

# Submodular Reranking with Multiple Feature Modalities for Image Retrieval

Fan Yang<sup>1</sup>(✉), Zhuolin Jiang<sup>2</sup>, and Larry S. Davis<sup>1</sup>

<sup>1</sup> University of Maryland College Park, College Park, MD, USA  
`{fyang, lsd}@umiacs.umd.edu`

<sup>2</sup> Noah's Ark Lab, Huawei Technologies, Hong Kong, China  
`zhuolin.jiang@huawei.com`

**Abstract.** We propose a submodular reranking algorithm to boost image retrieval performance based on multiple ranked lists obtained from multiple modalities in an unsupervised manner. We formulate the reranking problem as maximizing a submodular and non-decreasing objective function that consists of an information gain term and a relative ranking consistency term. The information gain term exploits relationships of initially retrieved images based on a random walk model on a graph, then images similar to the query can be found through their neighboring images. The relative ranking consistency term takes relative relationships of initial ranks between retrieved images into account. It captures both images with similar ranks in the initial ranked lists, and images that are similar to the query but highly ranked by only a small number of modalities. Due to its diminishing returns property, the objective function can be efficiently optimized by a greedy algorithm. Experiments show that our submodular reranking algorithm is effective and efficient in reranking images initially retrieved by multiple modalities. Our submodular reranking framework can be easily generalized to any generic reranking problems for real-time search engines.

## 1 Introduction

Numerous approaches have been proposed to improve the performance of content-based image retrieval (CBIR) systems. Most of them adopt a single feature modality such as bag-of-words (BoW) [1], Fisher vectors [2, 3] or vector locally aggregated descriptors (VLAD) [4]. Various extensions based on a single feature modality have been proposed, such as query expansion [5, 6], spatial verification [7] and Hamming embedding [8]. However, a single feature modality only captures one “view” of an image. Often, a lower-ranked but relevant retrieved image from one feature modality may be highly ranked by another modality. By fusing retrieval results from multiple feature modalities, we may discover both agreement and inconsistency among them to improve retrieval quality. Recent work combines multiple feature modalities for reranking by multi-modal graph-based learning [9],

---

**Electronic supplementary material** The online version of this chapter (doi:[10.1007/978-3-319-16865-4\\_2](https://doi.org/10.1007/978-3-319-16865-4_2)) contains supplementary material, which is available to authorized users.

query-specific graph fusion [10] or Co-Regularized Multi-Graph Learning [11]. In [9] requires a large number of queries to compute relevance scores for initially retrieved images, which is only suitable for large sets of queries. In [10], initial ranked lists were converted to undirected graphs, which were linearly combined without considering the inter-relationships between modalities. In [11] is a supervised learning based on image attributes, so it is not suitable for unsupervised reranking tasks.

We present a submodular objective function for reranking images retrieved by multiple feature modalities, which is very efficient and fully unsupervised. Submodularity [12] has been applied to various optimization problems in vision due to the availability of efficient approximate optimization methods based on its diminishing returns property - which means that as the incremental optimization algorithm proceeds, each item added to the evolving solution has less and less marginal value as the solution set grows. Our submodular objective function consists of two terms: an information gain term and a relative ranking consistency term. To compute the information gain, we first represent each initial ranked list as an undirected graph, where nodes are retrieved dataset images and edges represent similarities between images. The graph structure is then modeled as a transition matrix under the assumption of a random walk on a graph. Edge weights between nodes are converted to the probability of walking from a node to its neighbors. We select a subset of retrieved images by maximizing the information gain over the graph, which maximizes the mutual information between the selected subset and unselected nodes in the graph. The information gain takes pairwise relationships of retrieved images into consideration, and favors compact clusters of retrieved images which are similar to the query.

The relative ranking consistency term exploits the inter-relationships among multiple ranked lists obtained by different feature modalities. Specifically, if relative ranks between two images are consistent across multiple ranked lists, the ranking relationship between them is considered reliable and captured by our relative ranking consistency term. Additionally, our relative ranking consistency term encourages selecting images that are similar to the query but only found and highly ranked by a small number of modalities.

The final submodular objective function combines both the relationships among retrieved images from a single modality and the relative ranks of image pairs across different modalities, thereby improving initial retrieval results obtained by multiple independent modalities. Our approach only utilizes pairwise similarities between images in terms of appearance information without using any prior knowledge, hence it is fully unsupervised. Moreover, although we evaluate our submodular reranking algorithm on natural image retrieval, it only involves similarity graphs and initial ranked lists. Therefore, it can be easily extended to other generic retrieval tasks with multiple independent ranked lists returned by heterogeneous and non-visual features, such as audio and text. The main contributions of our work are summarized as follows:

- We address the problem of reranking natural images with multiple feature modalities by maximizing a submodular objective function, which is done by an efficient greedy algorithm.

- We model the image-level relationships for each modality as a graph and apply information gain theory to find the most similar images to the query. Only pairwise similarities between images are used to construct the graph. Our approach is unsupervised without using any label information.
- We propose a relative ranking consistency term to exploit the inter-relationships of multiple ranked lists across different modalities. The relative ranking consistency term effectively selects images that have consistent relative ranks across multiple modalities. It also discovers images that are similar to the query but only found by one or a few modalities.

## 2 Related Works

The majority of image retrieval approaches are based on a single feature modality. They usually adopt the bag-of-words (BoW) feature as an image representation, and then compute the similarities between a query image and dataset images for retrieval [1]. Many works focus on learning good feature representations for retrieval problems. Jégou *et al.* [4] proposed the vector locally aggregated descriptor as a compact representation. It achieved good results while requiring less storage compared to the BoW feature. Multi-VLAD [13] was later proposed to construct and match VLAD features of multiple levels from an image to improve localization accuracy. RootSIFT [14] was proposed to address the burstiness problem with standard BoW features. To compensate for the spatial information loss in the standard BoW-based approach, spatial verification [7] was proposed to match SIFT descriptors between images at the cost of extra storage space. Vocabulary trees [15] were proposed to improve efficiency in codebook construction and descriptor quantization by using hierarchical clustering. Contextual weighting [16] was further applied to vocabulary trees to increase the discriminative ability of visual words. Instead of quantizing a descriptor to a single visual word, assigning it to multiple words results in more discriminative BoW vectors and thus achieves better performance [17, 18]. Query expansion [5, 6, 14] has been widely applied to rerank initially retrieved images, where a small portion of top ranked images serve as additional queries and are fed into the retrieval system again to further explore similar images. Some improvements such as Hamming embedding with geometric constraints [8], dataset-side feature augmentation [14] and co-occurrences of visual words [19] have achieved state-of-the-art results.

Although a single feature modality can achieve good retrieval results, better performance is anticipated if retrieved results from multiple feature modalities are properly fused. This is because they usually describe images from complementary perspectives. Recent work on fusing multiple feature modalities for image retrieval has been proposed, such as multi-modal graph learning [9], query-specific graph fusion [10] and Co-Regularized Multi-Graph Learning [11]. In [9] proposed a graph-based learning algorithm to infer weights of modalities. However, it requires a large number of queries beforehand to estimate relevance scores of initially retrieved images, which is not feasible if only a small number of queries are available.

In [10] constructed a graph for each initial ranked list based on a single feature modality using k-reciprocal nearest neighbors. In [11] imposed intra-graph and inter-graph constraints in a supervised learning framework which requires image attribute information. However, image attributes are not always available and the training process may be time-consuming for larger graphs. In contrast, our reranking approach considers both image-level and modality-level relationships, and does not require any attribute information or label information.

Submodularity, as a discrete analog of convexity, is widely studied in combinatorial optimization [12] due to its diminishing returns property: adding an element to a smaller set contributes more than adding it to a larger set. Various submodular functions have been proposed and successfully applied to many vision applications, such as image segmentation [20, 21], dictionary selection/learning [22, 23], saliency detection [24], object recognition [25] and video hashing [26]. A few works applied submodular functions to diversified ranking [27–29], where elements in the reranked list are similar to the query but also diversified. For diversified ranking, submodular functions are designed to seek a trade-off between similarity and diversity. It should be noted that [27–29] are not similar to our submodular reranking, since we encourage elements in the reranked list to be similar to the query and homogenous rather than diversified.

### 3 Submodular Reranking

We formulate the reranking problem as selecting and rearranging a subset of retrieved images from initial ranked lists obtained from multiple modalities. Our submodular objective function utilizes similarities of pairs of images to exploit relationships between retrieved images within each modality. It also considers the relative ranking between retrieved images across multiple ranked lists.

#### 3.1 Preliminaries

**Submodularity.** Let  $\mathcal{V}$  be a finite set. A set function  $f : 2^{\mathcal{V}} \rightarrow \mathbb{R}$  is submodular if it satisfies  $f(\mathcal{S} \cup a) - f(\mathcal{S}) \geq f(\mathcal{T} \cup a) - f(\mathcal{T})$  for all  $\mathcal{S} \subset \mathcal{T} \subseteq \mathcal{V}, a \in \mathcal{V} \setminus \mathcal{T}$ . This is called the *diminishing returns property*: adding  $a$  to a small set has a bigger impact than adding it to a larger set. The gain of the function value  $f(\mathcal{S} \cup a) - f(\mathcal{S})$  is called the *marginal gain* of  $f$  when adding  $a$  to  $\mathcal{S}$ .

**Monotonicity.** A set function  $f : 2^{\mathcal{V}} \rightarrow \mathbb{R}$  is monotone (or non-decreasing) if for every  $\mathcal{S} \subseteq \mathcal{T} \subseteq \mathcal{V}$ ,  $f(\mathcal{S}) \leq f(\mathcal{T})$  and  $f(\emptyset) = 0$ .

#### 3.2 Information Gain with Graphical Models

**Graph Construction.** Given  $M$  feature modalities, we obtain  $M$  initial ranked lists of retrieved images for each query image. For efficient reranking, we select only the top  $K$  retrieved images from each ranked list. Note that the top  $K$  images are generally not the same across different modalities. Given an initial ranked list consisting of  $K$  retrieved images from modality  $m$ , we represent it as an undirected

graph  $\mathcal{G}_m = (\mathcal{V}_m, \mathcal{E}_m)$  where nodes  $v_m \in \mathcal{V}_m$  are images and  $e_m(i, j) \in \mathcal{E}_m$  denotes the edge that connects  $v_m(i)$  and  $v_m(j)$ . An affinity matrix  $\mathbf{A}_m \in \mathbb{R}^{K \times K}$  is used to represent the graph with the element  $a_m(i, j)$  corresponding to the edge weight of  $e_m(i, j)$ , which is the pairwise similarity between images  $v_m(i)$  and  $v_m(j)$ <sup>1</sup>. To facilitate the objective function construction (see Sect. 3.2), we do not include self-loops  $e_m(i, i)$  of nodes  $v_m(i)$  in the graph. Therefore,  $a_m(i, i)$  is set to 0. For notational convenience, we denote  $\mathcal{V}$  as the union of all nodes from the  $M$  undirected graphs, so that  $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2 \cup \dots \cup \mathcal{V}_M$ . We aim to select a subset of nodes  $\mathcal{S}$  from  $\mathcal{V}$  which are the most similar to the query image and arrange them in order to obtain the reranked result. Furthermore,  $\mathcal{U}$  denotes the set of images which are not selected, so that  $\mathcal{U} \cap \mathcal{S} = \emptyset$  and  $\mathcal{V} = \mathcal{S} \cup \mathcal{U}$ .

**Information Gain.** Given  $M$  graphs, we seek a method to combine them so that complementary modalities may help discover images similar to the query in a joint manner. Although the same graph construction is used for all ranked lists, pairwise similarities from different modalities are usually of incomparable scales, making a direct graph combination infeasible. To address this problem, we resort to information gain theory with graphical models [30], which is based on a simple probabilistic model.

We start from the random walk model on a graph  $\mathcal{G}_m$ . The random walk model can be interpreted as a Markov process: a walker stays at a node in the graph at time  $t$  and randomly walks to one of its neighboring nodes under some probability at time  $t + 1$ . The probability of “walking” between nodes is called the transition probability and is defined as  $\mathbf{P}_m = \mathbf{D}_m^{-1} \mathbf{A}_m$ , where  $\mathbf{D}_m \in \mathbb{R}^{K \times K}$  is a diagonal matrix with the diagonal element  $d_m(i, i) = \sum_j a_m(i, j)$ . The transition matrix  $\mathbf{P}_m$  is a row-stochastic matrix indicating the transition probabilities of a random walk on the graph.  $p_m(i, j)$  represents the conditional probability of walking from node  $v_m(i)$  to node  $v_m(j)$ , which indicates the similarity between  $v_m(i)$  and  $v_m(j)$  based on the observation of  $v_m(i)$ . With the transition matrix  $\mathbf{P}_m$ , edge weights are converted to probabilities. Then we adopt information gain as a direct measure of the value of information of our graphical models. We start from a single graph  $\mathcal{G}_m$ , and define the information gain as

$$F_m(\mathcal{S}) = H(\mathcal{V}_m \setminus \mathcal{S}) - H(\mathcal{V}_m \setminus \mathcal{S} | \mathcal{S}) \quad (1)$$

where  $\mathcal{S}$  is the subset we select from  $\mathcal{V}$ , and  $\mathcal{V}_m \setminus \mathcal{S}$  is the set  $\mathcal{V}_m$  with  $\mathcal{S}$  removed.  $H(\mathcal{V}_m \setminus \mathcal{S})$  is the entropy of unselected nodes in graph  $\mathcal{G}_m$ .  $H(\mathcal{V}_m \setminus \mathcal{S} | \mathcal{S})$  is the conditional entropy of remaining nodes on graph  $\mathcal{G}_m$  after we have observed  $\mathcal{S}$ . Specifically,  $H(\mathcal{V}_m \setminus \mathcal{S} | \mathcal{S})$  and  $H(\mathcal{V}_m \setminus \mathcal{S})$  are defined as

$$\begin{aligned} H(\mathcal{V}_m \setminus \mathcal{S} | \mathcal{S}) &= - \sum_{v \in \mathcal{V}_m \setminus \mathcal{S}, s \in \mathcal{S}} p_m(v, s) \log p_m(v | s) \\ H(\mathcal{V}_m \setminus \mathcal{S}) &= - \sum_{v \in \mathcal{V}_m \setminus \mathcal{S}} p_m(v) \log p_m(v) \end{aligned} \quad (2)$$

<sup>1</sup> Please see experiment section about how to compute pairwise similarities.

where  $p_m(v, s) = p_m(v|s)p_m(s)$ .  $p_m(v|s)$  is the transition probability of walking to a node  $v$  in graph  $\mathcal{G}_m$  when the walker is at node  $s$ .  $p_m(s)$  and  $p_m(v)$  are the marginal probabilities of nodes  $s$  and  $v$  being similar to the query from modality  $m$ .  $p_m(v|s)$  can be directly obtained from  $\mathbf{P}_m$ . To calculate the marginal probability  $p_m(v)$ , we use the normalized similarities between the query and retrieved images. We denote the similarities between the top  $K$  retrieved images and the query image from modality  $m$  as  $\mathbf{c}_m = (c_{m,1}, c_{m,2}, \dots, c_{m,K})^\top$ .  $\ell_1$  normalization is then applied to  $\mathbf{c}_m$  to obtain  $p_m(v) = c_{m,v}/|\mathbf{c}_m|_1$ .

We have the following proposition stating that the information gain with our graphical model is submodular.

**Proposition 1.**  $F_m : 2^{\mathcal{V}_m} \rightarrow \mathbb{R}$  is a submodular and monotone function.

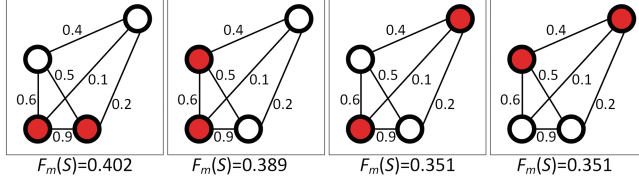
The proof is presented in the supplementary material.  $F_m$  is essentially the mutual information  $I(\mathcal{V}_m \setminus \mathcal{S}; \mathcal{S})$  capturing the mutual dependence between subset  $\mathcal{S}$  and unselected nodes  $\mathcal{V}_m \setminus \mathcal{S}$ , which measures how much  $\mathcal{S}$  is representative of the graph with respect to the query. That  $F_m$  is non-decreasing is obvious, because the addition of any node to  $\mathcal{S}$  always provides information or does not provide information at all, since “information never hurts”. Submodularity comes from the observation that the information gain of adding a node to  $\mathcal{S}$  becomes less in a later stage because it is more likely similar to elements in  $\mathcal{S}$  as  $\mathcal{S}$  grows.

To combine graphs, we need to determine the importance of each graph. Here we adopt the heuristic of simply summing up the information gains of the individual graphs to obtain the total information gain:

$$R(\mathcal{S}) = - \sum_m \left( \sum_{v \in \mathcal{V} \setminus \mathcal{S}} p_m(v) \log p_m(v) - \sum_{v \in \mathcal{V} \setminus \mathcal{S}, s \in \mathcal{S}} p_m(v, s) \log p_m(v|s) \right) \quad (3)$$

The information gain on a graph takes relationships between dataset images into account, so it propagates information about a dataset image to its neighbors, and better exploits dataset images that are similar to the query than simple pairwise comparisons. The combination seeks an agreement with respect to pairwise similarities derived from multiple modalities, so explores relationships of modalities to some extent. Note that since the top  $K$  images retrieved from different modalities may not be the same,  $p_m(v)$  and  $p_m(v|s)$  are set to 0 if an image is not included in graph  $\mathcal{G}_m$ , so it does not contribute to the objective function. An image discovered by most modalities contributes more to the information gain, therefore is selected to be in  $\mathcal{S}$  with greater chance.

Since  $F_m(\mathcal{S})$  is submodular and monotonically increasing, the linear combination of submodular functions,  $R(\mathcal{S})$ , is also submodular and non-decreasing. Since the information gain exploits the pairwise relationships between retrieved images, maximizing  $R(\mathcal{S})$  is equivalent to selecting a group of images that are similar to the query and closely related to each other. Intuitive examples are shown in Fig. 1.



**Fig. 1.** The importance of information gain for selecting nodes into subset  $\mathcal{S}$ . The number next to the edges is weight (similarity) between nodes. Red dots represent the selected subset  $\mathcal{S}$  while white dots are remaining nodes  $\mathcal{V}_m \setminus \mathcal{S}$ . The marginal probability of all nodes is set to  $1/4$ . Four cases of selection are presented, where the corresponding value of  $F_m(\mathcal{S})$  is shown under each sub figure. By computing the information gain, we observe that it prefers images that are closely related to each other to be selected into  $\mathcal{S}$ , resulting in a compact cluster. Therefore, relationships of dataset images are exploited to facilitate reranking (Color figure online).

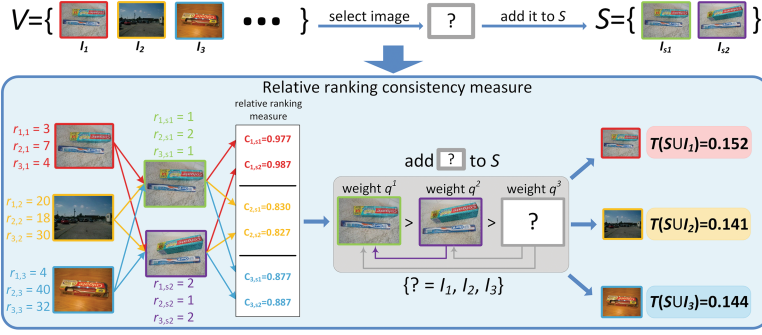
### 3.3 Relative Ranking Consistency

Simply summing up initial ranks obtained from different modalities for an image is not suitable, as a higher rank may be overly diluted by other lower ranks. Although complementary information from multiple modalities is used by integrating the  $F_m(\mathcal{S})$ , information gain does not completely utilize the inter-relationships between modalities. Additionally, it only considers pairwise similarities between images. However, the initial ranks of retrieved images from different modalities provide additional information that can further improve performance. For example, an image that is similar to the query and ranked lower by one modality may be ranked higher when it is perceived from a different perspective (*i.e.*, different modality). We propose a simple yet effective relative ranking consistency measure to model inter-relationships of multiple ranked lists.

Our measure is based on two criterion. First, relationships of relative ranks between retrieved images should be maintained. Images with similar ranks in the initial ranked lists from different modalities should also be ranked closely after reranking. Second, images with consistent ranks across multiple modalities should have their ranks preserved after reranking. An image that is similar to the query but highly ranked by only a smaller number of modalities should also be captured. In contrast to the information gain term, this relative ranking consistency measure models inter-relationships of modalities at a higher level: using ranks themselves rather than pairwise similarities between images.

Again, as in Sect. 3.2, we only consider the top  $K$  images from each ranked list and denote  $\mathcal{V}$  as the union of all retrieved images. Our goal is to select a subset of retrieved images  $\mathcal{S} \subseteq \mathcal{V}$ . We first define the *relative ranking* between a pair of images and then use it to measure the “inter-rank” consensus amongst multiple ranked lists.

Let  $\mathbf{r}_m \in \mathbb{R}^K$  denote the positions of the top  $K$  images in the initial ranked list by modality  $m$ ,  $\mathbf{r}_m = (r_{m,1}, r_{m,2}, \dots, r_{m,K})^\top$ , where  $r_{m,i}$  is the position of image  $I_i$  in the  $m$ -th ranked list. Smaller value means higher rank. The relative



**Fig. 2.** The effectiveness of the relative ranking consistency measure. The set  $\mathcal{V}$  contains  $K = 100$  images, from which we need to select an image into  $\mathcal{S}$ , which currently contains two images. Starting from initial ranks from the 3 modalities, we compute the relative ranking consistency measure between images in  $\mathcal{V}$  and  $\mathcal{S}$ . For illustration purposes, we only show the values of the relative ranking consistency measure for 3 images ( $I_1$ ,  $I_2$  and  $I_3$ ) in the set  $\mathcal{V}$ .  $I_1$  in  $\mathcal{V}$ , which is initially ranked close to images in  $\mathcal{S}$  across all modalities, has the largest relative ranking consistency  $\mathcal{C}$ . The relative ranking consistency of  $I_3$ , which is highly ranked by only a single modality, is larger than that of  $I_2$  in  $\mathcal{V}$ , which is lower ranked by all modalities. Therefore, the relative ranking consistency term favors adding  $I_1$  to  $\mathcal{S}$  as it produces the largest function value for  $T(\mathcal{S})$ . Then it favors adding  $I_3$  over  $I_2$ , which has the smallest function value. Our relative ranking consistency successfully captures inter-relationships amongst multiple ranked lists and uses them to select images.

ranking between two images is defined as

$$rr_m(v_i, v_j) = |r_{m, v_i} - r_{m, v_j}|, \quad v_i, v_j \in \mathcal{V} \quad (4)$$

where  $v_i$  and  $v_j$  correspond to images  $I_i$  and  $I_j$  in the graph representations. If either  $v_i$  or  $v_j$  is not included in the top  $K$  images by modality  $m$ ,  $rr_m(v_i, v_j)$  is set to  $K$ . The relative ranking considers the difference between ranks of retrieved images. Similarly, for modality  $m'$ , we also have the relative ranking,  $rr_{m'}(v_i, v_j)$ , of the same image pair in a different modality. On the one hand, the consensus between  $rr_m(v_i, v_j)$  and  $rr_{m'}(v_i, v_j)$  indicates that the rank relationship between  $v_i$  and  $v_j$  is reliable and should be maintained after reranking, which is related to the “consistency” between ranked lists. On the other hand, we also aim to discover images which are similar to the query but highly ranked by only a small number of modalities, thereby capturing the “distinctiveness” of specific modalities. To enforce both consistency and distinctiveness constraints, we define a relative ranking consistency measure across multiple ranked lists as

$$\mathcal{C}(v_i, v_j) = \frac{1}{Z} \sum_{m, m' \in M, m \neq m'} 1 - \frac{\min(rr_m, rr_{m'})}{K} \quad (5)$$

where  $Z = \frac{M(M-1)}{2}$  is a normalization factor corresponding to the number of all possible modality pairs. With this measure, if images  $I_i$  and  $I_j$  are ranked



similarly across multiple modalities, they will also have similar ranks in the reranked list, *i.e.*, they both will be selected and highly ranked in  $\mathcal{S}$  or both will be excluded from  $\mathcal{S}$ . This results from the constraint on relative ranking consistency. Now consider the situation in which an image  $I_i$  is ranked closely to a visually similar image  $I_j$  only in a small number of modalities. In this case, we still discover such similarity due to the use of the min function, and rank these images appropriately. If either  $v_i$  or  $v_j$  is not included in the top  $K$  images by modalities  $m$  and  $m'$ ,  $1 - \frac{\min(rr_m, rr_{m'})}{K} = 0$ , which indicates that these two images have disparate ranks and should contribute nothing to the objective function. Therefore, we take the inter-relationships amongst multiple ranked lists into account with respect to the relative ranking between two images. Several examples are shown in Fig. 2 with more explanations.

Finally, we define a set function based on the rank biased overlap (RBO) similarity [31], incorporating the aforementioned relative ranking consistency measure. RBO similarity was proposed in [31] but they did not observe or take advantage of its submodularity property. We extend the basic idea from [31] that highly ranked images should be more important than lower ranked images in our objective function. Suppose the images in  $\mathcal{S}$  are ordered and that the position of image  $I_i$  in the new ranked list is  $r_{v_i}$ . The relative ranking consistency term is defined as

$$T(\mathcal{S}) = (1 - q) \sum_{s=1}^{|\mathcal{S}|} q^s \cdot \frac{1}{s} \sum_{v_i, v_j \in \mathcal{S}, r_{v_i} < r_{v_j} = s} \mathcal{C}(v_i, v_j) \quad (6)$$

where the term  $\frac{1}{s} \sum_{v_i, v_j \in \mathcal{S}, r_{v_i} < r_{v_j} = s} \mathcal{C}(v_i, v_j)$  allows us to select the image  $v_j$  with new rank  $s$  and compute the average relative ranking measure between  $v_j$  and all other  $s$  images with higher new ranks than  $v_j$  (see Fig. 2).  $|\mathcal{S}|$  is the cardinality of  $\mathcal{S}$ . With the requirement that highly ranked images should have more weight in the objective function than lower ranked images, we introduce a weight parameter  $q$  for each image according to its new rank in  $\mathcal{S}$ .  $q$  controls the steepness of weight decay, so that a higher ranked image contributes more to the function value. Starting from the top ranked image with  $s = 1$ , the function assigns weight  $q^s$  to this image  $v_j$  and iteratively computes the average relative ranking between  $v_j$  and other higher ranked images  $v_i$  ( $r_{v_i} < r_{v_j}$ ). Maximizing this function leads to a subset of images  $\mathcal{S}$ , where images are highly ranked and similarly ranked with each other in the initial ranked list. Since at least two images are needed to compute the relative ranking consistency measure, a phantom item  $v_p$  is included into  $\mathcal{S}$  to select the first image. In practice, we use the query itself as the phantom with rank  $r_{v_p} = 0$ . Then we have the following proposition with the proof in the supplementary material.

**Proposition 2.**  $T : 2^{\mathcal{V}} \rightarrow \mathbb{R}$  is a submodular and monotone function if elements in  $\mathcal{S}$  are ordered with respect to a phantom item  $v_p \in \mathcal{S}$  and  $r_{v_p} = 0$ .

### 3.4 Optimization

Combining the information gain and relative ranking consistency terms, we obtain the final objective function  $Q(\mathcal{S}) = R(\mathcal{S}) + \lambda T(\mathcal{S})$  for the reranking

problem. The solution is obtained by maximizing the objective function:

$$\begin{aligned} & \max_{\mathcal{S}} R(\mathcal{S}) + \lambda T(\mathcal{S}) \\ & s.t. \quad \mathcal{S} \subseteq \mathcal{V}, |\mathcal{S}| \leq K_s \end{aligned} \quad (7)$$

where  $\lambda$  is a pre-defined weighting factor balancing the two terms.  $K_s$  is the largest number of selected images, which means we only select and rerank at most  $K_s$  images. Equation 7 is submodular and non-decreasing since it is a linear combination of submodular and non-decreasing functions. Direct optimization of Eq. 7 is a NP-hard problem, but it can be approximately optimized by a greedy algorithm. Starting from an empty set  $\mathcal{S} = \emptyset$ , the greedy algorithm iteratively adds a new element to  $\mathcal{S}$  which provides the largest marginal gain at each iteration, until  $K_s$  elements have been selected. Specifically, during each iteration, we search for an image  $a^* \in \mathcal{V} \setminus \mathcal{S}$ , which gives the largest combined marginal gain from the information gain and relative ranking consistency terms, add it to  $\mathcal{S}$  and set its rank to  $r_{a^*} = \rho^{cur}$ , where  $\rho^{cur}$  indicates the iteration step. The iteration terminates when  $|\mathcal{S}| = K_s$ . The reranked images are those from  $\mathcal{S}$ , and ranks are also obtained. We can tune  $K_s$  to control the efficiency and accuracy of the algorithm. The entire process is presented in Algorithm 1. The constraint on the number of reranked images leads to a uniform matroid  $\mathcal{M} = (\mathcal{V}, \mathcal{I})$ , where  $\mathcal{I}$  is the collection of subsets  $\mathcal{S} \subseteq \mathcal{V}$  satisfying the constraint that the number of reranked images is less than  $K_s$ . Maximizing a submodular function with a uniform matroid constraint yields a  $(1 - 1/e)$  approximation to the optimal solution [12].

To further accelerate the optimization, we adopt lazy evaluation [23] to avoid recomputing the function value for each node  $a^* \in \mathcal{V} \setminus \mathcal{S}$  during each iteration. The basic idea is maintaining a list of images with corresponding marginal gains in descending order. Only the top image is re-evaluated during each iteration. Other images are evaluated only if the top image does not remain at the top after re-evaluation. Lazy evaluation is based

on the diminishing returns property: the function value of an element cannot increase during iterations. The lazy greedy algorithm leads to a speed-up of more than 40, as we will show in the experiments.

---

**Algorithm 1.** Submodular Reranking

---

**Input:** Graphs  $\{\mathcal{G}_1, \dots, \mathcal{G}_M\}$ , initial ranked lists  $\{\mathbf{r}_1, \dots, \mathbf{r}_M\}$ ,  $K_s$  and  $\lambda$

**Output:** Reranked list  $\mathbf{r}$  and final retrieved images  $\mathcal{S}$

**Initialization:**  $\mathcal{S} \leftarrow \emptyset$ ,  $\rho^{cur} \leftarrow 0$ ,  $\mathbf{r} \leftarrow \mathbf{0}$

**while**  $|\mathcal{S}| < K_s$  **do**

$a^* = \arg \max_{\mathcal{S} \cup \{a\} \in \mathcal{V}} Q(\mathcal{S} \cup \{a\}) - Q(\mathcal{S})$

**if**  $Q(\mathcal{S} \cup \{a^*\}) \leq Q(\mathcal{S})$  **then**  
     break;

**end if**

$\rho^{cur} \leftarrow \rho^{cur} + 1 \quad \mathcal{S} \leftarrow \mathcal{S} \cup \{a^*\}; r_{a^*} \leftarrow$

**end**

---

## 4 Experiments

### 4.1 Experimental Setting

**Datasets.** We evaluate our submodular reranking algorithm on 4 public datasets: *Holidays* [8], *UKbench* [15], *Oxford* [7] and *Paris* [17]. The *Holidays* dataset

includes 1491 image from 500 categories, where the first image in each category is used as a query. The *UKbench* dataset contains 10200 images from 2550 objects or scenes. The *Oxford* and *Paris* datasets consist of 5062 and 6412 photos of famous landmarks in Oxford and Paris, respectively. Both datasets have 55 queries, where multiple queries are from the same landmark.

**Table 1.** Comparisons with state-of-the-art approaches. We use N-S score on *UKbench*, and mAP (in %) on other datasets. “-” means the results are not reported. Results using individual terms of our objective function are shown in the right-most columns.

Datasets	BoW [32]	GIST [33]	Color	<b>Ours</b>	[10]	[7]	[8]	[18]	[32]	[16]	[34]	[35]	[19]	IG	RRC
<i>Holidays</i>	77.2	35.0	55.8	<b>84.9</b>	84.6	-	75.1	83.9	-	78.0	82.1	76.2	61.4	83.9	73.1
<i>UKbench</i>	3.50	1.96	3.09	<b>3.78</b>	3.77	3.45	-	3.64	3.67	3.56	-	3.52	3.36	3.75	3.54
<i>Oxford</i>	67.4	24.2	8.5	<b>74.3</b>	-	66.4	54.7	68.5	81.4	-	78.0	75.2	41.3	68.5	33.0
<i>Paris</i>	69.3	19.2	8.4	<b>74.8</b>	-	-	-	-	80.3	-	73.6	74.1	-	64.6	39.2

**Evaluation Criteria.** Following [7, 8, 17], we use mean average precision (mAP) to evaluate retrieval performance on *Holidays*, *Oxford* and *Paris* datasets. For the *UKbench* dataset, we use N-S score [15] which is the average correct number of top 4 retrieved images.

**Features.** We use the visual words from [32] to construct BoW vectors except on *Holidays* dataset where we adopt Hessian affine + SIFT descriptor to construct 1M-dimension BoW vectors using single assignment and approximate k-means (AKM) [7]. Standard tf-idf weighting is used. For global representations, we use a 1192-dimension GIST feature [33] and a 4000-dimension HSV color feature with 40 bins for H and 10 bins for S and V components.

**Parameters.** The similarity between two BoW vectors is computed by cosine similarity. We use a Gaussian kernel to convert Euclidean distance  $d$  to a similarity by  $\exp(-d/\sigma)$  for GIST and color features.  $\sigma$  are empirically set to 0.34 and 0.14 respectively, and fixed in all experiments.  $q$  in Eq. 6 is set to 0.9 and  $\lambda$  in Eq. 7 is set to 0.01, both fixed in all experiments.  $K$  equals the number of dataset images in each dataset; while smaller value can be used for very large datasets.  $K_s = 1000$  for all datasets.

## 4.2 Results Comparisons

**Comparisons with State-of-the-art Approaches.** Our primary focus is a reranking algorithm that improves retrieval performance of multiple ranked lists obtained by multiple independent feature modalities. Although our implementation depends only on pairwise similarities without spatial verification and query expansion, the performance by our submodular reranking is comparable to other state-of-the-art approaches using a single modality, as shown in Table 1.

Since there are limited methods for reranking by fusion for natural image retrieval, we only compare our algorithm to [10], which is also an unsupervised reranking method using multiple feature modalities, as shown in Table 1. Note

that [11] is not directly comparable as it requires image attributes for learning. It is clear that our reranking algorithm outperforms [10], although we combine inferior individual modalities compared to [10]<sup>2</sup>. Results by our reranking are also comparable to other state-of-the-art approaches, even we only use pairwise similarities without any learning and post-processing techniques, such as query expansion and spatial verification. We improve the best single modality (BoW) by 10.0 %, 8.0 %, 10.2 % and 7.9 % on the four datasets, respectively. Additionally, without specifically inferring weight for each modality, our reranking algorithm is very robust against inferior modalities, such as the color feature on *Oxford* and *Paris*, which only achieves less than 9 % mAP. Although results on *Oxford* dataset by several approaches using a single modality [32, 34, 35] are better than those by our reranking algorithm, note that our reranking algorithm does not require SIFT descriptors or BoW vectors as [32, 34, 35] did, as long as we have pairwise similarities of pairs of images. Therefore, for the scenarios where original features cannot be stored and loaded efficiently due to limited resources, *i.e.*, mobile computing, our algorithm is more suitable than [32, 34, 35] for improving initial retrieval results. It is reasonable to expect that a higher accuracy might be obtained if we apply our reranking algorithm to fuse features which achieve better individual performance.

**Table 2.** Comparison of results by our reranking algorithm and other rank aggregation approaches. Runtime (in second) of reranking 1000 images for a single query using direct greedy optimization and lazy evaluation is shown in the right-most columns.

Datasets	Mean [36]	Median [37]	Geo-mean [37]	Robust [38]	Borda [37]	Ours	direct	lazy	speed-up
<i>Holidays</i>	59.2	71.7	76.4	71.5	59.2	<b>84.9</b>	16.5	0.40	41x
<i>UKbench</i>	2.89	3.47	3.50	3.33	2.89	<b>3.78</b>	55.7	1.34	42x
<i>Oxford</i>	18.6	34.7	40.5	35.6	18.6	<b>74.3</b>	38.3	0.74	52x
<i>Paris</i>	24.4	38.5	46.6	39.8	24.4	<b>74.8</b>	43.1	0.78	55x

**Comparisons of Individual Terms.** Our objective function consists of two terms: information gain and relative ranking consistency. These are complementary: the information gain term explores relationships between images and modalities at a fine level by using pairwise similarities, while the relative ranking consistency term exploits the inter-relationships between initial ranked lists in a coarser level as it only uses the ranks themselves. As shown in Table 1, by combining the two terms, our algorithm outperforms each individual term and achieves the best accuracy.

**Comparisons with Baselines.** We also compare the reranking accuracy of our reranking algorithm with other rank aggregation baseline approaches that combine multiple ranked lists. We use 5 rank aggregation approaches for comparison: mean rank aggregation [36], median rank aggregation [37], geometric mean

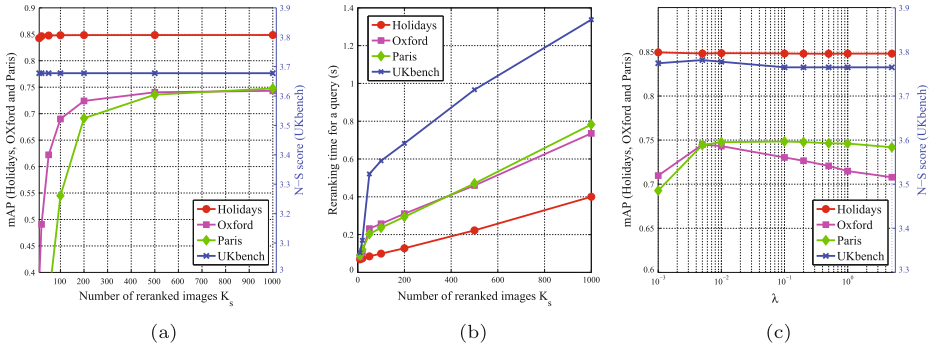
<sup>2</sup> In [10], BoW achieved 77.5 % mAP on *Holidays* and 3.54 N-S on *UKbench*, while color achieved 62.6 % and 3.17, respectively. N-S score by GIST is 2.21 on *UKbench*.

rank aggregation [37], robust rank aggregation [38] and Borda count [36, 37]. The results are shown in Table 2.

Our reranking algorithm outperforms all other rank aggregation approaches that do not as effectively use the inter-relationships amongst multiple ranked lists. The results by mean rank aggregation and Borda count are even much worse than those by a single modality (BoW), showing that a higher rank is overly diluted by other lower ranks. Incorporating the information gain and relative ranking consistency, our algorithm effectively exploits relationships of image pairs and multiple ranked lists at both a fine and a coarse level, leading to a higher retrieval accuracy.

### 4.3 Parameter Analysis

**Impact of  $K_s$ .** The parameter  $K_s$  controls the number of images to be reranked, which affects efficiency and reranking accuracy. Smaller  $K_s$  leads to fast convergence but may not discover images similar to queries but lower ranked since it discards a large number of initially retrieved images. We investigate the accuracy and execution time of our reranking with respect to  $K_s$ .



**Fig. 3.** (a) Change of mAP with respect to  $K_s$ . (b) Average reranking time for a single query with respect to  $K_s$ . (c) Change of mAP with respect to  $\lambda$ . Best view in color (Color figure online).

The retrieval accuracy in terms of mAP and average reranking time for a single query as  $K_s$  is varied are shown in Fig. 3(a), where  $K_s$  ranges from 10 to 1000. As we perform reranking on more images, the chance of discovering a similar but lower ranked image increases. Therefore, the mAP gradually improves. More specifically, the mAP rapidly increases as  $K_s$  increases from 10 to 500 for *Oxford* and *Paris* datasets. When more images are included in reranking after this point, the improvement of mAP is only incremental, showing that reranking images that are significantly lower ranked does not much benefit retrieval performance. In comparison, the mAP for *Holidays* and *UKbench* datasets reaches its highest value when  $K_s < 100$  and remains almost constant thereafter. Images in the *Oxford* and *Paris* datasets have significant variance and each query has a large number of similar dataset images that can be retrieved. Images similar to

the query can only be better discovered by a deeper inspection of initial ranked lists. In contrast, similar images in the *Holidays* and *UKbench* datasets are near-duplicates, and most queries have fewer than 10 similar images that are already highly ranked in the initial ranked lists. Therefore, only a smaller number of initially retrieved images need to be reranked.

To evaluate execution time, we calculate the average time spent to rerank  $K_s$  retrieved images for a single query in each dataset. From Fig. 3(b), it is not surprising that reranking a larger number of images takes more time. Nevertheless, our algorithm achieves sublinear time to rerank retrieved images for a single query with respect to  $K_s$ , showing the efficiency of the greedy algorithm with lazy evaluation. Furthermore, it takes the lazy evaluation less than 1.5 s on a desktop with 3.4 GHz CPU to rerank as many as 1000 images without any code optimization. Therefore, our reranking algorithm is scalable for large-scale image reranking tasks.

**Impact of  $\lambda$ .** In Eq. 7, we balance the information gain and relative ranking consistency by parameter  $\lambda$ . Since  $\lambda$  controls the importance of individual terms, it also affects the reranking accuracy. We investigate the change of reranking performance with respect to  $\lambda$ , as shown in Fig. 3(c). Our reranking algorithm is very robust: changing  $\lambda$  within a wide range does not affect the mAP too much, therefore we do not need to specifically tune  $\lambda$  to obtain good results. The change of mAP with respect to different  $\lambda$  is at most 5–6 %.

**Computational Complexity.** As stated in Sect. 3.4, we adopt a lazy evaluation approach to accelerate the optimization process. To show its effectiveness, we compare the reranking time for a single query by direct greedy optimization and lazy evaluation on the same machine, as shown in Table 2.

On all datasets, the lazy evaluation achieves more than a 40-fold speed-up compared to direct optimization. On the *Oxford* and *Paris* datasets, the lazy evaluation achieves more than a 50-fold speed-up. Therefore, our submodular reranking algorithm is very efficient and scalable for larger-scale reranking problems. With proper code optimization and parallel computing, our algorithm can be easily applied to reranking multiple ranked lists for real-time search engines.

## 5 Conclusions

We address the problem of reranking images that are initially ranked by multiple feature modalities by maximizing a submodular and monotone objective function. Our objective function is composed of an information gain term and a relative ranking consistency term. The information gain term utilizes relationships of initially retrieved images based on a random walk model on a graph. Based on this term, an image initially lower ranked but resembling other retrieved images that are similar to the query will have higher rank after reranking. The relative ranking consistency term measures the relative ranking between two initially retrieved images across multiple ranked lists. It maintains the consistency of relative ranks between two images during reranking, and also captures a high rank

of an image that is similar to the query but only discovered by one or a few modalities. The objective function can be efficiently maximized by a lazy greedy algorithm, leading to an ordered subset of initially retrieved images. Experiments show that our reranking algorithm improves overall retrieval accuracy and is computationally efficient.

**Acknowledgement.** This work was supported by the NSF EAGER grant: IIS1359900, Scalable Video Retrieval.

## References

1. Sivic, J., Zisserman, A.: Video google: a text retrieval approach to object matching in videos. In: ICCV, pp. 1470–1477 (2003)
2. Perronnin, F., Liu, Y., Sánchez, J., Poirier, H.: Large-scale image retrieval with compressed Fisher vectors. In: CVPR, pp. 3384–3391 (2010)
3. Douze, M., Ramisa, A., Schmid, C.: Combining attributes and fisher vectors for efficient image retrieval. In: CVPR, pp. 745–752 (2011)
4. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: CVPR, pp. 3304–3311 (2010)
5. Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A.: Total recall: automatic query expansion with a generative feature model for object retrieval. In: ICCV, pp. 1–8 (2007)
6. Chum, O., Mikulík, A., Perdoch, M., Matas, J.: Total recall II: query expansion revisited. In: CVPR, pp. 889–896 (2011)
7. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR, pp. 1–8 (2007)
8. Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 304–317. Springer, Heidelberg (2008)
9. Wang, M., Li, H., Tao, D., Lu, K., Wu, X.: Multimodal graph-based reranking for web image search. *IEEE Trans. Image Process.* **21**, 4649–4661 (2012)
10. Zhang, S., Yang, M., Cour, T., Yu, K., Metaxas, D.N.: Query specific fusion for image retrieval. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 660–673. Springer, Heidelberg (2012)
11. Deng, C., Ji, R., Liu, W., Tao, D., Gao, X.: Visual reranking through weakly supervised multi-graph learning. In: ICCV, pp. 2600–2607 (2013)
12. Nemhauser, G.L., Wolsey, L.A., Fisher, M.L.: An analysis of approximations for maximizing submodular set functions. *Math. Program.* **14**, 265–294 (1978)
13. Arandjelović, R., Zisserman, A.: All about VLAD. In: CVPR, pp. 1578–1585 (2013)
14. Arandjelović, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: CVPR, pp. 2911–2918 (2012)
15. Nistér, D., Stewénus, H.: Scalable recognition with a vocabulary tree. In: CVPR, pp. 2161–2168 (2006)
16. Wang, X., Yang, M., Cour, T., Zhu, S., Yu, K., Han, T.X.: Contextual weighting for vocabulary tree based image retrieval. In: ICCV, pp. 209–216 (2011)
17. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: improving particular object retrieval in large scale image databases. In: CVPR, pp. 1–8 (2008)

18. Jégou, H., Douze, M., Schmid, C.: On the burstiness of visual elements. In: CVPR, pp. 1169–1176 (2009)
19. Jégou, H., Chum, O.: Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 774–787. Springer, Heidelberg (2012)
20. Jegelka, S., Bilmes, J.: Submodularity beyond submodular energies: coupling edges in graph cuts. In: CVPR, pp. 1897–1904 (2011)
21. Kim, G., Xing, E.P., Li, F.F., Kanade, T.: Distributed cosegmentation via submodular optimization on anisotropic diffusion. In: ICCV, pp. 169–176 (2011)
22. Krause, A., Cevher, V.: Submodular dictionary selection for sparse representation. In: ICML, pp. 567–574 (2010)
23. Jiang, Z., Zhang, G., Davis, L.S.: Submodular dictionary learning for sparse coding. In: CVPR, pp. 3418–3425 (2012)
24. Jiang, Z., Davis, L.S.: Submodular salient region detection. In: CVPR, pp. 2043–2050 (2013)
25. Zhu, F., Jiang, Z., Shao, L.: Submodular object recognition. In: CVPR (2014)
26. Cao, L., Li, Z., Mu, Y., Chang, S.F.: Submodular video hashing: a unified framework towards video pooling and indexing. In: ACM Multimedia, pp. 299–308 (2012)
27. Tong, H., He, J., Wen, Z., Konuru, R., Lin, C.Y.: Diversified ranking on large graphs: an optimization viewpoint. In: KDD, pp. 1028–1036 (2011)
28. Zhu, X., Goldberg, A.B., Gael, J.V., Andrzejewski, D.: Improving diversity in ranking using absorbing random walks. In: HLT-NAACL, pp. 97–104 (2007)
29. He, J., Tong, H., Mei, Q., Szymanski, B.K.: GenDeR: a generic diversified ranking algorithm. In: NIPS, pp. 1151–1159 (2012)
30. Krause, A., Guestrin, C.: Near-optimal nonmyopic value of information in graphical models. In: UAI, pp. 324–331 (2005)
31. Webber, W., Moffat, A., Zobel, J.: A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.* **28**, 1–38 (2010)
32. Qin, D., Gammeter, S., Bossard, L., Quack, T., Gool, L.J.V.: Hello neighbor: accurate object retrieval with k-reciprocal nearest neighbors. In: CVPR, pp. 777–784 (2011)
33. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **42**, 145–175 (2001)
34. Qin, D., Wengert, C., Gool, L.V.: Query adaptive similarity for large scale object retrieval. In: CVPR (2013)
35. Shen, X., Lin, Z., Brandt, J., Avidan, S., Wu, Y.: Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking. In: CVPR, pp. 3013–3020 (2012)
36. Aslam, J.A., Montague, M.H.: Models for metasearch. In: SIGIR, pp. 275–284 (2001)
37. Dwork, C., Kumar, R., Naor, M., Sivakumar, D.: Rank aggregation methods for the web. In: WWW, pp. 613–622 (2001)
38. Kolde, R., Laur, S., Adler, P., Vilo, J.: Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* **28**, 573–580 (2012)



Computer Vision -- ACCV 2014

12th Asian Conference on Computer Vision, Singapore,

Singapore, November 1-5, 2014, Revised Selected

Papers, Part I

Cremers, D.; Reid, I.; Saito, H.; Yang, M.-H. (Eds.)

2015, XX, 727 p. 293 illus., Softcover

ISBN: 978-3-319-16864-7