Bayesian Visual Reranking

Xinmei Tian, Linjun Yang, *Member, IEEE*, Jingdong Wang, *Member, IEEE*, Xiuqing Wu, and Xian-Sheng Hua, *Member, IEEE*

***Abstract*—Visual reranking has been proven effective to refine text-based video and image search results. It utilizes visual information to recover "true" ranking list from the noisy one generated by text-based search. Visual reranking improves text-based search results by incorporating both textual and visual information. In this paper, we model the textual and visual information from the probabilistic perspective and formulate visual reranking as an optimization problem in the Bayesian framework, termed Bayesian visual reranking. In this method, the textual information is modeled as a likelihood, to reflect the disagreement between reranked results and text-based search results. The disagreement between two search results is named ranking distance. The visual information is modeled as a conditional prior, to indicate the ranking score consistency among visually similar samples which is called visual consistency in this paper. Bayesian visual reranking derives the best reranking results by maximizing visual consistency while minimizing ranking distance. It is a general framework and can unify several existing visual reranking methods. To model the ranking distance more precisely, we propose a novel pair-wise method which measure the ranking distance based on the disagreement in terms of pair-wise orders. For visual consistency, we study three different regularizers to mine the best way for its modeling. We conduct extensive experiments on both video and image search datasets. Experimental results demonstrate the effectiveness of our proposed Bayesian visual reranking.**

***Index Terms*—Visual reranking, video search, image search, visual consistency, ranking distance.**

I. Introduction

MOST of the frequently-employed video/image search engines are implemented for the "query by keyword" scenario. They are built by indexing and searching the associated textual information such as surrounding texts from the Web page, speech transcripts, closed captions, titles, URLs, and so on. However, due to the mismatch between the videos/images and the associated textual descriptions, the performance of text-based video/image search is yet unsatisfactory. Moreover, the performance of the state-of-the-art techniques for automatic speech recognition (ASR), video text detection and machine translation (MT) is still far from satisfactory for practical applications. Besides, the textual information cannot describe the video/image's rich content comprehensively and substantially. As a consequence, the essential visual information should be considered to improve the search performances. However, it has been acknowledged that pure content-based approaches [1] cannot work well, due to the semantic gap [2] between the low level visual features and the high level semantic concepts.

Visual reranking has been proposed in recent years. It is an integrated framework that aims to efficiently obtain effective search results. Figure 1 shows a typical process of visual reranking for video/image search. A list of text-based search results is first returned by using textual information only for efficiency. Then visual information is applied to reorder the initial result for refinement. As illustrated in Figure 1(a), after a query "Panda" is submitted, an initial result is obtained via a text-based search engine according to the relevance between the images' associated textual information and the query keyword. It is observed that text-based search often returns "inconsistent" results. Some visually similar images (and semantically close to each other meanwhile in most cases) are scattered in the ranking list, and frequently some irrelevant results are filled between them. For instance, in Figure 1(a), images 1, 2, 4, 6, 7 and 9 are all relevant and visually similar with each other while the irrelevant images 3, 5, and 8 are dissimilar from them. It is reasonably assumed that the visually similar samples should be ranked together. This is also coherent with human perception. Such a visual consistency pattern within the relevant samples can be utilized to refine the initial ranking list. For example, irrelevant images 3, 5, 8 will be demoted while the other relevant images are promoted to the front. A more satisfactory result will be obtained, as shown in Figure 1(b). Such a process of reordering the initial ranking list based on visual patterns is called content-based video/image search reranking, or visual reranking in brief.

Visual reranking incorporates both textual and visual cues to recover the "true" ranking list from the initial noisy one. As for textual cues, we mean that the text-based search result provides a good baseline for the "true" ranking list. Though the text-based search result is noisy, it still reflects partial facts of the "true" list and thus needs to be preserved to some extent. In other words, we should keep the correct information in it. The visual cues are introduced by taking visual consistency as a constraint that visually similar samples (images/video shots) should have close ranking scores and vice versa. Reranking is actually a trade-off between the two cues. It is worth emphasizing that this is actually the basic underlying assumption in many existing visual reranking methods [3], [4], [5], though it is not be explicitly stated.

In this paper, we model textual and visual cues from the probabilistic perspective within Bayesian framework. The textual cues are modeled as a likelihood which reflects the correlation between the reranked list and the initial one. The visual cues are modeled as a conditional prior which indicates the ranking score consistency within visually similar samples. In the Bayesian framework, reranking is formulated

Manuscript received May 12, 2010; revised Oct. 10, 2010.

X. Tian and X. Wu are with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China (email: xinmei@mail.ustc.edu.cn; wuxq@ustc.edu.cn).

L. Yang, J. Wang and X.-S. Hua are with Microsoft Research Asia, Beijing 100190, China (email: {linjuny, jingdw, xshua}@microsoft.com).

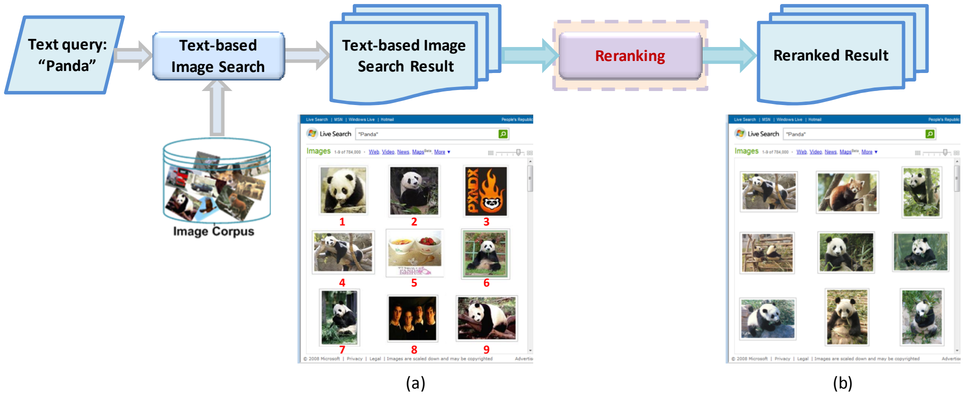


Fig. 1. An illustration of visual reranking. Firstly the text-based search engine returns the images/video shots related to the query "Panda" from textual cues and then the reranking process is applied to refine this result by mining visual information (visual cues). (a) and (b) show the top-9 ranked images in the text-based search results and the reranked results respectively.

as maximizing the product of the conditional prior and the likelihood. This is the reason why we named the proposed approach Bayesian visual reranking. As will be discussed in Section VIII-B, existing random walk-based methods [4], [5] can be unified into this framework.

A. Conditional Prior – Visual Consistency Regularizer

This paper models the conditional prior via a regularizer term. The widely used Laplacian and normalized Laplacian regularizers in the machine learning field can be directly used. However, both of them measure the visual consistency by approximating it pair-wisely. Specifically, for each sample, a set of pairs is formed between it and each of its visually similar neighbors. Then the overall consistency is measured by aggregating the individual consistency over each pair. A sample's consistency on a local area is multiple-wise instead of pair-wise since the consistency is a term defined over the whole neighboring samples instead of over each sample pair. From this point of view, the consistency approximated with pair-wise regularizers is not satisfactory enough.

A local learning regularizer is proposed in this paper to model the desired multiple-wise consistency. The consistency over a local area means that each sample has strong correlation with its neighbors. In other words, each sample's labeling information is partially embedded in its neighbors. Therefore, if we can deduce a sample's label from its neighbors precisely, this sample is regarded as locally consistent. The local learning regularizer is developed in such manner. For each sample, instead of calculating the consistency with each of its neighbors individually, the local learning regularizer considers the consistency with all of its neighboring samples simultaneously. In this regularizer, for each sample, a local model is first trained with its neighbors and then used to predict this sample's consistent ranking score. Finally, by minimizing the difference between the target ranking score and this locally predicted one, the desired multiple-wise consistency over the set of neighbors is guaranteed.

B. Likelihood – Ranking Distance

The likelihood is modeled via the ranking distance which estimates the disagreement between the two ranking lists before and after reranking. It is a crucial factor in visual reranking, which significantly affects reranking performance but has not been well studied yet.

Some existing visual reranking methods [4], [5] adopt the point-wise ranking distance. It simply sums the individual score difference for each sample in the two ranking score lists. However, such point-wise approach fails to capture the disagreement between two lists in terms of ranking accurately, as will be demonstrated in Section V-A. The essential way to model ranking distance is the list-wise method which takes the whole list as an "instance". However, this method is difficult to model and usually suffers from heavy computational cost [6]. On the other hand, since the ordinal information in a ranking list can be completely expressed by the ordering relationship on each sample pair, pair-wise ranking distance is introduced for approximation. The well-known Kendall's tau [7] is such a pair-wise ranking distance which directly counts how many pairs' order is reversed after reranking. However, the reranking process will be computationally intractable when Kendall's tau distance is adopted.

To tackle the problems above, a novel pair-wise ranking distance is proposed in this paper. For each pair of samples, we not only examine whether its order is preserved or reversed after reranking, but also consider to what extent its order is preserved or reversed. A term of preference strength is introduced to measure the degree of one sample ranked before the other. It is defined as the ranking score difference between the two samples in a pair. The change of preference strength can be utilized to measure this pair's order preservation degree. Penalty is given to those pairs whose preference strength is changed after reranking. The preference strength ranking distance is defined as the sum of the penalties over all pairs. With the preference strength ranking distance, Bayesian visual reranking can be solved efficiently with a closed-form solution.

The main contributions introduced in this paper are summarized as follows:

- We explicitly formulate visual reranking as a global optimization problem within the Bayesian framework. Many effective reranking methods can be developed under this framework for different applications.
- To find out the best visual consistency modeling method,

three regularizers are considered and evaluated experimentally.
- By investigating the effects of ranking distances in visual reranking, preference strength ranking distance is proposed from the pair-wise perspective with which Bayesian visual reranking can be solved efficiently.

The rest of this paper is organized as follows. We briefly review the existing visual reranking work in Section II. In Section III visual reranking is formulated in the Bayesian framework and the general Bayesian visual reranking model is derived. Three regularizers for visual consistency modeling are introduced in Section IV. In Section V, we discuss the ranking distance and propose the preference strength distance. Solutions to Bayesian visual reranking are given in Section VI. Different strategies for text prior utilization are presented in Section VII. The connections between Bayesian visual reranking, "learning to rank" and random walk-based methods are discussed in Section VIII. Experimental results and analyses on TRECVID 2005-2007 and Web image search datasets are given in Sections IX and X respectively. The conclusion is presented in Section XI.

The preliminary version of this paper was presented at the ACM Multimedia 2008 [8]. In this journal version, we have enhancement in four aspects: 1) we propose to use local learning regularizer to precisely model the visual consistency term in Bayesian visual reranking; 2) we evaluate six reranking algorithms derived under the Bayesian visual reranking framework via comprehensive experiments; 3) we conduct video search reranking experiments further on TRECVID 2005 dataset; 4) we collect a Web image search dataset and study the effectiveness of Bayesian visual reranking as well as other existing reranking methods for Web image search scenarios.

II. Related Work

Recently many methods [3], [4], [5], [9], [10], [11], [12], [13], [14] have been proposed for video/image search reranking, which can be divided into three categories: classification-based, clustering-based and random walk based.

The first category is classification-based [9], [10], [11]. This kind of methods simplifies reranking as a classification problem. There are normally three steps: (1) select the pseudo-positive and pseudo-negative samples from the initial text-based search results; (2) train a classifier using the selected samples; (3) reorder the samples according to the relevance scores predicted by the trained classifier. In the first step, pseudo relevance feedback (PRF) is usually utilized to select training samples. PRF is a concept introduced from text retrieval, which assumes that a fraction of the top-ranked documents in the initial search results are pseudo-positive [15]. Alternatively, [11] uses the query images or example video clips as positive samples. The pseudo-negative samples are selected from either the least relevant samples in the initial ranking list or the database with the assumption that few samples in the database are relevant to the query [9], [11]. In step (2), different classifiers, such as SVM [11], Boosting [10], Ranking SVM [9] and ListNet [16], can be adopted. Although the above classifiers are effective, sufficient training data are demanded to achieve satisfactory performance since a lot of parameters need to be estimated.

The second category is clustering-based. In [3], each sample is given a soft pseudo label according to the initial text search result, and then the Information Bottleneck principle [17] is adopted to find optimal clustering which maximizes the mutual information between the clusters and the labels. Reranked list is achieved by ordering the clusters according to the cluster conditional probability firstly and then ordering the samples within a cluster based on their local feature density estimated via kernel density estimation. This method achieves good performance on the named-person queries as shown in [3] while it is limited to those queries which have significant duplicate characteristic.

The third category is random walk-based methods [4], [5], [12]. In this kind of methods, a graph is constructed with the samples as the nodes and the edges between them being weighted by visual similarity. Then, reranking is formulated as random walk over the graph and the ranking scores are propagated through the edges. To leverage the text search result, a "dongle" node is attached to each sample with the value fixed to be the initial text ranking score. The stationary probability of the random walk process is adopted as the reranked score directly. In Section VIII-B we will show that random walk-based reranking can be unified into the proposed Bayesian visual reranking framework, while the pair-wise regularizer and point-wise ranking distance are involved.

There are also methods which incorporate auxiliary knowledge, including face detection [18], query example [11], [19], and concept detection [5], [20], [21], into visual reranking. Though the incorporation of auxiliary knowledge leads to the performance improvement, it is not a general treatment. They suffer from either limited applicability to the specific queries (face detection), the desire of the specific user interfaces (query example), or the limited detection performance and small vocabulary size (concept detection). In this paper, we only consider the general reranking problem which doesn't assume any auxiliary knowledge besides the visual information of samples and thus the reranking methods proposed in this paper can be applied to many tasks directly. Besides, there are also many paper focus on improving the diversity of the search result[22], [23], [24]. Diversity is important in image search. I. Cox et al. [25] shows that displaying diverse images to users can speed up search time in CBIR. However, the emphasis of this paper is on improving the relevance. A search result with high relevance can provide a good basis for improving diversity.

III. Bayesian Visual Reranking

Before formulating reranking, a few terms are defined below.

Definition 1: A ranking score list (score list in brief), $\mathbf{r} = [r_1, r_2, \cdots, r_N]^{\mathrm{T}}$ is a vector of the ranking scores, which corresponds to the sample set $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$.

Definition 2: A ranking list l is a permutation of $\mathcal{X}$ sorted by the ranking scores in descending order.

Definition 3: A reranking function is defined as

$$\mathbf{r} = f(\mathcal{X}, \bar{\mathbf{r}}), \tag{1}$$

where $\bar{\mathbf{r}} = [\bar{r}_1, \bar{r}_2, \cdots, \bar{r}_N]^{\mathrm{T}}$ is the initial ranking score list given by the text-based search. Permuting the samples according to this reranking function is called reranking.

The $\mathbf{x}_i$ is a vector which describes the visual content of the ith image. In this paper, the block-wise color moment feature is adopted. The r_i is the ranking score corresponding to sample $\mathbf{x}_i$. Reranking can generally be regarded as a mapping from the initial ranking list to the target ranking list. However, the ranking scores are widely adopted to represent the ranking list for convenience. For this reason, we define reranking on the score list instead of the ranking list, to chieve more flexibility [3], [4]. For the application scenarios where text-based search ranking scores are unavailable, such as Google image search reranking [26], the initial score list $\bar{\mathbf{r}}$ can be set according to the initial rank of samples, as detailed in Section VII.

The crucial problem in reranking is how to derive the optimal function (1). In this paper, we investigate the reranking problem from the probabilistic perspective and derive an optimal reranking function based on Bayesian analysis.

Supposing $\mathbf{r}$ is a random variable, reranking can be regarded as a process to derive the most probable score list given the initial one and the visual content of samples. From the probabilistic perspective, reranking is derives the optimum $\mathbf{r}^*$ with a maximum posterior probability given the samples $\mathcal{X}$ and the initial score list $\bar{\mathbf{r}}$,

$$\mathbf{r}^* = \arg\max_{\mathbf{r}} p(\mathbf{r}|\mathcal{X}, \bar{\mathbf{r}}). \tag{2}$$

According to Bayes' formula, the posterior is proportional to the product of the conditional prior probability and the likelihood,

$$p(\mathbf{r}|\mathcal{X}, \bar{\mathbf{r}}) \propto p(\mathbf{r}|\mathcal{X}) \times p(\bar{\mathbf{r}}|\mathcal{X}, \mathbf{r}), \tag{3}$$

where $p(\mathbf{r}|\mathcal{X})$ is the conditional prior of the score list given the visual content of samples. For instance, a small probability should be assigned to a score list in which visually similar samples have dissimilar rank scores. The $p(\bar{\mathbf{r}}|\mathcal{X}, \mathbf{r})$ is the likelihood, which expresses how probable the initial score list $\bar{\mathbf{r}}$ is given the "true" ranking score list $\mathbf{r}$. As will be discussed later, the likelihood can be estimated based on the ranking distance which represents the disagreement between $\mathbf{r}$ and $\bar{\mathbf{r}}$.

In most of the video/image search systems, $\bar{\mathbf{r}}$ is obtained by using the textual information regardless of the visual content. Therefore the conditional independency assumption of the visual information $\mathcal{X}$ and $\bar{\mathbf{r}}$ given the target score list $\mathbf{r}$ can be made,

$$p(\bar{\mathbf{r}}, \mathcal{X}|\mathbf{r}) = p(\bar{\mathbf{r}}|\mathbf{r}) \times p(\mathcal{X}|\mathbf{r}),$$

hence,

$$p(\bar{\mathbf{r}}|\mathcal{X}, \mathbf{r}) = p(\bar{\mathbf{r}}|\mathbf{r}). \tag{4}$$

Substituting (4) into (3) we obtain

$$p(\mathbf{r}|\mathcal{X}, \bar{\mathbf{r}}) \propto p(\mathbf{r}|\mathcal{X}) \times p(\bar{\mathbf{r}}|\mathbf{r}). \tag{5}$$

Replacing the posterior in (2) with (5), we formulate reranking as maximizing the product of a conditional prior and a likelihood, which is defined as Bayesian visual reranking.

Definition 4: Bayesian visual reranking is reranking using the function

$$f(\mathcal{X}, \bar{\mathbf{r}}) = \arg\max_{\mathbf{r}}\ p(\mathbf{r}|\mathcal{X}) \times p(\bar{\mathbf{r}}|\mathbf{r}), \tag{6}$$

where $\bar{\mathbf{r}}$ is the initial ranking score list, and $\mathcal{X}$ is the corresponding samples.

The conditional prior and the likelihood need to be estimated to complete the reranking function. In the following sections, we will show how to model the two terms.

A. The Conditional Prior

In visual reranking, visually similar samples are expected to have close ranking scores. This empirical prior knowledge can be modeled as the conditional prior in Bayesian visual reranking formulation. Specifically, we formulate the conditional prior as,

$$\begin{aligned} p(\mathbf{r}|\mathcal{X}) &= \frac{1}{Z}\exp(-\sum\nolimits_i \psi_i(\mathbf{r}, \mathcal{X})) \\ &= \frac{1}{Z}\exp(-\mathrm{Reg}(\mathbf{r}, \mathcal{X})), \end{aligned} \tag{7}$$

where $Z = \sum_{\mathbf{r}} \exp(-\sum_i \psi_i(\mathbf{r}, \mathcal{X}))$ is a normalizing constant and $\psi_i(\mathbf{r}, \mathcal{X})$ is the energy function defined over sample $\mathbf{x}_i$ for measuring the visual consistency on its neighboring local area. The energy over all samples is $\mathrm{Reg}(\mathbf{r}, \mathcal{X}) = \sum_i \psi_i(\mathbf{r}, \mathcal{X})$. Detail discussion on $\psi_i(\mathbf{r}, \mathcal{X})$ will be given in Section IV.

B. The Likelihood

As discussed before, the text-based search result is the basis for reranking, therefore the reranked results should preserve the useful information contained in this text prior. This knowledge is modeled in the likelihood term as

$$p(\bar{\mathbf{r}}|\mathbf{r}) = \frac{1}{Z}\exp(-c \times \mathrm{Dist}(\mathbf{r}, \bar{\mathbf{r}})), \tag{8}$$

where Z is the normalizing constant, c is a scaling parameter, and $\mathrm{Dist}(\mathbf{r}, \bar{\mathbf{r}})$ is the ranking distance representing the disagreement between the two score lists, which will be discussed in detail in Section V. With (7) and (8), the Bayesian visual reranking formulation in (6) is equivalent to minimizing the following energy function,

$$\mathrm{E}(\mathbf{r}) = \mathrm{Reg}(\mathbf{r}, \mathcal{X}) + c \times \mathrm{Dist}(\mathbf{r}, \bar{\mathbf{r}}). \tag{9}$$

The two terms on the right hand side of (9) correspond to the conditional prior in (7) and the likelihood in (8) respectively. The c is a trade-off parameter. In the following two sections, we will exploit the two terms respectively.

IV. Regularizer

For the regularizer term $\mathrm{Reg}(\mathbf{r}, \mathcal{X})$, various methods can be used to model the energy function $\psi_i(\mathbf{r}, \mathcal{X})$. With the visual consistency assumption, the widely used regularizers in semi-supervised classification and video annotation, Laplacian regularizer [27] and normalized Laplacian regularizer [28], can be directly utilized.

In both regularizers, a graph $\mathcal{G}$ is constructed with nodes being the samples and similar samples are linked by edges.

If two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ are linked, the weight w_{ij} on the edge between them is calculated by using the Gaussian radial basis function kernel $w_{ij} = \exp\{- \parallel \mathbf{x}_i - \mathbf{x}_j) \parallel^2 /(2\sigma^2)\}$, where σ is the scaling parameter; else, if two samples are not connected, $w_{ij} = 0$.

A. Laplacian Regularizer

In the Laplacian regularizer [27], the energy function $\psi_i(\mathbf{r}, \mathcal{X})$ is defined as

$$\psi_i(\mathbf{r}, \mathcal{X}) = \frac{1}{2} \sum\nolimits_j w_{ij}(r_i - r_j)^2. \tag{10}$$

It approximates the visual consistency of $\mathbf{x}_i$ from the pair-wise perspective, i.e., accumulating the weighted score difference between $\mathbf{x}_i$ and each of its neighbors $\mathbf{x}_j$.

With (10), the Laplacian regularizer is

$$\begin{aligned}\mathrm{Reg}_{\mathrm{Lap}}(\mathbf{r}, \mathcal{X}) &= \sum\nolimits_i \psi_i(\mathbf{r}, \mathcal{X}) \\ &= \sum\nolimits_i \big(\frac{1}{2} \sum\nolimits_j w_{ij}(r_i - r_j)^2\big) \\ &= \mathbf{r}^{\mathrm{T}}\mathbf{L}\mathbf{r}, \end{aligned} \tag{11}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the Laplacian matrix. The $\mathbf{W} = [w_{ij}]_{N\times N}$ and $\mathbf{D} = \mathrm{diag}(\mathbf{d})$ is the degree matrix with $\mathbf{d} = [d_1, d_2, \cdots, d_N]^{\mathrm{T}}$ and $d_i = \sum_j w_{ij}$.

B. Normalized Laplacian Regularizer

For the normalized Laplacian regularizer [28], $\psi_i(\mathbf{r}, \mathcal{X})$ is modeled in the similar way as (10) with normalized ranking scores,

$$\psi_i(\mathbf{r}, \mathcal{X}) = \frac{1}{2} \sum\nolimits_j w_{ij}\big(\frac{r_i}{\sqrt{d_i}} - \frac{r_j}{\sqrt{d_j}}\big)^2. \tag{12}$$

Then, the normalized Laplacian regularizer is

$$\begin{aligned}\mathrm{Reg}_{\mathrm{NLap}}(\mathbf{r}, \mathcal{X}) &= \sum\nolimits_i \psi_i(\mathbf{r}, \mathcal{X}) \\ &= \sum\nolimits_i \Big(\frac{1}{2} \sum\nolimits_j w_{ij}\big(\frac{r_i}{\sqrt{d_i}} - \frac{r_j}{\sqrt{d_j}}\big)^2\Big) \\ &= \mathbf{r}^{\mathrm{T}}\mathbf{L}_n\mathbf{r}, \end{aligned} \tag{13}$$

where $\mathbf{L}_n = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ and $\mathbf{I}$ is the unit matrix. The $\mathbf{W}$ and $\mathbf{D}$ are the same as that in the Laplacian matrix.

From (10) and (12), we can see that both Laplacian and normalized Laplacian regularizers approximate the ranking score consistency for each sample pair-wisely and have less ability to capture the multiple-wise ranking score consistency. As will be discussed later, local learning regularizer models the multiple-wise consistency by formulating the score estimation as a learning problem without heuristic assumptions.

C. Local Learning Regularizer

With the visual consistency assumption, the desired property of $\mathbf{r}$ is that: for each sample $\mathbf{x}_i$ and its neighbors, their ranking scores on $\mathcal{G}$ should be smooth enough. Smoothness is a term defined over the whole neighbor set, instead of over each of the samples separately. However, in both $\mathrm{Reg}_{\mathrm{Lap}}$ and $\mathrm{Reg}_{\mathrm{NLap}}$, only the individual consistency between $\mathbf{x}_i$ and each of its neighbors is considered while the consistency within the neighboring set is ignored.

To reveal the intrinsic multiple-wise consistency, we tackle this problem from the local learning perspective. If a sample's ranking score can be estimated from its neighbors, the multiple-wise consistency is guaranteed. From this point of view, we can model the ranking score consistency from the machine learning perspective. Specifically, for $\mathbf{x}_i$, we first learn the desirably consistent score $\hat{r}_i$ from its neighbors. By requiring the target r_i be close to $\hat{r}_i$, the multiple-wise consistency is guaranteed. The details are discussed in the following.

For each sample $\mathbf{x}_i$, a local model $o_i(\cdot)$ is trained with its neighboring samples set $\mathcal{N}(\mathbf{x}_i) = \{(\mathbf{x}_t^{(i)}, r_t^{(i)})\}_{t=1}^{n_i}$, where $\mathbf{x}_t^{(i)}$ is the tth nearest neighbor of $\mathbf{x}_i$ and n_i is the total number of its neighbors. A ranking score can be predicted by $o_i(\cdot)$, and then the energy function $\psi_i(\mathbf{r}, \mathcal{X})$ is derived as the local model's prediction loss,

$$\psi_i(\mathbf{r}, \mathcal{X}) = \big(r_i - o_i(\mathbf{x}_i)\big)^2.$$

Then, the local learning regularizer is

$$\mathrm{Reg}_{\mathrm{Local}}(\mathbf{r}, \mathcal{X}) = \sum\nolimits_i \psi_i(\mathbf{r}, \mathcal{X}) = \sum\nolimits_i \big(r_i - o_i(\mathbf{x}_i)\big)^2. \tag{14}$$

The task of the local model $o_i(\cdot)$ is to predict $\mathbf{x}_i$'s ranking score r_i from its neighbors accurately. Many approaches can be used as the local model. A linear one is adopted in [29]. However, due to the complexity of the real-world images, it is hard to predict the scores accurately by using simple linear model. To handle this difficulty, we propose to use a local kernel model by leveraging the strength of kernel methods. Since this is apparently a regression problem, the kernel ridge regression statistical model [30] which is well-known and simple to implement, is adopted in this paper.

In kernel ridge regression we define a kernel mapping function $\phi(\cdot)$ operating from input space $\mathcal{X}$ to a kernel space $\mathcal{F}$: $\phi : \mathbf{x} \in \mathcal{X} \mapsto \Phi(\mathbf{x}) \in \mathcal{F}$. The dependencies between $\mathcal{N}(\mathbf{x}_i)$and its score vector $\mathbf{r}^{(i)} = [r_t^{(i)}]^{\mathrm{T}}$ are modeled as

$$o_i(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}). \tag{15}$$

The cost function is

$$g(\mathbf{w}) = \sum_{t=1}^{n_i} \big(r_t^{(i)} - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_t^{(i)})\big)^2 + \lambda \parallel \mathbf{w} \parallel^2, \tag{16}$$

where λ is a coefficient to balance the capacity and complexity of this model.

Differentiating (16) w.r.t $\mathbf{w}$ and then equating it to zero, we obtain

$$\mathbf{w} = \Phi_i(\Phi_i^{\mathrm{T}}\Phi_i + \lambda\mathbf{I})^{-1}\mathbf{r}^{(i)},$$

where Φ_i denotes matrix $\big[\phi(\mathbf{x}_t^{(i)})\big]^{\mathrm{T}}$. Then, for $\mathbf{x}_i$, the score predicted by its local model $o_i(\cdot)$ is:

$$o_i(\mathbf{x}_i) = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_i) = \mathbf{k}^{\mathrm{T}}(\lambda\mathbf{I} + \mathbf{K})^{-1}\mathbf{r}^{(i)} = \beta_i^{\mathrm{T}}\mathbf{r}^{(i)}, \tag{17}$$

where $\beta_i^{\mathrm{T}} = \mathbf{k}^{\mathrm{T}}(\lambda\mathbf{I} + \mathbf{K})^{-1}$, $\mathbf{k}$ is a vector with $k_j = \phi(\mathbf{x}_i)^{\mathrm{T}}\phi(\mathbf{x}_j^{(i)}) = k(\mathbf{x}_i, \mathbf{x}_j^{(i)})$, and $\mathbf{K}$ is a matrix with $k_{mn} = \phi(\mathbf{x}_m^{(i)})^{\mathrm{T}}\phi(\mathbf{x}_n^{(i)}) = k(\mathbf{x}_m^{(i)}, \mathbf{x}_n^{(i)})$. As for kernel based methods, we only need to define the kernel function k without defining

TABLE I
A TOY EXAMPLE FOR RANKING DISTANCE

Ranking Score Lists	Samples				
	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
$\mathbf{r}^0$	1.0	0.9	0.8	0.7	0.6
$\mathbf{r}^1$	0.6	0.7	0.8	0.9	1.0
$\mathbf{r}^2$	1.5	0.7	0.8	0.9	1.0
$\mathbf{r}^3$	0.5	0.4	0.3	0.2	0.1

$\phi(\cdot)$ explicitly. The Gaussian kernel is adopted as the kernel function in this paper.

Substituting (17) into (14), we get the local learning regularizer

$$\begin{aligned}\mathrm{Reg}_{\mathrm{Local}}(\mathbf{r},\mathcal{X}) &= \sum\nolimits_i \big(r_i - o_i(\mathbf{x}_i)\big)^2 \\ &= \sum\nolimits_i \big(r_i - \beta_i^{\mathrm{T}}\mathbf{r}^{(i)}\big)^2 \\ &= \mathbf{r}^{\mathrm{T}}\mathbf{R}_{\mathrm{Local}}\mathbf{r}. \end{aligned} \tag{18}$$

The $\mathbf{R}_{\mathrm{Local}} = (\mathbf{I}-\mathbf{B})^{\mathrm{T}}(\mathbf{I}-\mathbf{B})$ is the local learning regularizer matrix and $\mathbf{B} = [b_{ij}]_{N\times N}$ where b_{ij} equals the corresponding element of β_i if $\mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i)$, otherwise $b_{ij} = 0$.

V. Ranking Distance

In this section we will analyze the issues in existing ranking distances and propose to measure the ranking distance from the pair-wise perspective. A toy example is given for illustration, which comprises five samples $\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4,\mathbf{x}_5\}$ and four ranking score lists $\{\mathbf{r}^0,\mathbf{r}^1,\mathbf{r}^2,\mathbf{r}^3\}$, as shown in TABLE I. Sorting the samples by their scores, the corresponding ranking lists are derived from $\mathbf{r}^0$, $\mathbf{r}^1$, $\mathbf{r}^2$ and $\mathbf{r}^3$ as

$$l^0 = \langle \mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4,\mathbf{x}_5\rangle,\ l^1 = \langle \mathbf{x}_5,\mathbf{x}_4,\mathbf{x}_3,\mathbf{x}_2,\mathbf{x}_1\rangle,$$
$$l^2 = \langle \mathbf{x}_1,\mathbf{x}_5,\mathbf{x}_4,\mathbf{x}_3,\mathbf{x}_2\rangle,\ l^3 = \langle \mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4,\mathbf{x}_5\rangle.$$

To measure the ranking distance between the score lists, one intuitive idea is to take each score list as an "instance" and then use list-wise approach. The list-wise ranking distance has been exploited in "learning to rank". For example, in [6], the distance of two score lists is defined as the cross entropy between the two distributions of permutations conditioned respectively on each of the score lists. However, this list-wise approach is computationally intractable since the number of permutations is $O(N!)$ where N is the number of samples. Therefore, researchers resort to other simpler ranking distance for efficiency.

A. point-wise ranking distance

The most direct and the simplest way to measure the ranking distance between two score lists is to compute the individual score difference of each sample respectively and then sum them up, so-called point-wise approach, as shown below,

$$\mathrm{Dist}_{\mathrm{Point}} = \sum\nolimits_i d(r_i,\bar{r}_i) = \sum\nolimits_i (r_i - \bar{r}_i)^2. \tag{19}$$

Such a point-wise approach has been applied in random walk reranking with a slightly different form, as will be detailed in Section VIII-B.

Point-wise ranking distance, however, fails to capture the disagreement between the score lists in terms of ranking order in some situations. Take the toy example in TABLE I for illustration. The distances between $\mathbf{r}^0$ and $\mathbf{r}^1$, $\mathbf{r}^2$, $\mathbf{r}^3$ computed using (19) are: $\mathrm{Dist}(\mathbf{r}^1,\mathbf{r}^0)$= 0.63, $\mathrm{Dist}(\mathbf{r}^2,\mathbf{r}^0)$= 0.70 and $\mathrm{Dist}(\mathbf{r}^3,\mathbf{r}^0)$= 1.12. The $\mathrm{Dist}(\mathbf{r}^3,\mathbf{r}^0)$ is the largest, however, in terms of ranking, the distance between $\mathbf{r}^3$ and $\mathbf{r}^0$ should be the smallest since l^3 is identical with l^0 while different from l^1 and l^2.

As the ranking information can be represented entirely by the pair-wise ordinal relations, the ranking distance between two score lists can be computed from the pairs, so-called pair-wise approach. Before further discussing pair-wise approaches, we first define the notation $\succ_{\mathbf{r}}$.

Definition 5: $\mathbf{x}_i \succ_{\mathbf{r}} \mathbf{x}_j$ is a relation on a pair $(\mathbf{x}_i,\mathbf{x}_j)$ if $r_i > r_j$, i.e., $\mathbf{x}_i$ is ranked before $\mathbf{x}_j$ in the ranking list l derived from $\mathbf{r}$.

All the pairs with $(\mathbf{x}_i,\mathbf{x}_j)$ satisfying $\mathbf{x}_i \succ_{\mathbf{r}} \mathbf{x}_j$ compose set $\mathcal{S}_{\mathbf{r}} = \{(i,j) : \mathbf{x}_i \succ_{\mathbf{r}} \mathbf{x}_j\}$. For any two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ either (i,j) or (j,i) belongs to $\mathcal{S}_{\mathbf{r}}$. Therefore, all the pair-wise ordinal relations are reflected in $\mathcal{S}_{\mathbf{r}}$.

The simplest pair-wise ranking distance could be defined as,

$$\mathrm{Dist}(\mathbf{r},\bar{\mathbf{r}}) = \sum\nolimits_{(i,j)\in\mathcal{S}_{\bar{\mathbf{r}}}} \delta(\mathbf{x}_j \succ_{\mathbf{r}} \mathbf{x}_i), \tag{20}$$

where $\delta(t)$ is a binary function defined as

$$\delta(t) = \begin{cases} 1, t & = \text{true} \\ 0, t & = \text{false} \end{cases}.$$

The basic idea of (20) is to count the number of pairs which disagree on the order relations in the two lists. The widely used Kendall's tau distance [7] is defined in this way. Using (20), $\mathrm{Dist}(\mathbf{r}^1,\mathbf{r}^0)$=10, $\mathrm{Dist}(\mathbf{r}^2,\mathbf{r}^0)$ =6 and $\mathrm{Dist}(\mathbf{r}^3,\mathbf{r}^0)$ =0. It really captures the differences between the ranking lists. However, the optimization problem of (9) with ranking distance (20) is computationally intractable. Below we will design a new pair-wise ranking distance with which the optimization problem of (9) can be solvable.

B. Pair-wise Ranking Distance

In reranking, not only the order relation but also the preference strength, which means the score difference of the samples in a pair, $r_i - r_j$ for pair $(\mathbf{x}_i,\mathbf{x}_j)$, is indicative. For example, given two pairs, one comprising two tigers with different relevance levels, and the other comprising a tiger and a stone. Obviously the preference strength is different for these two pairs. Changing the order of pair (tiger, tiger) is less sensitive than changing the order of pair (tiger, stone). Such information can be utilized in visual reranking and then we define a new pair-wise ranking distance, called preference strength distance

$$\begin{aligned}\mathrm{Dist}_{\mathrm{Pair}}(\mathbf{r},\bar{\mathbf{r}}) &= \sum\nolimits_{(i,j)\in\mathcal{S}_{\bar{\mathbf{r}}}} d\big((r_i,r_j),(\bar{r}_i,\bar{r}_j)\big) \\ &= \sum\nolimits_{(i,j)\in\mathcal{S}_{\bar{\mathbf{r}}}} \Big(1 - \frac{r_i - r_j}{\bar{r}_i - \bar{r}_j}\Big)^2. \end{aligned} \tag{21}$$

From (21) we can see that not only the ordinal relation but also the change of preference strength is considered in the preference strength ranking distance. For a pair, the ordinal relation is enhanced by a stricter criterion preference strength.

Only the preference strength is preserved, this pair of sample's relation is regarded as unchanged after reranking. With this distance, Bayesian visual reranking can be solved efficiently with closed-form solution.

VI. Solutions

With the three regularizers, Laplacian (Lap) in (11), normalized Laplacian (NLap) in (13) and local learning (Local) in (18), and two ranking distances, point-wise (Point) in (19) and pair-wise preference strength distance (Pair) in (21), six different reranking methods can be derived by combining them according to the Bayesian visual reranking framework in (9). We denote these six methods as Lap-Point, NLap-Point, Local-Point, Lap-Pair, NLap-Pair and Local-Pair respectively. In this section, we will give the solutions to these six methods. It is worth emphasizing that the Lap-Point is identical with GRF [27] and NLap-Point is identical with LGC [28]. GRF and LGC are two representative transductive learning methods in machine learning.

The three regularizers can be written in a unified form,

$$\mathrm{Reg}(\mathbf{r}, \mathcal{X}) = \mathbf{r}^{\mathrm{T}}\mathbf{R}\mathbf{r},$$

with certain matrix $\mathbf{R}$ for corresponding regularizers. Therefore, we only need to discuss the solutions with two different ranking distances.

Proposition 1: The solution of Bayesian visual reranking with point-wise distance (19) is

$$\mathbf{r} = c(\mathbf{R} + c\mathbf{I})^{-1}\bar{\mathbf{r}},$$

where $\mathbf{I}$ is the identity matrix.

Proof: Replacing the distance term in (9) with the point-wise distance, we get

$$\mathrm{E}(\mathbf{r}) = \mathbf{r}^{\mathrm{T}}\mathbf{R}\mathbf{r} + c\sum\nolimits_i (r_i - \bar{r}_i)^2.$$

The optimal solution $\mathbf{r}^*$ is obtained by minimizing $\mathrm{E}(\mathbf{r})$,

$$\begin{aligned}\mathbf{r}^* &= \arg\min_{\mathbf{r}} \ \mathbf{r}^{\mathrm{T}}\mathbf{R}\mathbf{r} + c\sum\nolimits_i (r_i - \bar{r}_i)^2 \\ &= \arg\min_{\mathbf{r}} \ \mathbf{r}^{\mathrm{T}}\mathbf{R}\mathbf{r} + c(\mathbf{r} - \bar{\mathbf{r}})^{\mathrm{T}}(\mathbf{r} - \bar{\mathbf{r}}). \qquad (22)\end{aligned}$$

Differentiating (22) w.r.t $\mathbf{r}$ and then equating it to zero, it gives

$$\begin{aligned}\mathbf{R}\mathbf{r} + c(\mathbf{r} - \bar{\mathbf{r}}) &= 0 \\ \mathbf{r} &= c(\mathbf{R} + c\mathbf{I})^{-1}\bar{\mathbf{r}}\end{aligned}$$

The solutions for Lap-Point, NLap-Point and Local-Point can be derived by replacing $\mathbf{R}$ with $\mathbf{L}$, $\mathbf{L}_n$ and $\mathbf{R}_{\mathrm{Local}}$ respectively. ■

Proposition 2: The solution of Bayesian visual reranking with the proposed pair-wise distance (21) is

$$\mathbf{r} = \frac{1}{2}(\mathbf{R} + c\mathbf{L}_{\mathbf{A}})^{-1}\tilde{\mathbf{c}},$$

where $\mathbf{L}_{\mathbf{A}}$ is a Laplacian regularizer matrix defined over the graph $\mathcal{G}_{\mathbf{A}}$ which has the same structure with $\mathcal{G}$ but the weight between nodes $\mathbf{x}_i$ and $\mathbf{x}_j$ is $|\alpha_{ij}|$ instead of w_{ij}. The $\tilde{\mathbf{c}} = 2c(\mathbf{A}\mathbf{e})$ where $\mathbf{e}$ is a vector with all elements equals 1 and $\mathbf{A} = [\alpha_{ij}]_{N\times N}$ is an anti-symmetric matrix with $\alpha_{ij} = 1/(\bar{r}_i - \bar{r}_j)$.

Proof: Replacing the distance term in (9) with the preference strength distance (21), the energy function is

$$\mathrm{E}(\mathbf{r}) = \mathbf{r}^{\mathrm{T}}\mathbf{R}\mathbf{r} + c\sum\nolimits_{(i,j)\in\mathcal{S}_{\bar{\mathbf{r}}}} \left(1 - \frac{r_i - r_j}{\bar{r}_i - \bar{r}_j}\right)^2.$$

The optimal solution $\mathbf{r}^*$ is obtained by minimizing $\mathrm{E}(\mathbf{r})$. Denote $\alpha_{ij} = 1/(\bar{r}_i - \bar{r}_j)$, then we can get

$$\begin{aligned}\mathbf{r}^* &= \arg\min_{\mathbf{r}} \mathbf{r}^{\mathrm{T}}\mathbf{R}\mathbf{r} + c\sum\nolimits_{(i,j)\in\mathcal{S}_{\bar{\mathbf{r}}}} \left(1 - \frac{r_i - r_j}{\bar{r}_i - \bar{r}_j}\right)^2 \\ &= \arg\min_{\mathbf{r}} \mathbf{r}^{\mathrm{T}}\mathbf{R}\mathbf{r} + c\sum\nolimits_{(i,j)\in\mathcal{S}_{\bar{\mathbf{r}}}} \Big(1 - \alpha_{ij}(r_i - r_j)\Big)^2 \\ &= \arg\min_{\mathbf{r}} \mathbf{r}^{\mathrm{T}}\mathbf{R}\mathbf{r} + c\sum\nolimits_{(i,j)\in\mathcal{S}_{\bar{\mathbf{r}}}} \alpha_{ij}^2(r_i - r_j)^2 \\ &\quad - 2c\sum\nolimits_{(i,j)\in\mathcal{S}_{\bar{\mathbf{r}}}} \alpha_{ij}(r_i - r_j) + \mathrm{const} \\ &= \arg\min_{\mathbf{r}} \mathbf{r}^{\mathrm{T}}\mathbf{R}\mathbf{r} + c\mathbf{r}^{\mathrm{T}}\mathbf{L}_{\mathbf{A}}\mathbf{r} - 2c\sum\nolimits_{(i,j)\in\mathcal{S}_{\bar{\mathbf{r}}}} \alpha_{ij}(r_i - r_j) \\ &= \arg\min_{\mathbf{r}} \mathbf{r}^{\mathrm{T}}(\mathbf{R} + c\mathbf{L}_{\mathbf{A}})\mathbf{r} - \tilde{\mathbf{c}}^{\mathrm{T}}\mathbf{r}. \qquad (23)\end{aligned}$$

Differentiating (23) w.r.t $\mathbf{r}$ and then equating it to zero, it gives,

$$\begin{aligned}2(\mathbf{R} + c\mathbf{L}_{\mathbf{A}})\mathbf{r} &= \tilde{\mathbf{c}} \\ \mathbf{r} &= \frac{1}{2}(\mathbf{R} + c\mathbf{L}_{\mathbf{A}})^{-1}\tilde{\mathbf{c}}. \qquad (24)\end{aligned}$$

The solutions for Lap-Pair, NorLap-Pair and Local-Pair can be derived by replacing $\mathbf{R}$ with $\mathbf{L}$, $\mathbf{L}_n$ and $\mathbf{R}_{\mathrm{Local}}$ respectively. ■

However, for Lap-Pair, since $\tilde{\mathbf{L}} = \mathbf{L} + c\mathbf{L}_{\mathbf{A}}$ has a zero eigenvalue, the solution of (24) is non-unique. To obtain a unique solution, we add a constrain $r_N = 0$ by replacing the last row of $\tilde{\mathbf{L}}$ with $[0, 0, \cdots, 0, 1]_{1\times N}$ to obtain $\breve{\mathbf{L}}$ and replacing the last element of $\tilde{\mathbf{c}}$ with zero to obtain $\breve{\mathbf{c}}$ respectively. Then, the solution is $\mathbf{r} = \frac{1}{2}\breve{\mathbf{L}}^{-1}\breve{\mathbf{c}}$.

VII. Utilization of Text-based Search Prior

As aforementioned, the text-based search prior provides information derived from the textual cues and thus should be well utilized. In Bayesian visual reranking, this text prior is involved as $\bar{\mathbf{r}}$ in the ranking distance term. Since $\bar{\mathbf{r}}$ reflects the ranking scores of the samples, the most direct way is to use the text-based search scores for it. However, in video search, the performance of the text baseline is often poor and text scores are mostly unreliable because of the inaccuracy and mismatch of ASR and MT from the video. Besides, in some situations, the text-based search scores are even unavailable. For example, when images are downloaded from Web search engines, we only know their ranks and cannot obtain their ranking scores. Therefore, alternative strategies are proposed to set $\bar{\mathbf{r}}$.

- Normalized Text Score (NTS)
  The initial scores $\bar{\mathbf{r}}$ can be assigned by normalizing the text scores $\mathbf{r}^{\mathrm{text}}$ into $[0, 1]$ as follows
  $$\bar{r}_i = \frac{r_i^{\mathrm{text}} - r_{\mathrm{min}}^{\mathrm{text}}}{r_{\mathrm{max}}^{\mathrm{text}} - r_{\mathrm{min}}^{\mathrm{text}}},$$
  where $r_{\mathrm{max}}^{\mathrm{text}}$ and $r_{\mathrm{min}}^{\mathrm{text}}$ are the maximal and minimal value in $\mathbf{r}^{\mathrm{text}}$.
- Normalized Rank (NRK)
  The normalized rank is widely used to estimate the sample's relevance probability [3], [9], [11], which will be employed to assign the initial scores as
  $$\bar{r}_i = 1 - \mathrm{RK}_i/N, \ i = 1, \cdots, N,$$
  where RK_i is the rank of $\mathbf{x}_i$ in text-based search result.

- Rank (RK)
  Different from NRK, the rank can be used directly as the initial score without normalizing by total sample number N,

$$\bar{r}_i = N - \mathrm{RK}_i,\ i = 1, \cdots, N.$$

VIII. Discussion

A. Connection to "Learning to Rank"

Firstly we define the ranking function analogical to reranking function.

Definition 6: A ranking function is defined as

$$\mathbf{r} = f(\mathcal{K}),$$

where $\mathcal{K} = \{\mathbf{k}_j\}$ is a set of features with $\mathbf{k}_j$ being extracted from the pair comprising the query q and the sample $\mathbf{x}_i$, and $\mathbf{r}$ is the target ranking score list.

The goal of most "learning to rank" methods [6], [31] is to learn a ranking function automatically from the training data,

$$f^* = \arg\max_f p\big(f|\{\mathcal{K}^i, \mathbf{r}^i\}\big), \tag{25}$$

and then predict the ranking score list of the samples under a test query q_t using the learned ranking function

$$\mathbf{r}^t = f^*(\mathcal{K}^t),$$

where $\mathcal{K}^t$ is the test feature set extracted from pairs of the test query q_t and samples, $\{\mathcal{K}^i, \mathbf{r}^i\}$ is the training data comprising m pre-labeled ranking lists for m queries $\{q_i\}$.

Reranking can be formulated as a learning to rank problem. Firstly a fraction of the initial ranking score list is selected based on some strategy as shown; then the selected fractions of the initial ranking list are used to learn an optimal ranking function; finally the reranked list can be achieved using the learned ranking function. This is actually the method used in [9], which adopts Ranking SVM to learn a pair-wise ranking function.

The problem (25) can be regarded as inductive learning to rank, which learns an explicit ranking function without utilizing the unlabeled data. In reranking, however, an explicit ranking function is not necessarily needed and what we desire is just the reranked score list. A more effective way should be to deduce the optimal ranking list from the training data directly without explicitly learning a ranking function as

$$\mathbf{r}^t = \arg\max_{\mathbf{r}} p\big(\mathbf{r}|\mathcal{K}^i, \{\mathcal{K}^i, \mathbf{r}^i\}_{i=1}^m\big), \tag{26}$$

which corresponds to the transduction paradigm in machine learning.

Rewriting the reranking objectives (2) as

$$\mathbf{r}^* = \arg\max_{\mathbf{r}} p\big(\mathbf{r}|\mathcal{X}, \{\mathcal{X}, \bar{\mathbf{r}}\}\big). \tag{27}$$

Since in reranking only one query is involved, the features are extracted from the samples regardless of the query. Except this the objectives (26) and (27) have the same form. We can see that reranking is actually transductive learning to rank with only one training sample, i.e., the initial ranking score list. From this perspective, the proposed Bayesian visual reranking can be applied as transductive learning to rank as well. In addition, any transductive learning to rank method which will be developed in the future can be used for reranking seamlessly.

B. Connection to Random Walk

The objective function of random walk-based reranking methods proposed by W. Hsu et al [4] and Y. Jing and S. Baluja [12] is derived as

$$\frac{\alpha}{2}\sum_{i,j} w_{ij}\Big(\frac{r_i}{d_i} - \frac{r_j}{d_j}\Big)^2 + (1-\alpha)\sum_i \frac{1}{d_i}(r_i - \bar{r}_i)^2, \tag{28}$$

from which we can see that random walk-based reranking actually has a similar objective function as Bayesian visual reranking. The two terms in the objective function (28) correspond to the pair-wise visual consistency regularizer and the normalized point-wise ranking distance respectively.

IX. Experiments on Video Search Dataset

In this section, we evaluated the proposed Bayesian visual reranking framework as well as the local learning regularizer and the pair-wise preference strength ranking distance on TRECVID which is a widely used video search benchmark.

A. Experimental Setting

We conducted experiments on automatic search task over the TRECVID 2005-2007 video search benchmark dataset [32], which consists of 508 videos with 143,392 shots. The data are collected from English, Chinese, and Arabic news programs, accompanied with ASR and MT transcripts in English provided by NIST [33]. The text-based search baseline we used in this paper is based on the Okapi BM-25 formula [34] using ASR/MT transcripts at shot level. For each of the 72 queries, 24 for each year, at most 1400 video shots are returned as initial text-based search result.

The low level visual feature we used in reranking are the 225-dimensional block-wise color moments extracted over 5x5 fixed grid partitions with each block described by 9-dimensional features [35]. When constructing the graph $\mathcal{G}$, each sample is connected with its K-nearest neighbors. The RK strategy for initial score is adopted and the parameters are globally set for all methods to achieve their best performance.

The performance is measured by the widely used non-interpolated Average Precision (AP) [33] which averages the precision values obtained when each relevant image occurs. We average the APs over all the 24 queries in each year to get the Mean AP (MAP) for overall performance measurement.

B. Performance Comparison

1) Comparison for Regularizers and Ranking Distance: We first compare the six methods derived under Bayesian visual reranking framework with the three regularizers and two ranking distances. The experimental results are summarized in Table II.

We analyze the results given in this table from two different views. The first view is for regularizer, to find out the best way for visual consistency modeling. We can see that, with the same ranking distance, no matter point-wise or pair-wise, the Local- algorithm outperforms the Lap- and NLap- algorithms on most cases over the three years. The only exception is that Local-point gives slightly worse performance than that

TABLE II
MAP COMPARISON BETWEEN THE SIX METHODS UNDER BAYESIAN VISUAL RERANKING FRAMEWORK

Method	TRECVID2005		TRECVID2006		TRECVID2007	
	MAP	Gain	MAP	Gain	MAP	Gain
Text Baseline	0.044	-	0.038	-	0.031	-
Lap-Point	0.045	2.27%	0.046	21.05%	**0.046**	**48.39%**
NLap-Point	0.049	11.36%	0.041	7.89%	0.040	29.03%
Local-Point	**0.053**	**20.45%**	**0.048**	**26.32%**	**0.046**	**48.39%**
Lap-Pair	0.049	11.36%	0.046	21.05%	0.046	48.39%
NLap-Pair	0.053	20.45%	0.043	13.16%	0.047	51.61%
Local-Pair	**0.058**	**31.82%**	**0.050**	**31.58%**	**0.048**	**54.84%**

TABLE III
MAP COMPARISON BETWEEN LOCAL-PAIR AND OTHER RERANKING METHODS

Method	TRCVID2005		TRCVID2006		TRCVID2007	
	MAP	Gain	MAP	Gain	MAP	Gain
Text Baseline	0.044	-	0.038	-	0.031	-
PRF-SVM	0.055	25.00%	0.042	10.53%	0.043	38.71%
VisualRank	0.051	15.91%	0.040	5.26%	0.033	6.45%
Local-Pair	**0.058**	**31.82%**	**0.050**	**31.58%**	**0.048**	**54.84%**

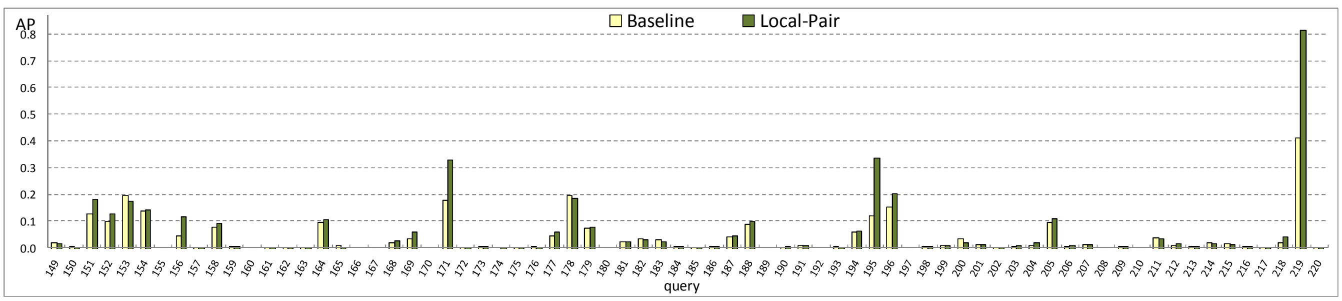


Fig. 2. Performance of Local-Pair and the text search baseline across all queries of TRECVID 2005-2007.

of Lap-point on TRECVID 2007. Generally speaking, the local learning regularizer is superior to both Laplacian and normalized Laplacian regularizers since it takes the multiple-wise correlations of the neighboring samples into consideration while the other two regularizers neglect it.

Then, we compare the two ranking distances with regularizers verifying. From Table II, we can see that Lap-Pair outperforms Lap-Point, NLap-Pair outperforms NLap-Point, and Local-Pair outperforms Local-Point consistently over three years. From this, we can conclude that pair-wise ranking distance performs better than point-wise ranking distance.

2) Comparison between Local-Pair and other reranking methods: From the above analyses, we already learned that Local-Pair method performs the best among the six. To further verify the effectiveness of Local-Pair, we need to compare it with other existing methods beyond the six introduced in this paper. Here, we compare Local-Pair with one typical classification-based method, SVM-PRF [11], and one well-known random walk-based method, VisualRank [12].

The results are given in Table III. Local-Pair gives better performance than both PRF-SVM and VisualRank consistently. For the PRF-SVM, we have tried several different pseudo-positive and pseudo-negative sample selection strategies and report the best one. However, due to the poor performance of the text baseline, too much noise is contained in the pseudo-positive samples which lead unsatisfactory reranking

TABLE IV
THE P VALUES OF PAIRED T-TEST BETWEEN LOCAL-PAIR AND OTHER METHODS

	Compared Methods	p
Local-Pair vs.	Text Baseline	0.0165
	Lap-Point	0.0076
	NLap-Point	0.0068
	Local-Point	0.0761
	Lap-Pair	0.0152
	NLap-Pair	0.0496
	PRF-SVM	0.0207
	VisualRank	0.0437

performance. For VisualRank, as discussed in Section VIII-B, it can be unified into Bayesian visual reranking framework with pair-wise regularizer and point-wise ranking distance. Local-Pair outperforms it since more powerful regularizer and ranking distance are utilized.

3) Performance of Local-Pair on each query: Besides the overall performance, we also investigated the effectiveness of Local-Pair over each query. Figure 2 shows the performance of Local-Pair across all the 72 queries over TRECVID 2005-2007. We can see that most of the queries benefit from Local-Pair after reranking and some queries show significant gain, such as *Query 156: Find shots of tennis players on the court*, *Query 171: Find shots of a goal being made in a soccer match*, *Query 195: Find shots of one or more soccer*

TABLE V
MAP COMPARISON OF DIFFERENT $\bar{\mathbf{r}}$ STRATEGIES

	TRECVID2005		TRECVID2006		TRECVID2007	
	MAP	Gain	MAP	Gain	MAP	Gain
Text Baseline	0.044	-	0.038	-	0.031	-
NTS	0.047	6.82%	0.039	2.63%	0.031	0.00%
NRK	0.050	13.64%	0.041	7.89%	0.047	51.61%
RK	**0.058**	**31.82%**	**0.050**	**31.58%**	**0.048**	**54.84%**

goalposts, *Query 196: Find shots of scenes with snow* and *Query 219: Find shots that contain the Cook character in the Klokhuis series*. In these queries, the relevant samples share high visual similarity, which is coherent with the visual consistency assumption. Remarkable improvements on these queries also demonstrate the effectiveness of the proposed visual consistency regularizer. On the other hand, these queries have better text baselines than the others, therefore more useful information is provided in the ranking distance term.

We can also see that the AP of some queries slightly degrade after reranking, such as *Query 153: Find shots of Tony Blair*, *Query 178: Find shots of US Vice President Dick Cheney* and *Query 200: Find shots of hands at a keyboard typing or using a mouse*. By further examining the data, we find that the relevant samples in these queries vary largely and the used low level feature is insufficient to represent the complex high level semantics. As a conclusion, Local-Pair presents stable performance improvements on most of queries with slight performance decrease on a few of them. This phenomenon further demonstrates the superiority of the local learning regularizer and pair-wise ranking distance.

To verify whether the improvement of Local-Pair is statistically significant, we further perform a statistical significance test. Here we conduct paired T-test between Local-Pair and all other methods. The p values are reported in Table IV. The T-test is conducted over 72 queries in TRECVID 2005-2007. From this result we can see that the improvement of Local-pair is statistically significant.

C. Text-based Search Prior and Parameter Sensitivity

In this section, we will first analyze the influence of different text prior utilization strategies presented in Section VII. Then, we will investigate the sensitivity of Bayesian visual reranking with respect to two important parameters, the K for graph construction and trade-off parameter c.

1) Text Search Prior: We discussed three different strategies for initial score list $\bar{\mathbf{r}}$ in Section VII. Different initial score strategies will give different effects to the reranking process. We investigate NTS, NRK and RK strategies by conducting experiments with Local-Pair reranking method for illustration.

As shown in Table V, RK and NRK, which only use the rank instead of text scores, outperform NTS on TRECVID 2005 - 2007 consistently. We argue the reason could be that the text-based search scores are not as reliable as rank. In addition, RK performs better than NRK. The reason could be as follows. In RK, for a pair of samples $(\mathbf{x}_i, \mathbf{x}_j)$, its preference strength is $j-i$ for all queries. In NRK, however, this preference strength is normalized by the number of samples N, i.e., $(j-i)/N$, which is different among queries. Based on the statistics, N

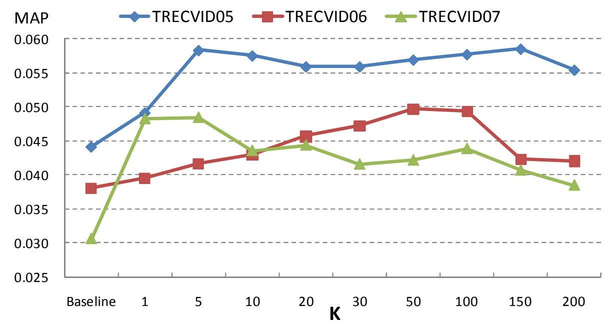


Fig. 3. The performance of Local-Pair with different K.

varies from 28 to 1400 in TRECVID 2005 - 2007. The optimal parameters, such as the tradeoff parameter c, vary according to the preference strength, as can be observed in the optimization objective. Since in our experiment the parameters are globally selected, it is more appropriate to assign each query with equal preference strength for pairs with same rank differences. Therefore, RK is much better in this situation.

2) The Number of Nearest-neighbors K: Below, we will analyze the sensitivity of important parameters K and c in Local-Pair. The RK is set as the default initial score strategy in the following experiments. The K and c are evaluated over: $K \in \{5, 10, 20, 30, 50, 100, 150, 200\}$, $c \in \{0.001, 0.01, 0.1, 1, 10, 100\}$. When studying the sensitivity of K, we conduct experiments with K fixed to certain value and record the experimental results. Experiments are repeated until each K has been tested. The evaluation procedure for c is similar with that for K.

The K is an important parameter when constructing the graph $\mathcal{G}$. A larger K can ensure more relevant samples connected to each other. However, the edges between relevant and irrelevant samples will be added too, which could degrade the performance because the score consistency between relevant and irrelevant samples is unnecessary. With a smaller K, the "incorrect" edges will be eliminated while some of the "correct" edges between relevant samples are also missed, which will weaken the necessary consistency.

Figure 3 shows the MAP-K curve. For TRECVID 2005, the MAP increases dramatically when K grows from 1 to 5. Then, it fluctuates between 0.055 and 0.058 when K grows from 10 to 200. For TRECVID 2006 and 2007, the MAPs increase with K growing and arrive at their peaks at around 50 and 5 respectively. Then, the MAPs decrease gradually when K grows larger than the peak point. Three datasets prefer different Ks. As analyzed from the data, the average numbers of relevant samples across queries are 41, 55 and 24 for TRECVID 2005 - 2007 respectively. We can observe that setting K around its average relevant sample number can

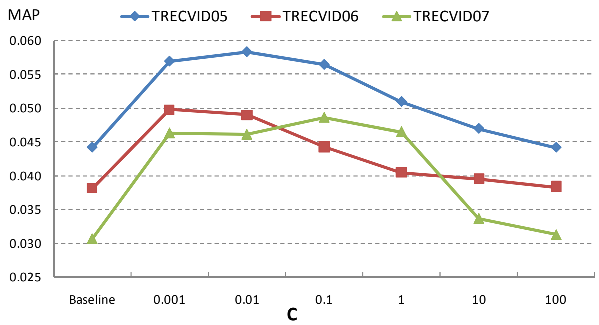


Fig. 4. The performance of Local-Pair with different c.

achieve a good, maybe not the best but at least moderate, performance. This provides a rough guideline for setting K empirically in practical applications.

3) The Trade-off Parameter c: The trade-off parameter c balances the effects of the two terms: consistency regularizer and ranking distance. A larger c indicates that more information is preserved from the text search baseline into the reranked list. When $c = \infty$, the reranked list will be the same as the initial one if all the pairs are used. A small c means that the visual consistency term plays a major role in reranking. When $c = 0$, the result would be totally dominated by the visual consistency regardless of the initial ranking score list at all.

Generally speaking, the optimal value of c is influenced by two factors, in direct proportion to the text-based search baseline and in inverse proportion to the quality of visual consistency. The text-based search baseline is already known for us, as illustrated in Figure 4. The visual consistency is hard to be measured numerically. However, intuitively a query with more relevant samples may have higher visual consistency. Therefore, we can use the average number of relevant samples per query to approximate visual consistency.

As illustrated in Figure 4, the performance varies with different c. We can see that the MAP increases with c growing and arrives at its peak at around $c = 0.01$ on both TRECVID 2005 and 2006 while on TRECVID 2007 the best c is around 0.1. When c increases to 100, the reranking performance is already very close to the baseline. For TRECVID 2005 and 2006, although the former has a higher baseline, its average relevant sample is less than the later. Therefore, the optimal c on these two years is close. TRECVID 2007 on one hand has the lowest text search baseline. On the other hand its average relevant samples per query are obviously less than that of TRECVID 2005 and 2006. Therefore, its optimal c is larger than that for the other two years. It can be concluded that the trade-off parameter c can be set according to the performance of text search baseline as well as the number of relevant samples.

D. Complexity Analysis

For a query, N images are returned by text-based search engine, and the dimension of feature $\mathbf{x}$ is d. The time complexities for Lap-Point/Pair, NLap-Point/Pair are $O(dN^2 + N^3)$. The time complexities for Local-Point/Pair are $O(dN^2+N^3+K^3N)$, where $K = |\mathcal{N}(\mathbf{x})|$ is the number of neighbors for Local classifier. Since K usually is much smaller than N, the complexities for Local-Point and Local-Pair can be regarded as $O(dN^2+N^3)$ approximately, which is comparable to Lap-Point/Pair and NLap-Point/Pair.

Besides theoretical analysis, we also test the time cost experimentally for the best performed algorithm Local-Pair. The algorithm is implemented using MATLAB and run on a server with 2.67GHz Intel Xeon cpu and 16GB memory in single thread. K is fixed to 30. By averaging the time cost of the reranking over all queries, we obtain that Local-Pair finishes the reranking process within about 1 second when $N = 1000$. Reducing N will largely decrease the cost time. For $N = 300$, it only takes 0.1 second for reranking. From the theoretical analysis and the statistical numbers discussed above, we can see that the efficiency of Local-Pair is acceptable for real applications.

Fig. 5. Example images for “Panda” with different relevance degrees.

X. Experiments on Web Image Search Dataset

In Section IX, we have demonstrated the effectiveness of Bayesian visual reranking in video search application. However, its effectiveness in image search remains unexamined. In this section, we further verify it by conducting experiments on a real Web image search dataset.

A. Web Image Search Dataset

This dataset consists of 73,340 images collected from three most popular commercial search engines, including Google [1], Live [2] and Yahoo [3]. We selected 29 queries from a commercial image search engine query log as well as popular tags from Flickr [4] . These queries cover a vast range of topics, including scene (sky, winter), objects (funny dog, grape, and panda), named person (George W. Bush), etc. For each query, at most top 1000 images returned by each of the three search engines are collected. For each image, its relevance degree with respect to the corresponding query is judged by three participants, on four levels, “Excellent”, “Good”, “Fair” and “Irrelevant”. To have a vivid visualization for the four relevance degrees, examples are given in Figure 5 to show their implications.

B. Experimental Setting

The text baselines are the initial search results returned by the three search engines. The low level feature used for reranking is also 225-dimensional block-wise color moments.

For the performance measurement, the AP used in the experiments on TRECVID dataset cannot be adopted here. The reason is that AP is only suitable for two relevance levels.

[1]http://images .google.com/
[2]http://images.live.com/
[3]http://images.yahoo.com/
[4]http://www.flickr.com

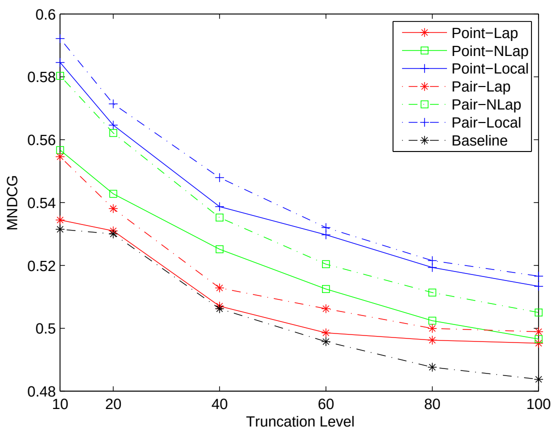


Fig. 6. MNDCG comparison for the six reranking methods on Live.

However we have four relevance levels for this Web dataset. The Normalized Discounted Cumulated Gain (NDCG) [36], which is a common measure used in information retrieval when relevance levels are more than two, is adopted here. For a given query, the NDCG score at position p in the ranking list l is calculated as

$$\mathrm{NDCG}@p = \frac{1}{Z}\sum\nolimits_{j=1}^{p}(2^{t_j}-1)/\log(1+j),$$

where t_j is the relevance degree of the jth image in l and Z is a normalization constant which is chosen to guarantee that the perfect ranking's NDCG@p is 1. The normalization constant Z is also called inverse perfect DCG, i.e., the inverse of DCG on the perfect search result.

Since we can't label the relevance for all the images in the index of the search engine for a given query, it is difficult to calculate Z. Here, we approximate the perfect search result by assuming that the top results returned by the three search engines comprise most, if not all, of the relevant images. Based on this strategy it should be fair to compare the performance of the three search engines. To evaluate the overall performance, we average the NDCGs over all queries to obtain the Mean NDCG (MNDCG).

C. Performance Comparison

As we discussed in Section IX-B, Local learning regularizer outperforms the other two regularizers and pair-wise ranking distance outperforms the point-wise distance. In this section, we further verify this conclusion on Web image search reranking. In addition, we also compare Local-Pair, with PRF-SVM and VisualRank to show the superiority of this reranking algorithm.

1) Comparison for Different Regularizers: Figure 6 gives the reranking result measured by MNDCG on Live for illustration. We can see that Local-Pair outperforms the other five algorithms.

By viewing Figure 6 to compare the regularizers, we can find that the Local-Pair outperforms Lap-Pair and NLap-Pair, and Local-Point outperforms Lap-Point and NLap-Point. From this observation, we get a rough conclusion that local learning regularizer is superior to the other two no matter which ranking

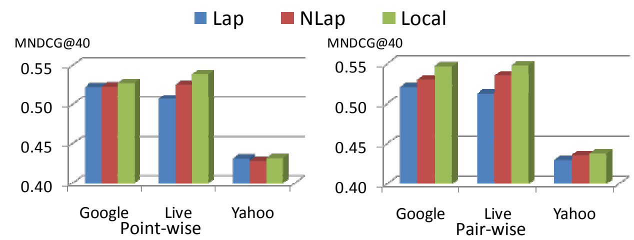


Fig. 7. MNDCG@40 comparison within different regularizers over the three search engines. The Local kernel regularizer performs the best among the three regularizers.

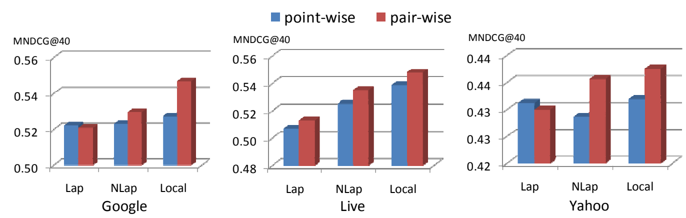


Fig. 8. MNDCG@40 comparison between point-wise and pair-wise ranking distances over the three search engines. The pair-wise ranking distance performs better than the point-wise one.

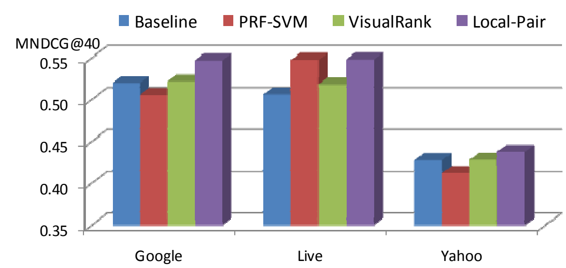


Fig. 9. MNDCG@40 comparison between Local-Pair and other two reranking methods as well as the text search baseline.

distance is adopted. To confirm this, we further conduct the experiments on other two search engines and the results are given in Figure 7. Due to the space limitation, we only illustrate the MNDCG@40 for comparison. We can clearly see that the local learning regularizer performs the best over all the three search engines consistently.

2) Comparison for Different Ranking Distances: Now, turn to view Figure 6 from the ranking distance comparison perspective. It shows that no matter which regularizer is adopted, the pair-wise ranking distance outperforms the point-wise one. In other words, Lap/NLap/Local-Pair achieve higher performance than Lap/NLap/Local-Point respectively. To further confirm this observation, experiments on other two search engines also have been done and the results are given in Figure 8. We can see that pair-wise ranking distance shows its superiority steadily. In summary, we can conclude that pair-wise ranking distance is better for web image search reranking than point-wise ranking distance.

As a conclusion, Bayesian visual reranking with pair-wise ranking distance and the local learning regularizer, i.e., Local-Pair, performs the best among the six variants. This finding is consistent with the experiments on the TRECVID dataset.

3) Comparison among Local-Pair, PRF-SVM and Visual-Rank: In the above we have verified that Local-Pair also performs the best among the six methods derived under Bayesian visual reranking. In this section, we will further confirm the superiority of Local-Pair by comparing it with PRF-SVM and VisualRank.

The performance in terms of MNDCG@40 of the three reranking methods as well as the text baseline is illustrated in Figure 9. We can see that PRF-SVM shows comparable reranking performance with Local-Pair on Live but its performance is not steady and its performance on Google and Yahoo even is worse than the baseline. For VisualRank, slight improvements are achieved over all three search engines. In contrast, Local-Pair improves the baseline steadily and outperforms both PRF-SVM and VisualRank. Up to now, we can get the conclusion that Local-Pair is effective for both video and image search reranking.

XI. Conclusion

In this paper, we propose a general framework, Bayesian visual reranking. It explicitly formulates visual reranking into a global optimization problem from the Bayesian perspective. Under this framework, a local learning based visual consistency regularizer and a pair-wise ranking distance are proposed to solve the problems existing in current pair-wise regularizers and point-wise ranking distance. The experiments conducted on the TRECVID 2005-2007 and Web image search datasets have demonstrated the effectiveness of the proposed Bayesian visual reranking framework as well as the local learning regularizer and the pair-wise ranking distance. This result encourages us to design more effective reranking methods under the Bayesian visual reranking framework in the future. For visual consistency term, we plan to embed semantic similarity and introduce distance metric learning to better model the visual consistency between images; for ranking distance term, we will mine precise and efficient list-wise ranking distances and incorporate them into Bayesian visual reranking objective function.

References

[1] H. S. Chang, S. Sull, and S. U. Lee, "Efficient video indexing scheme for content-based retrieval," *IEEE Trans. Circuits Systems for Video Technology*, pp. 1269–1279, 1999.

[2] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 1349–1380, 2000.

[3] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking via information bottleneck principle," *ACM Int'l Conf. Multimedia*, pp. 35–44, 2006.

[4] ——, "Video search reranking through random walk over document-level context graph," *ACM Int'l Conf. Multimedia*, pp. 971–980, 2007.

[5] J. Liu, W. Lai, X.-S. Hua, Y. Huang, and S. Li, "Video search re-ranking via multi-graph propagation," *ACM Int'l Conf. Multimedia*, pp. 208–217, 2007.

[6] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," *Int'l Conf. Machine Learning*, pp. 129–13, 2007.

[7] S. M. Kendall and J. D. Gibbons, "Rank correlation methods," *Edward Arnold: London*, 1990.

[8] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua, "Bayesian video search reranking," *ACM Int'l Conf. Multimedia*, pp. 131–140, 2008.

[9] Y. Liu, T. Mei, J. Tang, X. Wu, and X.-S. Hua, "Learning to video search rerank via pseudo preference feedback," *IEEE Int'l Conf. Multimedia & Expo*, pp. 297–300, 2008.

[10] R. Yan and A. G. Hauptmann, "Co-retrieval: A boosted reranking approach for video retrieval," *ACM Int'l Conf. Content-based Image and Video Retrieval*, pp. 60–69, 2004.

[11] R. Yan, A. G. Hauptmann, and R. Jin, "Multimedia search with pseudo-relevance feedback," *ACM Int'l Conf. Content-based Image and Video Retrieval*, pp. 238–247, 2003.

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

[13] S. Zhang, Q. Huang, G. Hua, S. Jiang, W. Gao, and Q. Tian, "Building contextual visual vocabulary for large-scale image applications," *ACM Int'l Conf. Multimedia*, pp. 501–510, 2010.

[14] L. Yang and A. Hanjalic, "Supervised reranking for web image search," *ACM Int'l Conf. Multimedia*, pp. 183–192, 2010.

[15] J. G. Carbonell, Y. Yang, R. E. Frederking, R. D. Brown, Y. Geng, and D. Lee, "Translingual information retrieval: A comparative evaluation," *Int'l Joint Conf. Artificial Intelligence*, pp. 708–715, 1997.

[16] Y.-H. Yang, P. T. Wu, C. W. Lee, K. H. Lin, W. H. Hsu, and H. H. Chen, "Contextseer: context search and recommendation at query time for shared consumer photos," *ACM Int'l Conf. Multimedia*, pp. 199–208, 2008.

[17] N. Slonim and N. Tishby, "Agglomerative information bottleneck," *Advances in Neural Information Processing Systerms*, pp. 250–255, 1999.

[18] T. Mei, X.-S. Hua, W. Lai, L. Yang, Z.-J. Zha, Y. Liu, Z. Gu, G.-J. Qi, M. Wang, J. Tang, X. Yuan, Z. Lu, and J. Liu, "Msra-ustc-sjtu at trecvid 2007: High-level feature extraction and search," *TREC Video Retrieval Evaluation Online Proceddings*, 2007.

[19] J. Meng, J. Yuan, Y. Jiang, N. Narasimhan, V. Vasudevan, and Y. Wu, "Interactive visual object search through mutual information maximization," *ACM Int'l Conf. Multimedia*, pp. 1147–1150, 2010.

[20] L. S. Kennedy and S.-F. Chang, "A reranking approach for context-based concept fusion in video indexing and retrieval," *ACM Int'l Conf. Content-based Image and Video Retrieval*, pp. 333–340, 2007.

[21] A. Natsev, A. Haubold, J. Tesic, L. Xie, and R. Yan, "Semantic concept-based query expansion and re-ranking for multimedia retrieval," *ACM Int'l Conf. Multimedia*, pp. 991–1000, 2007.

[22] R. van Zwol, V. Murdock, L. G. Pueyo, and G. Ramírez, "Diversifying image search with user generated content," *Multimedia Information Retrieval*, pp. 67–74, 2008.

[23] K. Song, Y. Tian, W. Gao, and T. Huang, "Diversifying the image retrieval results," *ACM Multimedia*, pp. 707–710, 2006.

[24] R. H. van Leuken, L. G. Pueyo, X. Olivares, and R. van Zwol, "Visual diversification of image search results," *WWW*, pp. 341–350, 2009.

[25] I. J. Cox, M. L. Miller, T. P. Minka, T. Papathomas, and P. N. Yianilos, "The bayesian image retrieval system, pichunter: Theory, implementation and psychophysical experiments," *IEEE Trans. Image Processing*, vol. 9, no. 1, pp. 20–37, January 2000.

[26] R. Fergus, P. Perona, and A. Zisserman, "A visual category filter for google images," *European Conf. Computer Vision*, pp. 242–256, 2004.

[27] X. Zhu, Z. Ghahramani, and J. D. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," *Int'l Conf. Machine Learning*, pp. 912–919, 2003.

[28] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," *Advances in Neural Information Processing Systerms*, pp. 321–328, 2003.

[29] M. Wu and B. Schölkopf, "Transductive classification via local learning regularization," *Int'l Conf. Artificial Intelligence and Statistics*, pp. 624–631, 2007.

[30] N. Cristianini and J. Shawe-Taylor, "An introduction to support vector machines and other kernel-based learning methods," *Cambridge University*, 2000.

[31] R. Herbrich, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regression," *Advances in Large Margin Classifiers*, pp. 115–132, 2000.

[32] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," *ACM Int'l Workshop on Multimedia Information Retrieval*, pp. 321–330, 2006.

[33] TRECVID, "Trecvid video retrieval evaluation. hppt://www-nlpir.nist.gov/projects/trecvid/."

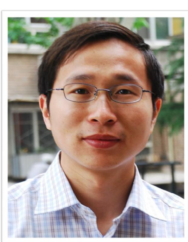
[34] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, M. Gatford, and A. Payne, "Simple, proven approaches to text retrieval," *Cambridge University Computer Laboratory Technical Report TR356*, 1997.

[35] W.-Y. Ma and H. J. Zhang, "Benchmarking of image features for content-based retrieval," *Conf. Record of the Thirty-Second Asilomar Conference on Signals, Systems & Computers*, pp. 253–257, 1998.

[36] K. Järvelin and J. Kekäläinen, "Ir evaluation methods for retrieving highly relevant documents," *ACM Special Interest Group on Information Retrieval*, pp. 41–48, 2000.

Xinmei Tian received the B.S. and Ph.D. from the University of Science and Technology of China in 2005 and 2010, respectively, both in Electronic Engineering and Information Science. From December 2007 to July 2008, she was a Research Intern with the media computing group at Microsoft Research Asia, Beijing. From August 2008 to December 2008, she was a Research Assistant with the School of Computing, the Hong Kong Polytechnic University. Her current research interests include computer vision, content-based video analysis, and image/video search reranking.

Linjun Yang is an associate researcher with the media computing group, Microsoft Research Asia. He received the M.S. degree from Fudan University in Computer Science and B.S. degree from East China Normal University in Electronics Engineering, in 2006 and 2001, respectively. He is currently pursuing PhD degree from Delft University of Technology, The Netherlands, in part-time. His current interests are in the broad areas of multimedia information retrieval, with focus on multimedia ranking and large-scale Web multimedia mining. He is a member of ACM and IEEE.



Jingdong Wang received the BSc and MSc degrees in automation from Tsinghua University, Beijing, China, in 2001 and 2004, respectively, and the PhD degree in computer science from the Hong Kong University of Science and Technology, Hong Kong, in 2007. He is currently an associate researcher at the Media Computing Group, Microsoft Research Asia. His areas of interest include machine learning, pattern recognition, multimedia computing, and computer vision. In particular, he has worked on kernel methods, semi-supervised learning, data clustering, image segmentation, image and video presentation, and organization and search. He is a member of the IEEE.

Xiuqing Wu received the B.S. degree from the University of Science and Technology of China, Hefei, in 1965. She is a Professor in the Department of Electronic Engineering and Information Science, University of Science and Technology of China. From 1985 to 1986, she was a Visiting Scientist in the Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA. Her research interests include intelligent information processing, multiresource data fusion, and digital image analysis.

Xian-Sheng Hua (M'04) received the B.S. and Ph.D. degrees from Peking University, Beijing, China, in 1996 and 2001, respectively, both in applied mathematics. Since 2001, he has been with Microsoft Research Asia, Beijing, where he is currently a Lead Researcher with the media computing group. His current research interests are in the areas of video content analysis, multimedia search, management, authoring, sharing, mining, advertising and mobile multimedia computing. He has authored or co-authored more than 180 publications in these areas and has more than 60 filed patents or pending applications. He is now an adjunct professor of University of Science and Technology of China, and serves as an Associate Editor of IEEE Trans. on Multimedia, Associate Editor of ACM Trans. on Intelligent Systems and Technology, Editorial Board Member of Advances in Multimedia and Multimedia Tools and Applications, and editor of Scholarpedia (Multimedia Category).