Retrieving Objects by Partitioning

Zhiyong Chen, Wei Zhang, Bin Hu*, Xiaochun Cao, Si Liu, and Dan Meng

Abstract—Retrieving objects from large image collection is challenging due to the so-called background-interference, i.e., matching between query object and reference images is usually confused by cluttered background, especially when objects are small. In this paper, we propose an object retrieval technique addressing this problem by partitioning the images. Specifically, several object proposals (hypothesis)are partitioned from the images by jointly optimizing their objectness and coverage. The proposal set with maximum objectness score and minimum redundancy is obtained. Therefore,the interference of cluttered background is greatly reduced. Next, the object retrieval are based on the partitioned objects, separately and independently to the background. Our method is featured by the fine partitioning, which not only removes interferences from background, but also significantly reduces the number of objects to index. In this way, the effectiveness and efficiency are both achieved, which ensures method's utility on big data. Subsequently, feature coding on partitioned objects generates much meaningful representation, and object level connectivity also introduces novel clues into the reranking through Random Walk. Extensive experiments on three popular object retrieval benchmark datasets (Oxford Buildings, Paris, Holiday) show the effectiveness of our method on retrieving objects.

Index Terms—Object Retrieval; Object Proposal; Partitioning; Reranking.

1 Introduction

Visual-based retrieval plays an increasing important role in the area of big data, which is to search a particular object from a large-scale image or video collection for the querying object. Different from similar image retrieval, object retrieval focuses on retrieving sub-image level object, which usually appears in different background context. Object retrieval is considered to be a more challenging problem, since targets usually only occupy small regions on images. Despite the difficulty, object retrieval is a fundamental problem with numerous applications on product search, archive video search, video organization, surveillance, protection of brand/logo use. In this paper, we study the problem of object retrieval that differs from traditional similar image retrieval.

In the past decades, there have been numerous studies on object retrieval, e.g., developing more advanced feature representations, quantization strategy, indexing structures, and post-processing techniques. Traditional approaches for object retrieval are usually based on the SIFT feature [1] and Bag-of-Word (BoW) model [2], where each image is represented as a holistic high-dimensional sparse vector. Indexed with inverted file structure, it can efficiently retrieve similar images in large scale dataset. Other works try to use graph model to cover the relationship among images. [3] augments the neighborhood graph with a bridge graph for approximate nearest neighbor search. [4] uses the query-driven iterated in neighborhood graph search to improve performance. [5] uses CCA and PSI to learn a similarity function and preserve the preference relations. Recent studies on aggregated feature representation, such as Fisher vector [6] and VLAD [7], improves the scalability by further compressing features into a more compact vector. Compared with BoW, aggregated feature, such as Fisher vector and VLAD, has a key advantage when Retrieving object on big data, as it only needs several bytes to represent one image, which is important for retrieval in the context of big data. Despite their success in similar image retrieval, these methods are not directly applicable for object retrieval. First, image level representation is ineffective in the context of object retrieval, since the relevancy is defined at object level. Coupling the target object with its background context into a unified feature vector makes it difficult to retrieve the objects. Obviously, a more effective feature representation is demanded to cope with the objects in reference images. Second, the spatial extent of the object is completely ignored in traditional graph-based re-ranking, which favors similar image (other than objects) clusters in the ranklist.

Figure 1 illustrates the "background interference", which becomes a big issue while retrieving and re-ranking objects. Although all of the four reference images contain the same query object, their VLAD representations are dissimilar, due to the interference of features from the background context. Traditional methods can not solve this problem, because the compressed feature vector does not distinguish the objects and background in an image. In this work, we address this problem by adapting image retrieval techniques to object retrieval.

We tackle the problem of "background interference" with partitioning. Although the image is difficult to retrieve as a whole, partitioned objects are still highly similar with the query (see Figure 1: the right column). The straightforward solution is by partitioning the reference image into object proposals [8]–[10] for further processing. This strategy addresses the background interference problem by separating the objects from their background. However, subsequent problems are still challenging. First, large number of object proposals (1k∼2k proposals per image) could flood the memory to index, which requires much larger memory and

- *Z. Chen, X. Cao and *B. Hu are with School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China (E-mail: caoxiaochun@iie.ac.cn; changingivan@gmail.com; bh@lzu.edu.cn).*
- *W. Zhang, S. Liu and D. Meng are with State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China (E-mail: wzhang34-c@my.cityu.edu.hk; fifthzombiesi@gmail.com; mengdan@iie.ac.cn).*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2016.2566664, IEEE Transactions on Big Data

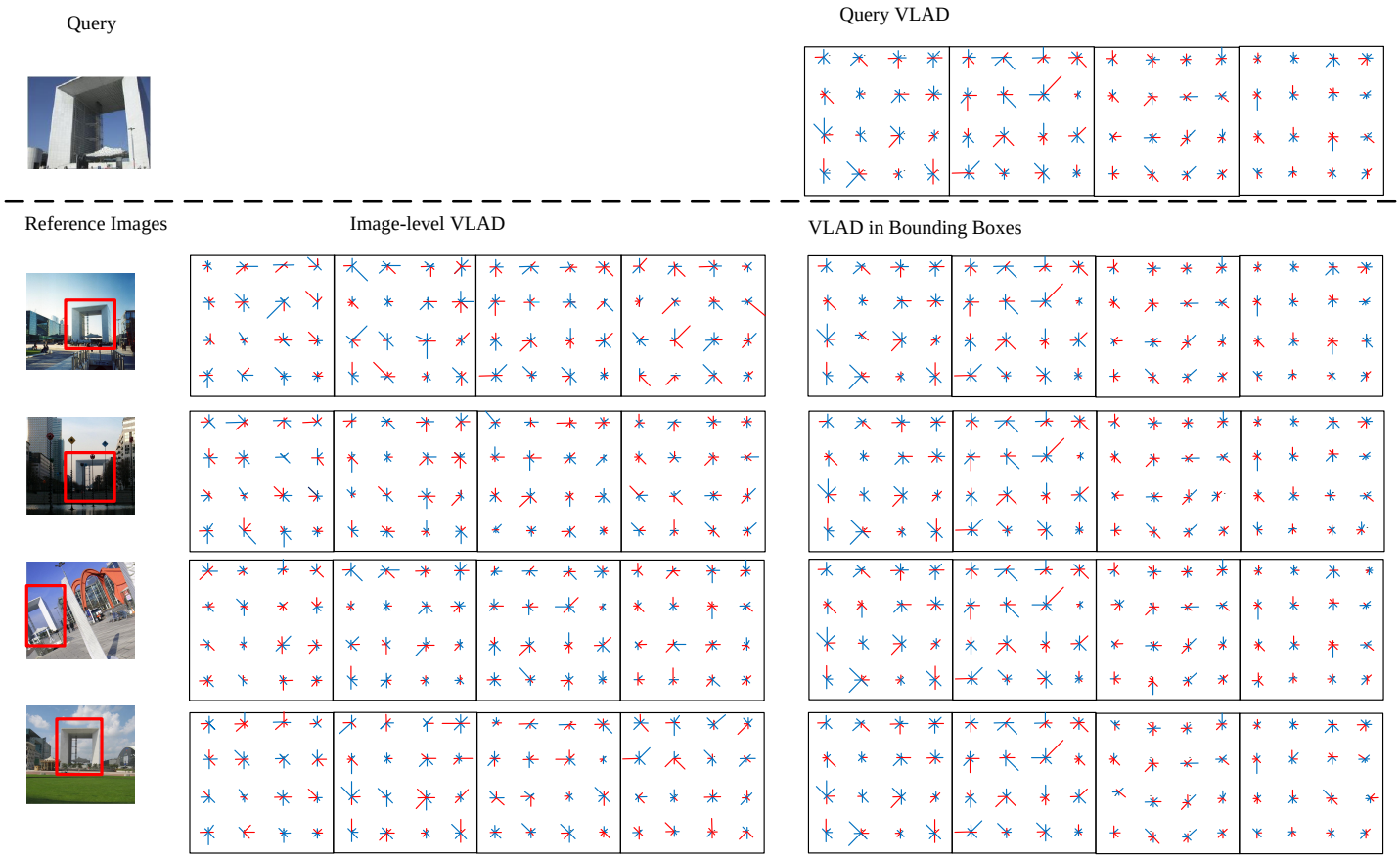


Fig. 1. Illustration of the "background interference". Each VLAD descriptors has 4 centroids (D=4x128). Top: a query object and its VLAD vector plotted in boxes. Each component of the VLAD vector is represented as a short line segment, just as in [7]. Blue and red indicate positive and negative components, respectively. Bottom-left: the query object (marked in rectangle) appears in four different reference images. Bottom-middle and right: extracted VLAD vectors from full image and the objects in rectangles, respectively. From Bottom-middle we can see that although these images contain same object, when background is included into feature coding, features from relevant images could become totally different.

time for retrieval. Second, excessive number of distracting object proposals could hurt the retrieval performance. As commonly noticed, the retrieval performance drops quickly as more irrelevant distracting images are added. In a typical image, most of the generated proposals are not relevant to the query object, which results in the similar effect of adding more distracting images.

Two issues need to be specially considered in our framework. 1) How to generate effective feature representation suitable for object retrieval. Generating compact yet effective representation is challenging. Previous studies on object-proposal partitioning guarantee a decent recall rate with excessive number of object candidates. However, even fine partition is required, which only generates a few (tens instead of thousands) proposals and covers as many objects as possible. 2) How to effectively re-rank in the context of object retrieval. Most of the traditional re-ranking methods have the cluster assumption [11] that emphasizes similar-image clusters. However, this assumption is invalid in object retrieval. New strategy that favors object-level clusters is studied in this paper.

Our contributions can be summarized as two folds. First, we propose an effective representation for object retrieval via partitioning. We jointly optimize the objectness and coverage to select most representative object proposals for each reference image to bypass the "background interference" problem. Second, we rerank the initial ranklist with an object-level graph, which features object-level rather than image-level pairwise similarities. Experimental results indicate that our fine partitioning and object level re-ranking method obtain better performance compared to the baseline.

The rest of this paper is organized as follows. We review related literatures in Section 2. Section 3 presents our partitioning method, and Section 4 discusses our feature representation and re-ranking method based on partitioning. Section 5 evaluates our method on several object retrieval datasets, and finally Section 6 concludes this paper.

2 Related Work

Our work is closely related to previous works on feature representation and visual reranking for object retrieval.

Feature Representation

Early object retrieval approaches adopt the Bag-of-Words (BoW) representation [2], where local features [1] are quantized according to a visual vocabulary. Popular methods usually construct a large [12]–[14] or hierarchical [15]–[17] vocabulary to improve the discriminative power of the BoW representation. Although the resultant representation is very high dimensional, the sparsity nature still enables efficient retrieval by indexing techniques such as inverted file. Other

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2016.2566664, IEEE Transactions on Big Data

works like [18] try to use online multi-label active learning to generate semantic concept representation.

Recent studies on aggregated feature vector further compress the local features into an even small vector, so that much larger dataset can be indexed and retrieved efficiently. Fisher vector [6] as well as its simplified version VLAD [7] are compact yet discriminative in describing the visual appearances of images, by encoding the derivatives of the generative model. These methods significantly reduce the memory cost for one single image (only tens of byte), which enables very large scale retrieval. Recent improvements for VLAD focus on vocabulary adaptation between datasets [19], multiple vocabulary representation [19], and different normalization techniques [19], [20].

Besides that, in order to improve the efficiency of retrieval system [21] considers multiple properties of local features to design the image index. All previous studies ignore the focus of objects, which suffers from the background interference problem. In this paper, we focus on object retrieval and propose a better representation to deal with background interference. Our solution leverages the object proposals that generate thousands of bounding boxes from an image. Previous studies on object proposals include objectness [22]: which combines tasks of localization and classification to obtain the "objectness", Selective search [23] [9], which use the segmentation and hierarchy structure to find objects, and BING [8], which is a simple yet efficient solution to calculate the objectness by leveraging the normalized gradient for fast computation. As saliency sometimes could also be considered as a clue of objects, recently, some methods have been proposed to solve the problem, such as [24], [25].

Visual Reranking

Previous works on visual reranking are mostly based on the following assumption [26]: the images with the dominant patterns are expected to be ranked higher than others. Under this assumption, various methods are proposed to seek the "dominant pattern".

Clustering based method [27] clusters the images in the ranklist to find the dominant patter. Assume that relevant images tend to be more similar to each other than to irrelevant ones.

Pseudo-relevance feedback [28], [29] assumes that the top-ranked documents are "pseudo-relevant" to the query, which can be treated as positive examples to estimate the dominant pattern. After each iteration, the visual pattern is updated according to the detected pattern.

Graph-based reranking [30] is another popular method motivated by PageRank [31]. This technique is usually based on a graph with nodes as the images and edges as the similarities between nodes. The dominant pattern here is discovered by propagating the relevancy among nodes to find the stationary probability.

Although there have been many visual reranking methods developed, all of them treats the dominant visual pattern as full image. Thus they are not suitable for object retrieval, where an object-level pattern is usually expected in the retrieved images. In this paper, we seek the object-level pattern for re-ranking by object partitioning.

3 Object Partitioning

In this section, we first discuss the background interference problem, and then propose our object partitioning method to address this problem. Figure 3 shows the framework of our retrieval system. First we partition the reference images to generate candidate object proposals (Section 3.2). Then we use a novel method (Section 3.3) to refine the partition.

3.1 Background Interference

Background interference stands for the feature degeneration occurred when background is considered in object retrieval. In natural images, an object, which only covers a part of the image, is likely to appear in diverse background. For state-of-the-art aggregated feature representation, such as VLAD [7], local features from an image are first quantized to visual words separately, and then the residual vectors are accumulated as the final representation. Combined with PCA, this representation significantly improves the scalability by compressing the local features into a compact vector. However, we argue that such representation is not suitable for object retrieval, since objects usually cover only some region of the image, and local features from the background confuse the representation. After compressing the local features into a compact vector, it becomes extremely difficult to separate the object from its background. Target object will be diluted in the ocean of background noise. As illustrated in Figure 1, the interference of background degenerates the VLAD representation adversely as more background is included. A single image-level representation is not adequate any more in the context of object retrieval. The single vector can be totally different from the query, although all of the reference images in Figure 1 contain the same querying object. As a result, the discrimination power is compromised due to background interference. However, if the VLAD vector encodes only the target object, the resultant representation is much more effective for retrieval / reranking. Next, we exploit to address the background interference problem through partitioning.

3.2 Partitioning with Object Proposals

As a good starting point, we adopt the well-studied object proposals to pre-partition each image into thousands of bounding boxes. This section discusses different techniques for generating object proposals.

There have been a number of previous studies on object proposal. The most famous works include Objectness [10], Selective Search [9] and BING [8]. Objectness [10] measures the likelihood of a bounding box containing a generic (category-independent) object. It combines multiple cues of edge density, color / texture contrast and consistency to evaluate the objectness score for an arbitrary image area. Selective Search [9] starts with the image over-segmentation, and gradually merges small regions to bigger ones. The benefit of this strategy is the hierarchical structure among proposals. However, such hierarchy is less helpful in our case, where direct object proposals at different sizes already gives a rather good start. Moreover, we also expect more proposals on "important" areas with multiple overlapping

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2016.2566664, IEEE Transactions on Big Data

objects. Recent work of BING [8] binarizes the normalized gradients of a bounding box for fast processing, which outperforms previous works on both effectiveness and efficiency. Thus in this work, we adopt BING to generate initial object proposals.

We follow the default setting of BING [8], and partition each reference images into thousands of candidate proposals. Before feeding into the selection algorithm in Section 3.3, we first pre-process the results to rule out unreliable proposals for retrieval. By observing the results generated by BING, we remove three types of bounding boxes into consideration: (1) Empty boxes. It does not help to include boxes that are with a few or no SIFT features inside. (2) Skinny boxes. Boxes are with extreme aspect ratios[1] are also removed in our implementation. (3) Too large boxes. Boxes covers most (80%) of the image area suffer less from background interference. Since we include the original image for indexing anyway, including too large boxes does not contribute too much to the retrieval.

Fig. 2. Example object proposals selected in a typical image. Top $k = 5$ object proposals are plotted for visualization.

3.3 Proposal Selection via Quadratic Optimization

After the image partitioning methodology introduced in Section 3.2, each image results in thousands of proposals. Although directly indexing all these proposals solves the "background interference" problem, it creates new problems in efficiency and memory load. In this section, we address this problem with a novel proposal selection via quadratic optimization.

Our method is motivated with two observations. On one hand, indexing highly overlapped bounding boxes does not effectively increase the information quantity. On the other hand, for a natural image, such as Figure 2, there are only tens of objects that could be possibly queried by a user. It is possible to select only a few representative proposals without decreasing too much in retrieval performance. Next, we propose an effective proposal selection strategy by jointly optimizing coverage and objectness.

Given n object proposal $\mathcal{P}^I$ extracted from the image I, we aim to select a subset (k) representative proposals, which covers salient objects in I as much as possible and reduces the object redundancy. Theoretically, there are $\binom{n}{k}$ possible combinations to choose from, which is computational infeasible. In this work, we pose this as an optimization problem, which favors high saliency, low overlap subset. Specifically, we propose to solve the following problem:

$$\mathbf{S}^* = \arg\max_{\mathbf{S}} \mathbf{W}^T\mathbf{S} - \frac{1}{2}\lambda\mathbf{S}^T\mathbf{O}\mathbf{S}, \qquad (1)$$

where $\mathbf{O}$ is a $n \times n$ overlap matrix encoding the overlap ratio for pairwise n proposals in each reference image. In our implementation, the *intersection over union* (IoU) is adopted to measure the overlapping ratio between two object proposals (bounding boxes). Therefore, $\mathbf{O}$ is a symmetry matrix with zero-entries along the diagonal. $\mathbf{W}$ is the n dimentional weight vector for the objectness score. In our case, $\mathbf{W}$ is assigned to the confidence score of the object proposals generated by BING, where higher value indicates more likelihood of being an object. $\mathbf{S}$ is the n dimensional indicator variable vector to optimize, in which the components corresponding to the selected proposals are assigned to one, and zeros otherwise. λ is the balance parameter, we set to 0.1 in our experiments.

Generally, the first term ensures a high objectness score of the selected objects, while the second term (with the minus sign) minimizes the overlap of the selected objects and encourage exploration to less salient areas. By optimizing the Eq. 1, we tradeoff between the objectness and coverageness with the parameter λ. However, directly optimizing this discrete quadratic optimization problem is NP-hard. In this paper, we relax the solution space of $\mathbf{S}$ to be continuous. That is, $\mathbf{S}_i, i = 1, 2, ..., n$ is relaxed to the range of $[0, 1]$, which indicates the probability of selecting the i-th proposal. With this relaxation, the problem becomes a standard quadratic optimization. Since our overlap matrix $\mathbf{O}$ is not positive semidefinite, no closed form solution exists in our case. Instead, we use gradient descent with multiple random initializations to solve Eq. 1. Then we binarize $\mathbf{S}$ by setting the top-k entries as one (zeros for others). Although other sophisticated methods (e.g., interior point, active set) exist, we find that our simple solution is efficient (usually convergence within 40 iterations) and works well in practice. Figure 2 shows an example of a typical image after the proposal selection. Most of the salient objects are captured with the top $k = 5$ proposals. By inspecting the example, we can see that our method prefer to select objects like person, ball and logo rather than the background grass and sky. Although there are some overlaps among selected proposals, our method covers most objects in the image. Note that determining the number of proposals (k) is a tricky problem. However in most cases, there are only tens of objects in a typical image, despite the variations among images. Therefore, we fix k for all images in our method.

By formulating the proposal selection as a joint optimization of objectness and coverageness, we capture most representative yet small overlapped proposals and reduce the number of object candidates from thousands to tens. This selection strategy could significantly reduce the cost of

1. Larger than 4 or smaller than 1/4

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2016.2566664, IEEE Transactions on Big Data

computational cost for each reference image as well as the memory requirement, which is crucial to real applications.

4 Object Retrieval with Partitioning

In this section, we will further exploit the object-level visual relationship among selected proposals for retrieval and reranking.

4.1 Object Retrieval via Partitioning

With the partitioning presented in Section 3, we manage to get only a small number of proposals from each image, while keeping the key objects for retrieval. VLAD is extracted separately for each of the top k objects.

During online retrieval, each partitioned object is retrieved independently without considering the image level information. That is, object level relevancy score is evaluated between the query and partitioned object proposals. The relevancy score for each image is simply taken as the max score of its all objects. Since our method is based on partitioning, we name our method after "pVLAD", which is short for partition VLAD.

4.2 Reranking based on Objects

Traditional ranking method based on image-level connectivity, which is asymmetric toward the task of object retrieval. In this section we will explore two new object-level ranking methods based on the relationship in a object matching graph. Following experiments show that our new ranking methods could improve the performance of final system compared with previous methods.

This section we first introduce a reranking method based on object-level partitioning and PageRank algorithm [31]. Similar to the construction of webs on the Internet, which connect each other by hyperlinks, we build our object matching graph and operate reranking method on it. Traditional graph-based ranking model, like [32] only uses the image-level similarity to construct the matching graph. Meanwhile we build object-level matching graph based on selected proposals in previous section. Rather than [32] uses the symmetry similarity measurement to build the graph model, we use the matching graph model introduced by [33]. Compared with establishing image-level matching graph, our graph truly cover the visual-link relationship among objects rather than images.

4.2.1 Object Graph

First, we use every selected proposals above as query to search the same object on the whole reference dataset after partitioning. The system returns the ranking list of reference proposals as well as its similarity score. After all proposals finish their search, we use these connections to construct the object matching graph: $\mathcal{G} = \{V, E\}$ where V are nodes representing query proposals, E are edges. For two nodes : $node\ i$ (query) , $node\ j$ (reference), one edge E_{ij} exists between them if and only if the reference proposal appears in the query one's retrieval ranking list, which means the graph may be asymmetry. Note that our graph is based on object-level, so one image could have several corresponding proposals nodes in the graph, such a design is robust toward multi-object appeared in single image.

4.2.2 Object based Reranking

With the object-level matching graph, we run our reranking method on it:

$$\mathbf{H_s}^{i+1} = \mu \mathbf{M_r} \mathbf{H_s}^{i} + (1-\mu)\mathbf{B}, \tag{2}$$

where $\mathbf{H_s}$ is the $n \times 1$ ranking score measuring the significance of each proposal, it is initialized as the retrieval scores; $\mathbf{M_r}$ is adjacent matrix of all proposals. μ is a damping factor, we set to 0.8 in practice, and $\mathbf{B}$ is a $n \times 1$ initial ranking score of these proposals. Compared with traditional reranking method, which directly uses the visual similarity as the element of this matrix, we use the sorted placement in corresponding query return list as a measurement of similarity between two proposals, that means for each m_{ij} in $\mathbf{M_r}$:

$$m_{ij} = \begin{cases} \frac{1}{t} & \text{if } j \in \text{return list } (i) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where i is the index of query proposal, j is the index of one reference proposal, and t is the rank of j-th proposal appeared in the i-th query return list. For example, if the 8-th proposal appears in the 7-th query on the 5-th placement then $m_{78} = 1/5$. Based on the object-level matching above, this design configures the visual relationship around those related proposals. Instead of directly using visual similarity, such as cosine similarity or Euclidean distance, using ranking placement as the similarity measurement provides another perspective to represent this visual relationship. We name our reranking as "obj-Rerank" since the reranking is based on object level connectivity.

4.3 Radius Sensitive Expansion

We further explore the object-level graph from another perspective. As illustrated in Figure 4, similar proposals are connected and closer in the object matching graph while different proposals are far from each other or disconnected. We use the $\mathcal{G}$ to expand our returned ranking list. This proposal matching graph provides a global similarity connection relationship for every candidate in reference dataset, and this global measurement could be complementary toward the special query search progress by combining local and global measurement together. We define r as the expansion search radius starting from nodes which are considered to expand the list. After observation from Figure 4 we found that only portions of those connected proposals are beneficial to the promotion of retrieval, which means the true positive proposals is sensitive to radius value of r : if r is too large, false positive would be included into return list, meanwhile if r is too small, true positive samples will be excluded. In the experiment section we will exploit the influence from different r values. We expand the initial query ranking list with its corresponding object graph neighbors in $\mathcal{G}$. Following experiments show that our method could improve the performance of retrieval to some extent. Here is the solution for the expansion:

$$\mathbf{R_i} = \mathbf{E}_{r_i} \oplus \mathbf{R_i}, \tag{4}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2016.2566664, IEEE Transactions on Big Data

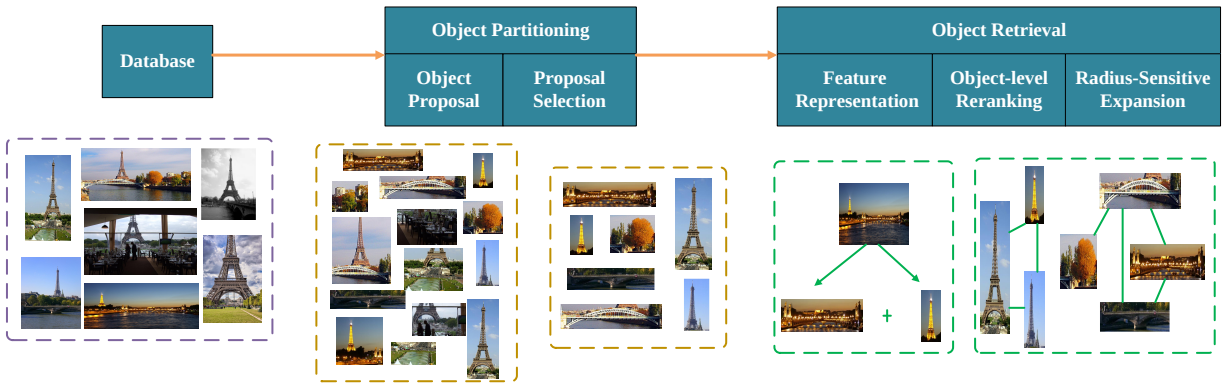


Fig. 3. The framework of our object retrieval method. Our framework is composed with two major steps: object partitioning and object retrieval. For object partitioning, we first generate candidate proposals, then we design a novel selector to choose most discriminative proposals from all candidates on each image (Section 3). In object retrieval, we consider three aspects. We obtain the VLAD for every selected proposal and combine them as the final representation of this image. Then we construct the object-level matching graph for each dataset. After retrieving query object based on the pVLAD, we exploit to rerank this return list in object-level (Section 4.2). We also expand this return list by considering the relationship among these proposals on the graph (Section 4.3).

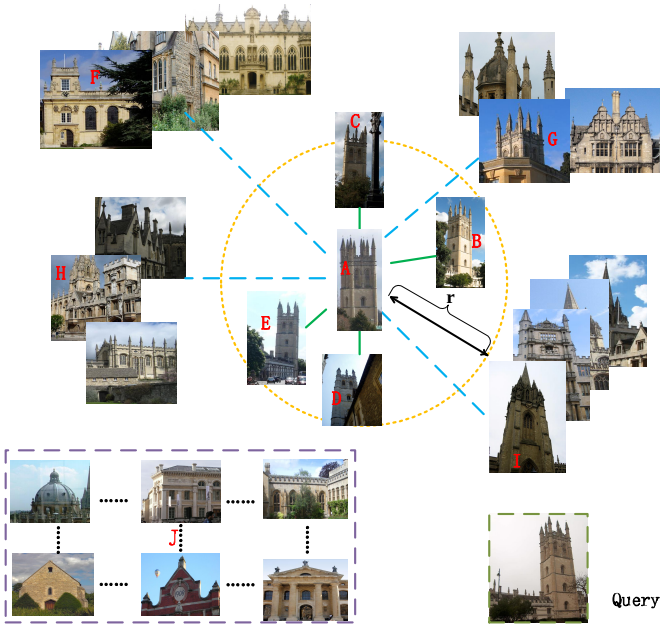


Fig. 4. Illustration of Radius Sensitive Expansion. Here is an example of sensitive radius in the object contain graph. A is the ranked object proposal node in the graph, while r is the sensitive radius, and $B \sim E$ are connected candidates within the r. Since r is the search radius begin from A, it's necessary to take $B \sim E$ into the consideration of the ranking process. $F \sim I$ are proposals connected with the reference node (A) but beyond the radius r. J are those proposal which has no relationship with the reference node in this graph.

where $\mathbf{R_i}$ is the i-th element in the initial ranking list, And $\mathbf{E}_{r_i}$ is a set of the nodes within the scope of the r_i (defined by Eq. 5). Then we define the $\oplus$ operation as obtaining the union of two set and unique the element in ranking list. That means if there are some nodes (proposals) in the object matching graph meeting this condition above, they will combine with the original element as the new retrieval element. In practice, r_i is decided by the $\mathbf{R_i}$ and its next direct neighbor in the ranking list:

$$r_i = \begin{cases} \mathbf{R_i} - \mathbf{R_{i-1}} & \text{if } \mathbf{R_{i-1}} \in \text{return list } (\mathbf{R_i}) \\ \mathbf{0} & \text{otherwise} \end{cases} \quad (5)$$

We use Euclidean distance to measure the r_i. From Eq. 5 we can see that the max radius is set as the difference between continuous two proposals in the retrieval list. In practice, this operation could be efficient by leveraging the object graph built above. We name this method as RSE, which is short for "Radius Sensitive Expansion". Note that, RSE is different from 'Discriminative query expansion' in [14], as they use enrich query to obtain positive and negative data and train a linear SVM to sort reference images.Radius Sensitive Expansion could be viewed as an exploration for the global similarity property of the object graph. Higher ranked candidates in ranking list will be re-considered with its global similar neighbors in the object graph. Following experiments show that expanding the ranking list by utilizing this property could make the result more complete.

5 Experiments

In this section, we evaluate our method on three popular object retrieval datasets. We start by introducing the datasets (Section 5.1) used in our experiments and the implementation details (Section 5.2), and then compare our methods with the baseline VLAD [7].

5.1 Datasets

Three publicly available object retrieval datasets are used in our evaluation. The mean Average Precision (mAP) is used as our evaluation metric throughout our evaluation. Table 1 summarizes the average area covered by each object on the query image for different datasets. In general, Oxford and Paris are included into our evaluation as object retrieval datasets, and Holiday is adopted as a performance comparison on similar image search.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2016.2566664, IEEE Transactions on Big Data

5.1.1 Oxford Buildings

The Oxford Buildings dataset [13] contains 5,062 images downloaded from Flickr using keyword search. The collection is 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 for evaluation. Unlike full-image retrieval, we only use the bounding-box object in these 55 images for queries.

5.1.2 Paris

The Paris dataset [12] consists of 6,412 images crawled from Flickr by searching for particular Paris landmarks, such as "Paris Eiffel Tower" and "Paris Triomphe". Similar to Oxford Buildings, this dataset also has 55 queries for evaluation.

5.1.3 Holiday

The Holiday dataset [34] contains 1,491 images for personal holidays photos, among which 500 images are used as queries and the rest are manually labeled as ground-truth. Since this dataset does not provide bounding box for the query, we manually label the bounding box for each query image by inspecting the common object (area) among ground-truth images.

TABLE 1
Average area(%) covered by the querying objects on the query images on different datasets.

Dataset	Oxford	Paris	Holiday
Average area ratio	26.7%	27.8%	40.5%

5.2 Implementation Detail

We adopt the SIFT descriptor [1] with Hessian Affine detector and VLAD [7] as the feature representation. A vocabulary with 64 centers are used to quantize local features. Unless explicitly specified, the radius (r) used in Radius Sensitive Expansion is set to 25, and the number of object proposals selected from each image (k) is set to 15 on our sensitivity tests (details in Section 5.3 and 5.5).

5.3 VLAD versus pVLAD

5.3.1 Different proposals generation

We first compare different proposals generation method in our pVLAD: BING, Selective Search [9], and Edge Boxes [35]. Table 2 shows mAP performance and time cost per image around them. From the table we can see that BING obtains higher performance and faster than other two methods. So we choose BING to generate proposals.

TABLE 2
Performance and time comparison for different proposal methods.

Method	BING	Selective Search	Edge Boxes
mAP	0.3583	0.3560	0.3073
Time (Second)	0.003	3.790	0.250

TABLE 3
Performance(mAP) comparison for different retrieval methods.

	Oxford	Paris	Holiday
VLAD	0.2912	0.3857	0.5019
Image-Rerank	0.3215	0.4046	0.5103
pVLAD	0.3583	0.4265	0.5207
obj-Rerank	0.3595	0.4331	0.5271
RSE	0.4095	0.4745	0.5303

5.3.2 Overall performance

We compare our retrieval method, pVLAD, with the standard VLAD [7] to evaluate our partitioning strategy. Table 3 compares the performance in mAP for different retrieval methods. As shown, our method outperforms VLAD on all three datasets. In particular, our method improves the mAP by 23.0% / 10.6% / 3.7% relatively to the baseline on Oxford / Paris / Holiday. The improvements on different datasets vary differently. For instance, pVLAD only slightly improves the performance on Holiday dataset, but obtain better performance on Oxford and Paris. This phenomenon can be explained with the size of objects in different datasets. From the Table 1 we notice that the bounding box for the query object covers larger area on Holiday than on Paris and Oxford. Generally, smaller area covered by the object results in more severe background interference. Since our partition-based method mainly addresses the background interference problem, it improves more on datasets with smaller objects, e.g., Oxford dataset. On the contrary, pVLAD only slightly improves the baseline on Holiday, which is more like a similar image retrieval dataset. Note that, in [36], authors extract deep conv-net features instead of VLAD from each proposal for the task of Classification and Retrieval, they obtain 0.887 mAP on Holiday, in the future we will explore to use deep feature in our method.

5.3.3 Number of proposals selected per image

We also test the sensitivity of the number of proposals selected (k) for each image. As shown in Figure 5, the performance goes up as we use more object proposals, and reaches the peak around $k = 10 \sim 20$. However, even more proposals do not introduce further improvement and finally hurt the performance. As k increases, it becomes more likely to include all possible objects in retrieval and reranking, and thus the recall could be improved. However, too large k will introduce too much noisy proposals for processing. We observe that for most of the nature images, there are only $5 \sim 10$ objects per image, which are likely to be queried by a user. Sampling too many object proposals will inevitably introduce severe noises, and thus hurt the retrieval precision. When k goes beyond 20, the newly introduced objects are mostly bounding boxes featuring less meaningful regions that are unlikely to be queried, we set $k = 15$ in our experiments.

5.4 Object-based Reranking

Here we evaluate our object-based reranking method. As shown in Table 3, we can observe that our reranking method (obj-Rerank) improves the mAP over pVLAD by 0.3% / 1.5% / 1.2% relatively on Oxford / Paris / Holiday.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2016.2566664, IEEE Transactions on Big Data

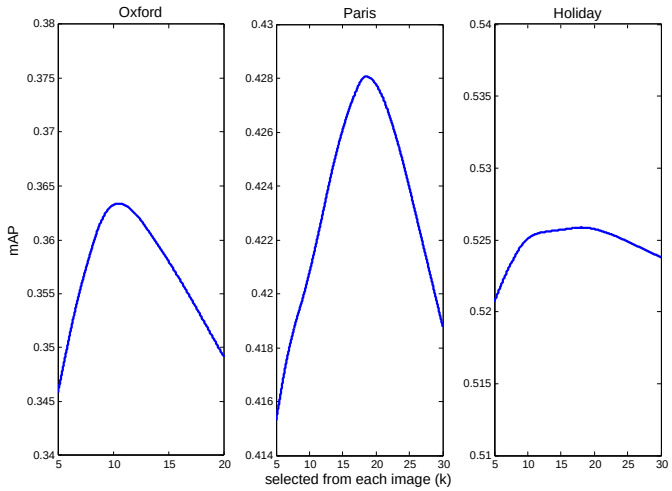


Fig. 5. Sensitivity test on number of object proposals adopted (k) in each reference image.

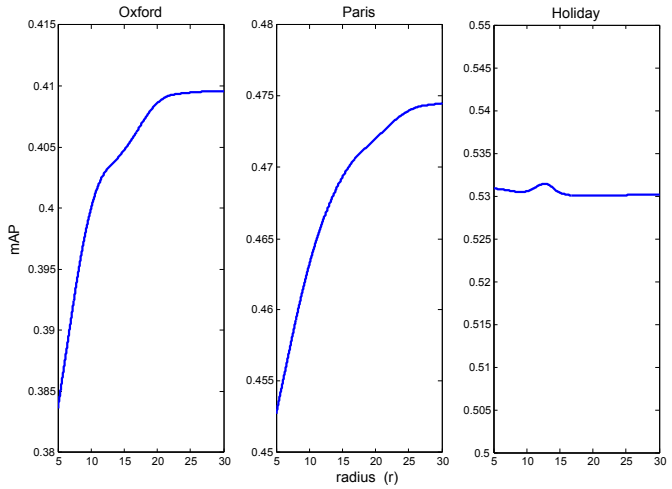


Fig. 6. Sensitivity test on the radius (r) in Radius Sensitive Ranking on different datasets.

Although the performance improvement is small, our obj-Rerank does boost novel results that were poorly ranked in the original list. Figure 8 shows several example images as well as their ranks in pVLAD and obj-Rerank. Since the object-level connectivity is modeled in our object graph based reranking, obj-Rerank gives much better results for small object. For example, the 3-rd image in the top-left example is boosted significantly (from 44 to 29) due to the object-level reranking. We also use the image-level feature to rerank the return lists, As shown in Table 3, the pVLAD could outperform Image-Rerank in all three datasets.

5.5 Radius Sensitive Expansion

We also test our Radius Sensitive Expansion (RSE) on the datasets. Table 3 (last row) shows that our RSE improves the performance significantly over pVLAD. For RSE, we obtain an improvement of 14.3% / 11.3% / 1.8% relative to pVLAD on Oxford / Paris / Holiday dataset, which shows the effectiveness of our method. Qualitative examples are given in Figure 9, where novel small objects (in yellow) are pulled from the object-graph based expansion, while for the original retrieval ranklist, these objects are easily missed due to the noise in partitioning (pVLAD) or background interference (VLAD).

TABLE 4
Performance(mAP) comparison with BoW methods.

Method	mAP	size of vocabulary
AKM [13]	0.618	1000k
RootSIFT [14]	0.881	1000k
HPM [37]	0.692	500k
[38]	0.738	1000k
pVLAD	0.358	1k

5.5.1 *Sensitivity test on the radius (r)*

We further explore the effect of the radius (r) used in our Radius Sensitive Expansion. Figure 6 plots the performance against different r. From the result in Fig.6 we can see that different r value in Radius Sensitive Expansion have different impact on the final result. Just as Fig.4 shown, when the value of r large, false positive may be merged into return list, when the value r is small, true positive may be excluded from the list. We test different radius ($r = 5 \sim 30$) on three datasets. We can see that on Oxford and Paris larger r gives better results and the performances become steady when $r > 25$. While for Holiday, the performance is not sensitive to r. This is because on Holiday each query only have about two ground-truth images on average (500 queries and 991 reference images), and the dataset is relatively small compared to Paris and Oxford. These proposals connect less similar nodes on the graph than other datasets. So varying r does not effect the performance too much.

5.6 Comparison with State-of-the-art BoW methods

Although pVLAD based on VLAD, we still compare our method with some BoW based methods. As Table 4 shown, BoW based methods usually need large vocabulary for better performance, meanwhile pVLAD only needs very much smaller vocabulary, which is important for retrieving object in the context of big data.

5.7 Scalability

In order to test the scalability of our method, we add another 1 million Flickr images from the YFCC100M dataset [39] as distracters. All performance are shown in Figure 7. From results we can see that our method could obtain better performance in large dataset.

6 Conclusion and Future Work

In this paper, we have presented a novel object retrieval framework to address the "background interference" problem by adapting state-of-the-art image retrieval techniques to object retrieval. Our solution performs in the manner of partitioning, that divides the whole image into tens of object candidates which are retrieved / reranked separately and independently to the background. More effective feature representation and reranking in the context of objet retrieval

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2016.2566664, IEEE Transactions on Big Data

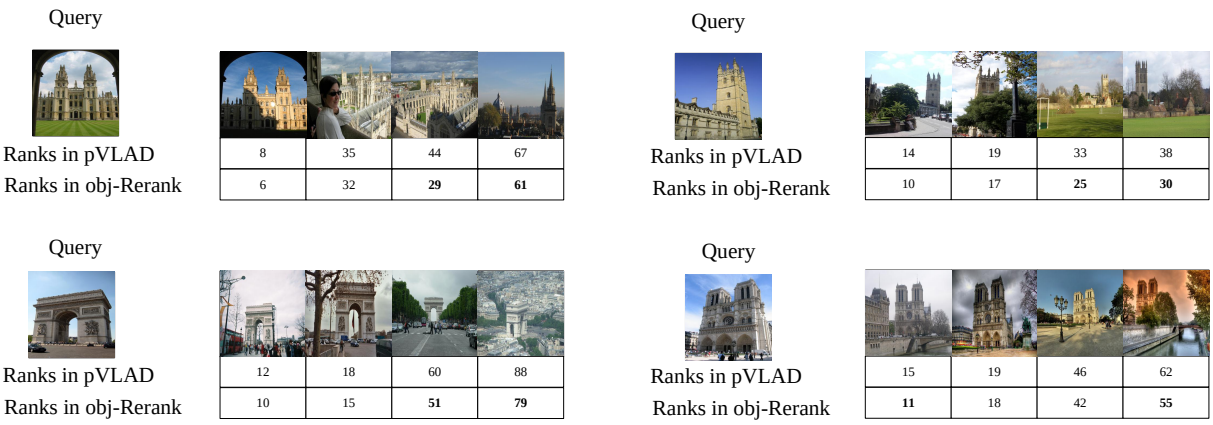


Fig. 8. The comparisons between pVLAD and obj-Rerank for four query images and their ranking results with the ranks plotted below. As shown, smaller objects receive more boosting in their ranks, due to the object level connectivity used in our object graph based reranking.

Fig. 9. Example query as well as the retrieved images by two different methods: pVLAD and RSE. Images with dashed yellow borders are results expanded through our RSE. Note that pink rectangles in the first column stand for the query object in image.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2016.2566664, IEEE Transactions on Big Data

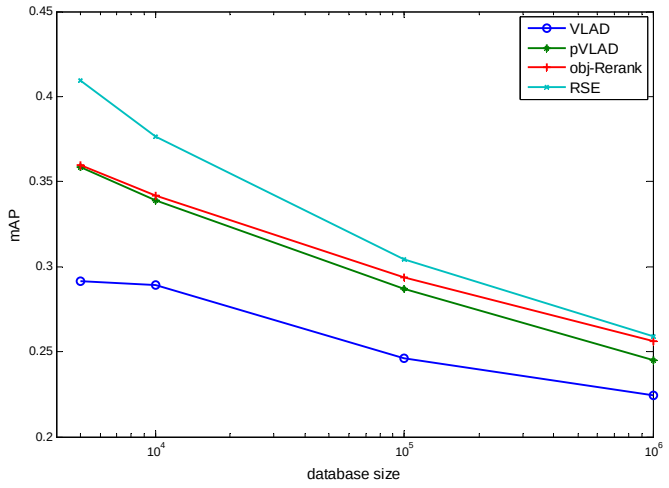


Fig. 7. Scalability test on different size retrieval dataset.

are developed with this partitioning. Experimental results on popular object retrieval datasets show a clear advantage of our method in object retrieval. In our future work, we plan to further exploit the object-level semantic relationship in the graph to capture more related instances for query.

7 Acknowledgment

This work was supported by National Natural Science Foundation of China (No.61572493, Grant U1536203), 100 Talents Programme of The Chinese Academy of Sciences, and Strategic Priority Research Program of the Chinese Academy of Sciences (XDA06010701).

References

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision (IJCV)*, vol. 60, no. 2, pp. 91–110, 2004.

[2] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *International Conference on Computer Vision*, vol. 2, Oct. 2003, pp. 1470–1477.

[3] J. Wang, J. Wang, G. Zeng, R. Gan, S. Li, and B. Guo, "Fast neighborhood graph search using cartesian concatenation," in *Multimedia Data Mining and Analytics*. Springer, 2015, pp. 397–417.

[4] J. Wang and S. Li, "Query-driven iterated neighborhood graph search for large scale indexing," in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 179–188.

[5] T. Yao, T. Mei, and C.-W. Ngo, "Learning query and image similarities with ranking canonical correlation analysis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 28–36.

[6] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[7] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 3304–3311.

[8] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[9] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders, "Segmentation as selective search for object recognition," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 1879–1886.

[10] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Analasis and Machince Intelligence*, vol. 34, no. 11, pp. 2189–2202, Nov. 2012.

[11] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X. sheng Hua, "Bayesian video search reranking," in *ACM International Conference on Multimedia*, 2008.

[12] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Computer Vision and Pattern Recognition (CVPR)*, 2008.

[13] ——, "Object retrieval with large vocabularies and fast spatial matching," in *Computer Vision and Pattern Recognition (CVPR)*, 2007.

[14] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2911–2918.

[15] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2006, pp. 2161–2168.

[16] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, and T. X. Han, "Contextual weighting for vocabulary tree based image retrieval," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 209–216.

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

[18] X.-S. Hua and G.-J. Qi, "Online multi-label active annotation: towards large-scale content-based video search," in *Proceedings of the 16th ACM international conference on Multimedia*. ACM, 2008, pp. 141–150.

[19] R. Arandjelovic and A. Zisserman, "All about vlad," in *Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1578–1585.

[20] J. Delhumeau, P.-H. Gosselin, H. Jégou, and P. Pérez, "Revisiting the vlad image representation," in *ACM international conference on Multimedia*, 2013, pp. 653–656.

[21] L. Zheng, S. Wang, Z. Liu, and Q. Tian, "Fast image retrieval: Query pruning and early termination."

[22] H. Harzallah, F. Jurie, and C. Schmid, "Combining efficient object localization and image classification," in *Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 237–244.

[23] I. Endres and D. Hoiem, "Category independent object proposals," in *European Conference on Computer Vision*. Springer, 2010, pp. 575–588.

[24] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *Image Processing, IEEE Transactions on*, vol. 22, no. 10, pp. 3766–3778, 2013.

[25] X. Cao, C. Zhang, H. Fu, X. Guo, and Q. Tian, "Saliency-aware nonparametric foreground annotation based on weakly labeled data," 2015.

[26] T. Mei, Y. Rui, S. Li, and Q. Tian, "Multimedia search reranking: A literature survey," *ACM Computing Surveys*, vol. 46, no. 3, pp. 38:1–38:38, Jan. 2014.

[27] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking via information bottleneck principle," in *ACM International Conference on Multimedia*, 2006, pp. 35–44.

[28] R. Yan, A. G. Hauptmann, and R. Jin, "Multimedia search with pseudo-relevance feedback," in *ACM International Conference on Image and Video Retrieval (CIVR)*, 2003.

[29] A. Amir, J. Argillander, M. Campbell, A. Haubold, G. Iyengar, S. Ebadollahi, and F. Kang, "Ibm research trecvid-2005 video retrieval system," in *TRECVID Workshop*, 2005.

[30] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking through random walk over document-level context graph," in *ACM International Conference on Multimedia*, 2007, pp. 971–980.

[31] P. Lawrence, B. Sergey, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford University, Technical Report, 1998.

[32] Y. Jing and S. Baluja, "Visualrank: Applying pagerank to large-scale image search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1877–1890, 2008.

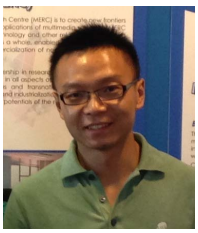
[33] J. Philbin and A. Zisserman, "Object mining using a matching graph on very large image collections," in *Computer Vision, Graphics & Image Processing (ICVGIP)*, 2008, pp. 738–745.

[34] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *European Conference on Computer Vision (ECCV)*, 2008, pp. 304–317.

[35] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 391–405.

[36] L. Xie, R. Hong, B. Zhang, and Q. Tian, "Image classification and retrieval are one," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 2015, pp. 3–10.
[37] Y. Avrithis and G. Tolias, "Hough pyramid matching: Speeded-up geometry re-ranking for large scale image retrieval," *International journal of computer vision*, vol. 107, no. 1, pp. 1–19, 2014.
[38] C.-Z. Zhu, H. Jégou, and S. Satoh, "Query-adaptive asymmetrical dissimilarities for visual object retrieval," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1705–1712.
[39] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "The new data and new challenges in multimedia research," *arXiv preprint arXiv:1503.01817*, 2015.

Zhiyong Chen is a Senior Master Student of the School of Information Science and Engineering, Lanzhou University, China. He received the B.E. degree from the School of Information Science and Engineering, Lanzhou University, in 2013. His current research interests include Image Retrieval, Object detection and Part classification.



Wei Zhang is a research assistant in Institute of Information Engineering, Chinese Academy of Sciences.He received the B.Eng. degree from the School of Computer Software and the M.Eng. degree from the School of Computer Science and Technology, Tianjin University, Tianjin, China, in 2008 and 2011, respectively. He received the Ph.D. degree from the Department of Computer Science, City University of Hong Kong, Kowloon,Hong Kong. He was a Former Member of the CV-TJU Laboratory, Tianjin University, from 2008 to 2011. His research interests include large scale video retrieval and digital forensic analysis.

Bin Hu is a Professor in the School of Information Science and Engineering, Lanzhou University, China; Dean of Technical Committee of Cooperative Computing, China Computer Federation; Director of China branch of Web Intelligence Consortium(WIC); Director of China branch of International Society for Social Neuroscience; and Adjunct Professor, Department of Computer Science, ETH, Zurich, Switzerland. His research fields are cognitive computing, context aware computing, and pervasive computing, and has published about 100 papers in peer reviewed journals, conferences, and book chapters. The works have been funded by quite a few famous international funds, e.g. EU FP7, HEFCE, U.K., NSFC, China, and industry. He has served more than 60 international conferences as a chair/pc member and offered about 40 keynotes/talks in high ranking conferences or universities, and has also served as editor/guest editor in about 10 peer reviewed journals in computer science.

Xiaochun Cao is a professor at the Institute of Information Engineering, Chinese Academy of Sciences. He is also a visiting professor in the School of Information Science and Engineering, Lanzhou University. He received the B.E. and M.E. degrees both in computer science from Beihang University (BUAA), China, and the Ph.D. degree in computer science from the University of Central Florida, USA, with his dissertation nominated for the university level Outstanding Dissertation Award. After graduation, he spent about three years at ObjectVideo Inc. as a Research Scientist. From 2008 to 2012, he was a professor at Tianjin University. He has authored and coauthored over 80 journal and conference papers. In 2004 and 2010, he was the recipients of the Piero Zamperoni best student paper award at the International Conference on Pattern Recognition. He is IEEE Senior Member.

Si Liu is an Associate Professor in Institute of Information Engineering, Chinese Academy of Sciences. She used to be a Research Fellow in Learning and Vision Re-search Group at the Department of Electrical and Computer Engineering, National University of Singapore. She obtained Ph.D. degree from Institute of Automation, Chinese Academy of Sciences (CASIA). She obtained Bachelor degree from Experimental Class of Beijing Institute of Technology (BIT). Her current research interests include Object Categorization, Object Detection, Image Parsing and Human Pose Estimation.

Dan Meng received the Ph.D. degree in computer science from the Harbin Institute of Technology,in 1995. He is currently a Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing. His research interests include high-performance computing and computer architecture. He is a Senior Member of the China Computer Federation.