A Bag-of-Objects Retrieval Model for Web Image Search*

Yang Yang[†‡], Linjun Yang[‡], Gangshan Wu[†], Shipeng Li[‡]

[†] State Key Laboratory for Novel Software Technology, Nanjing 210046, P.R. China
[‡] Microsoft Research Asia, Beijing 100190, P.R. China
charlie.yang.nju@gmail.com, {linjuny,spli}@microsoft.com, gswu@nju.edu.cn

ABSTRACT

Image search reranking has been an active research topic in recent years to boost the performance of the existing web image search engine which is mostly based on textual metadata of images. Various approaches have been proposed to rerank images for general queries and argue that, they may not necessarily be optimal for queries in specific domain, e.g., object queries, since the reranking algorithms are operated on whole images, instead of the relevant parts of images. In this paper, we propose a novel bag-of-objects retrieval model for image search reranking of object queries. Firstly, we employ a common object discovery algorithm to discover query-relevant objects from the search results returned by text-based image search engine. Then, the query and its result images are represented as a language model on the query-relevant object vocabulary, based on which the ranking function can be derived. As the common object discovery is unreliable and may introduce noises, we propose to incorporate the attributes of the discovered objects, e.g., size, position, etc., into the ranking function through a linear model, and the weights on the object attributes can be learned. The experiments on two subsets of Web Queries dataset comprising object queries demonstrate that our approach can significantly outperform the existing reranking methods on object queries.

Categories and Subject Descriptors

H.3.3 [**Information Systems Applications**]: Retrieval models

General Terms

Algorithms, Performance, Experimentation

*This work was performed when Yang Yang was visiting Microsoft Research Asia as a research intern.

MM'12, October 29–November 2, 2012, Nara, Japan.

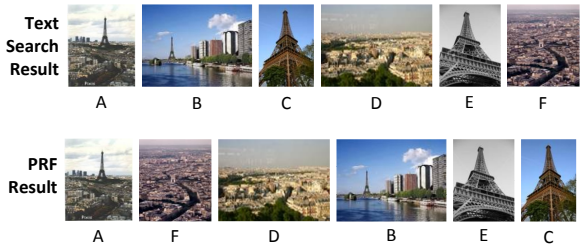

Figure 1: A reranking example for query "Eiffel Tower". The first row is the result by text based search and the second row is the result after reranking based on PRF assumption.

Keywords

Image search reranking, supervised reranking, bag-of-objects model, retrieval model

1. INTRODUCTION

As the major image search engines index and rank images mainly based on the surrounding text information, it usually leads to unsatisfactory results being returned to users. This is due to the mismatch between image content and the textual description. Image search reranking has been an active research topic in multimedia retrieval, aiming to refine the text-based search results based on cues from images' visual content.

The existing reranking approaches [6, 7, 28, 10, 11, 30, 31] are mostly based on the cluster assumption and PRF (Pseudo Relevance Feedback) assumption. The cluster assumption suggests that the relevant images are visually similar while irrelevant images have different appearances. This has been extensively adopted in various graph-based reranking algorithms based on the visual similarity graph [10, 11, 30, 31]. In the PRF assumption, the images ranked in the top of the text-based search result are regarded as pseudo-relevant, which can then be employed to learn a classifier [17] or multiple classifiers [36]. While these assumptions have been demonstrated as generally effective in the previous works [6, 7, 28], we can see that they are not sufficiently appropriate for object queries, for which users are likely to find images containing an object such as car. For such queries, the images which users are interested in may be partially related to the query. In other words, only some parts of these images are relevant to the query while the rest may not. Then, the cluster assumption or PRF assumption

which operates on the whole image cannot sufficiently capture useful information from the initial text-based search result.

Figure 1 shows an example for the query "Eiffel Tower". The upper row is the result from a text-based search engine. The lower row is the reranking result by PRF assumption where the first image is regarded as positive sample. In this case, two irrelevant images are boosted to the top, because image **F** and **D** have high visual similarity to image **A**.

We argue that the problem is mainly caused by the fact that the existing reranking approaches based on above assumptions usually employ the visual features on the whole image, such as histogram of visual words [3, 37]. However, for the object queries, i.e., the queries by which the user intends to search for images containing desired objects, including people/faces, logos, animals, buildings and industrial productions, images can be said relevant if only part of the image is about the query object. Hence, the assumptions operating on the whole image may be too rigid. For the above example, we may achieve a better performance if the reranking algorithm is performed on object level, so that images such as **B**, **C** and **E** containing the Eiffel Tower but with distinct viewpoints (or with different backgrounds) can still have a fair chance to be promoted.

Hence, we propose a bag-of-objects retrieval model to represent the query and its result images as a language model based on the containing object appearances. To make the models focus more on the query-relevant objects but be insensitive to noises or backgrounds, we represent the image and query language models on a query-relevant object vocabulary.

To construct such a query-relevant object vocabulary. we extend the iterative link analysis approach proposed in [15], which tries to mine the common objects in a set of images by a *PageRank*-like algorithm [15] on image regions. Since the text-based intial ranking can provide useful information on which image is more relevant than the others, we utilize this to improve the algorithm in [15] by considering the text-based ranking as a prior in *PageRank* to differentiate the images at different rank positions. Although the query-specific object discovery is generally effective, it may not be sufficiently reliable in complex circumstances and leads to irrelevant objects being discovered.

To estimate the relevance and confidence of discovered objects to the query, we compute a set of attribute scores based on the the position, size, and visual density of objects in containing images, etc. Then these attributes are integrated into the retrieval model so that a linear weighted ranking function is derived. The weights can be learned through RankSVM [13] from a human labeled training set.

The proposed approaches are evaluated on two subsets of the publicly available Web Queries dataset [17]. One comprises named person queries and the other comprises the other object queries. The results show that the bag-of-objects retrieval model outperform all the other reranking methods. It improves the result from the search engine by 39.52% in the term of Mean Average Precision (MAP) and the result from the state-of-the-art reranking method [36] by 6.84%.

The rest of this paper is organized as follows. After reviewing the related work on image search reranking, object-based image retrieval and common object discovery in Section 2, we describe the proposed bag-of-objects retrieval model with an illustrative example in Section 3. Section 4 presents the experimental results and analysis of our approach. In Section 5, we complete this paper with remarkable conclusions and a summary of future works.

2. RELATED WORK

In this section, we will review the related work on image search reranking, object-based image retrieval and common object detection and position the contribution of our paper with regard to these existing work.

Image Search Reranking. The existing image search reranking methods can generally be classified into unsupervised and supervised ones. The unsupervised reranking approaches are mostly based on assumptions on the structure of the initial text-based search result, among which the most well-known are cluster assumption and PRF (Pseudo-Relevance Feedback) assumption. The cluster assumption suggests that the relevant images are mostly visually similar while irrelevant images are not [30]. As the assumption can be naturally represented in a graph structure, various graph-based reranking methods are proposed to interpret this assumption from different viewpoints [30, 31, 11, 10]. Specifically, in these methods, a graph is firstly constructed with images as nodes and the edges are based on the visual similarity. Then image search reranking can be achieved by the propagation of ranking scores or positions on the graph. The main drawback of such approaches is that the visual similarity is difficult to estimate, and most of the approaches are based on the global visual features, which cannot handle object queries well, since the image similarity should be computed with regard to the relevant objects. The second is the PRF assumption, which rigidly assumes that the top-ranked images in text-based search result are relevant to the query. Based on this assumption a number of reranking models [33, 6, 7, 28, 21] are proposed, which learn a classifier by taking the top-ranked images as positive and then use this classifier to rank images. As the learned classifiers are mostly on the whole image without regard to objects in the image, these approaches will not perform sufficiently well for object queries.

Supervised reranking methods are proposed to introduce human labeling to train a reranking model which can better accord with users' perception. The different approaches differ mostly on how to derive the relevance features between the textual query and images which are in different modalities. In [35] and [17] the reranking features are usually manually designed, based on the domain knowledge of the authors on image search problem. While, in [36], the reranking features are automatically extracted by learning multiple classifiers, assuming the different importance for images with different text-based ranks. Our proposed approach follows the supervised reranking fashion, but with two major contributions to extend the existing work. First, instead of operating on the whole image as in the existing approaches, we build an object-based model so that the reranking process can be aware of the query-relevant objects. Second, in our work, we are not attempting to learn models to combine different reranking features, but targeting to learn parameters in the bag-of-objects retrieval model, where the parameters indicate the usefulness of different object attributes.

Object-based Image Retrieval. Object-based image retrieval is a well studied problem, for which the user usually provides a query image with the object of interest being

specified through e.g., a bounding box. Different approaches were proposed in the past years. In [26, 27, 23], objects on database images are manually labeled and then indexed by visual features such as color, texture and shape. In [9] images are firstly segmented into small regions and the query object is modeled based on the regions using Latent Semantic Analysis (LSA). In [34], images are represented as a bag of visual words, and then the language modeling approach to information retrieval is employed for ranking images. The authors argue that the object context is important, and thus visual words locating outside the object region should also be taken into consideration in the retrieval process, but with a discounted weight. The problem studied in our paper is different from object-based image retrieval, in which the query is specified by a keyword, instead of an example image with object of interest. Hence, our problem is more challenging because we need to infer the representation of the object of interest based on the user provided keyword query.

Common Object Discovery. Common object discovery is a recently investigated topic which aims to find the frequent objects in a group of images. The different approaches of common object discovery can be divided into two classes: segmentation based methods [16, 14, 8, 32] and bounding box based methods [15, 19, 4].

Segmentation based common object discovery, which is also known as co-segmentation, segments the frequent objects of each image simultaneously. Rother et al. [25], formulate the problem as minimizing an energy function consists of the Markov Random Field (MRF) smoothness for separating foreground and background and a histogram matching term guarantees foreground of images to be similar. In [16], the authors first segment each image into a group of super pixels and then segment the common foreground by a greedy expansion algorithm. The main drawbacks of these method are two-fold, limiting their application in out work: First, most of the co-segmentation methods are designed for images with clear foreground, which may not be effective on the web images with complicated background. Second, co-segmentation methods are very time consuming, which usually take hours when applied on hundreds of images.

Bounding box based methods aim to find the common object in the form of bounding box. These methods usually build a set of hypotheses regions of interest (ROI) based on the saliency map. In [4] 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. In [15] a link analysis algorithm is iteratively applied on all the ROIs to find the centrality as the common objects. For image search reranking, since the initial image ranking from the text-based search engine can provide useful information on the usefulness of images to the query, we extend [15] by incorporating such prior ranking information to build the query-relevant object vocabulary.

3. APPROACH

Our approach starts with the construction of query-relevant object vocabulary by mining the query-relevant objects from the images returned by the text-based search engine. Then, a bag-of-objects retrieval model is proposed to formulate the reranking problem based on the language modeling approach for information retrieval. To compute the ranking scores for each image based on the bag-of-objects retrieval model, we estimate the image and query models based on the results of query-relevant object discovery. Finally, we present the approach which learns the parameters in the retrieval model.

3.1 Query-relevant object vocabulary construction

The query-relevant objects are those different objects instances or different appearances of one object which are relevant to the query. For example, for the query "Eiffel Tower", the query-relevant objects may include the appearances of Eiffel Tower from different viewpoints or with different lighting conditions. For the query "Car", the query-relevant objects may comprise the different car instances such as "Audi A6" ad "BMW Q5".

The query-relevant object vocabulary is critical to our approach to serve as the foundation of the later processing. In this paper, we propose an algorithm based on link analysis [15] to discover those objects.

3.1.1 Algorithm overview

We first detect 30 ROIs with the highest saliency on each image using saliency object detection method proposed in [5], which are regarded as the hypotheses for query-relevant objects.

Our method is composed of two steps. In the first step, we select the qualified hypotheses that are highly confident to be query-relevant objects. The second step is to cluster the selected ROIs and use the clusters as query-relevant object vocabulary.

In the ROI selection step, the algorithm iteratively refines query-relevant ROI set until it becomes stable. In each iteration, the algorithm first recommends several representative ROIs which are considered to be the most query-relevant. These ROIs are called "hubs". Then, an ROI refinement procedure is applied on each image, where those ROIs with the highest similarity to the hubs are taken as query-relevant ROIs. The query-relevant ROIs selected by the second procedure is taken as the input of the next iteration.

3.1.2 ROI selection

In iteration t, the "hubs" are obtained using link analysis technique of *PageRank* [2]. Different from the hub-seeking procedure adopted in [15], the hubs selected in our method should not only be representative to an object, but also relevant enough to the query.

To achieve this, we construct an augmented bipartite graph $\boldsymbol{G}^{(t)}$ between $\boldsymbol{\mathcal{S}}^{(t-1)}$ and $\boldsymbol{\mathcal{C}}$, where $\boldsymbol{\mathcal{S}}^{(t-1)} = \{s_i^{(t-1)}\}$ denotes the ROIs selected in the last iteration, and $\boldsymbol{\mathcal{C}}$ is the image ranking list of the current query returned by the search engine. The *Page Rank* algorithm calculates the ranking score for each vertex in $\boldsymbol{G}^{(t)}$, where the ranking score on the vertex of an ROI intuitively shows the object confidence and query relevance. The augmented bipartite graph $\boldsymbol{G}^{(t)}$ is written as follows:

$$\boldsymbol{G}^{(t)} = \begin{bmatrix} \alpha \boldsymbol{G}_s & (1-\alpha)\boldsymbol{G}_d \\ \boldsymbol{G}_d^T & 0 \end{bmatrix}, \tag{1}$$

where $\boldsymbol{G}_s$ is a k-nearest neighbour (k-NN) self-similarity graph constructed on $\boldsymbol{\mathcal{S}}^{(t-1)}$. $\boldsymbol{G}_d$ is a bipartite graph constructed between $\boldsymbol{\mathcal{S}}^{(t-1)}$ and the ranking list $\boldsymbol{\mathcal{C}}$, where each image document is linked to all its containing ROIs appeared in $\boldsymbol{\mathcal{S}}^{(t-1)}$ with edge weight set to 1. In our experiment we set $\alpha = 0.8$.

For *PageRank* algorithm, the score vector $\boldsymbol{p}$ is updated as

follows:

$$\tilde{\boldsymbol{p}} = \beta \boldsymbol{G}\boldsymbol{p} + (1-\beta)\boldsymbol{I}. \quad (2)$$

Here, $\boldsymbol{I}$ is the vector for priori probabilities and $\boldsymbol{I} = [\boldsymbol{I}^k \ \boldsymbol{I}^d]$, where $\boldsymbol{I}_i^k = \frac{1}{|\boldsymbol{\mathcal{S}}^{(t-1)}|}$ and $\boldsymbol{I}_i^d = \frac{1}{\log(i+1)M}$ with the normalization term $M = \sum_i^{|\boldsymbol{\mathcal{C}}|} \frac{1}{\log(i+1)}$. The variable i indicates the ranking position of the image. Vector $\boldsymbol{I}^d$ is set in this way because we assume images with higher ranking from text-based search engine are more likely to be relevant to the query. In each iteration, the text-based ranking score of each image is propagated to all its linked ROIs through the bipartite graph $\boldsymbol{G}_d$.

After *PageRank* converges, we follow the hub-seeking method in [15] to find the ROIs which are diverse and also representative to the query by choosing the ROI vertices with local maxima of *PageRank* score.

The procedure of ROI refinement on each image is achieved by applying the *PageRank* on an augmented bipartite graph between the image's belonging ROIs and the hubs. From these ROIs, we selected the ones with the highest ranking score as potential object instances and use them as the input of the next iteration.

3.1.3 ROI Clustering

Since the "hubs" are regarded as typical for query-relevant objects, we treat each selected hub in the last iteration as the representation of a query-relevant object. Then, each selected ROI is assigned to its nearest hub if the distance is less than a threshold γ as the instance of the query-relevant object.

We argue that the ROI's distance to the hub is not accurate enough in measuring its confidence of belonging to a query-relevant object. Thus, we re-estimate this confidence by *PageRank*. We first construct a k-NN self-similarity graph among the ROIs within the cluster and apply the *PageRank* procedure. For an instance k_i of the object $\boldsymbol{k}$, we denote the score as $S(k_i, \boldsymbol{k})$. We call the set of discovered objects as the *object vocabulary*, and denote it as $\boldsymbol{\mathcal{K}}$.

3.1.4 An illustrative example

Figure 2 (a) shows an example process of query-relevant object discovery on the query "Arc de triumph". From each image, a group of ROI hypotheses is detected by the saliency object detection approach in [5]. We can see from the figure that the building of Arc de triumph is perfectly bounded by a certain ROI among the hypotheses. After applying iterative link analysis, these bounding boxes are selected due to similar visual appearance. On the contrary, the image of a leaf on the right has no selected ROI because it is not similar to any of the extracted hubs thus regarded as an outlier.

Figure 2 (b) shows the 4 object clusters extracted with respect to the selected hubs, where the ROI list is sorted by belonging confidence $S(k_i, \boldsymbol{k})$ in descending order. According to definition in the beginning of Section 3.1, query-relevant objects can be different object appearances as well as different objects. Due to the different view points and illuminating conditions, ROIs of "Arc de triumph" are clustered into 3 different categories. In Figure 2 (b), images of object 2 show the exact front of the arch but object 1 is taken from a different viewpoint. Object 3 comprises the images of query object in the night, and object 4 is irrelevant and showing another famous landmark "Arc de ceil". The reasons that the object 4 is detected as a hub are two-fold. First, several images of this arch exist in the search result, thus the appearance of this arc is regarded as frequent. Second, "Arc de ceil" is also an arch, and visually very similar to the query object, thus some images of "Arc de triumph" would also vote for this hub.

We can see that although above query-relevant object discovery is effective, it will also fail in some cases. Besides object 4, several ROIs not relevant to the query are also selected as instances of query-relevant objects, as shown in Figure 2 (b). Hence, we propose the bag-of-objects retrieval model to resolve this problem in the following.

3.2 Bag-of-objects retrieval model

Given the discovered query-relevant object vocabulary in the above, we can represent the images and the query as a bag-of-objects. Then the reranking problem can be formulated as a risk minimization following [18].

The ranking objective of an image $\boldsymbol{d}$ is related to the risk of returning it for a given query $\boldsymbol{q}$, which can be defined on the query and document language models:

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

Here, $a = \boldsymbol{d}$ means the action of returning the document $\boldsymbol{d}$ for the query $\boldsymbol{q}$, and $\boldsymbol{G}$ is the document collection in the database, r is the query-document relevance. θ_Q and θ_D are the language models for the query and the document, which are also called *query model* and *document model* respectively. L is the loss function which can usually be modeled by Kullback-Leibler divergence (KL divergence) between the query model and document model, written as follows:

$$\Delta(\theta_Q,\theta_D) = \sum_{i=1}^{M_{\mathcal{K}}} p(k_i|\theta_Q) \log \frac{p(k_i|\theta_Q)}{p(k_i|\theta_D)}. \quad (4)$$

Then, based on some derivations we can obtain the following ranking function:

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

where $\hat{\theta}_Q$ and $\hat{\theta}_D$ are the *the maximum a posteriori estimation* of the query and document models, and ξ_q is a constant which can be ignored for ranking. By sorting the image list with respect to the ranking function in Equation (5), we can get the reranked results.

In the following section, we are going to show how to estimate the *document model* in Section 3.3.1 and the *query model* in Section 3.3.2.

3.3 Document and query modeling

3.3.1 Document Model

In this paper, we assume our document model follows the following distribution

$$p(\boldsymbol{d}|\theta_D) \propto \prod_{i=1}^{M_d} p(d_i|\theta_D), \quad (6)$$

Figure 2: An example of Query dominant object extraction on the query "Arc de Triumph".

where d_i stands for the i_{th} object instance in image $\boldsymbol{d}$. We define

$$p(d_i|\theta_D) = \prod_{j=1}^{M_{\mathcal{K}}} p(\boldsymbol{k}_j|\boldsymbol{d})^{S(d_i,\boldsymbol{k}_i)\delta(d_i,\boldsymbol{k}_j)}, \tag{7}$$

where function $\delta(d_i, \boldsymbol{k}_j)$ indicates whether d_i is associated as an instance of object $\boldsymbol{k}_j$ in the query-relevant discovery and $S(d_i, \boldsymbol{k}_j)$ is a score obtained by the *PageRank* applied in Section 3.1.3, showing the confidence that the instance d_i belongs to object $\boldsymbol{k}_j$.

Then, the maximum-likehood estimation (MLE) of the *document model* is derived as follows:

$$p(\boldsymbol{k}_i|\theta_D) = \frac{\sum_{j=1}^{M_d} S(d_j, \boldsymbol{k}_i)\delta(d_j, \boldsymbol{k}_i)}{\sum_{j=1}^{M_d} S(d_j, \boldsymbol{k}_i)}. \tag{8}$$

3.3.2 Query Model

In this paper, the query is also modeled as a bag of objects. Since the text query itself does not have any visual information, we estimate the query language model θ_Q with respect to the query-relevant object vocabulary $\boldsymbol{\mathcal{K}}$ and the ranked image list $\boldsymbol{\mathcal{C}}$ from text-based search. The distribution of each query-relevant object for query $\boldsymbol{q}$ is denoted as $p(\boldsymbol{k}_i|\theta_Q)$, which we assume follows the following distribution:

$$\begin{aligned} p(\boldsymbol{q}|\theta_Q) &= p(\boldsymbol{\mathcal{K}}, \boldsymbol{\mathcal{C}}|\theta_Q) \\ &\propto \prod_{i=1}^{M_q} p(\boldsymbol{k}_i, \boldsymbol{\mathcal{C}}|\theta_Q), \end{aligned} \tag{9}$$

with

$$p(\boldsymbol{k}_i, \boldsymbol{\mathcal{C}}|\theta_Q) = p(\boldsymbol{k}_i|\theta_Q)^{S(\boldsymbol{k}_i,\boldsymbol{\mathcal{C}})}. \tag{10}$$

The function $S(\boldsymbol{k}_i, \boldsymbol{\mathcal{C}})$ computes the object relevance with respect to the initial rank list $\boldsymbol{\mathcal{C}}$. With the distribution described in Equation (9), the maximum likehood estimation of the query's object model θ_Q is then derived as follows:

$$p(\boldsymbol{k}_i|\hat{\theta}_Q) = \frac{S(\boldsymbol{k}_i, \boldsymbol{\mathcal{C}})}{\sum_{j=1}^{M_q} S(\boldsymbol{k}_j, \boldsymbol{\mathcal{C}})}. \tag{11}$$

In this paper, we calculate the object relevance $S(\boldsymbol{k}_i, \boldsymbol{\mathcal{C}})$ based on the attributes from each object. These attributes are proposed below to indicate the object relevance to the query. As the below attributes are extracted on each containing instance of the object, we calculate the score of each object based on the expectation and variance of the comprising instances' attribute scores, such that each attribute can capture not only the average information but also the variance of the instances. To reduce the impact of noisy ROIs, the score of each instance is weighted by its belonging confidence $S(k_i, \boldsymbol{k})$ calculated in Section 3.1.3.

- **Initial ranking:** As stated in PRF based methods, the initial ranking of each image is critical to its relevance. Motivated by this, we assume that if an object has a set of instance whose parent images are all highly ranked by text-based search engine, the object can be regarded as query-relevant with a high probability. On the contrary, if all the instance of the object are ranked at the bottom of the ranking list, the object is probably irrelevant to the query. In this paper, we calculate the ranking score of each image with respect to its ranking position as follows:

$$IR(x) = \frac{1}{\log(R(x)+1)}, \tag{12}$$

  where $R(x)$ stands for the ranking position of image x.

- **Initial ranking of neighborhood:** The information from the visual neighborhood can be propagated to improve the robustness of the estimation. Hence, we propose to use the initial ranking of neighborhood of

each instance. Which is calculated as follows:

$$NR(x_i) = \frac{1}{|nn(x_i;k)|} \sum_{x_j \in nn(x_i;k)} IR(x_j), \qquad (13)$$

where $nn(x_i;k)$ stands for the k-nearest neighborhood of object instance x_i.

- **Object size:** Intuitively, if an ROI region occupies a big part of an image, it is probably the topic of the image. On the contrary, if an ROI is small, it is likely to be a background object which is irrelevant to the topic of the image and so to the query. Also, a small ROI may not have reliable visual features to compute the distance correctly, which misleads the ROI to belong to the query-relevant object category. To avoid the influence of image size, we normalize the area of bounding box by it.

- **Object location:** An ROI locating on the center of the image intuitively tells that the photographer is taking a shot directly towards the object. Therefore, the distance from the ROI to the image center indicates the confidence that the object is representing the image. In this paper, we adopt the shift of the ROI to the image center along X and Y axis, as well as the euclidean distance of the ROI to the center.

- **Saliency:** We assume ROI with high saliency is more likely to be an object while low saliency suggests a probability to be an scene. Hence, we use the average saliency scores of an ROI as one dimensional attribute score to reflect the relevance of the ROI to be relevant to the object query.

- **Visual density:** We follow the assumption in [29] that relevant images have higher density than irrelevant images. Thus, we argue that relevant object instances may have higher density on visual appearance. For each instance, its density is calculated by the Kernel Density Estimation (KDE) [24] follows

$$p(x_i) = \frac{1}{|nn(x_i;k)|} \sum_{x_j \in nn(x_i;k)} \chi(x_i - x_j), \qquad (14)$$

where $nn(x_i;k)$ suggests the k-nearest neighborhood of region x_i, and $\chi(x)$ is a kernel function with $\chi(x) > 0$ and $\int \chi(x)dx = 1$.

The attributes scores computed using the above approaches are formed into a feature vector $\boldsymbol{L}_i^q$. Then, the function $S(\boldsymbol{k}_i, \boldsymbol{\mathcal{C}})$ can be written as follows:

$$S(\boldsymbol{k}_i, \boldsymbol{\mathcal{C}}) = \boldsymbol{W}^T \boldsymbol{L}_i^q, \qquad (15)$$

where W^T is the weight vector of the combination model.

We then substitute the Equation (15) into Equation (11), then $p(\boldsymbol{k}_i|\hat{\theta}_Q)$ becomes

$$\begin{aligned} p(\boldsymbol{k}_i|\hat{\theta}_Q) &= \frac{S(\boldsymbol{k}_i, \boldsymbol{\mathcal{C}})}{\sum_{j=1}^{|\boldsymbol{\mathcal{K}}|} S(\boldsymbol{k}_j, \boldsymbol{\mathcal{C}})} \\ &= \frac{\boldsymbol{W}^T \boldsymbol{L}_i^q}{\sum_{j=1}^{|\boldsymbol{\mathcal{K}}|} \boldsymbol{W}^T \boldsymbol{L}_j^q}. \end{aligned} \qquad (16)$$

Since the denominator of Equation (16) is mainly for normalization, it can be ignored for ranking. By integrating Equation (16) into the ranking function (5), we can get the final ranking function for the bag-of-objects retrieval model.

3.4 Learning

As the ranking function is a linear function we can naturally employ Ranking SVM to learn the parameters. Ranking SVM [13] is an adaptation of the classification SVM to the ranking problem. It decomposes the rankings into a set of pair-wise preferences and then reduces the ranking learning problem into the classification of pairs. The optimization problem of Ranking SVM is

$$\begin{aligned} \min_{\boldsymbol{W}} \quad & \frac{1}{2}\boldsymbol{W}^T\boldsymbol{W} + C\sum \varepsilon_{x,y}^i \\ \text{s.t.} \quad & \forall q_i, k_x \succ k_y : R(x;q_i) - R(y;q_i) \geq 1 - \varepsilon_{x,y}^i \\ & \forall x, y, i : \varepsilon_{x,y}^i \geq 0. \end{aligned} \qquad (17)$$

The problem can be efficiently solved using SMO (Sequential Minimal Optimization) or cutting-plane algorithm. In this paper, we particularly employ the software provided in [12] for the learning process.

4. EXPERIMENTS

To demonstrate the effectiveness of our proposed approach, we perform an experimental study on two subsets of a publicly available dataset, comprising the queries of objects and the queries for people. Various baseline approaches including the result from the search engine, the existing unsupervised and supervised reranking methods are compared to show the superiority of our proposed approach.

4.1 Experimental Steps

4.1.1 Dataset

To make our experiment as reproducible as possible, we employ a publicly available Web Queries dataset[1] for evaluating our approach and comparing with the baseline approaches. The dataset contains totally 353 representative image search queries, covering a wide range of topics including products, celebrities, animals etc. Then, these queries are issued to an image search engine to collect top ranked image results. Finally 71478 images are obtained in total. A binary relevance between label each query and the retrieved images is provided by human as the ground-truth.

As proposed approach is designed for object queries, we construct two subsets of the Web Queries dataset by selecting those object queries. The WEB_QRY_OBJECTS dataset comprises object queries including landmarks, products, flags and logos. To more comprehensively evaluate our approach, the WEB_QRY_HUMAN dataset is constructed by selecting those named person queries. Finally, the WEB_QRY_OBJECTS dataset is consisted of 101 queries with 19586 images, while WEB_QRY_HUMAN dataset contains 103 queries and 20398 images.

4.1.2 Experimental Settings

To demonstrate the effectiveness of our method, we compare it with different baseline approaches, including text-based search engine (*"Text-baseline"*) and the Bayesian reranking(*"Bayesian"*) [30], pseudo relevance feedback rerank-

[1] `http://lear.inrialpes.fr/~krapac/webqueries/webqueries.html`

Table 1: The performance comparison of various reranking methods WEB_QRY_OBJECTS.

Methods	MAP	NDCG@10	NDCG@25	NDCG@40
Text-baseline	0.582	0.674	0.649	0.667
Bayesian	0.647 (+11.17%)	0.739 (+9.64%)	0.717 (+10.48%)	0.698 (+4.65%)
PRF	0.743 (+27.66%)	0.850 (+26.11%)	0.830 (+27.89%)	0.809 (+21.29%)
Letorr	0.746 (+28.18%)	0.809 (+20.03%)	0.833 (+28.35%)	0.812 (+21.74%)
Query-relative	0.750 (+28.87%)	0.859 (+27.45%)	0.839 (+29.28%)	0.817 (+22.49%)
L^2	0.760 (+30.58%)	0.856 (+27.00%)	0.838 (+29.12%)	0.830 (+24.44%)
Proposed method	0.812 (+39.52%)	0.847 (+25.67%)	0.844 (+30.05%)	0.848 (+27.14%)

Table 2: The performance comparison of various reranking methods on WEB_QRY_HUMAN.

Methods	MAP	NDCG@10	NDCG@25	NDCG@40
Text-baseline	0.603	0.762	0.725	0.688
PRF	0.620(+2.89%)	0.761(-0.16%)	0.716(-1.18%)	0.682(-0.92%)
Bayesian	0.675(+11.94%)	0.876(+14.96%)	0.819(+12.97%)	0.772(+12.21%)
Letorr	0.680(+12.84%)	0.819(+7.57%)	0.784(+8.19%)	0.755(+9.70%)
Query-relative	0.640(+6.23%)	0.761(-0.04%)	0.737(+1.71%)	0.714(+3.78%)
L^2	0.721(+19.69%)	0.888(+16.62%)	0.848(+17.01%)	0.813(+18.13%)
Proposed method	0.764(+26.70%)	0.861(+12.99%)	0.853(+17.66%)	0.827(+20.20%)

ing(*"PRF"*) [33], supervised-reranking(*"Loterr"*) [35], the query-relative classifier (*"Query-relative"*) [17] and the L^2 reranking (*"L^2"*) [36]. For PRF reranking, top ranked images are selected as positive samples while the negative samples are sampled following [36]. When evaluating Bayesian reranking, the pair-wise ranking distance and the best performing local learning consistency is adopted.

For the supervised reranking approaches including L^2 reranking, superived-reranking, query-relative classifier, and ours, ranking models are trained by RankSVM[13]. 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.

For each image in our dataset, we extract the Pyramid Histogram Of visual Words (PHOW) described in [1] as the visual feature representation of images. Firstly, SIFT[22] descriptors are extracted on the points based on dense sampling in 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 based on the k-means quantization. Finally, each image is represented by a spatial pyramid histogram of the quantized visual words. 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 the papers.

To better deal with the images in "WEB_QRY_HUMAN" dataset, we employ a face detector embedded in OpenCV[2] to obtain facial ROIs as a supplement to ROIs sampled from saliency map.

4.1.3 Evaluation Measure

The ranking performance is measured by Averrage 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. NDCG is defined as follows:

$$\text{NDCG@}k = \frac{\text{DCG@}k}{\text{IDCG@}k}, \tag{18}$$

where

$$\text{DCG@}k = \sum_{i=1}^{k} \frac{2^{r_i} - 1}{\log_2(i+1)}. \tag{19}$$

In Equation (19), r_i is the ground truth labeling of relevance for the i_{th} image, while k is the truncation level.

4.2 Experimental Results

4.2.1 Performance Comparison

Table 1 shows the performance comparison of the different reranking on the WEB_QRY_OBJECT dataset, in terms of MAP, NDCG@10, NDCG@25 and NDCG@40. It is obvious that all reranking methods can outperform the text-baseline in the term of MAP, for example L^2 reranking improves the text-baseline by 30.58%, and our method improves it by 39.52%. This demonstrates that the reranking approaches are generally effective to boost the image search ranking performance. Among all the evaluated reranking methods, our method outperforms the other 5 reranking methods. Specifically, it can improve the Superived-reranking, query-relative classifier and L^2 reranking by 19.41%, 26.88% and 6.84%. This suggests the general effectiveness of our approach on object queries.

Table 2 compares all the evaluated methods on the dataset WEB_QRY_HUMAN. We can see that the baseline reranking methods are much less effective on this dataset comparing to WEB_QRY_OBJECT, because of the specification of human images. For example, the state-of-the-art method L^2 reranking can only achieve an improvement of 19.69% on MAP, while it can increase the performance on WEB_QRY_OBJECT dataset by 30.58%. This is because

[2] `http://opencv.willowgarage.com`

Table 3: The performance of each individual object attribute.

Attribute mean	Initial ranking	Initial ranking of neighborhood	Saliency	Size	Location	Visual density
MAP	0.730	0.764	0.733	0.711	0.749	0.763
Attribute var.	Initial ranking	Initial ranking of neighborhood	Saliency	Size	Location	Visual density.
MAP	0.716	0.749	0.721	0.713	0.740	0.754

Table 4: The performance by leaving one attribute out.

Attribute mean	Initial ranking	Initial ranking of neighborhood	Saliency	Size	Location	Visual density
MAP	0.780	0.764	0.780	0.778	0.771	0.760
Attribute var.	Initial ranking	Initial ranking of neighborhood	Saliency	Size	Location	Visual density.
MAP	0.787	0.772	0.779	0.789	0.777	0.768

global features extracted from the whole image have a limited ability in identifying people. Our method achieves an improvement of 26.70% in the term of MAP, with the help of query-relevant objects and face detector. The proposed method improves the Superived-reranking by 12.35% and L^2 reranking by 5.96%. This experiment shows that our method is able to address the limitation of existing methods on human queries.

We can see from both Table 1 and Table 2 that our method has no improvement compared to L^2 reranking in terms of NDCG@10. One of the reasons is that in the training process, we select the optimal parameter C for Ranking SVM by validating MAP on the validation set.

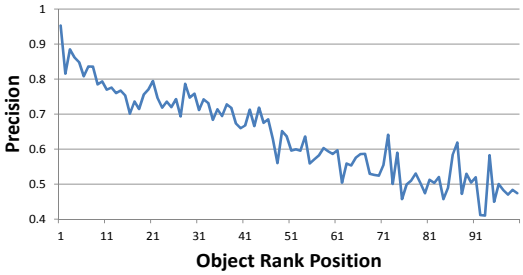

Figure 3: The average quality of objects on different rank positions.

4.2.2 *Object Ranking Analysis*

In this section, we evaluate the performance of object relevance prediction in our approach. To study the correlation between the prediction sore and the object relevance, we first calculate the precision of each object's containing images, which roughly indicates the object relevance. Then, the precisions of objects on the same ranking position from different queries are averaged. From Figure 3, we can see that the precisions on top positions are higher than those of bottom positions. This suggests that highly relevant objects are boosted by the proposed prediction model in Equation (15), while irrelevant objects are suppressed by a low score.

4.2.3 *Object Attribute Analysis*

In this section, we will analyse the usefulness of the 6 proposed object attributes individually. We measure the importance of each attribute using two strategies. One is to evaluate the performance of each attribute feature individually. The other is to evaluate the performance of the proposed approach by leaving each attribute feature out from the model once a time. We can observe from Table 3 and 4 that the attribute "Visual density" and "Initial ranking of neighborhood" are the two most important features. The importance of "Initial ranking of neighborhood" suggests that the ranking voted by visual neighbors is more reliable than the ranking of the image itself. Besides, attribute "Object Location" plays a very important role due to its high performance. It is because human tends to locate the important objects at the center of an image.

4.2.4 *Performance Analysis over Different Queries*

By comparing our result with that of L^2 reranking, 74% of queries in WEB_QRY_OBJECTS and 71% of queries in WEB_QRY_HUMAN is improved in terms of MAP. Figure 4 shows several examples of the search result by L^2 reranking and our method. The query "Champions League" improves because most of the "UEFA" logos lie on a small region of images. L^2 reranking with global PHOW feature cannot deal with such case. But in our method, this logo can be easily detected by query-relevant object discovery. The query "Aircraft Carrier" shows a failure case of our method. It is because the images for this query show too much different postures of the aircraft carrier, and our query-relevant object discovery method fails to detect any ROIs in common.

5. CONCLUSION AND FUTURE WORK

Image search reranking has been studied for several years and various approaches have been developed recently to boost the performance of text-based image search engine for general queries. In this paper we argue that there is no single method which can fit well all queries and the research on image search reranking requires a new methodology developing specific models for queries in different domains.

This paper serves as a first attempt on this direction. We observe that the existing reranking methods which operate on the whole image may not perform sufficiently well for object queries where only part of an image is required to be relevant. Motivated by that, we propose a novel bag-of-objects retrieval model for reranking images for object queries. The retrieval model is developed based on language modeling techniques for information retrieval. By discovering common objects relevant to the query from the search results returned by the text-based search engine, we can build object language models for the query and images on this query-relevant object vocabulary. The attributes of discovered objects are computed to represent how relevant and confident the objects are to the query, which are incorporated with weights into the ranking function, to address the possible unreliability and noises in the query-relevant objects. Finally, a learning to rank approach is employed to learn the

weights on object attributes from human labeled data. The experimental results on two subsets of Web Queries dataset demonstrate that the proposed approach can improve 6.84% compared to the state-of-the-art reranking approaches.

We believe that this is a right and promising direction for further advancing image search reranking. Regarding this general direction as well as the specific work in this paper, we envision the following future works. 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 [20], 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. Third, we hope to identify and address the system challenges so as to most efficiently integrate this algorithm into a real-world image search engine.

6. REFERENCES

[1] A. Bosch, A. Zisserman, and X. Muoz. Image classification using random forests and ferns. In *CVPR*, pages 1–8, 2007.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.

[3] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, page 22, 2004.

[4] T. Deselaers, B. Alexe, and V. Ferrari. Localizing objects while learning their appearance. *ECCV*, pages 452–466, 2010.

[5] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun. Salient object detection by composition. In *ICCV*, pages 1028–1035, 2011.

[6] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *ICCV*, volume 2, pages 1816–1823, 2005.

[7] M. Fritz and B. Schiele. Decomposition, discovery and detection of visual categories using topic models. In *CVPR*, pages 1–8, 2008.

[8] D. Hochbaum and V. Singh. An efficient algorithm for co-segmentation. In *ICCV*, pages 269–276, 2009.

[9] L. Hohl, F. Souvannavong, B. Merialdo, and B. Huet. Enhancing latent semantic analysis video object retrieval with structural information. In *ICIP*, volume 3, pages 1609–1612, 2004.

[10] W. Hsu, L. Kennedy, and S. Chang. Video search reranking through random walk over document-level context graph. In *ACM Multimedia*, pages 971–980, 2007.

[11] Y. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. *IEEE Trans. on PAMI*, 30(11):1877–1890, 2008.

[12] T. Joachims. Making large-scale svm learning practical. *Advances in Kernel Methods Support Vector Learning*, pages 169–184, 1999.

[13] T. Joachims. Training linear svms in linear time. In *ACM SIGKDD*, pages 217–226, 2006.

[14] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, pages 1943–1950, 2010.

[15] G. Kim and A. Torralba. Unsupervised Detection of Regions of Interest using Iterative Link Analysis. In *NIPS*, 2009.

[16] G. Kim, E. Xing, L. Fei-Fei, and T. Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 169–176, 2011.

[17] J. Krapac, M. Allan, J. Verbeek, and F. Juried. Improving web image search results using query-relative classifiers. In *CVPR*, pages 1094–1101, 2010.

[18] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *ACM SIGIR*, pages 111–119, 2001.

[19] Y. Lee and K. Grauman. Object-graphs for context-aware category discovery. In *CVPR*, pages 1–8, 2010.

[20] L. Li, H. Su, E. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. *NIPS*, 2010.

[21] Y. Liu, T. Mei, X. Hua, J. Tang, X. Wu, and S. Li. Learning to video search rerank via pseudo preference feedback. In *ICME*, pages 297–300, 2008.

[22] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[23] E. Oomoto and K. Tanaka. Ovid: Design and implementation of a video-object database system. *IEEE Trans. on KDE*, 5(4):629–643, 1993.

[24] E. Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

[25] C. Rother, T. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs. In *CVPR*, volume 1, pages 993–1000, 2006.

[26] S. Sav, G. Jones, H. Lee, N. O'Connor, and A. Smeaton. Interactive experiments in object-based retrieval. *Image and Video Retrieval*, pages 1–10, 2006.

[27] S. Sav, H. Lee, A. Smeaton, N. O'Connor, and N. Murphy. Using video objects and relevance feedback in video retrieval. 6015:353–364, 2005.

[28] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *ICCV*, pages 1–8, 2007.

[29] X. Tian, Y. Lu, L. Yang, and Q. Tian. Learning to judge image search results. In *ACM Multimedia*, pages 363–372, 2011.

[30] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X. Hua. Bayesian video search reranking. In *ACM Multimedia*, pages 131–140, 2008.

[31] X. Tian, L. Yang, X. Wu, and X. Hua. Visual reranking with local learning consistency. *Advances in Multimedia Modeling*, pages 163–173, 2010.

[32] S. Vicente, V. Kolmogorov, and C. Rother. Cosegmentation revisited: Models and optimization. *ECCV*, pages 465–479, 2010.

[33] R. Yan, A. Hauptmann, and R. Jin. Multimedia

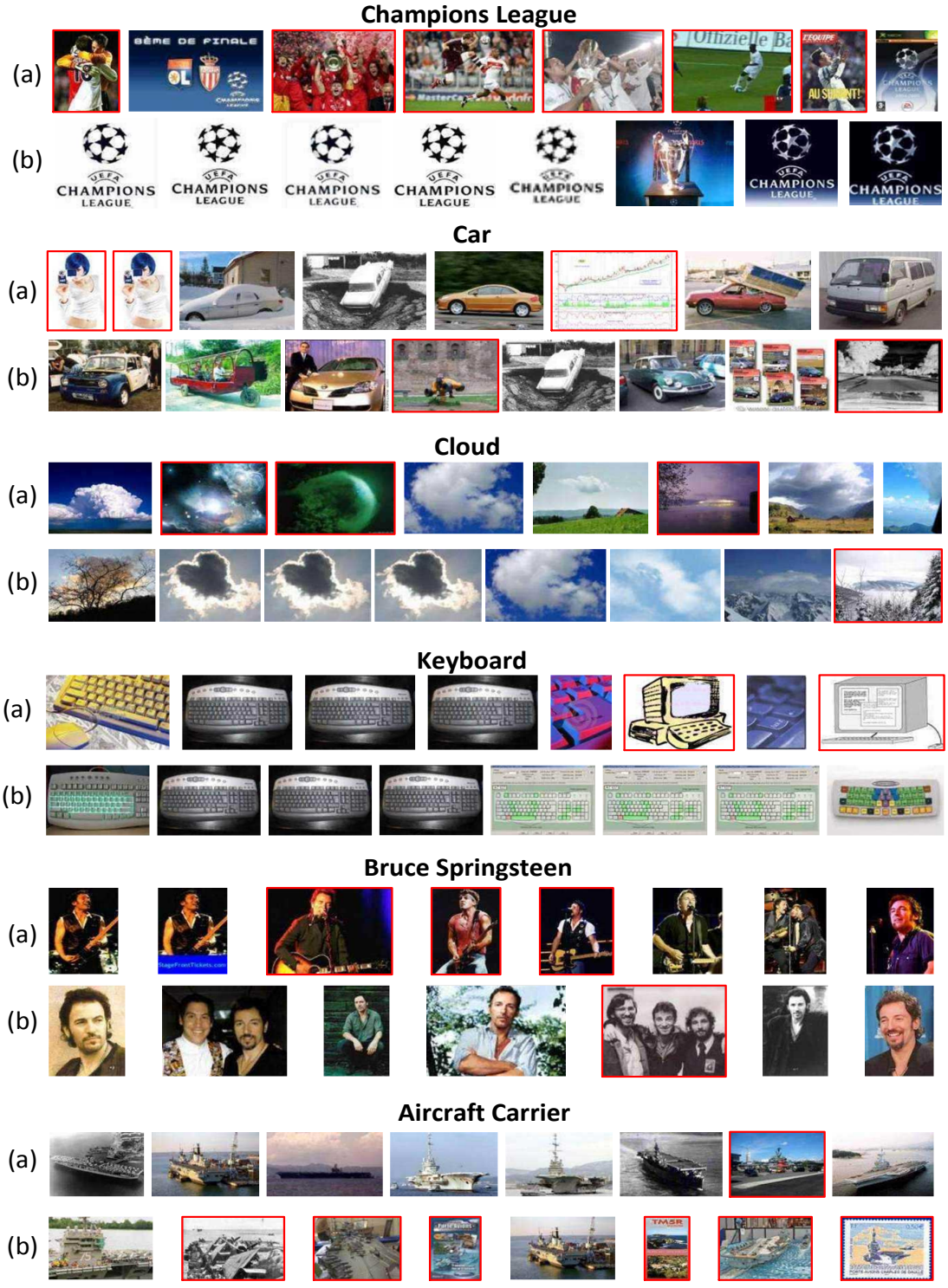

Figure 4: Six sample results of (a) L^2 reranking and (b) the proposed method. Irrelevant images are marked by red rectangle.

search with pseudo-relevance feedback. *Image and Video Retrieval*, pages 649–654, 2003.

[34] L. Yang, B. Geng, Y. Cai, A. Hanjalic, and X. Hua. Object retrieval using visual query context. *IEEE Trans. on Multimedia*, (99):1–1, 2011.

[35] L. Yang and A. Hanjalic. Supervised reranking for web image search. In *ACM Multimedia*, pages 183–192, 2010.

[36] L. Yang and A. Hanjalic. learning from search engine and human supervision web image search. In *ACM Multimedia*, pages 1365–1368, 2011.

[37] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li. Descriptive visual words and visual phrases for image applications. In *ACM Multimedia*, pages 75–84, 2009.