Image Relevance Prediction Using Query-Context Bag-of-Object Retrieval Model

Yang Yang, Linjun Yang, *Member, IEEE*, Gangshan Wu, and Shipeng Li, *Fellow, IEEE*

***Abstract*—Image search reranking and image research result summarization are two effective approaches which enhance text-based image search results using visual information. Since the existing approaches optimize search relevance in terms of average performance, they usually cannot achieve satisfactory results for some particular classes of queries, like "object queries," which is defined as the queries with the intent of searching for some kinds of objects. One possible reason is that the generic approaches such as [40], [43], [46] are mostly built based on the global statistics of images as features while ignoring the fact that the relevance between the image and the query sometimes depends on an image patch instead of the whole image. In this paper, we therefore design a novel bag-of-object retrieval model to predict image relevance, which is particularly effective for object queries. First, we construct an object vocabulary containing query-relative objects by mining frequent object patches from the result image collection of the expanded query set. After representing each image as a bag of objects, our retrieval model can be derived from a risk-minimization framework for language modeling. To demonstrate the effectiveness of the proposed model, this paper also present two related applications: for image search reranking, we adopt a supervised framework to combine multiple ranking features from different assumptions; for image search result summarization, we propose a two-step ranking process which optimizes not only representativeness but also image attractiveness. The experimental results show that the proposed methods can significantly outperform the existing approaches.**

***Index Terms*—Common object discovery, image search reranking, object vocabulary, web image search.**

I. Introduction

MOTIVATED by the fact that text-based image search engines may provide the users with noisy search results, image search reranking and image search result summarization are proposed from different aspects to improve result quality and search experience using visual information. Image search reranking boosts the relevant images to the top of the search result list while suppressing the irrelevant ones to the bottom, such that the user can access satisfying images in top positions. Image search result summarization discovers a small group of images that are both relevant and representative to the query, providing the user with an overview of the search result list. For both of the approaches, predicting image relevance serves as a key technology and is considered as one of the most challenging problems.

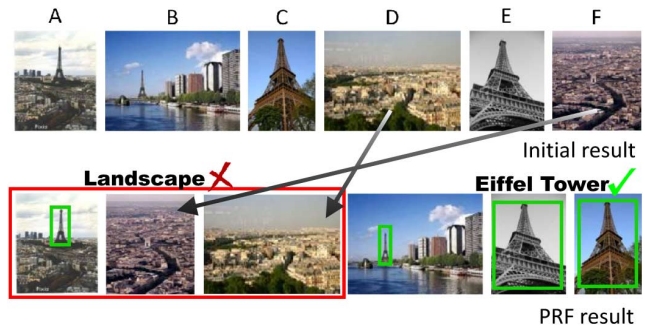


Fig. 1. Example of PRF reranking process for query "Eiffel tower." The irrelevant image **D** and **F** are boosted because of the appearance of landscape while the images are expected to be compared using the regions of the Eiffel Tower.

Two assumptions have been formerly proposed to estimate the relevance of the images: the cluster assumption and the pseudo relevance feedback (PRF) assumption. The cluster assumption suggests that relevant images usually have close visual appearance while irrelevant ones are regarded as noise thus different with each other. PRF assumption regards the images ranked to the top of the text-based search result as pseudo-relevant, which can be employed to train a classifier [23] or multiple classifiers [46]. In this paper, we focus on a typical category of queries, named "object queries", where the user intends to find images containing the desired objects, including landmarks, products, vehicles, animals and people. While the two assumptions have been demonstrated generally effective in existing reranking approaches [10], [12], [35], we find that they are not sufficiently effective to deal with such kind of queries.

The first problem is that, for images retrieved by object queries, usually some parts of the image are relevant to the object query, while the others are not. For example, Fig. 1 shows an simple PRF reranking process for the query "Eiffel Tower" where image **A** is assumed as pseudo-relevant while the other images are ranked with respect to their visual similarities to **A**. Unfortunately, two irrelevant images, image **D** and **F** are boosted to the top because image **A**, **D** and **F** are telling the landscapes of the Paris city, although image **A** shows the

Manuscript received October 18, 2013; revised March 10, 2014; accepted May 16, 2014. Date of publication May 29, 2014; date of current version September 15, 2014. This work is supported by the NSFC of China under Grant 61321491, the 863 Program of China under Grant 2011AA01A202, and the National Special Fund under Grant 2011ZX05035-004-004HZ. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xiao-Ping Zhang.

Y. Yang and G. Wu are with the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China (e-mail: charlie.yang.nju@gmail.com; gswu@nju.edu.cn).

L. Yang is with the Microsoft Corporation, Redmond, WA 98052 USA (e-mail: linjuny@microsoft.com).

S. Li is with the Microsoft Research Asia, Beijing 100080, China (e-mail: spli@nju.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2014.2326836

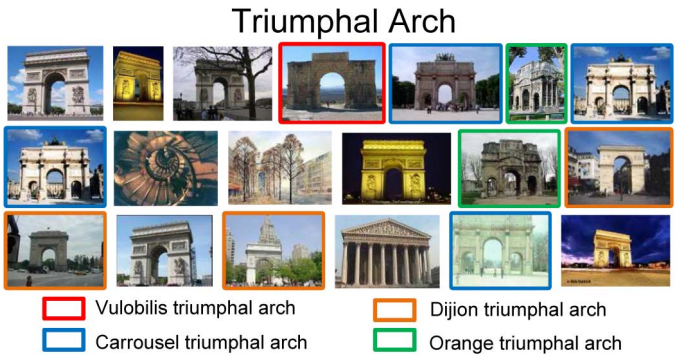


Fig. 2. An example of ambiguous objects: the search result for the query "Triumphal Arch" involves a group of other arch buildings, which are marked as different colors in the figure. We can see that these buildings are visually similar with the correct arch building.

Eiffel Tower. This failure is captured by our previous work in [48], where we believe it is caused by the fact that the existing reranking approaches usually employ image features extracted from the whole image, such as histogram of visual words [6], [50] and it is too rigid for object queries. As a matter of fact, the image could be considered relevant only if one part of the image is relevant, thus the images should be compared using the corresponding object regions. Motivated by this, we have proposed a novel Bag-of-Object retrieval model in [48] which represents the query and result images into a language model using object appearance. In order to focus on the valuable object categories and suppress the background noise, we employed a common object discovery (COD) algorithm to pick up the query-relevance ROIs and construct a query-relevant object vocabulary.

In this paper, we discover another problem that text-based search engines may sometimes be confused by concepts with overlapping key words but different meanings. This is referred to as "the ambiguity problem". Fig. 2 shows some image search results of the query "Triumphal Arch" which is referred to as the famous arch building in Paris. We can see that multiple different arch buildings are cluttered in the result image collection such as "Dijion Triumphal Arch" locating in east France and "Volubilis Triumphal Arch" in Monaco. This problem may have negative effect to the quality of object vocabulary. First, because these ambiguous objects are visually close to each other, they cannot be simply detected as outliers. Second, we can observe from Fig. 2 that some ambiguous objects are also similar with the relevant images. Therefore, the main challenge here it to successfully identify such images so as to prevent them from being boosted. Motivated by such circumstance, this paper proposes to extend the query-relevant object vocabulary presented in our previous work [48] into query-context object vocabulary. We first expand the text query into an expanded query set with most of the ambiguous keywords included. Then a COD algorithm is applied on the result images gathered using all keywords in the expanded set. Comparing to query-relevant object vocabulary, this approach can collect more supporting samples for ambiguous objects, such that they are more likely to be organized as discriminant object categories after applying the COD algorithm. Consequently, the ambiguous images in the search result can be correctly recognized as irrelevant objects other than being assigned to a relevant object category. Besides, we also propose a very effective assumption based on Normalized Google Distance (NGD) for the approximation of query language model, so as to suppress these irrelevant object categories.

We also present two applications in this paper to demonstrate the effectiveness of the proposed retrieval model. For image search reranking, we integrate the risks calculated using different assumptions while the combination model is learned through Ranking SVM [18] from human-labeled data on representative queries. For image search result summarization, the representative images are selected through a two-step selection process. Besides representativeness, we also consider object layout in image quality prediction so as to improve the user experience.

The proposed approaches are evaluated on a subset of the Web Queries dataset [23], comprising object queries. The proposed image search reranking method improves the result from the search engine by 43.64% in the term of Mean Average Precision (MAP), while the proposed summarization approach obtains 56% best votes in the user study.

The rest of the paper is organized as follows. After reviewing the related works on corresponding fields in Section II, we introduce the proposed retrieval model in Section III. In Section IV, we demonstrate the two applications based on the proposed retrieval model. Section V presents the experiments on the two applications and analyzes the experimental results. The last section concludes this paper with some remarks on the future work.

II. Related Work

In this section, we are going to review the previous research works on common object discovery, object-based retrieval models, image search reranking and image search result summarization.

1) Common Object Discovery: Recently, common object discovery is widely discussed in computer vision. This kind of techniques aims to find the frequent objects in a given image collection. The proposed approaches can be generally classified into two categories: segmentation-based methods [22], [19], [13], [42] and bounding box-based methods [21], [25], [7].

Segmentation-based common object discovery segments the objects on multiple images simultaneously. Therefore, it is also named as "Co-segmentation". In [34], the segmentation problem is formulated as the minimization of an energy function, taking the Markov Random Field (MRF) smoothness and a histogram-matching-based consistency into consideration. In [22], the authors propose to decompose each image into super pixels at first and employ a greedy expansion algorithm to associate the super pixels into object regions. Two serious drawbacks of these methods keep us from adopting them in our work. First, most of the co-segmentation methods are not sufficiently effective for web images with complicated background. Second, our scenario requires dealing thousands of result images at a time but co-segmentation-based methods usually take hours.

Bounding box-based methods make use of the result of salient object detection in the form of regions of interest (ROIs). In [7],

a conditional random field (CRF) is built for all the candidate ROIs and the common object discovery problem is transformed as finding an optimal configuration in the CRF. [21] describes a iterative algorithm which alternatively finds the ROI exemplars and decide the foreground ROI for each image. It is employed in this paper due to its high time efficiency because relevance prediction should be performed online.

2) Object-Based Image Retrieval: Object-based image retrieval (OBIR) is a typical kind of content-based image retrieval (CBIR), through which the user intends to find object images. According to the form of queries, it can be divided into interactive OBIR and automatic OBIR. For interactive OBIR, the user is supposed to offer the object region manually, usually in the form of bounding-box. For example, image representations in [14] are based on image segmentations and the objects are modeled based on the regions using Latent Semantic Analysis (LSA). In [44], the authors represent the images as a bag of visual words, and employ language modeling to derive the ranking function. The authors also argue that the object context is as important as the object itself, thus visual words locating outside the object region should also be taken into consideration. For automatic OBIR, the query object region is supposed to be detected automatically by the search engine. Recently, a number of works utilize multi-instance learning (MIL) to model the retrieval process. For instance, in [29] and [27], the users are required to provide multiple query images at a time and model the queries as a multi-instance classifier. To achieve this, the authors of [29] propose a new kind of multi-instance Support Vector Machine (SVM) with a convex training method, while in [27] the authors bring in a novel feature representation scheme using identified evidence ROI. The scenario studied in our paper is different from object-based image retrieval because the search result here is retrieved by keywords. This problem is more challenging because we need to discover the underlying object categories from the noisy search results with no supervision and the relevance of each object category should be predicted as well.

3) Image Search Reranking: As a typical post-process for web image search, image search reranking can serve either with human interaction or automatically. Interactive reranking is also known as "relevance feedback", for which the user labels a set of relevant images after initial search. For example, [38] and [37] focus on building effective classifiers as ranking functions, while the authors of [1] propose a discriminant feature embedding based on labeled images to semantically represent the low-level features.

The automatic reranking approaches mostly make use of the ranking score from text-based search engine and image visual similarity, among which the cluster assumption and PRF (Pseudo-Relevance Feedback) assumption are the most famous. Cluster assumption suggests that the relevant images have visual content in common while irrelevant images are not similar with each other. As pair-wise image similarity can be naturally interpreted as a graph, various reranking methods [40], [41], [16], [15] are proposed from different aspects to implement this assumption using graph structure. The main drawback of such approaches on object queries is the inaccurate measurement of the visual similarity. As claimed in Section I, image similarity may be effected by noise background when images are compared using global features. In PRF assumption, the top-ranked images in text-based search result are rigidly assumed as relevant. Similar to relevance feedback, classification is encouraged by this assumption. Thus a number of approaches [43], [10], [12], [35], [31] propose to learn one or more classifiers, discriminating the top-ranked images and sampled negatives. Because the learned classifiers are mostly based on the whole image other than the objects, these approaches will also incur the ineffectiveness for object queries. Besides directly deriving ranking functions using the above assumptions, the approaches in [28] and [39] adopt these assumptions to predict the difficulty of reranking and judge whether the reranking algorithm should be applied to the query.

Some recent methods suggest that the ranking model trained on a representative query set through human labeling can be adapted to other unlabeled queries, named "Supervised Reranking". Such a ranking model combines multiple ranking features into relevance score and the existing approaches differ mostly on how to design the features. For example, [45] and [23] manually design the ranking features based on empirical domain knowledge. On the contrary, in [46], the results of multiple PRF classifiers trained on different feedback levels are adopted as ranking features. In this paper, the demonstrated reranking approach also adopts a weight model to fuse the risks derived by different assumptions, following the supervised fashion.

4) Image Search Result Summarization: Numerous previous works are proposed to automatically summarize a given image collection from different aspects [36], [20], [8], [49], [33], in which image relevance and diversity are always taken into account as two key priors. The authors of [36] focus on scene summarization and develop a greedy algorithm to solve the proposed optimization on diversity and coverage. In [20] the author clusters landmark images using K-Means for each image group and select image exemplars based on visual word coherence, while [8] represents scene images with visual words and iteratively select images with maximal coverage to the most informative visual words. Besides diversity and relevance, the authors of [30] introduce image quality as another informative prior so as to improve the user experience. The search result summarization approach in this paper implements the three priors using object appearance. Typically, we adopt object size and location to indicate the image quality.

III. Approach

The first part of this section introduces how to construct the proposed object vocabulary, where the proposed retrieval model is based. In the second part, we present the bag-of-objects retrieval model using the constructed object vocabulary.

A. Query-Context Object Vocabulary

The entire pipeline of the proposed object vocabulary construction approach is illustrated in Fig. 3. To build the object vocabulary, we first enrich the query by associating some highly frequent text terms. Then, we apply a common object discovery method on the result image collection to discovery the underlying object categories. Finally, each object category is modeled by a linear classifier trained using multi-class support vector machine (SVM).

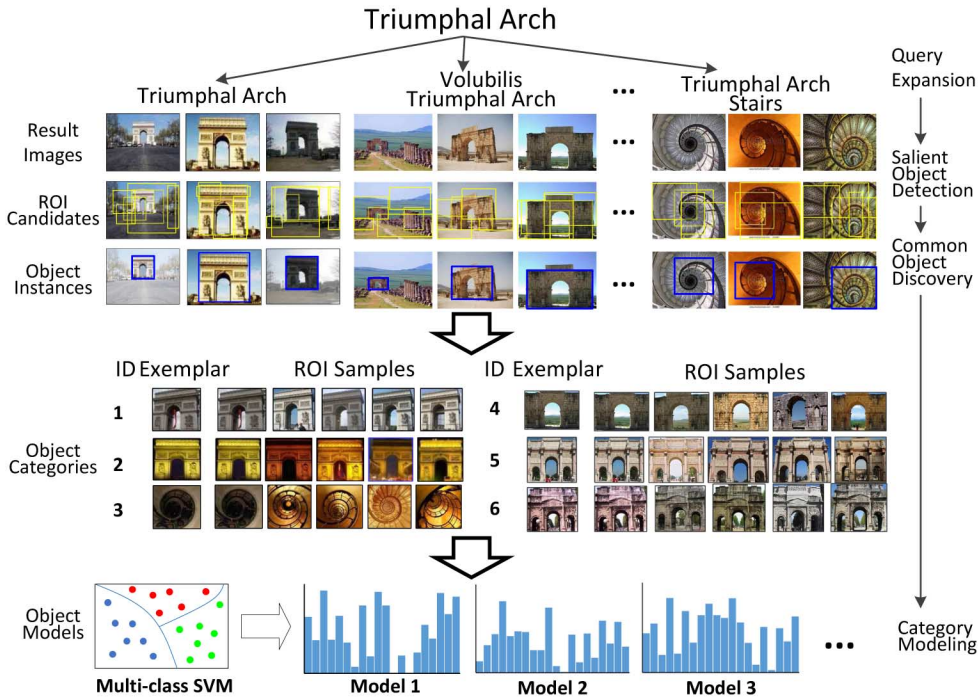


Fig. 3. Pipeline of the proposed object vocabulary construction approach, illustrated with query "Triumphal Arch" as an example.

1) Query Expansion: Given a query q the goal of this step is to discover a set of key words $\mathcal{Q}' = \{q'_i\}$, which are semantically related to the query. We assume such kind of keywords should be frequent in the surrounding text of the result images. Therefore, it is natural to design the selection criterion following TF-IDF [4]. Assume query q returns a set of images $\mathcal{D} = \{d\}$, the confidence $f(t;q)$ for text term t to be chosen is calculated as follows:

$$f(t;q) = \frac{\sum_{d\in\mathcal{D}} \omega_d f(t;d)}{\sum_{d\in\mathcal{D}} \omega_d} \tag{1}$$

$$f(t;d) = \frac{\langle d,t\rangle}{\sum_{t'\in d}\langle d,t'\rangle} \log \frac{|\mathcal{D}|}{|\{d' : \langle d',t\rangle > 0\}|}. \tag{2}$$

Equation (2) is the standard form of TD-IDF calculation, where $\langle d,t\rangle$ stands for the frequency of t in image d's surrounding text. $f(t;d)$ indicates the confidence that text term t is related to image d. The motivation of Eq. (1) is to suppress the noisy terms from irrelevant images using the discount weight w_d. According the fact that images with lower rankings are more likely to be irrelevant than top images, ω_d should be of negative correlation to image d's ranking position. In this paper, ω_d is calculated as $\omega_d = \frac{1}{\log(1+r_d)}$, where r_d is the ranking position of d. In order to suppress the noise caused by daily-used words and phrases, we employ a stop-word list to eliminate such terms as well as eliminating all the verbs, which are regarded as non-informative for object queries. Then all the text-terms are ranked according to $f(t;q)$ and the expended query set $\mathcal{Q}_f$ is generated by selecting the top terms. To further complement the expanded query set, we also use the related terms suggested by the search engine Bing, denoting as $\mathcal{Q}_b$. At last, $\mathcal{Q} = \mathcal{Q}_b \cup \mathcal{Q}_f \cup \{q\}$ is adopted as the final expansion query set. Fig. 4 shows the query expansion result for the query "Notre Dame". While frequency-based expansion can capture a number of related terms, search engine Bing brings in some extra ambiguous keywords such as "Notre Dame Football", enabling the categorization of the football player in image **C** and the team flag in image **F**.

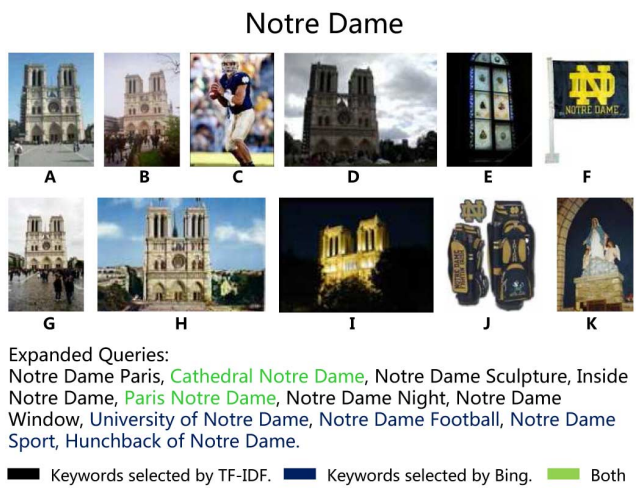


Fig. 4. An example of query expansion. Upper part: An overview of the search result for the query "Notre Dame." Lower part: The query expansion result for the query "Notre Dame."

After $\mathcal{Q}$ is constructed, we issue all queries in $\mathcal{Q}$ to a search engine and gather the result images. For the convenience of following descriptions, we denote the result image collection for the original query q as $\mathcal{D}$ while the result image collection for $\mathcal{Q}$ is denoted as $\mathcal{D}'$.

2) Object Category Mining: Each object in the vocabulary is supposed to be either an instance of the query object or an

appearances from a certain aspect or with different illumination conditions. For example, if the query is "car", the object can be "BMW Q5" or "Audi A6". For query "Arc de triomphe", the objects can be photos taken from the front side, left side, or taken at night.

The upper part of Fig. 3 shows the procedure of object category mining. At the very beginning, a group of salient ROIs are extracted from each image using the salient object detection approach in [9]. Each image d is then represented as a bag of ROIs, denoted as $\mathbf{d} = \{r_i^d\}$. Then the true object ROIs can be further located using common object discovery. In this paper, we employ the method described in [21] to achieve this, which alternatively executes the exemplar seeking procedure and the ROI refinement procedure until convergence. In the exemplar seeking procedure, the foreground ROIs detected in the last iteration are clustered using a MeanShift-like approach and the cluster centers are adopted as exemplars to represent the foreground objects. The ROI refinement procedure selects one ROI from each image bag as the new foreground candidate according the selected ROI exemplars. In [21], an augmented bipartite graph is constructed between the exemplars and all the ROIs in the image bag. Then the PageRank [3] algorithm is applied to the graph and the ROI with the highest PageRank score is selected. In this paper, we extract at most 100 ROIs from each image using the saliency object detection method proposed in [9] and make a modification on the exemplar seeking procedure. Instead of using MeanShift, the exemplars are discovered by Affinity Propagation proposed in [11], which is considered more effective and parameter-free.

After the common object discovery procedure, the foreground ROI of each image is confirmed and all the foreground ROIs are organized into several categories by assigning it to the nearest hub. In convenience, we denote the i_{th} category as c_i and the vocabulary can be expressed as $\mathcal{C} = \{c_i\}$.

3) Object Category Modeling: The task of this step is to train a model for each object category, such that the category can provide a prediction score for each image, indicating the confidence that the image is containing the object. To achieve this, we first cluster the foreground ROIs detected in the last iteration by assigning each ROI to its nearest exemplar. Then, we adopt the ROIs' visual features as training samples and train a SVM classifier for each category using the one-versus-all strategy, where the category's belonging ROIs are taken as positive samples while ROIs from the other categories are used as negative samples. For a given ROI r, we denote the confidence score predicted by the classifier of category c_i as $g_i(r)$, and the prediction score $g_i(\mathbf{d})$ for image bag $\mathbf{d}$ is calculated as $g_i(\mathbf{d}) = \max_{r\in\mathbf{d}} g_i(r)$.

B. Object-Based Retrieval Model

With the trained object vocabulary, we can represent the images and the query itself in the form of bag-of-objects. The proposed retrieval model in this paper follows the risk minimization framework proposed in [24], where the ranking objective of image d is estimated by the risk of returning d to the given query q. The risk is formulate as the expectation of loss function L over the space of query language model θ_Q and document language model θ_D, expressed as follows:

$$\begin{aligned} R(\boldsymbol{d};\boldsymbol{q}) &= R(a=\boldsymbol{d}|\boldsymbol{q},\mathcal{D}) \\ &= \sum_{r\in 0,1}\int_{\theta_Q}\int_{\theta_D} L(\theta_Q,\theta_D,r)\times p(\theta_Q|\boldsymbol{q}) \\ &\quad \cdot p(\theta_D|\boldsymbol{d})p(r|\theta_Q,\theta_D)d\theta_Q d\theta_D, \end{aligned} \tag{3}$$

where $a = \mathbf{d}$ stands for the action of return d and $\mathcal{D}$ is the image document collection in the database. Here, we assume that the loss function $L(\theta_Q,\theta_D,r)$ does not depend on the query-image relevance and define L as the Kullback-Leibler divergence, written as follows:

$$\Delta(\theta_Q,\theta_D) = \sum_{i=1}^{|\mathcal{C}|} p(c_i|\theta_Q)\log\frac{p(c_i|\theta_Q)}{p(c_i|\theta_D)}. \tag{4}$$

After some derivations, the risk $R(\boldsymbol{d};\boldsymbol{q})$ can be expressed as follows:

$$R(\boldsymbol{d};\boldsymbol{q}) \propto -\sum_{i=1}^{|\mathcal{C}|} p(c_i|\hat{\theta}_Q)\log p(c_i|\hat{\theta}_D) + \xi_q, \tag{5}$$

where $\hat{\theta}_Q$ and $\hat{\theta}_D$ are the *the maximum a posteriori estimation* of the query language model and document language model. ξ_q is a constant which can be simply ignored. Finally, the relevance of image document $\boldsymbol{d}$ can be predict using the inverted the risk. In the following sections, we are going to illustrate the estimation of document language model θ_D and query language model θ_Q, which are requested in Equation (5).

1) Document Language Model: We assume the document language model to follow the following distribution:

$$p(\mathbf{d}|\theta_d) = \prod_{i=1}^{|\mathcal{C}|} p(c_i|\theta_d)^{r(c_i,\mathbf{d})} \tag{6}$$

where c_i stands for the i_{th} object in the object vocabulary. $r(c_i,\mathbf{d})$ is the confidence that image $\mathbf{d}$ contains the object c_i. Here, we define r as follows:

$$r(c_i,\mathbf{d}) = \max\{0, g_i(\mathbf{d})\}. \tag{7}$$

Then, by applying MLE to equation (6), the document language model is derived as follows:

$$p(c_i|\hat{\theta}_d) = \frac{r(c_i,\mathbf{d})}{\sum_{j=1}^{|\mathcal{C}|} r(c_j,\mathbf{d})}. \tag{8}$$

2) Query Language Model: Because the queries in this paper is given in the form of key words, the object-based language model cannot be directly derived based on visual appearance. However, the query language model can be approximated if we predict the relevance of each category using some effective assumptions. We assume that the query language model follows the distribution below:

$$p(\mathbf{q}|\theta_q) = \prod_{i=1}^{|\mathcal{C}|} p(c_i|\theta_q)^{s(c_i,\mathbf{q})} \tag{9}$$

where the score $s(c_i, \mathbf{q})$ indicates the relevance of object c_i to query $\mathbf{q}$. By MLE, the query language model is then derived as follows:

$$p(c_i|\hat{\theta}_q) = \frac{s(c_i, \mathbf{q})}{\sum_{j=1}^{|\mathcal{C}|} s(c_j, \mathbf{q})}. \tag{10}$$

To predict the relevance score $s(c_i, \mathbf{q})$, we propose five assumptions from different aspects such as PRF and visual density. Each assumption can approximate the query language individually and the assumptions can also be combined using the approach presented in Section IV-A. The proposed assumptions are illustrated below with the calculation methods of $s(c_i, \mathbf{q})$ described in details.

- **Pseudo Relevance Feedback:** If the top k images in $\mathcal{D}$ are assumed as pseudo relevant, the prediction score for each object category is calculated by accumulating all positive responses from the k pseudo relevant images, formulated as follows:

$$s_{\mathrm{PRF}}(c_i, \mathbf{q}) = \sum_{d_j \in \mathcal{D}} max\{0, g_i(d_j)\delta(t_j \leq k)\} \tag{11}$$

  where t_j denotes the ranking position of the j_{th} image in the result list from the search engine and $g_i(d_j)$ stands for the prediction score of image j predicted by the classifier of the i_{th} object category.
- **Nearest Neighbor Hard Voting:** According to the cluster assumption, if the object has more "sponsors" in the result image collection $\mathcal{D}$, it is more likely to be relevant. Here, we gather the sponsors of each object by assigning each image to its closest object category. The calculation of relevance score $s(c_i, \mathbf{q})$ is formulated as follows:

$$\begin{aligned} s_{\mathrm{HV}}(c_i, \mathbf{q}) &= \sum_{d \in \mathcal{D}} \delta(NN(d) = c_i) \\ NN(d) &= \arg\max_{c \in \mathbf{C}} r(c_i, d) \end{aligned} \tag{12}$$

  where (C) stands for the object vocabulary and the function $NN(d)$ returns the nearest object category for image d.
- **Visual Density:** This assumption is also based on the cluster assumption. Instead of counting the number of "sponsors", it evaluate whether the "sponsors" are close to each other. The relevance score here is estimated by kernel density estimation (KDE), formulated as follows:

$$\begin{aligned} s_{\mathrm{VD}}(c_i, \mathbf{q}) &= \frac{1}{|\mathbf{h_{c_i}}|^2} \sum_{x \in \mathbf{h_{c_i}}} \sum_{y \in \mathbf{h_{c_i}}} \langle x, y \rangle \\ \mathbf{h}_{c_i} &= \{d \in \mathcal{D} | NN(d) = c_i\} \end{aligned} \tag{13}$$

  where $\langle x, y \rangle$ stands for the visual similarity of image x and y, and $\mathbf{h}_{c_i}$ is the collection of the images assigned to object category c_i.
- **Saliency, Size and Location:** These three assumptions are inferred from the fact that the relevant images for object queries usually have a single foreground with clear object appearance. Therefore, if a category is mostly associated with such images, it is intuitively relevant. First, saliency suggests the "significance" of the image region, where highly salient ROIs always lead to a clear foreground object. Second, if an object occupies the most of the image area, it is probably an important part of what image tells. Last, the location of the object directly reflects whether the photographer is intentionally taking it, because in that case, the camera would be definitely aimed down to the object. Here, we can generate three different prediction scores respectively based on saliency, size and location. After the result images are categorized, the prediction score for each object category is calculated by averaging the attributes from its belonging ROIs. For object size, the ROI's size of each image is normalized by the image size before averaging. For object location, we calculate the L1 distance between the ROI center and the image center.
- **Normalized Google Distance:** This assumption aims to solve the ambiguous problem mentioned in Section I by using side information. Normalized Google Distance (NGD) [5] measures the relationship between two text terms according to their coherence in the search engine. Based on a huge amount of data, NGD can partially indicate the semantic distance between two text terms. The calculation of NGD is expressed as follows:

$$NGD(x, y) = \frac{\max\{\log f(x), f(y)\} - \log f(x, y)}{\log M - \min\{\log f(x), \log f(y)\}} \tag{14}$$

  where $f(x)$ stands for the frequency of term x while $f(x, y)$ stands for joint frequent of text term x and y. The constant M stands for the total amount of document being indexed by the search engine. To predict the relevance of each object category, we first calculated the NGDs between expended queries in $\mathcal{Q}$ and the original query q. Then, for each category, $s(c_i, \mathbf{q})$ is calculated as follows:

$$s_{\mathrm{NGD}}(c_i, \mathbf{q}) = \frac{\sum_{j \in \mathbf{h}_{c_i}} NGD(q^j, c_i)\langle q, h_i \rangle}{\sum_{j \in \mathbf{h}_{c_i}} \langle q, h_i \rangle} \tag{15}$$

  where h_i is the exemplar for the i_{th} category while q^j stands for the query where image j is from.

IV. Applications

In order to illustrate the effectiveness of the proposed object-based retrieval model, we demonstrate two applications in this paper: image search reranking and image search result summarization.

A. Image Search Reranking

Since image search reranking sorts the result images only based on image relevance, the search results can be reranked using the retrieval risk in equation (3) as ranking score. However, to generate more precise ranking results, it is wiser to take the advantages of different assumptions on the query language model. That is, to solve the image search reranking problem in a supervised way. According to [47], the ranking function of supervised method can be formulated as follows:

$$\begin{aligned} f &= h \circ \mathbf{g}(\mathbf{d}, \mathcal{D}, \mathbf{I}, q) \\ &= h(\mathbf{g}(\mathbf{d}, \mathcal{D}, \mathbf{I}, q)) \end{aligned} \tag{16}$$

where $\mathbf{g} = \{g_k\}$ are referred as ranking features, which is learned from the result images' visual features as well as the rankings return by the text-based search engines, partially indicating the image's relevance to query. In this paper, the ranking features $\{g_k\}$ are implemented by the risks calculated by different assumptions on the query language models. Inspired by [46], we generate 200 different risks from the PRF assumption by enumerating k from 1 to 200.

Function h in equation (16) is a query-independent model which integrates the ranking features to produce the final ranking scores. Since it is impossible to assign a specific model for each query, h is trained from a set of representative queries using human-labeled image relevance. Usually, h is assumed as a linear model, hence the ranking function can be derived as

$$f = -\sum_i \omega_i R_i(\mathbf{d};\mathbf{q}) \tag{17}$$

where $R_i(\mathbf{d};\mathbf{q})$ stands for the retrieval risk calculated using the i_{th} assumption. The calculation of optimal weight vector $\boldsymbol{\omega}$ can be reduced to a learning-to-rank problem [18]. In this paper we adopt a learning-to-rank model called Ranking SVM, which adapts the well-known SVM classifier to the task of ranking. It decompose the rankings into a set of ordered pairs and the ranking problem is reduced to pair-wise classification problems. The optimization objective of Ranking SVM is formulated as follows:

$$\begin{aligned} \min \quad & \omega^T\omega + C\sum \varepsilon^i_{x,y}, \\ s.t. \quad & \forall q_i, k_x \prec k_y : \omega x - \omega y > 1 - \varepsilon^i_{x,y}, \\ & \forall x, y, i : \varepsilon^i_{x,y} > 0 \end{aligned} \tag{18}$$

where k_x and k_y is the manually labeled relevance for image x and y. The optimization can be efficiently solved using SMO (Subsequence Minimal Optimization) or cutting-plane algorithm. In this paper, we employ the software provided in [17] for our ranking process.

B. Image Search Result Summarization

The task of image search result summarization is to recommend k representative images as an overview the image search result. According to the definition of "representativeness" in previous work on result summarization [8], [49], [33], the selected images should not only be relevant to the query but also diverse enough to cover all concepts presented by the image search results. In this section, we introduce a novel summarization method based on the proposed object retrieval model. With the constructed object vocabulary, the summarizing images are delivered by following steps: First, all object categories in the vocabulary are ranked according to the proposed criterion based on relevance and diversity before the top k categories are selected. For each category, the images are ranked based on both relevance and attractiveness.

1) Category Selection: To accurately estimate the relevance of each object category, we export the weight vector ω trained for image search reranking, and combine the relevance scores from different assumptions. The relevance of the i_{th} object category is calculated as follows:

$$rel(c_i;q) = \sum_i \omega_i s_i(c_i;q \tag{19}$$

where $s_i(c_i;q)$ stands for the relevance score predicted by the i_{th} assumptions.

To guarantee the result diversity, we adopt Non-Maximum Suppression (NMS) in the selection process. For each category c_i, the ranking score is calculated by the distance between c_i and the closest category which has higher relevance than c_i, expressed as follows:

$$\begin{aligned} rel(c_i;q) &= \min_{c_j \in P_{c_i}} \{\langle c_i, c_j\rangle\}, \\ P_{c_i} &= \{c_j | rel(c_j;q) > rel(c_i;q)\}. \end{aligned} \tag{20}$$

The term $\langle c_i, c_j\rangle$ in above equation stands for the distance between category c_i and c_j, which is approximated by the visual appearance of their ROI exemplars in this paper. After that, the categories are sorted in descending order and the top k categories are selected for the next step.

2) Image Selection: For each category, the goal of this step is to select the most suitable image as its representation. For an image b and an object category c_i, the representativeness of b to c_i is measured by the prediction score of b on category c_i, with comparison to the prediction scores on other categories. We denote the score of b predicted by the model of the i_{th} category as $g_i(b)$, and the representativeness $rep(b;c_i)$ is calculated as follows:

$$rep(b;c_i) = g_i(b) - \max_{j\neq i}\{\max\{0, g_j(b)\}\}. \tag{21}$$

Besides representativeness, we realize that the attractiveness of the image is also an important factor for user experience. For example, Fig. 5 shows three images of the Triumphal Arch in the dark with the same illumination condition. Image **B** and **C** illustrate two images which are unsatisfying to the user: image **B** presents the arch in a very small region while image **C** locates the Triumphal Arch on its very corner. On the contrary, the users would definitely prefer image A because the arch is located on image center while occupying most of the image area. Motivated by this, we assign each image with an attractiveness score based on object location and size, expressed as follows:

$$attr(b) = e^{-\frac{d}{\sqrt{S}}} \cdot \frac{S_f}{S}, \tag{22}$$

Fig. 5. Three example images of "Triumphal Arch." The users usually prefer image **A** and feel uncomfortable when seeing image **B** and **C**.

where d represents the normalized L1 distance between the ROI center and image center. S_f stands for the size of the foreground ROI while S denotes the image size.

Combining both representativeness and object appearance, the final preference score of each image is calculated as follows:

$$pref(b, c_i) = \alpha \frac{rep(b, c_i)}{\overline{rep}(c_i)} + (1 - \alpha) \frac{attr(b, c_i)}{\overline{attr}(c_i)} \tag{23}$$

where $\overline{rep}(c_i)$ and $\overline{apr}(c_i)$ are the average of representativeness and object appearance within category c_i, playing as normalization terms. α is the factor to trade-off between representativeness and image appearance. In our experiment, α is empirically set to 0.2.

V. Experiments and Results

To evaluate the effectiveness of the proposed retrieval model, we test the two demonstrated applications on a subset of a publicly available dataset, comprising the queries of objects. For image search reranking, we compare our approach with various baseline approaches including the results from the search engine, unsupervised and supervised methods. For image search result summarization, we design a user study to evaluate the summarized search result by the proposed methods as well as four baseline methods.

A. Image Search Reranking

1) Dataset and Settings: We employ the "Web Queries" dataset[1] which is publicly available on the web to make our experiment reproducible. This query set consists of 353 queries, covering topics including celebrities, animals, products, landmarks, etc. The result images are collected by searching these queries using a existing image search engine and 71478 images are finally downloaded. Each image in the dataset is assigned with a binary label indicating the relevance to the query as the ground-truth. As the proposed approaches are specifically designed for object queries, we construct a subset of "Web Queries" by selecting the object queries in the dataset. The constructed subset consists of 19586 image and 101 queries, including landmarks, products, flags, and logos. With the above query set, the proposed query expansion method brings 673 extra queries to the expansion set. Then, we crawl the result images of these queries from the Bing image search engine, where for each query, we download the top 50 images from the search result list. Finally, 21047 images are successfully downloaded. The images from the object subset and the expansion set are adopted to train the object vocabulary.

The proposed reranking method is compared with multiple baseline approaches, including text-based search engine (*"Text-baseline"*) and the Bayesian reranking (*"Bayesian"*) [40], pseudo relevance feedback reranking (*"PRF"*) [43], supervised-reranking (*"Loterr"*) [45], the query-relative classifier (*"Query-relative"*) [23] and the L^2 reranking ("L^2") [46]. For PRF reranking, top ranked images are selected as positive samples while the negative samples are sampled following [46]. When evaluating Bayesian reranking, the pair-wise ranking distance and the best performing local learning consistency are adopted. For the supervised reranking approaches including L^2 reranking, superived-reranking, query-relative classifier, and our approach, the ranking models are trained by RankSVM [18]. To better evaluate these approaches, we randomly split the dataset into 10 folds and employ the cross validation strategy to train and evaluate different queries in a round robin way. In each round, 8 of 10 folds are used for training, one for parameter validation, and the rest one is used for evaluation.

We extract the Pyramid Histogram Of visual Words (PHOW) described in [2] as the visual feature representation for all the images in the dataset. Firstly, Scale-Invariant Feature Transform (SIFT) [32] descriptors are extracted on the points densely sampled from different image pyramids. Specifically, 4 SIFT descriptors on 4 different scale levels are computed on each sampled point. Then, descriptors are quantized into visual words using a codebook trained by k-means clustering. Finally, each image is represented by a spatial pyramid histogram of the quantized visual words. For all the baselines methods, the PHOW feature is extracted after resizing each image to a size of 320×240. For our method, the PHOW feature is extracted from each ROI resized to a size of 160×120. We adopt the histogram intersection kernel on PHOW feature for the computation of image similarity, which has shown generally good performance for object recognition. As a special case, we adopt linear kernel for SVM in L^2 reranking and query-relative classifier, as suggested in [23] and [46].

2) Performance Measurement: The ranking performance is measured by Average Precision (AP) and Normalized Discounted Cumulative Gain (NDCG), which are widely used to evaluate search and ranking methods. AP is defined as average of precisions at various recall levels, and MAP is the average of APs among all queries. The calculation of NDCG is expressed as follows:

$$\mathrm{NDCG@}k = \frac{\mathrm{DCG@}k}{\mathrm{IDCG@}k} \tag{24}$$

where k denotes the truncate level and $\mathrm{DCG@}k$ is formulates as follows:

$$\mathrm{DCG@}k = \sum_{i=1}^{k} \frac{2^{r_i} - 1}{\log_2(i+1)}. \tag{25}$$

In the equations above, r_i stands for the ground truth labeling of relevance for the i_{th} image, while $\mathrm{IDCG@}k$ serves as a normalization term.

3) Comparison With Conventional Methods: Table I shows the performance comparison of the different reranking methods on the adopted dataset, in the terms of MAP, NDCG@10, NDCG@25 and NDCG@40. It is obvious that all reranking methods outperform the text-baseline in MAP, for example PRF reranking improves the text-baseline by 27.66%, while Bayesian reranking improves it by 11.17%. It suggests that the visual reranking approaches are generally effective in boosting the image search ranking performance. We can also observe from the table that all supervised reranking methods have better performance than the unsupervised ones, for example "Letorr" outperforms Bayesian reranking by 15.30% and query-relative classifier outperform it by 15.92%. The above improvement on MAP shows the fact that employing human supervision to

[1]The dataset can be found at http://lear.inrialpes.fr/krapac/webqueries/webqueries.html. It was mentioned in [23] for the first time.

TABLE I
THE PERFORMANCE COMPARISON OF VARIOUS RERANKING METHODS

Methods	MAP	NDCG@10	NDCG@25	NDCG@40
Text-baseline	0.582	0.674	0.649	0.667
Bayesian [41]	0.647 (+11.17%)	0.739 (+9.64%)	0.717 (+10.48%)	0.698 (+4.65%)
PRF [43]	0.743 (+27.66%)	0.850 (+26.11%)	0.830 (+27.89%)	0.809 (+21.29%)
Letorr [45]	0.746 (+28.18%)	0.809 (+20.03%)	0.833 (+28.35%)	0.812 (+21.74%)
Query-relative [23]	0.750 (+28.87%)	0.859 (+27.45%)	0.839 (+29.28%)	0.817 (+22.49%)
L^2 [46]	0.760 (+30.58%)	0.856 (+27.00%)	0.838 (+29.12%)	0.830 (+24.44%)
Proposed method	**0.836 (+43.64%)**	**0.878 (+30.27%)**	**0.884 (+36.21%)**	**0.873 (+30.88%)**

TABLE II
THE PERFORMANCE OF DIFFERENT EXPANSION STRATEGIES

Methods	MAP
No query expansion[48]	0.812
$\mathcal{Q}_f$ only	0.829 (+2.09%)
$\mathcal{Q}_f + \mathcal{Q}_b$	0.836 (+2.96%)

integrate the ranking features from different assumptions are more effective than using any single assumption. Among all the evaluated reranking methods, our method outperforms the other five evaluated reranking methods. Specifically, it can improve the Superived-reranking, query-relative classifier and L^2 reranking by 12.06%, 11.47% and 10.00% respectively. Such a large improvement suggests that the proposed object-based retrieval model is generally effective in dealing with object queries.

4) Comparison Between Different Object Vocabularies: The motivation of this experiment is to specify how much improvement is brought by the query-context object vocabulary proposed in this paper, comparing to our previous work in [48] and how much improvement is brought by the complement terms from Bing, mentioned in Section III-AI. Table II shows the ranking performance of 1) query-relevant object vocabulary (our previous method in [48]), 2) query-context vocabulary using frequency-based query expansion ($\mathcal{Q}_f$ only) and 3) query-context using both frequency-based query expansion and complementary from Bing ($\mathcal{Q}_f + \mathcal{Q}_b$) in the term of MAP. From Table II, we can observe that the two methods using query-context object vocabulary outperform query-relevant object vocabulary by 2.09% and 2.96% respectively. These two improvements indicates that ambiguous images which are hard to be handled by query-relevant object vocabulary are successfully suppressed using query-context object vocabulary. Comparing the performance of query-context object vocabulary with different query expansion strategies, we can observe that the object vocabulary constructed using $\mathcal{Q}_f + \mathcal{Q}_b$ outperforms the one using $\mathcal{Q}_f$ only. It implies that the key words suggested by Bing play as a good complementary for those collected based on frequency.

5) Single Assumption Validation: The above result comparisons show that the proposed reranking approach is generally effective. Since our approach integrates the advantages from different kind of assumptions, it is necessary for us to further explore the contribution of each assumption category by evaluating its ranking performance. For each assumption category, we learn a ranking model using Ranking SVM and evaluate it using the round-robin cross validation process. As a complementary, a leave-one-out strategy is also adopted where we eliminate the ranking feature belonging to the assumption and evaluate the others. The performance of each assumption is compared and shown in Table III in the term of MAP while the results by the leave-one-out strategy are shown at the second row. The assumption of PRF achieved a performance of 77.62%, for a further investigation, we present a curve in Fig. 6 to show the MAP performance of each truncate level. We can observe from the figure that the MAP performance generally grows with the truncate level and becomes stable after the truncate level 70. It conforms the fact that the query language model becomes more representative by associating more result images. While the highest MAP achieved by a single truncate level is only 68.76%, the integration of multiple truncate levels outperform improves it by 7.81%. The assumptions of hard voting and visual density outperforms the text-baseline by 32.13% and 38.32%. It indicates that both the amount and quality of the "sponsors" predict reasonable query language models and make effort to the final result. According to Table III, the context information including object saliency, location and size also improves the final rankings. We believe the effectiveness of such context is due to the human habit on taking photos, people trends to locate the desired object on the distinct position. In Table III, we can observe that the assumption on NGD makes the highest contribution, it suggest that the ambiguous problem mentioned in Section I can be solved using extra information from the search engine other than visual features.

6) Failure Analysis: Here we analyze several typical cases where our approach fails. Our approach achieves an MAP of 0.41 on the query of "French Stadium". We believe the low performance is because few of the images are taken outside the stadium, while most of them are taken indoor, telling various indoor sights of the stadium. Thus, the employed ROI based common object discovery method cannot detect any objects from them, and the proposed retrieval model consequently fails to predict correct image relevance. The query "Pear" returns a low MAP of 0.29. We can observe from Fig. 7 that its result images includes various concepts like "drawing of pear", "pear trees", "pear pies". Although the expanded query set has covered these concepts, the downloaded result images comprise too much gestures of pears and pear trees. Without consistent foreground appearance, the common object discovery method fails to category them, leading the object vocabulary less effective than expected.

TABLE III
THE PERFORMANCE OF EACH INDIVIDUAL OBJECT ATTRIBUTE

Assumptions	PRF	Hard Voting	Visual Density	Sz., Loc. and Sal.	NGD
MAP	0.771	0.769	0.805	0.755	0.811
MAP leave-one-out	0.817	0.822	0.803	0.828	0.824

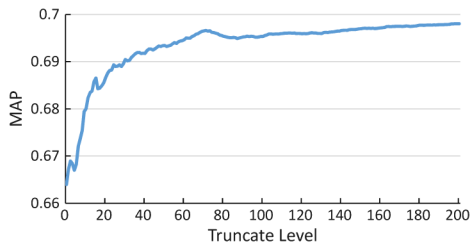


Fig. 6. MAP performance of the PRF assumption on different truncate levels.

B. Image Search Result Summarization

1) Dataset and Settings: This experiment adopts the same dataset of the reranking experiment. Considering the workload limit of the user study, we randomly select 50 queries from the dataset, and the sampled dataset consists of 9632 images in total. The weight vector ω for different query language models is exported from the Ranking SVM model trained in the reranking experiment. Because the reranking performance is evaluated in a round-robin validation framework, for each query, we export ω from the fold where the query plays as a test sample.

We compare our summarization approach with following baseline approaches:

- *TOP_5_T*. In this baseline, the top 5 images from the text-based search result are taken as the summarization.
- *TOP_5_R*. Similar with above, the search result is summarized using the top 5 images of the refined result list using the proposed reranking method.
- *APSP*. This baseline applies Affinity Propagation (AP) [11] on the result image collection and the exemplars are ranked by the confidence suggested in [11].
- *REP_ONLY*. This baseline is a degraded version of our approach, where in image ranking process, only the representativeness is considered.

2) User Study: To quantify the summarization performances of mentioned approaches from the user's perspective, we invite nine participants with diverse social backgrounds to take part in our user study, including 7 males and 2 females.

We design an user interface for the user study process, where the users can navigate through the queries and compare the summarization results. The screen capture of our user interface is shown in Fig. 8. At the left of the screen, we show the entire search result image collection of the query such that the users can scan the result images for a complete understanding of query's concept. The summarized result from all evaluated methods are listed on the right. In order to avoid the user's preference on a certain approach, we hidden the approach names and randomly shuffle the presentation positions of different summarization results each time. Below each summarization result, we ask the participants following questions:

- **Relevance**: How many images in the summarization are relevant to the query? (0-5 points)
- **Diversity**: How many different object appearances are there in the summarization? (1-5 points)
- **Appearance**: How many images are of proper appearances to you? (0-5 points)
- **Best Votes**: Within all the compared summarizations, is this one the most satisfactory to you? (yes/no)

3) Results and Analysis: The ratings from all participants are gathered and Table IV shows a summarization of the user study. The first three columns shows the average scores of the corresponding questions while the last column shows the frequency that the summarization result is voted as the best.

We can see from Table IV that *TOP_5_R* outperforms *TOP_5_T* on relevance. It implies again that the proposed object retrieval model is effective in prediction image relevance. However, these two approaches generate much less diverse result than other approach thus obtain few best votes, suggesting that image search result summarization is necessary because the users are not able to collect enough information about the query when looking at the top result images. We can also observe from Table IV that *APSP* get a much lower rating on diversity than *REP_ONLY* and the proposed method. The reason of the low performance is two-fold. First, without concentration on objects, the *APSP* method might select non-informative images of simple textures and colors, because of their unique appearances. Second, images with a same object located on different positions are probably regarded as different concepts if images are simply compared using global appearance. Therefore, the proposed methods are more flexible in dealing with object queries. Comparing to *REP_ONLY*, the full version of the proposed method get higher scores on image appearance as well as the frequency of best ratings. This implies that the summarization can give a better impression to the user if object layout is considered.

Notice that the proposed method is not the best one on relevance and diversity. As a matter of fact, sometimes the proposed approach has to trade off between the three key criterion to avoid bad ratings from the users. For example, the proposed method selects the images with better appearance while taking the risk that some of the images may not be representative for its category. On the contrary, *TOP_5_R* only focus on image relevance but obtains low diversity score because there are too many duplicate images in the summarization. From Table IV, we can observe that our approach gets the highest "Best Votes", implying such a trade-off is generally effective from the users' perspectives.

Fig. 9 shows the summarization results of two queries, comparing *APSP*, *REP_ONLY* and the proposed method. For the query "Windows Logo", APSP generates a less diverse result because it collects three images with a same version of windows

Fig. 7. Comparison of L^2 reranking and the proposed approach on the top 15 result images of some example queries. The first four queries presents the successful rankings of our approach, while the last two illustrate the failures. From our observation, the proposed approach can outperform L^2 reranking if the query objects are successfully mined. On the contrary, the employ common object discovery method may probably fail due to inconsistent object appearances.

Fig. 8. Screen capture of the user study GUI used in the experiment for image search result summarization.

TABLE IV
COMPARISON OF VARIANT SEARCH RESULT SUMMARIZATION METHODS BY THE CONDUCTED USER STUDY

Method	Rel.	Div.	App.	Best Votes
TOP_5_T	4.50	3.2	3.92	4%
TOP_5_R	4.76	3.04	3.72	6%
APSP	4.38	3.54	4.18	20%
REP_ONLY	4.48	4.22	3.70	14%
Proposed	**4.56**	**4.08**	**4.50**	**56%**

logo. Since images are compared with global appearance, this method treats images with the same object but different locations as different kind of images. On the contrary, object-based methods capture different versions of windows logo. We can

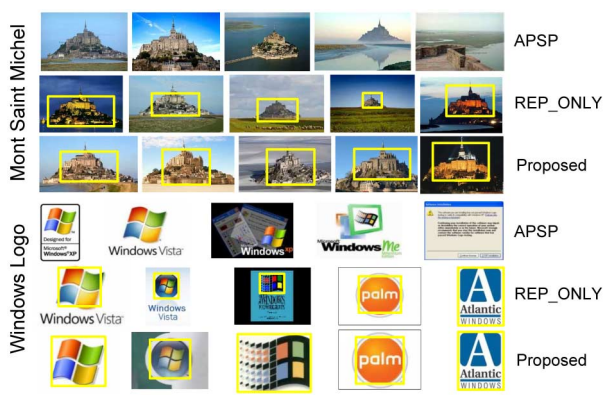


Fig. 9. Two examples of summarization results with comparison of: 1) APSP; 2) REP_ONLY; and 3) the proposed method. For REP_ONLY and the proposed method, the localized object ROI is marked yellow.

also observe that the proposed method selects images with more clear objects presented. For example, in "Mont Saint Michel", *REP_ONLY* picks up two images taken far away from the island in the summarization, in which the castle is not clearly presented, while the full version provides the objects with normal size. For "Windows Logo", the full version selects an image presenting the logo of Windows 98 only while *REP_ONLY* find one with lots of text descriptions. According to the two comparisons above, we can conclude that images with better object layout can make the summarization more comprehensible.

VI. Conclusion

Relevance prediction is the one of biggest challenges in image search reranking and image search result summarization. We observed that existing assumptions operating on the whole image cannot sufficiently deal in predicting image relevance for object queries. Motivated by this observation, we proposed a novel bag-of-object retrieval model to give a more accurate prediction of relevance. We associated a group of high frequency text terms as expanded query set and issue them the Bing image search engine. Then we constructed the query-context object vocabulary by applying the employed common object discovery approach on the result image collection. With the query language model delivered by given assumptions, the retrieval model was derived via a risk minimization framework.

Based on the proposed retrieval model, this paper also gave two solutions to image search reranking and image search result summarization. Since all previous image search rerankng method builds one generic model to all kinds of queries, this paper serves as the first attempt to deal the queries from a specific domain. We suggested that the retrieval risk delivered by a single assumption is not enough and proposed to integrate the risks from different assumptions where the combination model is trained using learning-to-rank methods from human labeled data. To provide satisfying summarized search result, we proposed a two-step ranking process. We considered both relevance and diversity in ranking object categories and the object layout was taken into account while selecting the most representative image for each category.

We believe that focusing on object queries is a promising direction for further advancing image search reranking and we envision the work in the future as follows: First, we will systematically classify queries into different domains regarding the possibility of image search reranking, and then develop algorithms to solve them respectively. Second, motivated by the object bank image representation [26], we may combine the object vocabulary discovered for the query and the objects from the collection to seek a more comprehensive representation of images and queries. Finally, we hope to identify and address the system challenges so as to most efficiently integrate this algorithm into a real-world image search engine.

References

[1] W. Bian and D. Tao, "Biased discriminant euclidean embedding for content-based image retrieval," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 545–554, Feb. 2010.

[2] A. Bosch, A. Zisserman, and X. Muoz, "Image classification using random forests and ferns," in *Proc. IEEE 11th Int. Conf. Comput. Vision*, 2007, pp. 1–8.

[3] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Comput. Netw. ISDN Syst.*, vol. 30, no. 1-7, pp. 107–117, 1998.

[4] G. Chowdhury, *Introduction to Modern Information Retrieval*. London, U.K.: Facet Publishing, 2010.

[5] R. L. Cilibrasi and P. M. Vitanyi, "The Google similarity distance," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 370–383, Mar. 2007.

[6] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Workshop Statist. Learn. Comput. Vision ECCV*, 2004, vol. 1, p. 22.

[7] T. Deselaers, B. Alexe, and V. Ferrari, "Localizing objects while learning their appearance," in *Proc. ECCV*, 2010, pp. 452–466.

[8] J. Fan, Y. Gao, H. Luo, D. A. Keim, and Z. Li, "A novel approach to enable semantic, and visual image summarization for exploratory image search," in *Proc. 1st ACM Int. Conf. Multimedia Inf. Retrieval*, 2008, pp. 358–365.

[9] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun, "Salient object detection by composition," in *Proc. 2011 IEEE Int. Conf. Comput. Vision*, 2011, pp. 1028–1035.

[10] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from Google's image search," in *Proc. 10th IEEE Int. Conf. Comput. Vision*, 2005, vol. 2, pp. 1816–1823.

[11] B. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[12] M. Fritz and B. Schiele, "Decomposition, discovery and detection of visual categories using topic models," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2008, pp. 1–8.

[13] D. Hochbaum and V. Singh, "An efficient algorithm for co-segmentation," in *Proc. IEEE 12th Int. Conf. Comput. Vision*, 2009, pp. 269–276.

[14] L. Hohl, F. Souvannavong, B. Merialdo, and B. Huet, "Enhancing latent semantic analysis video object retrieval with structural information," in *Proc. Int. Conf. Image Process.*, 2004, vol. 3, pp. 1609–1612.

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

[16] Y. Jing and S. Baluja, "Visualrank: Applying pagerank to large-scale image search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1877–1890, Nov. 2008.

[17] T. Joachims, "Making large-scale SVM learning practical," *Advances Kernel Methods Support Vector Learn.*, pp. 169–184, 1999.

[18] T. Joachims, "Training linear SVMs in linear time," in *Proc. ACM SIGKDD*, 2006, pp. 217–226.

[19] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *Proc. 2010 IEEE Conf. Comput. Vision Pattern Recog.*, 2010, pp. 1943–1950.

[20] L. S. Kennedy and M. Naaman, "Generating diverse, and representative image search results for landmarks," *World Wide Web*, pp. 297–306, 2008.

[21] G. Kim and A. Torralba, "Unsupervised detection of regions of interest using iterative link analysis," in *Proc. NIPS*, 2009.

[22] G. Kim, E. Xing, L. Fei-Fei, and T. Kanade, "Distributed cosegmentation via submodular optimization on anisotropic diffusion," in *Proc. IEEE Int. Conf. Comput. Vision*, 2011, pp. 169–176.

[23] J. Krapac, M. Allan, J. Verbeek, and F. Juried, "Improving web image search results using query-relative classifiers," in *Proc. 2010 IEEE Conf. Comput. Vision Pattern Recog.*, 2010, pp. 1094–1101.

[24] J. Lafferty and C. Zhai, "Document language models, query models, and risk minimization for information retrieval," in *Proc. ACM SIGIR*, 2001, pp. 111–119.

[25] Y. Lee and K. Grauman, "Object-graphs for context-aware category discovery," in *Proc. 2010 IEEE Conf. Comput. Vision Pattern Recog.*, 2010, pp. 1–8.

[26] L. Li, H. Su, E. Xing, and L. Fei-Fei, "Object bank: A high-level image representation for scene classification and semantic feature sparsification," in *Proc. NIPS*, 2010, pp. 1378–1386.

[27] W.-J. Li and D.-Y. Yeung, "Localized content-based image retrieval through evidence region identification," in *Proc. 2009 IEEE Conf. Comput. Vision Pattern Recog.*, 2009, pp. 1666–1673.

[28] Y. Li, B. Geng, D. Tao, Z.-J. Zha, L. Yang, and C. Xu, "Difficulty guided image retrieval using linear multiple feature embedding," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1618–1630, Dec. 2012.

[29] Y.-F. Li, J. T. Kwok, I. W. Tsang, and Z.-H. Zhou, "A convex method for locating regions of interest with multi-instance learning," *Mach. Learn. Knowl. Discovery Databases*, pp. 15–30, 2009, Springer.

[30] R. Liu, L. Yang, and X.-S. Hua., "Image search result summarization with informative priors," in *Computer Vision – ACCV 2009*, ser. Lecture Notes in Computer Science. Berlin, Germany: Springer, 2010, pp. 485–495.

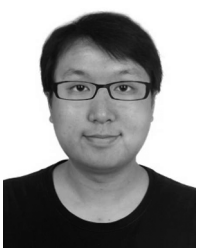
[31] Y. Liu, T. Mei, X. Hua, J. Tang, X. Wu, and S. Li, "Learning to video search rerank via pseudo preference feedback," in *Proc. 2008 IEEE Int. Conf. Multimedia Expo.*, 2008, pp. 297–300.

[32] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[33] R. Raguram and S. Lazebnik, "Computing iconic summaries of general visual concepts," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recog. Workshop*, 2008, pp. 1–8.

[34] C. Rother, T. Minka, A. Blake, and V. Kolmogorov, "Cosegmentation of image pairs by histogram matching-incorporating a global constraint into MRFs," in *Proc. 2006 IEEE Comput. Soc. Conf. Comput. Vision Pattern Recog.*, 2006, vol. 1, pp. 993–1000.

[35] F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting image databases from the web," in *Proc. IEEE 11th Int. Conf. Comput. Vision*, 2007, pp. 1–8.
[36] I. Simon, N. Snavely, and S. M. Seitz, "Scene summarization for online image collections," in *Proc. IEEE 11th Int. Conf. Comput. Vision*, 2007, pp. 1–8.
[37] D. Tao, X. Tang, X. Li, and Y. Rui, "Direct kernel biased discriminant analysis: A new content-based image retrieval relevance feedback algorithm," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 716–727, Aug. 2006.
[38] D. Tao, X. Tang, X. Li, and X. Wu, "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1088–1099, Jul. 2006.
[39] X. Tian, Y. Lu, L. Yang, and Q. Tian, "Learning to judge image search results," in *Proc. ACM Multimedia*, 2011, pp. 363–372.
[40] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X. Hua, "Bayesian video search reranking," in *Proc. ACM Multimedia*, 2008, pp. 131–140.
[41] X. Tian, L. Yang, X. Wu, and X. Hua, "Visual reranking with local learning consistency," *Advances Multimedia Model.*, pp. 163–173, 2010.
[42] S. Vicente, V. Kolmogorov, and C. Rother, "Cosegmentation revisited: Models and optimization," in *Proc. ECCV*, 2010, pp. 465–479.
[43] R. Yan, A. Hauptmann, and R. Jin, "Multimedia search with pseudo-relevance feedback," *Image Video Retrieval*, pp. 649–654, 2003.
[44] L. Yang, B. Geng, Y. Cai, A. Hanjalic, and X. Hua, "Object retrieval using visual query context," *IEEE Trans. Multimedia*, vol. 13, no. 6, pp. 1295–1307, Dec. 2011.
[45] L. Yang and A. Hanjalic, "Supervised reranking for web image search," in *Proc. ACM Multimedia*, 2010, pp. 183–192.
[46] L. Yang and A. Hanjalic, "Learning from search engine and human supervision web image search," in *Proc. ACM Multimedia*, 2011, pp. 1365–1368.
[47] L. Yang and A. Hanjalic, "Learning to rerank web images," *IEEE Multimedia*, vol. 20, no. 2, pp. 13–21, Apr.-Jun. 2013.
[48] Y. Yang, L. Yang, G. Wu, and S. Li, "A bag-of-objects retrieval model for web image search," in *Proc. ACM Multimedia*, 2012, pp. 49–58.
[49] Y. H. Yang, P. T. Wu, C. W. Lee, K. H. Lin, W. H. Hsu, and H. H. Chen, "Contextseer: Context search, and recommendation at query time for shared consumer photos," in *Proc. ACM Multimedia*, 2008, pp. 199–208.
[50] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li, "Descriptive visual words and visual phrases for image applications," in *Proc. ACM Multimedia*, 2009, pp. 75–84.



Yang Yang received the B.S degree from Nanjing University, Nanjing, China, in 2009. He is currently working towards the Ph.D. degree from the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China.

He served as a Research Intern with both the Multimedia Computing Group, Microsoft Research Asia, Beijing, China, and the Multimedia Search Team, Microsoft Bing, Beijing, China, from 2011 to 2012. He was also with the Bing Multimedia Team, Seattle, WA, USA, from September 2012 to June 2013. His research interests include multimedia search ranking, content-based image retrieval, and duplicate detection.

Linjun Yang (M'08) received the B.S. degree from East China Normal University, Shanghai, China, in 2001, the M.S. degree from Fudan University, Shanghai, China, in 2006, and the Ph.D. degree from Delft University of Technology, Delft, The Netherlands, in 2012.

Since 2006, he has been with Microsoft Research Asia, Beijing, China, where he is currently an Associate Researcher with the Media Computing Group. His current interests are in the broad areas of multimedia information retrieval, with a focus on multimedia search ranking and large-scale Web multimedia mining. He received the Best Paper Award from ACM Multimedia 2009 and the Best Student Paper Award from the ACM Conference on Information and Knowledge Management 2009. He is a member of ACM.

Gangshan Wu received the B.Sc., M.S., and Ph.D. degrees from the Department of Computer Science and Technology, Nanjing University, Nanjing, China, in 2000, 1991, and 1988, respectively. He is currently a professor of the Department of Computer Science and Technology, Nanjing University, Nanjing, China. His current research interests include multimedia content analysis, multimedia information retrieval, and digital museum.

Shipeng Li (M'97–SM'09–F'11) was a member of the technical staff for the Multimedia Technology Laboratory, Sarnoff Corporation (formerly David Sarnoff Research Center and RCA Laboratories), Princeton, NJ, USA, from October 1996 to May 1999. He joined Microsoft Research Asia (MSRA), Beijing, China, in May 1999. He is currently a Principal Researcher and Research Manager of the Media Computing Group, MSRA, Beijing, China. He also serves as the Research Area Manager coordinating the multimedia research activities at MSRA. He has been actively involved in research and development in broad multimedia areas. He has made several major contributions adopted by MPEG-4 and H.264 standards. He invented and developed the world first cost-effective high-quality legacy HDTV decoder in 1998. He started P2P streaming research at MSRA as early as August 2000. He led the building of the first working scalable video streaming prototype across the Pacific Ocean in 2001. He has been an advocate of scalable coding format and is instrumental in the SVC extension of the H.264/AVC standard. He first proposed the 694; Media 2.0 concepts that outlined the new directions of next generation internet media research (2006). He has authored and coauthored more than 200 journal and conference papers, and holds over 90 U.S. patents in image/video processing, compression and communications, digital television, multimedia, and wireless communication.