Multimedia Tools and Applications (2019) 78:13247–13261
https://doi.org/10.1007/s11042-018-6427-1

Instance-level object retrieval via deep region CNN

Shuhuan Mei[1,2] · Weiqing Min[2] · Hua Duan[1] · Shuqiang Jiang[2,3]

Received: 31 July 2017 / Revised: 6 February 2018 / Accepted: 20 July 2018 /
Published online: 13 September 2018
© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract

Instance retrieval is a fundamental problem in the multimedia field for its various applications. Since the relevancy is defined at the instance level, it is more challenging comparing to traditional image retrieval methods. Recent advances show that Convolutional Neural Networks (CNNs) offer an attractive method for image feature representations. However, the CNN method extracts features from the whole image, thus the extracted features contain a large amount of background noisy information, leading to poor retrieval performance. To solve the problem, this paper proposed a deep region CNN method with object detection for instance-level object retrieval, which has two phases, i.e., offline Faster R-CNN training and online instance retrieval. First, we train a Faster R-CNN model to better locate the region of the objects. Second, we extract the CNN features from the detected object image region and then retrieve relevant images based on the visual similarity of these features. Furthermore, we utilized three different strategies for feature fusing based on the detected object region candidates from Faster R-CNN. We conduct the experiment on a large dataset: INSTRE with 23,070 object images and additional one million distractor images. Qualitative and quantitative evaluation results have demonstrated the advantage of our proposed method. In addition, we conducted extensive experiments on the Oxford dataset and the experimental results further validated the effectiveness of our proposed method.

Keywords Faster R-CNN · Deep learning · Instance-level object retrieval · Instre

✉ Hua Duan
huaduan59@163.com

Shuhuan Mei
shuhuan.mei@vipl.ict.ac.cn

Weiqing Min
weiqing.mei@vipl.ict.ac.cn

Shuqiang Jiang
sqjiang@ict.ac.cn

[1] College of Mathematics and Systems Science, Shandong University of Science and Technology, Qingdao, 266590, China

[2] Key Lab of Intelligent Information Processing, Institute of Computing Technology, CAS, Beijing, 100190, China

[3] University of Chinese Academy of Sciences, Beijing, 100049, China

1 Introduction

Object retrieval is an important task in computer vision. In recent years, researchers have achieved great success in the field of object retrieval. For example, Jegou et al. [18] combined hamming embedding and weak geometric consistency to extract robust visual features for object retrieval. Albert et al. [13] extracted a global and compact fixed-length representation for each image and aggregats many region-wise descriptors for higher retrieval performance. Based on the level of the target object class, object retrieval can be roughly divided into two groups: The first group is category-level object retrieval [35], where an image in the dataset is deemed to be similar to the query image if they share the same object class. The other group is the instance-level object retrieval [29], where an image is considered to match the query if they contain the same object. The instance-level object retrieval is harder that the retrieval methods need to encode the local and detailed information in order to tell two images apart, e.g., the algorithm should be able to detect the differences between the Coca-Cola bottle and Mirinda bottle although they have similar shapes. In this paper, we focus on the instance-level object retrieval.

There are some existing methods for image retrieval at the instance level, such as SIFT based work [24, 37]. Recent advances [3, 29, 42] show that Convolutional Neural Networks (CNN) offer an attractive alternative for image search representations with small memory. Their success is mainly due to the computational power of GPUs and the use of very large annotated datasets [32]. Using the CNN layer activations as off-the-shelf image descriptors [7, 34] appears very effective and is adopted in many tasks [11, 12]. For example, Babenko et al. [3] proposed the use of Fully Connected (FC) layer activations as descriptors, while convolutional layer activations are later shown to have superior performance [2, 21, 39]. However, for the CNN-based method, these features are extracted from the global image, the feature contains a large amount of background information,which can affect the performance of object retrieval.

In order to solve this problem, we can detect the area of the instance, and then extract the features from the region where the instance is located. For that solution, we proposed a deep region Faster R-CNN method for an instance-level object retrieval. As shown in Fig. 1, our provided instance-level object retrieval system mainly consists of two components: offline Faster R-CNN training and online instance retrieval. For the offline Faster R-CNN training,

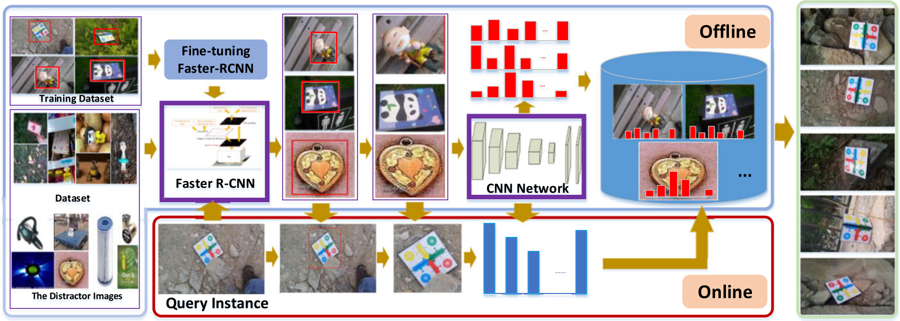


Fig. 1 The pipeline of our proposed instance-level object retrieval system

the instance retrieval system requires two basic components that allow the user to search for what they want to get the search results. First, we need to train the Faster R-CNN to better locate the objects we use for instance retrieval. Second, we extract the CNN features for the detected object image region. For the online instance retrieval, we use the trained Faster R-CNN to detect the region with the object, and then extract the CNN features from the detected regions. The retrieval result of the index picture is obtained based on the visual similarity of these features. Furthermore, we adopt and compare three different strategies including connection, mean-pooling and max-pooling to fuse the features of the detected regions from the instance images.

The contributions of the proposed approach can be summarized as follows:

- We proposed a deep region CNN method with object detection by combining Faster R-CNN and CNNs for instance-level object retrieval.
- We utilize and compare different strategies of feature fusion based on the detected object region candidates.
- We conducted the experiment on two object datasets and the experimental results have validated the effectiveness of our proposed method.

The remainder of the paper is structured as follows. Section 2 introduces the related work, Section 3 presents the methodology of this paper, including Faster R-CNN based object region extraction and CNN based Representation. Section 4 includes the performed experiments on both Instre and Oxford datasets as well as the comparison to other state of the art sift-based instance search systems and CNN-based search systems. Finally, Section 5 draws the conclusions of this work.

2 Related Work

Our work is closely related to the following three research areas: (1) Convolutional Neural Networks (CNNs), (2) object detection CNNs and (3) instance-level object retrieval.

2.1 Convolutional neural networks(CNNs)

Recently, CNNs are recognized as a mainstream approach and can be successfully applied into several tasks, such as image classification [22, 38], object detection [11, 31] and image retrieval [28, 39]. Compared with traditional visual methods, CNNs can extract richer semantic information based on the deep learning architecture in a large number of labeled data. There is comparatively less work on CNN-based descriptors for instance retrieval compared to large-scale image classification. Razavian et al. [34] evaluated the performance of CNN model of [22] on a wide range of tasks including instance retrieval, and showed initial promising results. Babenko et al. [3] found that the CNN representations can be compressed more effectively than their sift based method for large-scale instance retrieval. Chandrasekhar et al. [4] proposed a hash method by changing the sparse high-dimensional CNN representation to very compact representations for large scale image retrieval. These work shares similarities with all the former in the usage of convolutional features of a pre-trained CNN. However, we choose to use a state-of-the-art object detection CNN to extract region-based convolutional features for instance retrieval.

2.2 Object detection CNNs

Many works have proposed CNN-based object detection pipelines. Girshick et al. presented R-CNN [11],where instead of full images, the regions of an object proposal algorithm were used as inputs to the network. At test time, fully connected layers for all windows were extracted and used to train a bounding box regressor and classifier. Since then, great improvements to R-CNN have been released, both in terms of accuracy and speed, such as SPP-net [15] and Fast R-CNN [10]. Ren et al. [3, 34] introduced Faster R-CNN [31], which removed the object proposal dependency of former object detection CNN systems by introducing a Region Proposal Network. Recently there have been some works on object detection, such as YOLO v2 [30] and SSD [23]. In this work, we take advantage of the end-to-end self-contained object detection architecture of Faster R-CNN to extract region features for more robust instance search. We further utilize and compare different feature fusing strategies on top-ranked detected regions.

2.3 Instance-level object retrieval

Image instance-level retrieval is the discovery of images from a database representing the same object or scene as the one depicted in a query image. State-of-the-art image instance retrieval pipelines consist of two major blocks: first, a subset of images similar to the query are retrieved from the database, next, geometric consistency checks are applied to select the relevant images from the subset with high precision. The first step is based on the comparison of global image descriptors: high-dimensional vectors with up to tens of thousands of dimensions representing the image content. Better global descriptors are key to improving retrieval performance and have been the objective of much recent interest from the multimedia research community with work on specific applications such as digital documents [8], mobile visual search [5, 41], distributed large scale search [17, 19] and compact descriptors for fast real-world applications [9, 20]. Some recent works [16, 25] further utilized attention based methods for image retrieval and achieved better retrieval results. Other works such as [14, 33] moved beyond instance-level retrieval and the goal is to retrieve images that share the same semantics as the query image. In this paper, we detect the region of the object on the image by Faster R-CNN, and then select the extracted region features. Compared to the global CNN based method, the features extracted by region-CNN lead to better performance.

3 Methodology

This paper explores instance-level object retrieval from images using image regions detected by an object detection CNN. The framework of our provided instance-level object retrieval system is shown in Fig. 1, and it has two phases, i.e., offline Faster R-CNN training phase and online instance retrieval phase. Next, we describe these two phases in details.

3.1 Offline faster R-CNN training

3.1.1 Fine-tuning Faster R-CNN

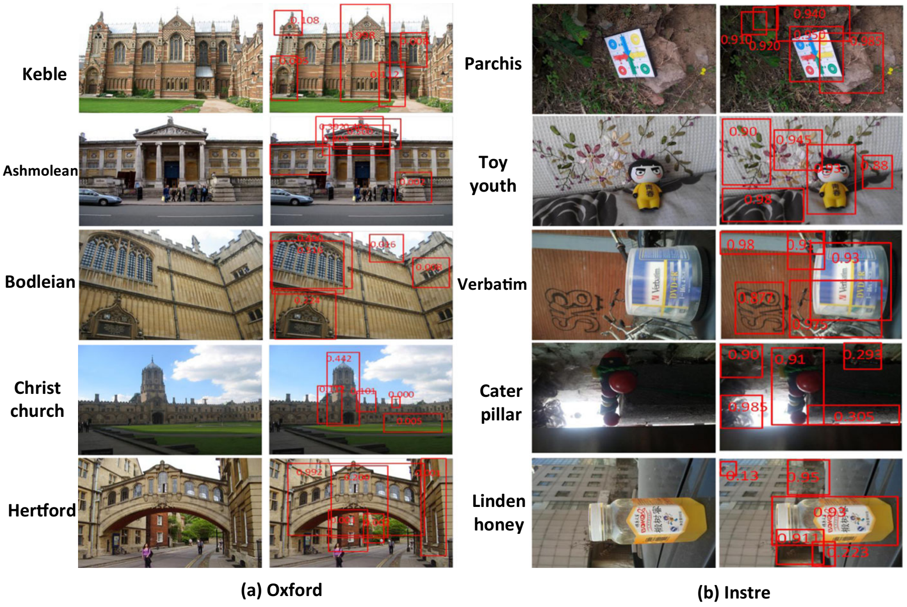
For the object dataset, it has more diverse intra-class instance variations, cluttered and complex background. In order to overcome these difficulties, we explore the suitability of Faster

R-CNN [31] to obtain better feature representation for better instance retrieval performance. Particularly, we choose the following fine-tuning pattern. The initial two convolution layers have the unchanged weights, and the weights of all subsequent layers are updated. By changing the convolutional features, RPN proposals and fully connected layers to make it more adaptable to the query instance. The resulting fine-tuned networks are to be used to extract better features that are conductive to retrieval. To train the RPN, we assign each anchor a binary tag (not the target). We assign a positive label to two types of anchor: (i) an anchor (perhaps less than 0.7) that overlaps with a ground truth (GT) bounding box with the highest IoU (Intersection-over-Union) (ii) an anchor that overlaps an IoU greater than 0.7 with any GT bounding box. Note that a GT bounding box may assign positive tags to multiple anchors. We assign a negative label to all GT bounding boxes with an IoU ratio of less than 0.3 of the anchor. Non-positive and negative anchor has no effect on training objectives. With these definitions, we follow the multitasking loss in Fast R-CNN to minimize the objective function. We define the loss function for an image as

$$L(p_i, t_i) = \frac{1}{N_{cts}} \sum_i L_c ks(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \qquad (1)$$

Here, i is the index of an anchor in a mini-batch and p_i is the predicted probability of anchor i being an object. The ground-truth label p_i^* is 1 if the anchor is positive, and is 0 if the anchor is negative. t_i is a vector representing the 4 parameterized coordinates of the predicted bounding box, and t_i^* is that of the ground-truth box associated with a positive anchor.The classification loss $L_c ls$ is log loss over two classes (object vs. not object).

After fine-tuning faster R-CNN, we utilize the fine-tuned faster R-CNN to get the bounding boxes of the images and the corresponding score for each bounding box. Figure 2 shows



(a) Oxford

(b) Instre

Fig. 2 The detected top-5 bounding boxes and their corresponding scores via Faster R-CNN for both datasets (**a**) Oxford and (**b**) Instre

some examples, including detected top-5 bounding boxes and their corresponding scores for both datasets (a) Oxford and (b) Instre.

Next, we then use the CNN model to extract the deep visual features from the detected object region.

3.1.2 CNN based feature representation

CNNs [22] are mainly used to identify two-dimensional images of displacement, scaling and other forms of twist invariance. Since CNN's feature detection layer is learned through the training data, the use of CNN avoids explicit feature extraction and implicitly learns features from the training data. Moreover, due to the neuron weights on the same feature map surface, the network can learn in parallel, which is a major advantage of CNNs.

As mentioned above, we focus on leveraging convolutional networks for feature representation. We adopt the well known architecture [22] as our basic framework. More specifically, we select the top-K bounding boxes with higher scores for each image in the database. According to its coordinates to take out this piece of image,we use the VGG-16 network trained on Imagenet [36] to extract 4096-D features from the seventh Fully-Connected layer (FC7).

3.2 Online instance retrieval

Given the query instance image, we extract the visual features based on the fine-tuned Faster R-CNN and CNN. The similarity is computed to return retrieved results. Generally, we selected the bounding box with the highest score for each image as the target region to extract CNN features for the instance retrieval. However, the detected region with the highest score diverges from the actual instance object region. As a result, this method may lose some useful information. Figure 2 shows some examples. For both the oldman and the parchis classes, their highest scoring region is not the object we need to retrieve, resulting in the reduction in the performance. By observing the highest three of the detect results, we found that the instance object will generally be in the top K score's region, so these corresponding regions of the features will be as much as possible to increase the correct information. Therefore, according to the selection of the number of bounding boxes, we adopt the following three strategies to fuse the features from top K's regions for each instance image, namely (1)concatenation, (2)mean-pooling and(3)max-pooling. For concatenation, we fuse features from different regions by simply connecting their corresponding 4096-D features.

Through the above-mentioned methods, we get the corresponding search results for each query instance image. By Faster R-CNN, we effectively reduce the affect from the background of the image. At the same time, through the CNN method, we get the discriminative semantic information of the detection areas. Finally, we preserve the information of the image as much as possible by selecting top-K bounding boxes, which reduces the effect due to poor detection results on some categories.

4 Experiments

In this section, we firstly describe the experimental setting including the dataset and implementation details. We then evaluate the performance of the proposed method qualitatively and quantitatively.

4.1 Datasets

We validate our method on two object datasets including Oxford105k and Instre.

Oxford105k: This dataset [27] consists of 5,062 Oxford landmark images and additional 100,000 images collected from Flickr. These 5,062 landmark images have been manually annotated to generate a comprehensive ground truth for 11 different landmarks, each represented by 5 possible queries. This gives a set of 55 queries over which an object retrieval system can be evaluated. The 100,000 images are disjoint with the 5,062 images and are used as distractors to test the retrieval performance when the dataset scales to larger size.

Instre: This dataset [40] consists of two subsets:Instre-S and Instre-M [40]. INSTRE-S contains 200 single-label classes and INSTRE-M is designed for multiple objects. In this work, we select Instre-S in our experiment. Instre-S dataset contains 23,070 images in total and each image is provided with object location annotations. In addition, there are additional one million distractor images crawled from Flickr, and are also used to test the retrieval performance when the dataset scales to larger size.

4.2 Experimental setup

For the Oxford dataset, only the query image is marked with the ground truth, and in order to compare with other CNN methods, we only use the 55 markup images for fine tuning. For the Instre dataset, we randomly select 75 images from each class to form the training set for fine-tuning Faster R-CNN. After Faster R-CNN fine-tuning, we used the trained Faster R-CNN to detect regions from the dataset, and then extract the features of detected regions with higher scores using CNNs. In our experiment, we choose to extract the 4096-D feature using the VGG-16 network. All the experiments were run in an Nvidia Titan X GPU. Similar to [40], we select Mean Average Precision (mAP) as the evaluation metric. As the mean of the average precision scores for each query, mAP has been proven to have especially good discrimination and stability.

4.3 Evaluation on Oxford105k

In order to compare our method with existing methods on this dataset, we consider the following baselines for comparison:

- CNN based method (CNN). We directly use the VGG16 network to extract the 4096-D features for all the images.
- Fine-tuning CNN method (F-CNN). In this baseline, we first fine-tune the VGG16 model, pretrained on the Imagenet dataset using the train dataset, which is the same for training Faster RCNN. In the case of Oxford,we modify the output layers in the network to return 11 class probabilities. After fine-tuning, we follow the steps described in the baseline CNN to extract the visual features from all the images.
- CNN+full cropped [45].
- VLAD-intra [1].

4.3.1 Experimental Results

The results are shown in Fig. 3. As our method, we use the trained Faster R-CNN to detect regions from the dataset, and then extract the 4096-D features of detected region with the highest score for each image. From these comparison results, we can see that (1) The performance of F-CNN is better than CNN. This is because fine-tuned CNN is suitable for

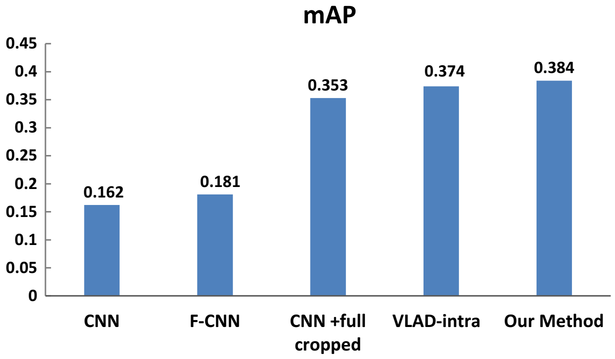


Fig. 3 The performance of all different methods for instance retrieval in Oxford dataset

the current task. (2) The performance of our method achieves the best performance. This is because our method can more accurately extract the information of the object, and thus improve the retrieval performance. Because the detected region with the highest score of the object from Faster R-CNN may not be accurate, adding the features from the regions with higher scores can enhance the performance of the retrieval results. In order to verify it, we show different fusion strategies based on different number of candidate regions from each image. The results are shown in the Fig. 4. We can see that (1) The performance of concatenation strategy is better than both mean-pooling and max-pooling method. The probable reason is that the concatenation method preserves more object information than these two strategies (2) These methods all achieves the best performance when K = 2. In this case, the concatenation strategy achieves the best mAP, that is 0.404. That means K = 2 achieves the balance between the correct object information and the background noise.

4.4 Evaluation on Instre

We consider the following baselines for comparison:

- Spatial Coding (SC) [43]. False SIFT matches can be removed by checking the composed 3-D spatial maps. We alleviate the sensitivity to image rotation through rotating the query image 20 times by 18 degrees to generate new queries for query expansion.

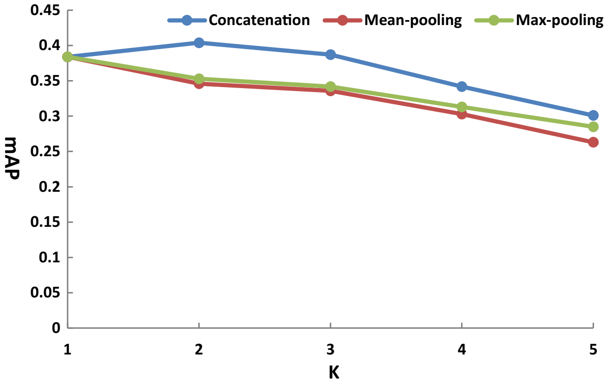


Fig. 4 The mAP scores over K candidate regions for different fusion strategies

- Geometric Coding (GC) [44]. It improves spatial coding in rotation invariancy.
- Combined-Orientation-Position Consistency (COP) [6]. COP employs a graph model to model the mutual spatial consistency of each two candidate SIFT matches.
- Hamming Embedding + Weak Geometric Consistency (HE+WGC) [18]. HE assigns each SIFT with a binary signature to encode the localization of it within the Voronoi cell and WGC exploits a Hough scheme to vote for quantized transformation. In our experiment, we use the signature length of 64 and Hamming threshold of 22. For CNN based experiment,we explore two method.One is Off-the-shelf CNN features method,the others is fine-tune method.
- CNN based method (CNN). In this section, we assess the performance using the VGG16 network for instance retrieval. We use the VGG16 network to extract the 4096-D features for all the images.
- Fine-tuning CNN method (F-CNN). In this baseline, we first fine-tune the pretrained network using the train dataset, which is the same with training Faster RCNN. Particularly, we choose to fine-tune the VGG16 model, pretrained on the Imagenet dataset. In the case of Instre,we modify the output layers in the network to return 200 class probabilities. After fine-tuning, we follow the steps described in the baseline CNN to extract the visual features from all the images.

4.4.1 Experimental results

The results are shown in Fig. 5. As our method, we used the trained Faster R-CNN to detect regions from the dataset, and then extract the 4096-D features of detected region with the highest score for each image. From these comparison results, some observations and analysis are included as follows:(1) The performance of COP and HE-WGC is better than CNN. The reason is that the background from the images in the Instre dataset is an important interference. Through the local feature extraction, the construction of visual dictionary, generation of original BOF features, introduction of TF-IDF weights HE-WGC can possibly extract features from the object regions and greatly reduced the interference from the background. (2) The performance of F-CNN is better than CNN. This is because the fine-tuned VGG16 model can more accurately extract the information of the instance object, and thus improve the retrieval performance. (3) The performance of our method achieves the best performance. Because the sift method constructs the vector by constructing the vector of the feature points, and then matches the vector so that the image must satisfy enough

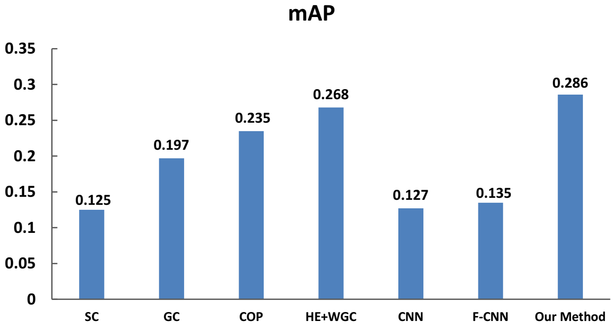


Fig. 5 The performance of all different methods for instance retrieval in Instre dataset

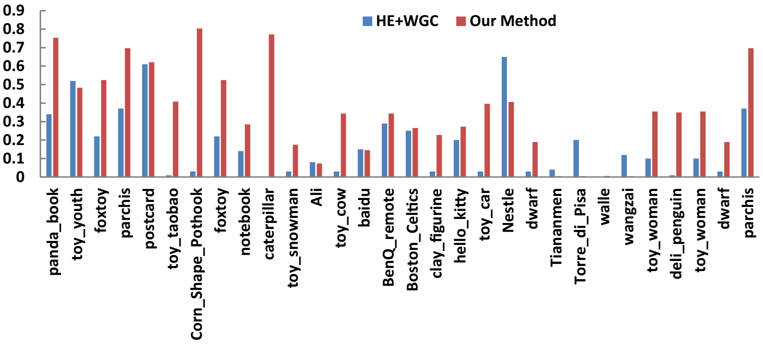


Fig. 6 The mAP performance on selected 30 object classes

texture, otherwise the constructed vector discriminant is not too large and can cause false matching. CNN extracts features from the entire target area, and there is no such limitation.-Considering that HE+WGC and our method achieve better performance than other baselines, we have a more detailed analysis of the experimental results from these two methods for different classes. Figure 6 further presents the mAP performance on randomly selected 30 object classes. On most classes, our method gives the best performance.

Because the detected region with the highest score of the object from Faster R-CNN may not accurate, adding the features from the regions with higher scores can enhance the performance of the retrieval results. In order to verify it, we show different fusion strategies based on different number of candidate regions from each image (Fig. 4). The results are shown in the Fig. 5. We can see that (1) The performance of concatenation strategy is better than both mean-pooling and max-pooling method. The probable reason is that the concatenation method preserves more object information than these two strategies (2) These methods all achieves the best performance when $K = 2$. That means $K = 2$ achieves the balance between the correct object information and the background noise.

Finally, we qualitatively evaluate the retrieval results from our methods when K = 1, 2 and 3 since they obtain better retrieval results than other baselines. Figure 7 shows some example results. As expected, results obtained with Top2 features achieve competitive results compared to Top1 and Top3, which suggests that connecting the features from the first 2 regions with higher scores is an effective solution.

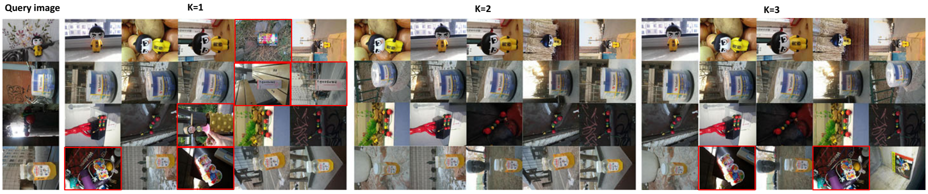


Fig. 7 Three image retrieval examples using our method when K=1, 2 and 3, respectively

5 Discussions

Considering the background noise of the image will hurt the performance of the instance-level object retrieval. In order to reduce the effect, we detected the object region and directly extracted the features from the object region for retrieval. However, we cannot guarantee that the region with the highest score is the object region. Therefore, we obtain the best performance by combining features from top-K regions results. However, there is a balance between the number of regions and the performance. In our experiment, when K>2, there is more noise included in the detected bounding box with the increase of K. Therefore, many results have the drop.

Another point to notice is that in our experiment,instead of using the ZF network from Faster R-CNN, we selected the VGG-16 network to extract visual features. The reason is that VGG-16 has more layers than the ZF network, the performance of VGG-16 network is generally better than the ZF network, and thus improved the performance of instance-level retrieval. We conducted the experiment. The accuracy of ZF network is 26.9% while our adopted VGG-16 is 28.6%. There is about 2 percent improvement compared with the ZF network.

In addition, we designed a method to combine the regional visual features and predicted class scores for object retrieval. The accuracy is 32.2% and there is about 1 percent improvement compared with our previous method. The experiment verified the effectiveness of introducing the predicted class information. Therefore, we can explore such class information to improve the performance of object retrieval in the future.

6 Conclusion

This paper has presented an instance-level object retrieval method using CNN features from an object detection CNN. It provides an effective strategy that uses fine-tuned Faster R-CNN features to describe images. We have shown that it has the capacity of improving the performance compared with traditional SIFT based method and CNN based global feature extract method. This work can be extended in the following three directions. The first direction is to use our existing framework for instance retrieval with multiple objects in one image. As the second direction, we plan to adjust existing solution for mobile instance retrieval. For example, Panda et al. [26] proposed a mobile instance retrieval method by reducing the visual index size. The third direction is to apply our method into different areas, such as instance-level food retrieval and clothes retrieval.

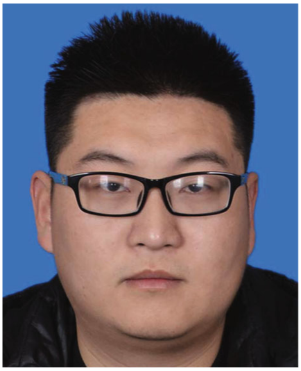
Acknowledgement This work was supported in part by the National Natural Science Foundation of China (61532018,61322212, 61602437, 61672497, 61472229 and 61202152), in part by the Beijing Municipal Commission of Science and Technology (D161100001816001),in part by Beijing Natural Science Foundation (4174106), in part by the Lenovo Outstanding Young Scientists Program, in part by National Program for Special Support of Eminent Professionals and National Program for Support of Top-notch Young Professionals, and in part by China Postdoctoral Science Foundation (2016M590135, 2017T100110). This work was also supported in part by Science and Technology Development Fund of Shandong Province of China (2016ZDJS02A11 and ZR2017MF027), the Taishan Scholar Climbing Program of Shandong Province, and SDUST Research Fund (2015TDJH102).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

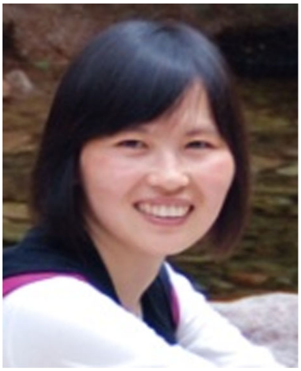
1. Arandjelovic R, Zisserman A (2013) All about VLAD. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1578–1585
2. Babenko A, Lempitsky V (2015) Aggregating local deep features for image retrieval. In: Proceedings of the IEEE international conference on computer vision, pp 1269–1277
3. Babenko A, Slesarev A, Chigorin A, Lempitsky V (2014) Neural codes for image retrieval. In: European conference on computer vision, pp 584–599. Springer, Berlin
4. Chandrasekhar V, Lin J, Morere O, Veillard A, Goh H (2015) Compact global descriptors for visual search. In: Data compression conference (DCC), 2015, pp 333–342. IEEE
5. Chen DM, Girod B (2015) A hybrid mobile visual search system with compact global signatures. IEEE Transactions on Multimedia 17(7):1019–1030
6. Chu L, Jiang S, Wang S, Zhang Y, Huang Q (2013) Robust spatial consistency graph model for partial duplicate image retrieval. IEEE Transactions on Multimedia 15(8):1982–1996
7. Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T (2014) Decaf: a deep convolutional activation feature for generic visual recognition Icml, vol 32, pp 647–655
8. Duan LY, Ji R, Chen Z, Huang T, Gao W (2014) Towards mobile document image retrieval for digital library. IEEE Transactions on Multimedia 16(2):346–359
9. Duan LY, Lin J, Wang Z, Huang T, Gao W (2015) Weighted component hashing of binary aggregated descriptors for fast visual search. IEEE Transactions on multimedia 17(6):828–842
10. Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448
11. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 580–587
12. Gong Y, Wang L, Guo R, Lazebnik S (2014) Multi-scale orderless pooling of deep convolutional activation features. In: European conference on computer vision, pp 392–407. Springer, Berlin
13. Gordo A, Almazán J, Revaud J, Larlus D (2016) Deep image retrieval: Learning global representations for image search. In: European conference on computer vision, pp 241–257. Springer, Berlin
14. Gordo A, Larlus D (2017) Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In: IEEE Conference on computer vision and pattern recognition (CVPR)
15. He K, Zhang X, Ren S, Sun J (2014) Spatial pyramid pooling in deep convolutional networks for visual recognition. In: European conference on computer vision, pp 346–361. Springer, Berlin
16. Hoang T, Do TT, Le Tan DK, Cheung NM (2017) Selective deep convolutional features for image retrieval. In: Proceedings of the 2017 ACM on Multimedia Conference, pp 1600–1608
17. Hong R, Li L, Cai J, Tao D, Wang M, Tian Q (2017) Coherent semantic-visual indexing for large-scale image retrieval in the cloud. IEEE Trans Image Process 26(9):4128–4138
18. Jegou H, Douze M, Schmid C (2008) Hamming embedding and weak geometric consistency for large scale image search. Computer Vision–ECCV 2008:304–317
19. Ji R, Duan LY, Chen J, Xie L, Yao H, Gao W (2013) Learning to distribute vocabulary indexing for scalable visual search. IEEE Transactions on Multimedia 15(1):153–166
20. Jiang YG, Wang J, Xue X, Chang SF (2013) Query-adaptive image search with hash codes. IEEE transactions on Multimedia 15(2):442–453
21. Kalantidis Y, Mellina C, Osindero S (2016) Cross-dimensional weighting for aggregated deep convolutional features. In: European conference on computer vision, pp 685–701. Springer, Berlin
22. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
23. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) Ssd: Single shot multibox detector. In: European conference on computer vision, pp 21–37. Springer, Berlin
24. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
25. Noh H, Araujo A, Sim J, Han B (2016) Image retrieval with deep local features and attention-based keypoints. arXiv:1612.06321
26. Panda J, Brown MS, Jawahar CV (2013) Offline mobile instance retrieval with a small memory footprint, pp 1257–1264
27. Philbin J, Chum O, Isard M, Sivic J, Zisserman A (2007) Object retrieval with large vocabularies and fast spatial matching. In: IEEE conference on computer vision and pattern recognition, 2007, pp 1–8
28. Radenović F, Tolias G, Chum O (2016) Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In: European conference on computer vision, pp 3–20. Springer, Berlin

29. Razavian AS, Sullivan J, Carlsson S, Maki A (2014) Visual instance retrieval with deep convolutional networks. arXiv:1412.6574
30. Redmon J, Farhadi A (2016)
31. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp 91–99
32. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252
33. Sang J, Xu C, Liu J (2012) User-aware image tag refinement via ternary semantic analysis. IEEE Transactions on Multimedia 14(3):883–895
34. Sharif Razavian A, Azizpour H, Sullivan J, Carlsson S (2014) Cnn features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 806–813
35. Sharma G, Schiele B (2015)
36. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
37. Sivic J, Zisserman A et al (2003) Video google: a text retrieval approach to object matching in videos. In: Iccv, vol 2, pp 1470–1477
38. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
39. Tolias G, Sicre R, Jégou H (2015) Particular object retrieval with integral max-pooling of cnn activations. arXiv:1511.05879
40. Wang S, Jiang S (2015) Instre: a new benchmark for instance-level object retrieval and recognition. ACM Transactions on Multimedia Computing Communications, and Applications (TOMM) 11(3):37
41. Xie Y, Jiang S, Huang Q (2013) Weighted visual vocabulary to balance the descriptive ability on general dataset. Neurocomputing 119:478–488
42. Zheng L, Yang Y, Tian Q (2016) Sift meets cnn: a decade survey of instance retrieval. arXiv:1608.01807
43. Zhou W, Lu Y, Li H, Song Y, Tian Q (2010) Spatial coding for large scale partial-duplicate web image search. In: Proceedings of the 18th ACM international conference on Multimedia, pp 511–520. ACM
44. Zhou W, Li H, Lu Y, Tian Q (2013) Sift match verification by geometric coding for large-scale partial-duplicate web image search. ACM Transactions on Multimedia Computing Communications, and Applications (TOMM) 9(1):4
45. Zisserman A (2014) Triangulation embedding and democratic aggregation for image search. In: Computer vision and pattern recognition, pp 3310–3317



Shuhuan Mei received the B.E. degree from Shandong University of Science and Technology, Qingdao, China, in 2015 and is pursuing the M.E. degree in Shandong University of Science and Technology, Qingdao, China. His current research interests include multimedia retrieval and applications.

Weiqing Min received the B.E. degree from Shandong Normal University, Jinan, China, in 2008 and M.E. degree from Wuhan University, Wuhan, China, in 2010, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, in 2015, respectively. He is currently an Assistant Professor with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences. His current research interests include location based multimedia analysis, understanding and applications. He has co-authored 10+ academic papers in prestigious international conference and journals, including ACM Multimedia, IEEE Trans. on Multimedia, IEEE Multimedia Magazine, etc. He is the reviewer of some international journals including IEEE Trans. on Cybernetics, IEEE Multimedia Magazine, Nerocomputing, Multimedia tools and application, etc. He is the recipient of the Best Paper in ACM TOMM 2016 and the Best Paper in IEEE Multimedia Magazine 2017.



Hua Duan is an associate professor at Shandong University of Science and Technology. She obtained her PhD in applied mathematics from Shanghai Jiaotong University in 2008. Her research interests are in the areas of Petri nets, Process Mining, and Machine Learning.

Shuqiang Jiang (IEEE SM'08) is a professor with the Institute of Computing Technology, Chinese Academy of Sciences(CAS), Beijing and a professor in University of CAS. He is also with the Key Laboratory of Intelligent Information Processing, CAS. His research interests include multimedia processing and semantic understanding, pattern recognition, and computer vision. He has authored or coauthored more than 100 papers on the related research topics. He was supported by the New-Star program of Science and Technology of Beijing Metropolis in 2008, NSFC Excellent Young Scientists Fund in 2013, Young top-notch talent of Ten Thousand Talent Program in 2014. He won the Lu Jiaxi Young Talent Award from Chinese Academy of Sciences in 2012, and the CCF Award of Science and Technology in 2012. He is the senior member of IEEE and CCF, member of ACM, Associate Editor of IEEE Multimedia, Multimedia Tools and Applications. He is the general secretary of IEEE CASS Beijing Chapter, vice chair of ACM SIGMM China chapter. He is the general chair of ICIMCS 2015, program chair of ICIMCS2010, special session chair of PCM2008, ICIMCS2012, area chair of PCIVT2011, publicity chair of PCM2011, web chair of ISCAS2013, and proceedings chair of MMSP2011. He has also served as a TPC member for more than 20 well-known conferences, including ACM Multimedia, CVPR, ICCV, ICME, ICIP, and PCM.