information from images. Since it is not reasonable to compare the accuracy directly between the memory-efficient methods and accuracy-focused methods, we show results of these two categories separately in Table 4.

In Table 4, we present results without any post-processing, *i.e.*, query expansion, re-ranking, or spatial verification, for fair comparison. We also show the results of our approach using both CaffeNet [15] and VGG-19 [26], to clearly compare the performance with

Fig. 5 Comparison of qualitative results between SPoC and ours. The leftmost images are queries, and right images are the results in order. In each row, the top images represent the result of SPoC, and the bottom images show the result of our method. We mark the correctness of a result image with a color bar above the image. Green and red bars indicate true positives and false positives, respectively. Blue bar indicates "junk" images of the $Oxford5k$ dataset, which contain the correct building to the query but with high occlusion or distortion. We also mark the regions with a high voting score on our result images. Our method works well for the queries where SPoC finds correct images well (top row). Moreover, our method shows good performance even for the queries where SPoC cannot retrieve correct images well (middle and bottom rows).

Table 4 Comparison of accuracy of the state-of-the-art search methods. Methods in the improved descriptors category generate memory-efficient compact features, and methods in the improved similarity measurement category propose new similarity measures utilizing multiple regional features. Results marked with an * are achieved with retrained or fine-tuned networks.

Methods	*Holidays* mAP	*Oxford 5k* mAP	*UKB* recall@4
Improved descriptors			
Neural code [2]	0.793*	0.557*	0.890*
SPoC [1]	0.784	0.657	0.915
CroW [16]	0.851	0.708	-
R-MAC [30]	-	0.669	-
DeepIR [6]	**0.891***	**0.831***	-
Improved similarity measurements			
Off-the-shelf [22]	0.843	0.680	0.911
ONE [31]	0.887	-	0.968
Ours with CaffeNet	0.907	0.735	0.957
Ours with VGG-19	**0.917**	**0.768**	**0.970**

others. Among those methods, neural code and off-the-shelf used the networks with the CaffeNet structure, and other methods used the VGG structure.

Methods in the improved descriptors category show great performance with a single global feature. Neural code [2] improves the descriptor for each benchmark by retraining the networks. SPoC [1], CroW [16], and R-MAC [30] enhance descriptors with their own aggregation methods. They implicitly embed spatial information into the descriptors with centering prior, spatial weighting, and max-pooling, respectively. While other methods use the networks originally trained for image classification, DeepIR [6] trained the networks based on image similarities. With this search-specialized network, it shows the best performance among the methods in the first category, and even shows better performance than the methods in the second category for some cases. Since these approaches propose single compact descriptors, all of them are highly memory-efficient.

As the second category utilizes additional information from images, techniques in the category can show better accuracy than that of the first category on the condition of using equally trained networks. Fig. 5 shows qualitative comparisons between ours and SPoC. In the second category, Off-the-shelf [22] uses an average $L2$

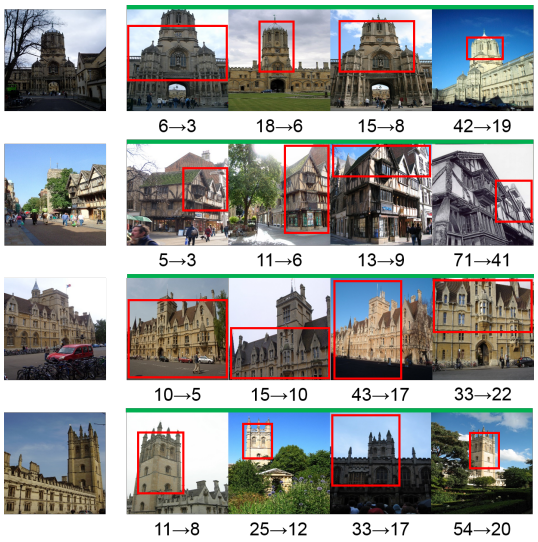
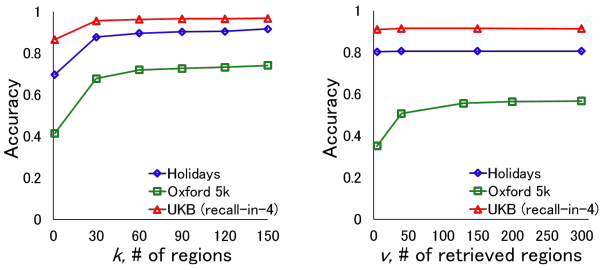


Fig. 6 This figure shows that using our relation term $RelS(\cdot,\cdot)$ improves ranking of true positive images, resulting in a higher accuracy.

distance, and ONE [31] utilizes NBNN for similarity measure. Comparing the results with the same network structures, our rank-based voting achieves better accuracy than others across all the benchmarks. Also, compared to the ONE method using 220 regions per image, our method achieves higher accuracy with 68% of the regions, i.e., 150 regions per image. Furthermore, our method does not require any additional datasets during the search process, while ONE requires additional categorized feature sets for the similarity measure due to the nature of NBNN.

Discussions. While we compared different methods in two different categories, they can be combined together, since their goals, improving image descriptors and similarity measurements, are complementing each other. As shown in Table 3, our method consistently improves the accuracy over various descriptors by using them as multiple regional features instead of single global features. Also, DeepIR [6] leverages the network trained with triplet ranking loss. Since the ranking loss is evaluated based on image similarity ranks, this network brings improvement to image search than the general networks trained with the classification loss. While our method currently utilizes the networks trained with the classification loss, it can be further improved by leveraging the network trained with the ranking loss. Nonetheless, our method even with the classification loss shows higher accuracy than DeepIR in two of three tested benchmarks.

Effects of relation term. We analyze the effect of relation term in our voting method. In order to study the effect attributed by the relation term $RelS(\cdot,\cdot)$, we



(a) Accuracy according to k (b) Accuracy according to v

Fig. 7 (a) shows the accuracy according to k, the number of representative regions in each image. Accuracy increases as k increases, but its rate becomes smaller as k becomes larger. (b) shows the accuracy as a function of v, the number of retrieved regions from each query region. We also observe the similar trend to that of k.

compare search accuracy w/ and w/o using $RelS(\cdot,\cdot)$ (Table. 1). We found that using the term improves accuracy, especially for difficult cases, where incorrect matches occur frequently. Since it gives higher weights to more confident matches, impacts of incorrect matches relatively decrease. Therefore, it improves the final ranking of correct images appropriately as shown in Fig. 6.

4.5 Scene completion as an application

We apply our image search method to the application of scene completion that utilizes similar images identified from an image database [7]. Although we use a small size of the dataset, our search method effectively finds visually similar images to an input query. Fig. 2 shows two example images of the application based on our image search method.

4.6 Analysis

Our similarity measurement method uses two parameters k and v, which are the number of representative regions in an image and the number of retrieved regions from a query region, respectively.

Fig. 7 shows retrieval accuracy as a function of k and v, individually. Interestingly, all data represent a similar tendency, where accuracies rapidly increase in early phases and slowly keep increasing until they reach the stable points. As the results indicate non-decreasing shapes, our method can show its best performance by simply assigning large values to k and v. Nonetheless, larger k and v need more time and memory resources. As a result, we set $k = 150$ and $v = 250$ for our experiments as a practical balance between the time and memory requirements.

5 Conclusion and Future Work

We have introduced a novel image search method, which uses representative regions with an associated similarity measure, to improve the search accuracy. Our similarity measure incorporates the spatial relation of regions and improves the confidence of the matching. We observed that the proposed algorithm achieves the best accuracy among the methods using equally trained networks.

There are many interesting directions to further improve the performance for image search. Our method showed accuracy improvement by adopting the state-of-the-art region proposal and pre-trained CNN features. Fortunately, our method is well modularized, so there is a high chance to be combined with new cutting-edge object localization methods or new discriminative features. Second, our method achieved higher accuracy by utilizing the inclusion relationship of regions. Embedding a high-level spatial relation into the similarity measure can further increase the search accuracy.

Acknowledgements

This work was supported in part by MSIP/IITP R0126-16-1108, MI/KEIT 10070171, and DAPA/DITC (UC160003D).

References

1. Babenko, A., Lempitsky, V.: Aggregating local deep features for image retrieval. In: CVPR, pp. 1269–1277 (2015)
2. Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.: Neural codes for image retrieval. In: Computer Vision–ECCV 2014, pp. 584–599. Springer (2014)
3. Bell, S., Zitnick, C.L., Bala, K., Girshick, R.: Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. Computer Vision and Pattern Recognition (CVPR) (2016)
4. Cheng, M.M., Zhang, Z., Lin, W.Y., Torr, P.: Bing: Binarized normed gradients for objectness estimation at 300fps. In: CVPR (2014)
5. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE conference on computer vision and pattern recognition (2014)
6. Gordo, A., Almazan, J., Revaud, J., Larlus, D.: Deep image retrieval: Learning global representations for image search. arXiv preprint arXiv:1604.01325 (2016)
7. Hays, J., Efros, A.A.: Scene completion using millions of photographs. ACM Transactions on Graphics (SIGGRAPH 2007) **26**(3) (2007)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
9. Heo, J.P., Lin, Z., Shen, X., Brandt, J., eui Yoon, S.: Shortlist selection with residual-aware distance estimator for k-nearest neighbor search. In: CVPR (2016)
10. Jégou, H., Chum, O.: Negative evidences and cooccurences in image retrieval: The benefit of pca and whitening. In: ECCV (2012)
11. Jégou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: European Conference on Computer Vision (2008)
12. Jégou, H., Douze, M., Schmid, C.: On the burstiness of visual elements. In: CVPR (2009)
13. Jegou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. IEEE transactions on pattern analysis and machine intelligence **33**(1), 117–128 (2011)
14. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: CVPR (2010)
15. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
16. Kalantidis, Y., Mellina, C., Osindero, S.: Cross-dimensional weighting for aggregated deep convolutional features. arXiv preprint arXiv:1512.04065 (2015)
17. Kemelmacher-Shlizerman, I.: Transfiguring portraits. ACM Trans. Graph. **35**(4) (2016)
18. Krähenbühl, P., Koltun, V.: Geodesic object proposals. In: Computer Vision–ECCV 2014, pp. 725–739. Springer (2014)
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105 (2012)
20. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: CVPR
21. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2007)
22. Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: an astounding baseline for recognition. In: Computer Vision and Pattern Recognition Workshops (2014)
23. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems 28, pp. 91–99 (2015)
24. Samii, A., Měch, R., Lin, Z.: Data-driven automatic cropping using semantic composition search. Comput. Graph. Forum **34**(1), 141–151 (2015)
25. Van de Sande, K.E., Uijlings, J.R., Gevers, T., Smeulders, A.W.: Segmentation as selective search for object recognition. In: ICCV. IEEE (2011)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)
27. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: ICCV (2003)
28. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: Exploring photo collections in 3d. In: SIGGRAPH Conference Proceedings, pp. 835–846 (2006)
29. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR (2015)
30. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of cnn activations. arXiv preprint arXiv:1511.05879 (2015)
31. Xie, L., Hong, R., Zhang, B., Tian, Q.: Image classification and retrieval are one. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, pp. 3–10. ACM (2015)