

Visual Reranking through Weakly Supervised Multi-Graph Learning

Cheng Deng¹, Rongrong Ji², Wei Liu³, Dacheng Tao⁴, and Xinbo Gao¹

¹Xidian University, Xi'an, China

²Xiamen University, Xiamen, China

³IBM Watson Research Center, Armonk, NY, USA

⁴University of Technology, Sydney, Australia

{chdeng.xd, jirongrong, wliu.cu, dacheng.tao, xbgao.xidian}@gmail.com

Abstract

Visual reranking has been widely deployed to refine the quality of conventional content-based image retrieval engines. The current trend lies in employing a crowd of retrieved results stemming from multiple feature modalities to boost the overall performance of visual reranking. However, a major challenge pertaining to current reranking methods is how to take full advantage of the complementary property of distinct feature modalities. Given a query image and one feature modality, a regular visual reranking framework treats the top-ranked images as pseudo positive instances which are inevitably noisy, difficult to reveal this complementary property, and thus lead to inferior ranking performance. This paper proposes a novel image reranking approach by introducing a Co-Regularized Multi-Graph Learning (Co-RMGL) framework, in which the intra-graph and inter-graph constraints are simultaneously imposed to encode affinities in a single graph and consistency across different graphs. Moreover, weakly supervised learning driven by image attributes is performed to denoise the pseudo-labeled instances, thereby highlighting the unique strength of individual feature modality. Meanwhile, such learning can yield a few anchors in graphs that vitally enable the alignment and fusion of multiple graphs. As a result, an edge weight matrix learned from the fused graph automatically gives the ordering to the initially retrieved results. We evaluate our approach on four benchmark image retrieval datasets, demonstrating a significant performance gain over the state-of-the-arts.

1. Introduction

Visual reranking, *i.e.*, refining an image ranking list initially returned by a textual or visual query, has been intensively studied in image search engines and beyond [1]. Under such a circumstance, the initial ranking list is reordered



Figure 1. The user gap: given the positive labels as shown in light blue rectangles, how can the machine interpret the user's labeling intention?

by exploiting some inherent visual similarity measure, typically accomplished through learning a refined ranking function from accurately or noisily labeled instances. Such labeled instances are gathered via relevance judgments with respect to the query, and these judgments can be determined by an automatic scheme, *e.g.* top-ranked images as pseudo positive (relevant) instances [2], or a manual scheme, *e.g.* positive instances specified by a user [3] or click data [4]. To make relevance judgements as sufficient as possible, various methods have been investigated, ranging from a straightforward means like query expansion [5][6] to sophisticated skills such as recently developed query-relative classifier learning [7] and irrelevant image elimination [8].

On one hand, images are similar in terms of diverse visual cues including sharing similar colors or textures, containing near-duplicate objects, and reflecting close semantic concepts. The latest reranking work [9] addressed the multi-cue issue associated with visual similarities by adopting multiple modalities of visual features and fusing them in a query adaptive manner, which allows a flexible learning mechanism and demonstrates better performance than the state-of-the-art methods.

On the other hand, a successful reranking also relies on the credibility of involved labeled instances. These labels are known to be noisy in usual, due to different factors including ambiguous querying, undefined user intention, and

subjectivity in user’s judgements [4]. For example, taking the top-ranked images as pseudo positive instances tends to be unreliable because there may exist false positive samples (*a.k.a.* outliers) in the top list.

To this end, label denoising is advocated as an emerging focus in the literature. [2] assigned pseudo positive labels to a few top-ranked images, and then selected sparse smooth eigenbases of the normalized graph Laplacian, that is built on the working set of higher ranked images, to filter out the outliers and hence achieve the reliable positive labels. [8] extracted the textual clues hidden among particular search results and then exploited them as augmented queries for reranking, which can filter out the outliers as well.

Nevertheless, even provided with sufficient positive instances for labeling, how to discover the user’s search intention remains open, which is referred to as the “user gap” issue [6]. For example, given a single positive label, as shown in Figure 1(a), what is the user’s actual search purpose? Dog, flower, or flower arrangement shaped like dog? For another example displayed in Figure 1(b), an unsophisticated user is less likely to describe the content and emotions of Picasso’s paintings. Instead, a higher-level mining to discover the underlying semantic attributes of images, if feasible, could help a lot.

It turns out that such a semantic mining step is specially valuable for enhancing the reranking performance. To this end, we propose an image attribute driven learning framework, namely *Weakly Supervised Multi-Graph Learning*, to address visual reranking. In this scenario, the (pseudo) labeled instances are not directly used as seed labels for reranking, but undergo a selective procedure like the scenario of weakly supervised learning [10][11]. The mined image attributes are subsequently leveraged to learn a refined ranking function by applying proper off-the-shelf learning algorithms such as multi-view learning and graph-based semi-supervised learning.

We follow a state-of-the-art method [9] for multi-feature fusion based visual reranking, upon which we conduct graph-based learning rather than straightforward feature learning, aiming at capturing the intrinsic manifold structure underlying the images to be reranked. In our proposed framework, multiple retrieved image sets stemming from different modalities of visual features are expressed into multiple graphs, which are aligned and then fused towards learning an optimal similarity metric across multiple graphs for sensible reranking. The weakly supervised learning driven by image attributes can yield critical graph anchors within each graph, which enable the effective alignment and fusion across multiple graphs. Extensive experimental results shown on four benchmark image datasets, *i.e.*, **Oxford**, **Paris**, **INRIA Holidays** and **UKBench**, bear out that the proposed reranking approach outperforms the state-of-the-arts [9][12][13] by a significant margin in terms of ro-

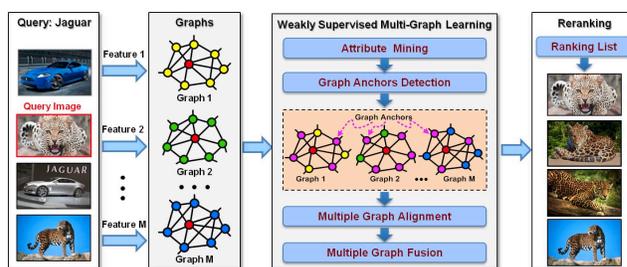


Figure 2. Our proposed image reranking approach.

business and accuracy.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 presents the proposed visual reranking framework based on multi-graph learning. Section 4 describes the graph anchor seeking procedure using weakly supervised learning via co-occurred attribute mining. Section 5 gives the experimental results, and Section 6 concludes the paper.

2. Related Work

Graph-based Reranking. Graph-based image reranking methods have shown promising performance recently. It targets at refining the initial ranking list by propagating the initial rank scores of seed (or anchor) nodes to the other nodes in a graph [2][8]. In [14], the video reranking process was modeled as a modified PageRank over a set of graphs to propagate the final ranking scores. Wang et al. [15] integrated multiple feature modalities into a graph-based learning algorithm for image reranking. Zhang et al. [9] proposed a graph-based query specific fusion approach, where multiple retrieval sets from different visual cues were merged and reranked by link analysis on a fused graph.

Attributes. Learning image attributes for object recognition has been well established in [16][17][18]. In [17], semantic attributes and non-semantic attributes were learned for recognizing objects within categories and across categories. Moreover, attributes can be used as mid-level features for scene recognition [19], face recognition [20], and image retrieval [21]. There also exist some recent endeavors aiming to discover attributes interactively [22] or from noisy web-crawled data [23].

Weakly Supervised Learning. Weakly supervised learning methods have been extensively studied in the recent literature. For instance, [10] unified weakly supervised learning into undirected graphical models for object recognition; [11] learned object categories in a weakly supervised manner for object recognition. In [24], weakly supervised information was integrated with latent SVMs to conduct object localization. In [25], accurate semantic segmentation with a multi-image model was achieved by supplementing weakly supervised classes to training images.

3. Reranking via Multi-Graph Learning

Notations. Given a query image I_q and its initial top- N ranking list $\mathcal{I} = \{I_i\}_{i=1}^N$. Let D^m be the feature dimension for the m -th visual feature channel ($m \in \{1, \dots, M\}$) whose corresponding feature set for \mathcal{I} is denoted as $\mathbf{X}^m = \{\mathbf{x}_i^m\}_{i=1}^N$. For the m -th feature channel, we construct a weighted undirected graph $G^m = (V^m, E^m, \mathbf{w}^m)$, where each node in V^m corresponds to an image. $E^m = \{(i, j)\}$ is the edge set, and $\mathbf{w}^m \in \mathbb{R}^{N \times N}$ is edge weight matrix where each w_{ij}^m represents the edge weight over (i, j) to be learned. The aggregated weight matrix $\widetilde{\mathbf{W}}$ from \mathbf{w}^m will be used as the final scores to refine the initial ranking results.

Let $Z = \{z_i\}_{i=1}^N$ be a binary indicator vector to label all images, where $z_i = 1$ means that the image I_i is a graph anchor and $z_i = 0$ otherwise, with anchor feature set $\widetilde{\mathbf{X}}^m$

$$\widehat{\mathbf{X}}^m = Z \circ \mathbf{X}^m, \quad \widetilde{\mathbf{X}}^m = \{\hat{\mathbf{x}}_k^m\}_{k=1}^A. \quad (1)$$

where “ \circ ” is the indicator operation to describe the procedure of graph anchor selection, A is the anchor number, and $\hat{\mathbf{x}}^m$ is the anchor feature vector for the m -th feature channel.

Problem Formulation. The diagram of our approach is shown in Figure 2. We analyze multi-graph learning based on two intuitions. First, we consider *intra-graph* constraints where the distribution agreement on “anchor-to-anchor” and on “query-to-anchor” should be maximized. Second, we introduce *inter-graph* constraints where the pairwise distribution between pairs of anchors and non-anchors across graphs should behave consistently. These two intuitions are respectively formulated as the objective function and the regularizer in a learning framework, called **Co-Regularized Multi-Graph Learning (Co-RMGL)**. Co-RMGL can be interpreted as a multiple graphs fusion algorithm via graph anchor alignment, as described in Eq. (4).

Given a set of graph anchor¹ for the m -th feature channel, the intra-graph learning aims to obtain a new edge weight matrix \mathbf{w}^m to minimize the distances between the query and the anchors, as well as between pairwise anchors. Here, because the whole graph can be approximated as a set of overlapped linear neighborhood patches, we instead exploit the locally linear reconstruction (LLR) method (like [26]) to describe the distance constraints encoded into the weight matrix. This results in the following objective function for intra-graph learning:

$$\mathcal{Q} = \sum_{i=1}^A \left\| \hat{\mathbf{x}}_i^m - \sum_{j \neq i}^A w_{ij}^m \hat{\mathbf{x}}_j^m \right\|_2^2 + \left\| \mathbf{x}_q - \sum_{i=1}^A w_{qi}^m \hat{\mathbf{x}}_i^m \right\|_2^2, \quad (2)$$

where the first term is the reconstruction error of a given anchor $\hat{\mathbf{x}}_i^m$ using other anchors $\hat{\mathbf{x}}_j^m$ ($j \neq i$), and the second term is the reconstruction error of the query \mathbf{x}_q using all anchors $\hat{\mathbf{x}}_i^m$.

¹The selection of anchors will be detailed later in Section 4.

To achieve the inter-graph learning, we impose the following inter-graph regularizer for the m -th feature channel:

$$\mathcal{R} = \sum_{i=1}^A \left(\left\| \hat{\mathbf{x}}_i^m - \sum_{k=1}^{N-A} w_{ik}^m \mathbf{x}_k^m \right\|_2^2 - \left\| \hat{\mathbf{x}}_i^{m'} - \sum_{k=1}^{N-A} w_{ik}^{m'} \mathbf{x}_k^{m'} \right\|_2^2 \right), \quad (3)$$

where m' is another feature channel ($m' \neq m$). Similar to Eq. (2), LLR is used again to measure the distribution consistency between pairs of anchors and non-anchors. We denote \mathbf{W} a weight matrix to be learned among all M feature channels by concatenating all column vectors of \mathbf{w}^m , *i.e.*, $\mathbf{W} = [\mathbf{w}^1 | \mathbf{w}^2 | \dots | \mathbf{w}^M]$. By unifying Eq. (3) and Eq. (2), with non-negative and normalized constraints, we derive the overall objective function as follows:

$$\begin{aligned} \mathcal{J}(\mathbf{W}, \lambda, \gamma) = & \min_{\mathbf{W}} \sum_{m=1}^M \mathcal{Q}(\mathbf{w}^m) + \lambda \sum_{m' \neq m}^M \mathcal{R}(\mathbf{w}^{m'}) \\ & + \gamma \sum_{m=1}^M \|\mathbf{w}^m\|_1, \quad \text{s.t.} \sum_j w_{ij} = 1, w_{ij} \geq 0, \end{aligned} \quad (4)$$

where $\|\cdot\|_1$ is the ℓ_1 -norm of a matrix, $\lambda > 0$ balances the effect of the disagreement between inter graphs, $\gamma > 0$ controls the sparsity of the edge weight matrix.

Optimization. Let $\mathcal{H}(\mathbf{w}^m) = \mathcal{Q}(\mathbf{w}^m) + \lambda \mathcal{R}(\mathbf{w}^m)$ be a smooth function; the gradient $\nabla \mathcal{H}(\mathbf{w}^m)$ can be directly derived. Since $\mathcal{J}(\cdot)$ only involves a very simple non-smooth portion (*i.e.*, the ℓ_1 -norm penalty), we adopt the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [27], as shown in Algorithm 1, to minimize $\mathcal{J}(\cdot)$. The learned weight matrix \mathbf{w}^m are merged into $\widetilde{\mathbf{W}}$ as the final similarity scores to perform reranking.

Efficiency Analysis. It has been proved in [28] that Algorithm 1 can achieve an $O(1/\epsilon)$ convergence rate for a desired accuracy loss ϵ . As for the time complexity, the main computational cost in each iteration comes from calculating the gradient $\nabla \mathcal{H}(\mathbf{v}_t)$, which costs $O((D^m)^2 M)$ time.

Correlation to Existing Graph Alignments. As far as we know, this is the first time to conduct multi-graph alignment with graph anchors. For a better alignment, as shown in Figure 2, the common anchors across multiple graphs are retained as many as possible, while uncommon ones are omitted. Our formulation is general enough to unify several existing graph fusion techniques developed for reranking and beyond [2][9]. Now, the remaining problem is how to seek graph anchors, as addressed by the following section.

4. Weakly Supervised Anchor Seeking

Issues. Graph anchors are first introduced for graph-based SSL in [29] where K-Means clustering centers are used for graph anchors. While the most straightforward approach is to treat the pseudo labeled instances as anchors,

Algorithm 1: Co-RMGL for Multi-Graph Learning.

1 **Input:** $\{\mathbf{X}^m\}_{m=1}^M, \mathbf{w}_0, \lambda, \gamma$.
2 **Initialization:** set $\theta_0 = 1, \mathbf{v}_0 = \mathbf{w}_0$.
3 **for** $t = 0, 1, 2, \dots$ until convergence of \mathbf{w}_t **do**
4 Compute $\nabla\mathcal{H}(\mathbf{v}_t^m)$.
5 Set $\nabla\mathcal{H}(\mathbf{v}_t) = [\nabla\mathcal{H}(\mathbf{v}_t^1) | \dots | \nabla\mathcal{H}(\mathbf{v}_t^M)]$.
6 Compute Lipschitz constant $L = \lambda_{max}(\nabla\mathcal{H}(\mathbf{v}_t))$,
 where λ_{max} is the largest eigenvalue.
7 Perform the generalized gradient update step:
$$\mathbf{w}_{t+1} = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{2} \left\| \mathbf{w} - \left(\mathbf{v}_t - \frac{1}{L} \nabla\mathcal{H}(\mathbf{v}_t) \right) \right\|_2^2 + \frac{\gamma}{L} \|\mathbf{w}\|_1. \quad (5)$$

 Set $\theta_{t+1} = \frac{2}{t+3}$.
8 Set $\mathbf{v}_{t+1} = \mathbf{w}_{t+1} + \frac{1-\theta_t}{\theta_t} \theta_{t+1} (\mathbf{w}_{t+1} - \mathbf{w}_t)$.
9 **end**
10 **Output:** \mathbf{w} .

three issues remain open: (1) not all labels are reliable enough to be regarded as positive; (2) user’s search intention cannot be reflected to refine the original noisy labels; and (3) the expensive human labor prevents the scalability of user’s labeling, as previously discussed in Section 1.

Our Inspiration. Our solution is to discover the intrinsic attributes among the labeled instances, upon which we seek a better anchor set. This is accomplished by mining discriminative attributes from all attribute vectors of the initially retrieved results, via the cutting-edge image descriptors like Clasemes [30] or ObjectBank [31]. The mined attributes are then utilized to select top-ranked images with the maximum responses as the target anchors.

Mining Co-occurred Attributes. We introduce an effective yet efficient attributes discovery scheme based on Aprior [32] over the attribute vectors detected from all retrieved results, which, as shown in our subsequent experiments, has superior performance over the straightforward attribute vectors intersection scheme.

Formally speaking, we use Clasemes to derive the middle-level attribute vector set \mathcal{A} from the initial retrieval set \mathcal{I} . In attribute mining, let \mathcal{S} be a set of co-occurred attributes, we first transform all possible attribute combinations into a transaction database $\mathcal{D} = \{\mathbf{T}_1, \dots, \mathbf{T}_U\}$ offline with an occurrence thresholding to binarize the appearance/disappearance of a given attribute vector. The *support* and *confidence* of a co-occurred attribute set $\mathcal{A} \subseteq \mathcal{S}$ is respectively defined as

$$\operatorname{sup}(\mathcal{A}) = \frac{|\{\mathbf{T} \in \mathcal{D} | \mathcal{A} \subseteq \mathbf{T}\}|}{|\mathcal{D}|}, \quad (6)$$

Algorithm 2: Weakly Supervised Anchor Seeking.

1 **Input:** Iteration $t = 0$, maximal iteration T , initial retrieval set \mathcal{I} , attribute vector set \mathcal{A} , and \mathcal{Q}_c .
2 **while** $t < T$ **do**
3 Mining co-occurred attributes \mathcal{S} by using Eq. (6)
 and Eq. (7).
4 $t++$.
5 **end**
6 **for** co-occurred attributes \mathcal{S} **do**
7 Computing associated discriminative vector \mathcal{C}
 based on Eq. (8).
8 Setting the images having maximum responses
 with \mathcal{C} as anchors.
9 **end**
10 **Output:** graph anchor set $\tilde{\mathbf{X}}^m = \{\tilde{\mathbf{x}}_i^m\}_{i=1}^A$.

$$\operatorname{conf}(\mathcal{A} \Rightarrow \mathcal{B}) = \frac{\operatorname{sup}(\mathcal{A} \cup \mathcal{B})}{\operatorname{sup}(\mathcal{A})} = \frac{|\{\mathbf{T} \in \mathcal{D} | (\mathcal{A} \cup \mathcal{B}) \subseteq \mathbf{T}\}|}{|\{\mathbf{T} \in \mathcal{D} | \mathcal{A} \subseteq \mathbf{T}\}|}. \quad (7)$$

Here, \mathcal{B} is another co-occurred attribute set. The confidence is regarded as a maximum likelihood estimation of the conditional probability that \mathcal{B} is true if \mathcal{A} is true. Then, Aprior is used to derive discriminative attributes (at most 50). Some visualized examples can be seen in Figure 8.

Weakly Supervised Anchor Seeking. Given the mined attributes, we then select A images with the maximum responses as the anchors for graph alignment and fusion in Section 3. Intuitively, we utilize the co-occurred attributes to generate an associated discriminative vector, with which the images having the maximum responses are found as the anchors. The associated discriminative vector \mathcal{C} is described as

$$\mathcal{C}^m = \operatorname{SimDis}_{\{K\}}(\mathcal{S}_k^m, \mathcal{Q}_c), \quad I_{\text{anchor}}^m = \operatorname{MaxRes}(\mathcal{C}^m, \mathcal{A}), \quad (8)$$

where K is the number of \mathcal{S}^m in m -th feature channel, \mathcal{S}_k^m is k -th co-occurred attribute pattern, and \mathcal{Q}_c is the query’s attribute vector. “SimDis” operator means to find the most similar \mathcal{S}_k^m with \mathcal{Q}_c , and “MaxRes” operator is used to select the graph anchors with the maximum responses. In order to implement the operation in Eq. (8), \mathcal{S}_k^m should first be expanded to the full-dimension of the attribute vector. Algorithm 2 gives a procedure, namely **Weakly Supervised Anchor Seeking (WSAS)**, to yield the desirable anchor set.

Efficiency Analysis. Note that both Clasemes based attribute description and its transaction set are done offline. Subsequently, the computational cost is mainly spent on the online calculation of co-occurred attribute set and the maximum responses, which totally require $O(Ld)$ time. Note that L is the number of the labeled instances, and d is the dimensionality of co-occurred attributes.

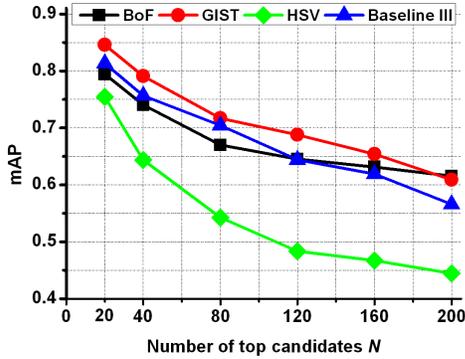


Figure 3. mAP on Oxford5k with numbers of top N images.

5. Experiments

Datasets. We evaluate our method on four popular datasets: Oxford [33], Paris [34], INRIA Holidays [35], and UKBench [36]. Oxford and Paris respectively contain 5,062 and 6,412 images, which are all provided with ground truth by manual annotation. INRIA includes 1,491 relevant images of 500 scenes or objects, where the first image is used as a query in each group. 100,000 and 1M images randomly downloaded from *Flickr* are respectively added as distractors to form the Oxford105k and INRIA 1M, which test the performance of our reranking approach. UKBench contains 10,200 images that always show the same object.

Evaluation Criteria. We use mean Averaged Precision (mAP) to evaluate the performance on the first three datasets, while the performance measure on the UKBench is the average number of correct returning in top-4 images, denoted as Ave. Top Num.

Baselines. To comprehensively evaluate our proposed scheme, four intermediate alternatives are treated as baselines: (I) individual three feature channels without learning the edge weight matrix \mathbf{w}^m ; (II) individual feature channels with learning the edge weight matrix \mathbf{w}^m ; (III) direct fusion into a final weight matrix on Baseline I; (IV) direct fusion into a final weight matrix on Baseline II.

Features. Following the state-of-the-art setting in multi-feature fusion based reranking, we design the following feature channels.

- *BoF*: We use dense SIFT descriptor [37] computed from 16×16 sized image patches with a step-size of 8 pixels using VLFeat library². Then, 1,024-dimensional visual words is constructed with 1M descriptors. We use 1×1 , 2×2 , 3×1 sub-regions to compute a BoF as the final image feature.
- *GIST*: We calculate a 960-dimensional GIST [38] descriptor for each color image. The images are resized to 32×32 and then orientation histograms are computed

²<http://www.vlfeat.org/>

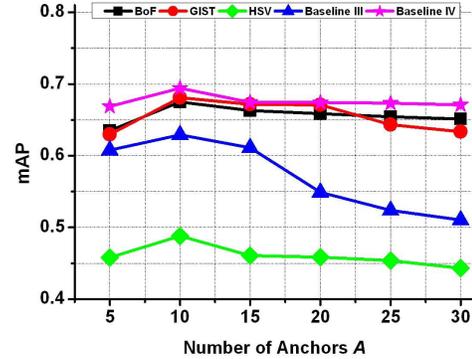


Figure 4. mAP on Oxford5k with different numbers of anchors A .

ed on a 4×4 grid. Three scales (8, 8, and 4) are used as the number of orientations.

- *HSV*: We generate a 2,000-dimensional HSV color histogram feature. Specifically, we use 20, 10, and 10 bins for H , S , V respectively.

Selecting Labeling Instances. As described in Section 4, our scheme is a generic approach for labeling instances selection with either supervised manner (user labeling [3][4]) or unsupervised manner (pseudo positive [1][2]), both tested as follows: (1) unsupervised manner: the top- L images from the initial ranking list as labeling instances. (2) supervised manner: the manually selected L images from the initial ranking list as labeling instances.

For both case, we run our weakly supervised anchor learning as in Algorithm 2 to come up with an extended and purified label set, i.e., graph anchors.

Parameter Tuning. In our method, the top- N dataset candidates for the query image I_q are considered to evaluate reranking performance. In the objective function learning, we use $\lambda = 0.1$ and $\gamma = 0.05$, which demonstrate that these setting can yield the best performance [8].

We first evaluate the performance of our approach given different numbers of top dataset candidates N on Baseline III. Figure 3 shows the performance on Oxford5k when we change N . when N becomes larger, the mAP of each feature channel and fusion continues to decrease. With N increasing from 20 to 200, the mAP of fusion drops from 0.81 to 0.56. The mAP of direct fusion is lower than the one of GIST-based reranking since the complementary properties of different feature channels are not exploited.

For all other datasets, the mAP also decreases with the increase of N . We use the same setting (N from 20 to 200) in all these datasets. Specially, since the queries in UKBench only have three relevant images, the performance on this dataset drops slowly with the increase of N . In the subsequent experiments, without specification, we fix $N = 200$ for all datasets but UKBench.

The number of anchors A is directly related to the accuracy and scalability of our scheme. Figure 4 shows the mAP

Table 1. Comparisons on Oxford5k, Oxford105k, and Paris.

Dataset	Our scheme	[13]	[12]	[33]	[5]
Oxford5k	0.843	0.814	0.849	0.647	0.827
Oxford105k	0.802	0.767	0.795	0.541	0.767
Paris	0.834	0.803	0.824	0.689	0.805

Table 2. Comparisons on INRIA, INRIA 1M, and UKBench.

Dataset	Our scheme	[13]	[12]	[33]	[35]
INRIA	0.847	- ¹	0.758	-	0.848
INRIA 1M	0.794	-	-	-	0.77
UKBench	3.75	3.67	-	3.45	3.64

varies with A on Oxford5k under the case of supervised labeling selection when $N = 200$. The mAP of fusion, both Baseline III and Baseline IV, reaches the maximum when $A = 10$, and then decreases with the increase of A , since more irrelevant images are added. Considering computational efficiency and retrieval accuracy, we set $A = 10$ on Oxford105k, Paris, INRIA, and INRIA 1M, in all of which similar phenomena are observed. For UKBench, $A = 3$.

Comparison Results. Figure 5 and Figure 6 illustrate the mAP of the Baseline III and Baseline IV on four datasets. It shows that the ranking performance is steadily improved when we incrementally add the designed components as in Section 3 and Section 4 into the multi-graph learning framework. In Figure 5 and Figure 6, “graph alignment” refers to the direct fusion after aligning multiple graphs guided by a set of anchors that are determined by clustering centers; while “anchor learning” also refers to direct fusion where multiple graphs are aligned through a set of anchors that are selected by attribute intersection.

Figure 5 and Figure 6 together further compare the performance of unsupervised and supervised labeling instances selections on four datasets when the number of labeling instances $L = 30$. We find that both labels selection criteria achieve relatively good performance with either unsupervised or supervised, which demonstrates that our method is generalized and compatible for different labeling instances selection schemes. In addition, the mAP of supervised criteria is improved by nearly 2% over the unsupervised one.

In weakly supervised attribute learning, there are two methods to select anchors. One is to directly calculate a histogram intersection operation over two attribute vectors (AI), named **Co-RMGL+AI**. The other is selection of top-ranked images with maximum responses as anchors via attribute mining (AM), named **Co-RMGL+AM**. We also verify “Baseline III+Classemes” which directly use Classemes to align graphs on Baseline III without attribute-based anchor selection. Figure 7 shows the performance comparison for these two methods on all datasets under

¹“-” means the mAP is not reported in the corresponding methods.

Table 3. Average query time of individual stage on Oxford5k.

Stage	BoF ¹	GIST	HSV	Anchor selection	Graph fusion
Avg. Time (s)	1.18	1.12	1.04	2.11	0.1

Table 4. Addition memory cost and average query time per dataset.

Dataset	Oxford5k	Paris	INRIA	UKbench
Memeory (GB)	0.01	0.02	0.02	0.01
Avg. Time (s)	2.6	2.8	3.4	4.6

the premise of supervised labeling instances selection. As shown in Figure 7, we conclude that both Co-RMGL+AM and Co-RMGL+AI significantly improve the performance over Baseline III and “Baseline III+Classeme”, while Co-RMGL+AM significantly outperforms Co-RMGL+AI in all datasets. Moreover, as we can see from Figure 7, our scheme makes great improvement by comparison with [9] in all datasets except UKBench.

We further compare our LLR based metric with unsupervised distance metric learning (UDML) for the stage of intra-graph learning, the latter of which learns similarity metrics in individual feature channels, potentially with a fusion operation to achieve reranking. Figure 7 compares the performance of our method and UDML, which shows our methods, including Baseline III, Co-RMGL+AI, and Co-RMGL+AM, significantly outperform this alternative and therefore prove our correctness.

Table 1 and Table 2 show the comparisons of our scheme with other state-of-the-art schemes on all datasets. Since most of these methods are visual word based image retrieval, their results are measured by only one feature channel. The results of our approach are among the best on Oxford105k, Paris, INRIA 1M and UKBench. On Oxford and INRIA, our scheme is next only to the best case. Nevertheless, on Oxford105k and INRIA 1M, our approach only decreases by 6.2% and 4.8%, respectively. In contrast, the best competing methods drops 9.2% and 6.4%, respectively, which indicates that our approach is more robust to large-scale datasets.

Computation Cost. The average search time for a query depends on the scale of dataset. Table 3 lists the averaged times for each operation steps on Oxford5k dataset. Table 4 shows an overview over the total memory overhead per dataset and the average query time overhead for each query, as tested in Intel Xeon with 3.47GHz CPU and 24G memory. Both Table 3 and Table 4 indicate that our scheme achieves relatively high time efficiency.

Case Study. Figure 8 shows some visualized results of our Co-RGML+AM reranking on Oxford5k, INRIA and Paris respectively, with comparisons with Baseline III and

¹We pre-compute and store all of these features offline.

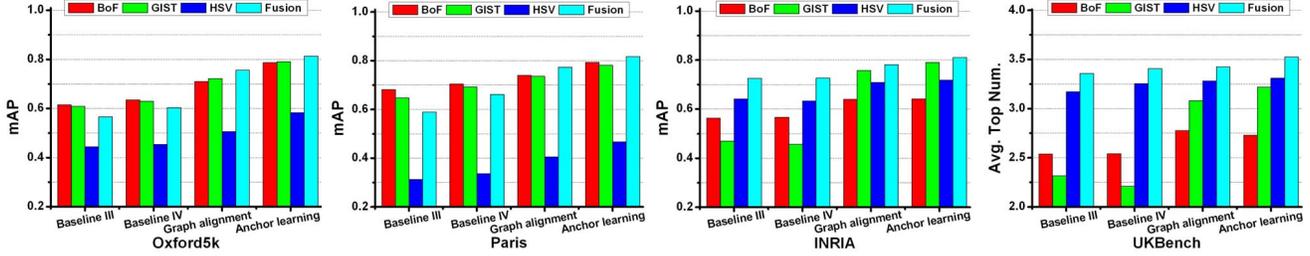


Figure 5. Comparisons mAP of four datasets under “unsupervised labeling instances selection” when we set $A = 10$, in which each group respects different stages of our approach, such as Baseline III, Baseline IV, graph alignment and anchor learning (Co-RMGL+AI).

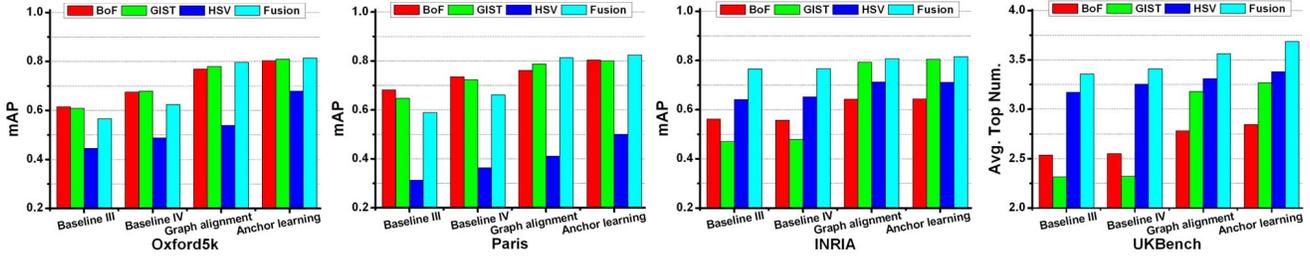


Figure 6. Comparisons mAP of four datasets under “supervised labeling instances selection” when we set $A = 10$, in which each group respects different stages of our approach, such as Baseline III, Baseline IV, graph alignment and anchor learning (Co-RMGL+AI).

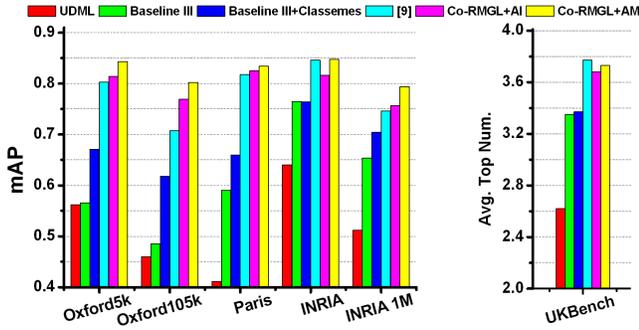


Figure 7. Comparison of Co-RMGL+AM with UDML, Baseline III, Baseline III+Classeme, and Co-RMGL+AI.

[9]. It is obvious that our approach is superior to Baseline III and [9] since our scheme has great ability to rank the relevance images in front for simple object images as well as complex scene images. For example, for the query “hertford” on Oxford5k, our scheme can rank relevance images into top 9, because it can automatically refine the initial labeling instances and extend the mined relevance to other images that can not be recognized by previous methods.

6. Conclusion

In this paper we propose a novel visual reranking approach through performing weakly supervised multi-graph learning. The contributions of our proposed approach primarily lie in: (1) the reranked result is yielded by integrating distinct modalities of visual features; (2) the reranking task is formulated as a multi-graph learning paradigm in-

corporating intra-graph and inter-graph constraints; (3) the graph anchors are intelligently sought via weakly supervised learning; (4) automatic multiple graph alignment and fusion are achieved by means of the graph anchors. Extensive experimental results bear out that our approach is not only effective but also efficient, and significantly outperforms the state-of-the-arts.

Acknowledgement We want to thank the helpful comments and suggestions from the anonymous reviewers. This research was supported partially by the National Natural Science Foundation of China (Nos. 61125204, 61101250 and 61373076), the Program for New Century Excellent Talents in University (NCET-12-0917), and the Program for New Scientific and Technological Star of Shaanxi Province (No. 2012KJXX-24).

References

- [1] W. Hsu and L. Kennedy and S.-F. Chang. Reranking methods for visual search. *IEEE Multimedia*, 2007.
- [2] W. Liu and Y. Jiang and J. Luo and S.-F. Chang. Noise resistant graph ranking for improved web image search. *CVPR*, 2011.
- [3] X. Tian and D. Tao and X.-S. Hua and X. Wu. Active reranking for web image search. *IEEE TIP*, 2010.
- [4] V. Jain and M. Varma. Learning to re-rank: Query-dependent image reranking using click data. *WWW*, 2011.
- [5] O. Chum and A. Mikulík and A. Perdoch and J. Matas. Total recall II: Query expansion revisited. *CVPR*, 2011.
- [6] X. Wang and K. Liu and X. Tang. Query-specific visual semantic spaces for web image re-ranking. *CVPR*, 2011.
- [7] J. Krapac and M. Allan and J. Verbeek and F. Juried. Improving web image search results using query-relative classifiers. *CVPR*, 2010.
- [8] J. Lu and J. Zhou and J. Wang and T. Mei and X.-S. Hua and S. Li. Image search results refinement via outlier detection using deep contexts. *CVPR*, 2012.
- [9] S. Zhang and M. Yang and T. Cour and K. Yu and D. N. Metaxas. Query specific fusion for image retrieval. *ECCV*, 2012.

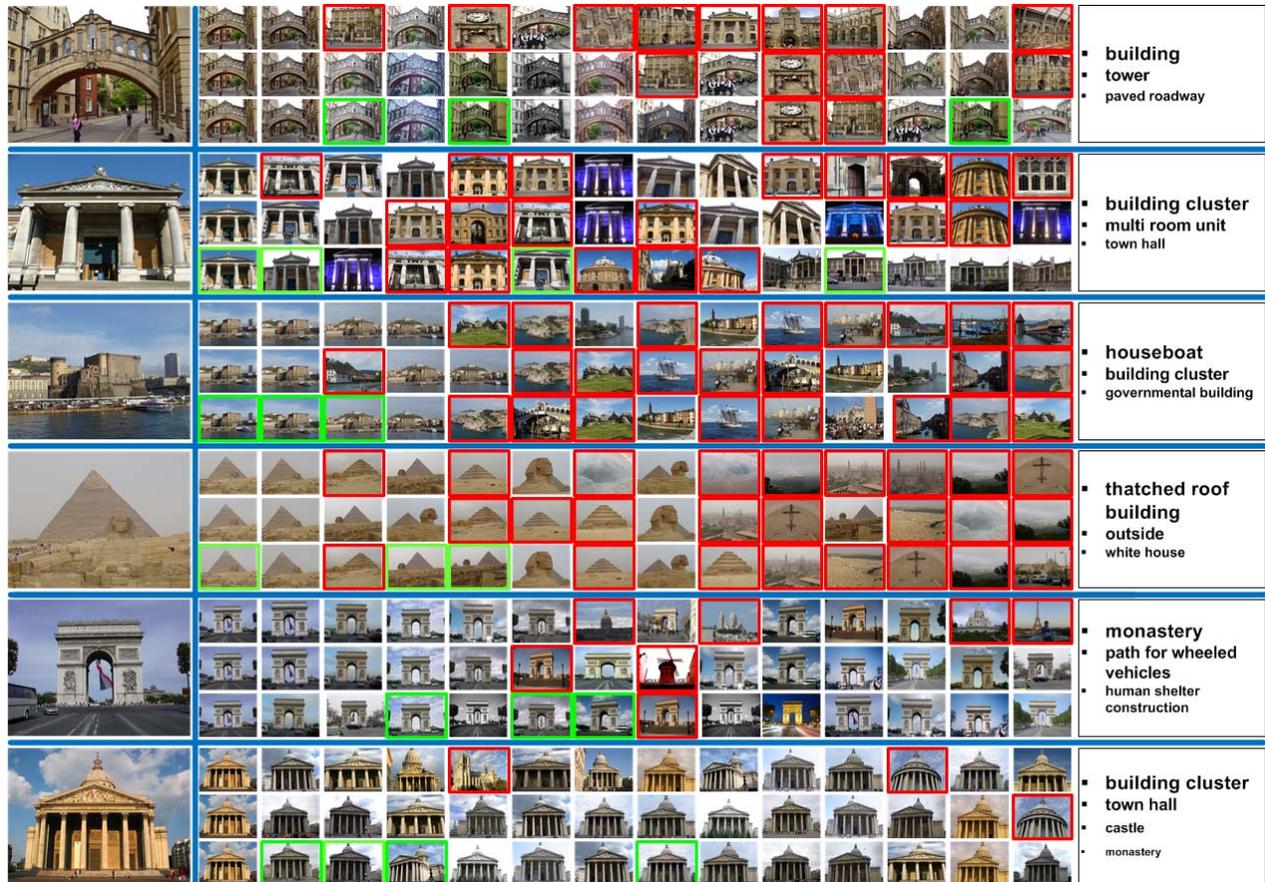


Figure 8. Visual results of the image reranking results among Baseline III (first row in each query), [9] (second row in each query), and Co-RMGL+AM (third row in each query) on Oxford, INRIA and Paris (red rectangle indicates the irrelevance images with the query, and green rectangle represents part of learned graph anchors. The last column lists some important attributes mined by our proposed WSAS).

- [10] D. Crandall and D. Huttenlocher. Weakly supervised learning of part-based spatial models for visual object recognition. *ECCV*, 2006.
- [11] R. Fergus and P. Perona and A. Zisserman. Weakly supervised scale-invariant learning of models for visual recognition. *IJCV*, 2007.
- [12] A. Mikulík and M. Perd'och and O. Chum and J. Matas. Learning a fine vocabulary. *ECCV*, 2010.
- [13] D. Qin and S. Gammeter and L. Bossard and T. Quack and L. VanGool. Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. *CVPR*, 2011.
- [14] J. Liu and W. Lai and X.-S. Hua and Y. Huang, and S. Li. Video search reranking via multi-graph propagation. *ACM MM*, 2007.
- [15] M. Wang and H. Li and D. Tao and K. Lu and X. Wu. Multimodal graph-based reranking for web image search. *IEEE TIP*, 2012.
- [16] V. Ferrari and A. Zisserman. Learning visual attributes. *NIPS*, 2007.
- [17] A. Farhadi and I. Endres and D. Hoiem and D. Forsyth. Describing objects by their attributes. *CVPR*, 2009.
- [18] N. Kumar and A. C. Berg and P. N. Belhumeur and S. K. Nayar. Attribute and simile classifiers for face verification. *ICCV*, 2009.
- [19] D. Parikh and K. Grauman. Relative Attributes. *ICCV*, 2011.
- [20] C. Lampert and H. Nickisch and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. *CVPR*, 2009.
- [21] M. Douze and A. Ramisa and C. Schmid. Combining attributes and Fisher vectors for efficient image retrieval. *CVPR*, 2011.
- [22] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. *CVPR*, 2011.
- [23] T. Berg and A. Berg and J. Shih. Automatic attribute discovery and characterization from noisy web data. *ECCV*, 2010.
- [24] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. *ICCV*, 2011.
- [25] A. Vezhnevets and V. Ferrari and J. M. Buhmann. Weakly supervised semantic segmentation with a multi-image model. *ICCV*, 2011.
- [26] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000.
- [27] A. Beck and M. Teboulle. A fast iterative shrinkage thresholding algorithm for linear inverse problems. *SIAM J. Image Science*, 2009.
- [28] X. Chen and Q. Lin and S. Kim and J. Carbonell and E. Xing. Smoothing proximal gradient method for general structured sparse regression. *Ann. Appl. Stat.*, 2012.
- [29] W. Liu and J. He and S.-F. Chang. Large graph construction for scalable semi-supervised learning. *ICML*, 2010.
- [30] L. Torresani and M. Szummer and A. Fitzgibbon. Efficient object category recognition using classemes. *ECCV*, 2010.
- [31] L.-J. Li and H. Su and E. P. Xing and L. Fei-Fei. Object Bank: A high-level image representation for scene classification & semantic feature sparsification. *NIPS*, 2010.
- [32] T. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. *VLDB*, 1994.
- [33] J. Philbin and O. Chum and M. Isard and J. Sivic and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. *CVPR*, 2007.
- [34] J. Philbin and O. Chum and M. Isard and J. Sivic and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. *CVPR*, 2008.
- [35] H. Jégou and M. Douze and C. Schmid. Hamming embedding and weak geometry consistency for large scale image search. *ECCV*, 2008.
- [36] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. *CVPR*, 2006.
- [37] S. Lazebnik and C. Schmid and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognition natural scene categories. *CVPR*, 2006.
- [38] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 2001.