Non-parametric Kernel Ranking Approach for Social Image Retrieval

Jinfeng Zhuang, Steven C.H. Hoi
School of Computer Engineering
Nanyang Technological University
Singapore 639798
{zhua0016, chhoi}@ntu.edu.sg

ABSTRACT

Social image retrieval has become an emerging research challenge in web rich media search. In this paper, we address the research problem of text-based social image retrieval, which aims to identify and return a set of relevant social images that are related to a text-based query from a corpus of social images. Regular approaches for social image retrieval simply adopt typical text-based image retrieval techniques to search for the relevant social images based on the associated tags, which may suffer from noisy tags. In this paper, we present a novel framework for social image re-ranking based on a non-parametric kernel learning technique, which explores both textual and visual contents of social images for improving the ranking performance in social image retrieval tasks. Unlike existing methods that often adopt some fixed parametric kernel function, our framework learns a non-parametric kernel matrix that can effectively encode the information from both visual and textual domains. Although the proposed learning scheme is transductive, we suggest some solution to handle unseen data by warping the non-parametric kernel space to some input kernel function. Encouraging experimental results on a real-world social image testbed exhibit the effectiveness of the proposed method.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.6 [**Computing Methodologies**]: Artificial Intelligence

General Terms

Algorithms, Experimentation

Keywords

Social Image Retrieval, Non-parametric Kernel Learning, Visual Ranking, Semidefinite Programming, Convex Optimization

CIVR '10, July 5-7, Xi'an China

1. INTRODUCTION

Along with the popularity of digital imaging devices and the advances of Internet technologies, more and more digital images and photos have been uploaded, shared, and distributed on the World Wide Web (WWW). Today, the volume of multimedia data has accounted for a significant portion of the entire WWW data. Web image search thus has become an active yet very challenging research problem [6, 15, 22].

One major difficulty for web image search is that most images are usually not annotated with proper tags, and many of them are even completely unlabeled. In addition, even for those annotated images, many of the associated tags are usually *noisy*, *irrelevant*, and typically *incomplete* for describing the semantic contents of the images. This poses a challenge for typical text-based web image search approaches used by most existing web search engines, which simply apply the regular text-based indexing and retrieval techniques to the web image search task.

One possible way to improve existing text-based web image search solutions is to explore the content-based image retrieval (CBIR) techniques [21, 28], which have been extensively studied in multimedia communities over the past decade. While CBIR techniques are able to retrieve images based on low-level visual contents, a critical drawback of CBIR lies in the difficulty of providing/identifying an appropriate query example to describe the user's search intention. This is probably why text-based image query paradigm remains the most popular approach even though their resulting performance is often far from perfect.

Instead of directly applying CBIR techniques, another approach to improving the text-based image retrieval is to investigate automatic image annotation techniques [9, 23, 30, 32], which aims at automatically annotating a web image with a set of relevant tags. While automatic image annotation has been actively studied and often demonstrated to improve the performance of text-based image retrieval in small-scaled datasets, it remains unclear if these techniques are able to scale effectively for real large-scale applications, especially for WWW images.

Recently, there is some emerging study for improving a text-based web image retrieval task by implicitly exploring the visual content without using query image examples [16]. They proposed a visual re-ranking method, termed "VisualRank", which is an extension of PageRank [3] by exploiting the visual information in the web image ranking task. Although encouraging results had been shown for improving regular text-based web image search, the VisualRank ap-

proach adopts a rigid visual similarity measure approach, which may suffer from the well-known challenge of semantic gap between low-level visual features and high-level semantic concepts [21]. This motivates our study in this paper, which aims to improve VisualRank by a novel non-parametric kernel ranking scheme.

Besides, unlike previous studies on web image retrieval, our study considers *social images*, which are uploaded and shared by web users in social web sites. In contrast to generic web images, social images have rich user-generated contents, including tags provided by web users. Despite the good quality tags, noise remains an unavoidable issue for social image retrieval. In this paper, our goal is to improve the performance of text-based social image retrieval by mining rich tag and visual contents through a novel non-parametric kernel based ranking approach.

In particular, our framework differs from the previous approaches that often assume some fixed kernel function for similarity measure. We propose to learn an optimized kernel from the rich annotated social images, aiming to bridge the semantic gap towards effective image similarity measure. To the best of our knowledge, this is the first kernel learning approach that learns to optimize a kernel for a web image ranking task.

As a summary, the key contributions of this paper include:

- We propose a novel kernel based ranking approach for social image ranking using the state-the-of-the-art non-parametric kernel learning, which aims to learn an optimized kernel from both textual tags and visual contents of social images;
- We present a fast algorithm for non-parametric kernel ranking, which can efficiently learn non-parametric kernels for improving the "VisualRank" performance in a text-based social image retrieval task;
- We conduct an empirical study on a large-scale social image testbed for evaluating the performance of the proposed non-parametric ranking algorithm on real social image retrieval tasks.

The rest of this paper is organized as follows. Section 2 reviews some preliminaries of the VisualRank [16] framework, in which a similarity matrix plays a central role. Section 3 presents the proposed non-parametric kernel learning technique for learning kernels from social images, and studies its application for improving text-based social image retrieval tasks. Section 4 discusses our empirical study, Section 5 reviews related work, and Section 6 sets out our conclusions.

2. PRELIMINARIES

2.1 Problem Setting

A social image often contains rich contents, including user-generated tags, visual content, rating, etc. In our approach, we simplify the social image representation by considering only visual content and user-generated tags. Specifically, a social image is represented as $z_i = (x_i, t_i)$, where $x_i \in \mathcal{X}$ denotes the vector of its visual content, and $t_i \in \mathcal{T}$ denotes the vector of its associated tags. We further assume that the query space $\mathcal{Q}$ shares the same vocabulary space with the tag space $\mathcal{T}$, i.e., $\mathcal{T} := \mathcal{Q}$.

Consider a social image retrieval task, given some text-based query $q \in \mathcal{Q}$, our goal is to retrieve and rank the relevant social images from a social image repository $\mathcal{D} = \{z_i = (x_i, t_i), i = 1, \ldots, n\}$. Specifically, let us denote by $\mathcal{R}(q)$ and $\overline{\mathcal{R}}(q)$ the set of relevant and irrelevant social images related to query q respectively, the goal of the social image retrieval task is to find some ranking function

$$f : \mathcal{X} \times \mathcal{T} \rightarrow \mathcal{R}$$

such that $f(z_i) > f(z_j)$ for any two images $z_i \in \mathcal{R}(q)$ and $z_j \in \overline{\mathcal{R}}(q)$.

Most of the existing approaches for web image retrieval by current web search engines often formulate the ranking function f by adopting some text-based retrieval models to rank the web images according to their associated tags. While these approaches are efficient by taking advantage of effective text indexing and search techniques, their ranking performance is not always satisfied due to the noisy textual content that is common for WWW images. To address this challenge, the recent study in [16] has proposed a "Visual Reranking" method for improving the web image search performance, which motivates our study in this paper. Next we briefly review the Visual Ranking method.

2.2 VisualRank for Web Image Search

The key idea of VisualRank [16] was adapted from the intuition of PageRank [3], i.e., images similar to a relevant image are also relevant. Specifically, suppose we have a similarity matrix K, where K_{ij} is the visual similarity between image x_i and x_j. If the entry K_{ij} has a large value, we deem that x_i supports the importance of x_j to a large extent. Therefore, by exploiting the visual information between pairs of images, we re-rank the images and hopefully improve the accuracy of retrieval results of text-based methods.

Let $VR \in \mathbb{R}^n$ be the ordering score of n social images. A larger score indicates the image is more relevant, thus resulting a higher position in the ranking result. The VisualRank (VR) is iteratively defined as:

$$VR = K \times VR,$$

where K is a normalized matrix such that $0 \leq K_{ij} \leq 1$ and each column sums to 1. When K is symmetric, it becomes a doubly stochastic matrix [25]. From a random walk perspective, K_{ij} measures the probability of z_i randomly travels to z_j. Typically, a damping factor λ_d is introduced to incorporate some prior ranking score P:

$$VR = \lambda_d K \times VR + (1 - \lambda_d)P, \quad \text{where } P = \left[\frac{1}{n}\right]_{n \times 1}$$

The success of VisualRank highly relies on a good similarity matrix K. The regular approach [16] is to simply measure the similarity/distance matrix based on visual contents $\mathcal{X}$ (either local or global visual features) extracted from the images. Such an approach falls short when the visual similarity is not very effective, which is primarily due to the well-known semantic gap between low-level visual features and high-level semantic concepts. To address the limitation of the regular VisualRank approach, in this paper, we present a novel non-parametric kernel ranking approach, which aims to learn an effective visual kernel K that best preserves the semantics by exploiting both visual and textual contents of social images.

3. NON-PARAMETRIC KERNEL RANKING FOR SOCIAL IMAGE RETRIEVAL

In this section, we present a novel non-parametric kernel ranking method for social image retrieval. The basic idea is to improve VisualRank by learning an optimized kernel that encodes both visual and tag information.

3.1 Learning Semantics-Preserving Kernels

The goal of our task is to learn a kernel $K \in \mathbb{S}_+^{n\times n}$ for measuring similarity $k(z_i, z_j)$ between any two social images z_i and z_j, which best preserves the semantics by exploring both textual and visual contents.

The first concern of our kernel learning task is how to effectively explore the textual contents of social images. One approach to this challenge is to optimize K by maximizing the dependance between the two domains $\mathcal{X}$ and T. To this end, we suggest to employ the Hilbert-Schmidt Independence Criterion (HSIC) [11] for the dependence measure between $\mathcal{X}$ and $\mathcal{T}$, which is defined below:

DEFINITION 1 (EMPIRICAL HSIC[11]). *Let us denote by $Z = \{(x_1, t_1), \ldots, (x_n, t_n)\}$ a collection of n independent observations drawn from the joint distribution $\mathcal{X}$ and $\mathcal{T}$, an estimator of HSIC is given by*

$$\text{HSIC}(Z, \mathcal{X}, \mathcal{T}) = \frac{1}{(n-1)^2} tr\, KHK_tH,$$

where K and K_t refers to the kernel matrices in visual space and textual space respectively, and $H = I - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$.

The HSIC measures the dependence between two random variables $x \in \mathcal{X}$ and $t \in \mathcal{T}$ by computing the square of the norm of the cross-covariance operator over the domain $\mathcal{X} \times \mathcal{T}$ in Hilbert space. One can show the norm vanishes on the condition that x and t are independent. A large value of HSIC indicates a strong dependence with respect to the choice of the kernels. If we ignore the centering matrix H, it coincides with the concept of kernel target alignment [8], which has been successfully applied in many recent machine learning studies [29].

In addition to the tag information, we expect the kernel K should also encode the information of visual features. Combining the empirical HSIC and the visual feature concern, we formulate the semantics-preserving kernel learning problem as a non-parametric kernel learning task:

$$\begin{aligned} \max_K \quad & \text{tr}\, KHK_tH \\ \text{s.t.} \quad & K_{ii} + K_{jj} - 2K_{ij} \le d_{ij}^2 \;\; \forall (i,j) \in \mathcal{N}, \\ & K \subset \mathbb{S}_+^{n\times n}, \end{aligned} \tag{1}$$

where $\mathbb{S}_+^{n\times n}$ denotes the space of positive semi-definite cones, and $\mathcal{N}$ is a set of all pairs (i, j) where x_i and x_j are close to each other according to their visual distance, which is empirically obtained by checking if they are among each other's k nearest neighbors based on some distance metric defined on the visual space $\mathcal{X}$.

Remark. Unlike a regular fixed kernel (e.g. an RBF kernel) defined on visual space X, the above formulation optimizes the kernel K by considering two factors: (1) exploring the textual contents through K_t and (2) encoding the visual information by enforcing the constraints according to visual distances d_{ij}, in a unified learning framework to effectively overcome the semantic gap challenge.

3.2 Efficient Kernel Learning Algorithm

The optimization task of (1) is a Semi-definite Programming (SDP) problem [1], which in general can be solved by existing SDP solvers. However, the time complexity of solving an SDP problem by a standard SDP problem is often as high as $O(n^6)$. Such intensive computation cost prohibits its application to large-scale real applications. To address this challenge, we propose an efficient algorithm that adopts a special SDP solver to explore the problem structures.

First of all, we notice that it may be too strong to enforce all distance constraints satisfied in the optimization problem (1). We reformulate the optimization by introducing slack variables ξ (which is somewhat similar to the soft margin SVM approach) as follows:

$$\begin{aligned} \min_{K,\xi} \quad & -\text{tr}\, HKHK_t + \frac{C}{2}\sum \xi_{ij}^2 \\ \text{s.t.} \quad & K \in \mathbb{S}_+^{n\times n}, \text{tr}\, KK \le B, \\ & \text{tr}\, KA^{ij} \le d_{ij}^2 + \xi_{ij},\; \forall (i,j) \in \mathcal{N} \end{aligned} \tag{2}$$

where ξ penalizes the violation of the distance constraints, and A^{ij} is a matrix of size $n \times n$ with only four nonzero elements, i.e., $A_{ii}^{ij} = A_{jj}^{ij} = 1$ and $A_{ij}^{ij} = A_{ji}^{ij} = -1$. In the above, we also introduce a normalization constraint $trKK \le B$ to avoid K being too large, where B is a positive constant.

To solve the optimization, we introduce dual variables α for the slack variables ξ and arrive at the partial Lagrangian (refer to [1] for standard techniques):

$$\begin{aligned} L(K;\alpha) \;=\; & -\text{tr}\, KHK_tH + \frac{C}{2}\sum \xi_{ij}^2 \\ & + \sum \alpha_{ij}(\text{tr}\, KE^{ij} - d_{ij}^2 - \xi_{ij}) \end{aligned} \tag{3}$$

Taking derivatives of $L(K;\alpha)$ with respect to ξ_{ij}'s and set them to zero, we can rewrite the optimization in an equivalent form:

$$\begin{aligned} \max_\alpha \min_K \quad & J(K,\alpha) \\ \text{s.t.} \quad & K \in \mathbb{S}_+^{n\times n},\; \text{tr}\, KK \le B \end{aligned} \tag{4}$$

where the objective function $J(K,\alpha)$ is formulated as:

$$\begin{aligned} J(K,\alpha) \;=\; & \text{tr}\,((\sum_{ij}\alpha_{ij}A^{ij} - HK_tH)K) \\ & - \sum_{ij}\alpha_{ij}d_{ij}^2 - \frac{1}{2C}\sum_{ij}\alpha_{ij}^2 \end{aligned} \tag{5}$$

Next we solve the above optimization using an iterative gradient projection approach, i.e.,(1) fixing α to update K, and then (2) fixing K to update α.

First of all, by fixing α, we can find K by reducing the optimization to the following

$$\max_K \text{tr}\, AK \;\; s.t. \;\; K \succeq \mathbf{0}, \;\; \text{tr}\, KK \le B, \tag{6}$$

where $A = \sum_{ij}\alpha_{ij}A^{ij} - HK_tH$. This optimization can be resolved by making use of the following theorem [34]:

THEOREM 1. *Consider the following optimization with a symmetric matrix A and a positive constant B*

$$\max_K tr\, AK \;\; s.t. \;\; K \succeq \mathbf{0}, \;\; tr\, KK \le B, \tag{7}$$

the optimal solution K^* can be expressed as the following

closed-form solution:

$$K^* = A_+ \sqrt{\frac{B}{tr\,(A_+ A_+)}} \tag{8}$$

where $A_+ = \pi_+(A)$ that projects A onto the positive semi-definite space, i.e., $\pi_+(A) = V\Sigma^+ V^\top$, where $\Sigma^+ = max(\Sigma, 0)$, V is the matrix of A's eigenvectors and Σ is a diagonal matrix containing A's eigenvalues.

Second, when K is fixed, we can update α by computing the gradient:

$$\alpha^{t+1} = \alpha^t + \eta(\text{tr } KA^{ij} - \frac{1}{C}\alpha_{ij}^t - d_{ij}^2)$$

where η is a learning rate that can be either a constant or may be changed at each step. Finally, we summarize the key steps of the proposed efficient algorithm in Algorithm 1.

Algorithm 1 Iterative Algorithm for Fast Kernel Learning.

Input: Training data set Z, kernel function k_t on $\mathcal{T} \times \mathcal{T}$, pairwise constraint matrix A, parameters C and B;
Output: K^* and α.

1: Construct the matrix HK_tH by kernel k_t;
2: Initialize α;
3: Set $A = HK_tH - \sum_{ij} \alpha_{ij}^t A^{ij}$;
4: Compute the closed-form solution of K^* using (8);
5: Determine a learning rate η;
6: Update $\alpha^{t+1} \leftarrow \alpha^t + \eta(\text{tr } K^* A^{ij} - \frac{1}{C}\alpha_{ij}^t - d_{ij}^2)$;
7: Repeat steps 3-6 until convergence.

Since the problem is convex, it can be shown that such an iterative gradient projection converges to the correct solution according to convex optimization theory [2].

3.3 Extension of Handling Unseen Data

The kernel learned above is purely non-parametric, which in general cannot handle unseen data directly. To address this limitation, we can extend the kernel $\widehat{K}$ to handle unseen data by some modification. Specially, we can adopt some kernel warping approach by warping the norm in Hilbert space using a regularizer depending on training data (both labeled and unlabeled) [27].

Let $\mathcal{H}$ denote the original Hilbert space reproduced by kernel function $k(\cdot,\cdot)$, and $\widetilde{H}$ denote the deformed Hilbert space. In [27], the authors assume the following relationship between the two Hilbert spaces:

$$\langle f, g\rangle_{\widetilde{\mathcal{H}}} = \langle f, g\rangle_H + f^\top G g$$

where $f(\cdot)$ and $g(\cdot)$ are two functions, $f = (f(x_1), ..., f(x_N))$ evaluates the function $f(\cdot)$ for both labeled and unlabeled data, and G is the distance metric that captures the geometric relationship among all the data points. The deformation term $f^\top Gg$ is introduced to assess the relationship between the functions $f(\cdot)$ and $g(\cdot)$ based on the observed data. Given an input kernel k_x, the explicit form of the new kernel function $\tilde{k}$ can be derived as below:

$$\tilde{k}(x,y) = k_x(x,y) - \kappa_y^\top (I + \widehat{L}K_x)^{-1}\widehat{L}\kappa_x, \tag{9}$$

where K_x is the kernel matrix evaluated by the input kernel k_x, L is chosen to be the graph Laplacian $\widehat{L} = \text{diag}\{\widehat{K}\cdot\mathbf{1}\} - \widehat{K}$ with respect to the learned transductive kernel $\widehat{K}$.

3.4 Implementation Issues: Further Speedup

Algorithm 1 makes the non-parametric kernel ranking method feasible for medium-scale data sets. However, in a real web-scale application, the size of training data n can be very large, typically in millions or billions scale. To make our ranking scheme practical to real applications, we further speed up our solution by the following strategies.

First, we notice that the main computation of the algorithm is the projection step by computing eigen-decomposition. Since the target kernel K is often not full rank [34], it is possible to adopt some low rank matrix to approximate $\widehat{K} = VPV^\top$, where V is of size $n \times m$, $m \ll n$. Specifically, let L be the graph Laplacian of $\mathcal{X}$ using n nearest neighbors, we choose V to be the m eigenvectors corresponding to the m smallest eigenvalues of L. Using the low-rank representation of K, the main computation, i.e., the eigen-decomposition can be reduced from $\mathcal{O}(n^3)$ to $\mathcal{O}(m^3)$.

Second, the above kernel learning scheme in general learns a kernel matrix for all the social images from the database. In practice, we can speed up the solution by considering a query-dependent re-ranking approach, which is commonly adopted in web image retrieval, such as VisualRank [16], by considering that precision is often a more concern than recall in a web image retrieval task. Specifically, we employ a two-level social image retrieval scheme by first retrieving the relevant images using *tag* based ranking, and then applying the proposed non-parametric kernel ranking technique for re-ranking the subset of the social images.

4. EXPERIMENTS

In this section, we empirically evaluate the performance of our non-parametric kernel ranking method for social image retrieval tasks.

4.1 Experimental Testbed

We create a large-scale social image data set with about 1,000,000 social images crawled from Flickr [1]. These social images contain rich information, including user-generated tags. In our text-based retrieval baseline, only associated tags are engaged for a social image retrieval task.

To form a benchmark testbed for performance evaluation, we pick a set of text based queries that covers a large range of generic WWW image search, including *animals*, *plants*, *humans*, *landmarks*, *natural sceneries*, and *human-made objects*. Figure 1 shows a list of sample images, and Table 1 summarizes the statistics of the related social images in the evaluation set. For the set of queries, we invited 10 staff to manually label the relevance of the top retrieved social images for each query to indicate if the retrieved image is relevant to the query.

Table 1: The statistics of images in our evaluation.

	#Images	#Tags	#Tags/#Image
training	25,819	557,160	21.58

Table 2 shows the statistics of the list of queries with their top retrieved and labeled images. We evaluate the

[1]http://www.flickr.com/

Figure 1: List of sample images related to the queries used in our experimental testbed.

performance of the ranking methods by examining if the re-ranking results are improved according to the ground truth.

4.2 Compared Methods and Experimental Setup

We adopt the VisualRank framework for social image retrieval and compare the following kernels:

- Text: the original text-based retrieval baseline method. We simply compute the similarity between a text-based query and the tags of some image. No re-ranking algorithm is performed for this baseline;
- K-tag: The kernel $k(x_i, x_j)$ is computed by a linear kernel among the tags with TF-IDF weighting;
- Linear: a linear kernel on visual features, $k(x_i, x_j) = x_i^\top x_j$;
- Poly: a polynomial kernel on visual features $k(x_i, x_j) = (1 + x_i^\top x_j)^d$;
- RBF: an RBF kernel on visual features, $k(x_i, x_j) = \exp(-\|x_i - x_j\|^2/\sigma)$;
- NPK: the proposed non-parametric kernel on visual features.

The hyper-parameters of the above kernels are tuned by a set of validation queries. All experiments were conducted on a Windows PC with 3.4GHz 32bit CPU and 3GB RAM.

Table 2: The statistics of the queries in the data set.

TestQuery	#RelDoc	TestQuery	#RelDoc
eiffel tower	919	Barack Obama	88
great wall	57	Big Ben	200
red car	477	pyramid	229
airplane	771	Ferrari	401
lily	452	bicycle	1475
banana	124	sunflower	798
fruit orange	266	strawberry	715
singing	945	panda	1512
lion	1308	sheep	487
elephant	739	eagle	664

4.3 Feature Representation and Performance Evaluation Metric

For each social image in our data set, we must extract features from the tag and image domains.

For textual features, we use *normalized tf-idf* feature for each tag in the vocabulary. Assume each tag occurs at most one time for each image, the j-th dimension for t_i is simplified:

$$t_{ij} = \frac{\delta_{ij} idf_j}{\sqrt{\sum_{j=1}^{d_1} (\delta_{ij} idf_j)^2}},$$

here δ_{ij} indicates whether the tag t_{ij} appears in the tag t_i of image x_i, $d_1 = |\mathcal{T}|$ is the vocabulary size. We filtered out non-English and extremely low frequent ones. The inverse document frequency is $idf_j = \log(r_j)$, where r_j is the number of images tagged by t_j.

For visual features, we extract four kinds of effective global features that have been used in previous CBIR studies [33].

- **Grid Color Moment:** An image is partitioned into 3×3 grids. For each grid, we extract three kinds of color moments: mean, variance and skewness in each color channel (R, G, and B), respectively. Thus, an 81-dimensional grid color moment vector is used;
- **Local Binary Pattern:** The local binary pattern [24] is defined as a gray-scale invariant texture measure, derived from texture in a local neighborhood. We adopt a 59-dimensional LBP histogram vector;
- **Gabor Wavelet Texture:** Each image is first scaled to 64×64 pixels, and the Gabor wavelet transform [19] is applied on the image with 5 levels and 8 orientations, resulting in 40 subimages. For each subimage, 3 moments are calculated: mean, variance and skewness. Thus, a 120-dimensional Gabor feature vector is used;
- **Edge:** The edge orientation histogram is extracted from an image. We first convert the image to a gray image and then employ a Canny edge detector to obtain the edge map, which is then quantized into 36 bins of 10 degrees each. Besides, an additional bin is used to count the number of pixels without edge. Hence, a 37-dimensional vector is used for shape features.

In total, a 297-dimensional vector is used to represent the visual features for each image in the data sets.

Table 3: The top-k re-ranking accuracy (%) of different kernels used in the VisualRank framework.

	Text	K-tag	Linear	Poly	RBF	NPK
Top3	51.7	61.7	50.0	50.0	65.0	**70.0**
Top5	51.0	58.0	46.0	52.0	60.0	**66.0**
Top10	46.0	59.0	47.0	52.0	58.5	**63.0**
Top20	45.3	59.0	49.8	50.8	60.8	**68.3**
Top50	48.4	55.8	48.8	50.6	59.0	**65.0**
Top100	50.4	51.7	48.0	50.1	56.5	**62.9**

Finally, for performance evaluation metric, we follow the study of VisualRank[16] and employ the Top(k) accuracy, which is the percentage $\Pr(k)$ of relevant images ranked within the top k returned images by the retrieval model. Therefore, Top(k) accuracy evaluates the possibility that a user would locate the relevant images on the first k returned images by an image search engine.

4.4 Experiment I: Retrieval Accuracy

The experimental results of average top-k retrieval accuracy are summarized in Table 3. From the results, we can draw several observations.

First of all, among all compared re-ranking methods, we found that not every approach can always achieve the improvements over the text-based retrieval baseline. In particular, only K-tag, RBF, and NPK achieved significant improvements over the baseline, while the "Linear" and "Poly" approaches only achieved slight improvement or even degrade the performance in some cases. These results again validate that it is very important to choose an appropriate kernel for the VisualRank framework in the re-ranking task.

Second, comparing with two parametric kernels, i.e., a tag based "K-tag" kernel and a visual based "RBF" kernel, we found that both of them can improve over the baseline approach considerably, in which the visual based "RBF" performed slightly better. These results show that it is effective by applying the VisualRank framework to resolve the social image retrieval problem if the kernel is chosen appropriately.

Third, among all approaches, it is clear to see that the proposed NPK approach achieved the best performance in all cases. Specifically, for top 3 retrieved results, the accuracy of the original text based retrieval baseline is about 51.7%, however, the proposed NPK approach is able to boost the retrieval accuracy to 70%, which is considerably better than other parametric kernels, e.g. K-tag of 61.7% and RBF of 65%. These results again validate the proposed NPK technique is effective and significant for improving the re-ranking performance when applying the VisualRanking framework for the social image retrieval tasks.

To better understand the above encouraging empirical results, we here discuss several reasons. First of all, we note that the reason that the original text-based retrieval baseline does not always achieve perfect results is primarily due to the noisy or incomplete annotated tags with the social images. Hence, by applying the VisualRank approach, we are able to improve the search results by re-ranking the top retrieved results based on mining their pairwise similarity relationship. Further, similar to conventional CBIR methods, the *semantic gap* issue is a common challenge for visual similarity measure. This challenge may hinder the success of the VisualRank framework if it adopts a regular parametric kernel (e.g. polynomial or RBF). However, by applying the proposed NPK learning technique, we are able to narrow down the semantic gap, and thus could significantly boost the re-ranking performance.

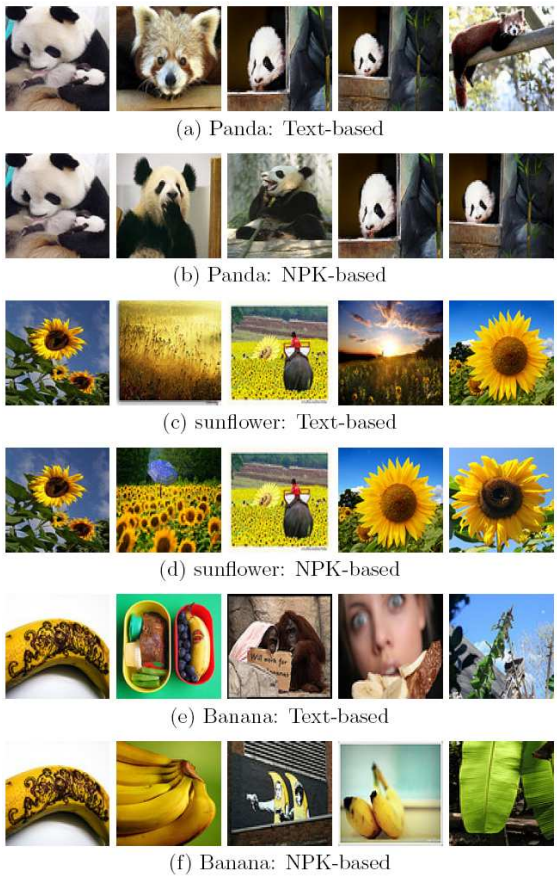

Figure 2: Qualitative re-ranking performance by the proposed NPK learning method.

Finally, we also show some examples of qualitative ranking results in Figure 2. Comparing with the text-based baseline approach, the proposed NPK approach mostly can return more relevant images in the top retrieved examples. These results again validate the proposed NPK ranking approach is able to improve over the baseline.

4.5 Experiment II: Computation Cost

The retrieval time is crucial for IR systems. In our non-parametric learning framework, we propose a fast solution that avoids general semi-definite programming solvers. In this section, we evaluate the time cost of the following solvers:

- **SDP**: We solve (5) using general SDP solvers. However, the time complexity of such an SDP solver is often $O(n^6)$, which prohibits the learning task scalable to real applications. Here we adopt the approximation method in Section 3.4 (see also [29]).

- **NPK**: The proposed method described in Algorithm 1.

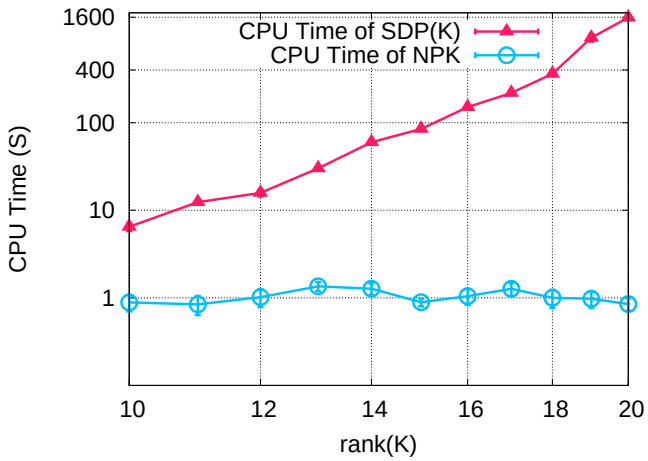

Figure 3: Log-scaled computation time of SDP and NPK on the query *Big Ben* which contains 200 images with the low-rank approximation scheme.

Figure 3 plots the time cost of these two methods by varying the rank of the resulting kernel. From the result, we have two observations: (1) The NPK method took only about 1 second for a data set with around 200 images, which is practically efficient for an online application. (2) It is clear that the time complexity of SDP is not efficient, which increases exponentially and thus cannot scale to a real application. These results again show that the proposed NPK ranking solution is practically efficient and scalable for real social image retrieval applications.

5. RELATED WORK

Our work is related to several research topics: content-based annotation and retrieval, text-based social image retrieval, and kernel learning techniques.

Content-based methods for auto-image annotation and retrieval have been extensively studied in multimedia area [21, 31]. The earlier studies on CBIR have focused on investigating effective low-level features for image representation. While various features have been proposed, effective CBIR methods remain a challenging research problem till now, which is primarily due to the well-known semantic gap between the low-level features and high-level semantic concepts. To overcome the semantic gap challenge, many research efforts have been paid on improving the interactive image retrieval performance using relevance feedback techniques [26, 14]. Our work is in general related to CBIR as we also exploit the visual contents of images, but is also different in that we focus on improving text-based image retrieval performance without explicit query image example while regular CBIR often focuses on solving example-based image retrieval tasks without text.

In addition to CBIR, multimedia researchers have actively explored content-based methods for auto-image annotation. By annotating an image with some machine learning model automatically, we are able to enable a large amount of unlabeled images indexed and searchable by existing text based image search engines. A variety of techniques have been proposed for auto-image annotation in recent years [9, 30, 23, 32]. In general, auto-image annotation can be viewed as an intermediate task for a generic web image retrieval task. Our work is however different from the annotation-based retrieval approach, in that our focus is not on improving the intermediate annotation performance, but for optimizing the retrieval performance directly.

Text-based image retrieval aims to search for similar images relevant to a text-based query, in which each image is represented by a set of tags/keywords similar to a document. Thus text-based image retrieval often closely relates to document information retrieval, e.g., learning to rank [17, 5, 4, 7, 10], etc. As one example of text-based web image retrieval, social image retrieval [22, 6, 16, 10, 12, 9] has been actively studied recently. The goal of social image retrieval is to learn a model that outputs an order over a set of documents. For our proposed algorithm, we try to refine the results of some ranking model instead of the ranking framework, though it could be plugged into some kernel-based discriminative ranking model, like [10]. A lot of methods have been proposed for improving regular text-based web image retrieval performance, including re-ranking by relevance model [22], clustering based re-ranking [6], and the VisualRank approach by modifying PageRank [16]. Our work is mainly motivated by the VisualRank approach for overcoming its limitation through a kernel based learning approach, where a similarity matrix plays a central role for propagating the ranking scores among images.

Our algorithm also closely relates to the machine learning subject, i.e., *kernel learning* [20], which has been actively studied in machine learning community. Different from traditional parametric kernel combinations [20], researchers have proposed *non-parametric kernel learning* (NPK) algorithms, of which the output is just the kernel matrix [18, 13, 34]. It is known that each positive semi-definite kernel matrix K (p.s.d.) corresponds to the evaluation of some underlying Mercer kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Therefore, instead of using explicit kernel functions, it is more powerful to learn a non-parametric kernel in a transductive setting, such as the social image re-ranking task.

The key merit of NPK is that it is more powerful and flexible for fitting diverse patterns in a real complicated application. It often produces the state-of-the-art empirical results for many applications, such as classification and clustering, etc. To the best of our knowledge, no existing study in machine learning has addressed social image retrieval by optimizing kernel in the re-ranking task. Our work focuses on adapting the recent advances of non-parametric kernel learning techniques with applications to the social image ranking task. However, due to the p.s.d. constraint over K, the learning often results in a semi-definite program [1], which has the time complexity of $O(N^6)$. Such intensive computation makes NPK prohibitive for large-scale application. Very recently, Zhuang et. al. [34] proposed an iterative algorithm which speeds up the kernel learning significantly. Based on the previous work, this paper proposed a fast NPK ranking algorithm to resolve our social image re-ranking task.

6. CONCLUSIONS

This paper presented a novel non-parametric kernel ranking approach for overcoming the challenge of social image retrieval. Unlike the previous studies that may considerably suffer from the semantic gap by some rigid visual similarity function for web image re-ranking, our new approach aims

to overcome the semantic gap issue by learning an effective kernel from mining both textual tags and visual contents in a unified learning framework. Encouraging experimental results on a large-scale social image testbed validated the efficacy of the proposed method. Future work will investigate more effective visual features to improve the performance and explore other user-generated contents of social images in the ranking task.

Acknowledgements

The work was supported by Singapore NRF Interactive Digital Media R&D Program, under research grant NRF2008IDM-IDM004-006, Singapore MOE tier-1 Research Grant (RG67/07), and Microsoft MCE research grant.

7. REFERENCES

[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[2] S. Boyd and L. Xiao. Least-squaures covariance matrix adjustment. *SIAM Journal of Matrix Anal. Appl.*, 27(2):532–546, 2005.

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

[4] C. J. C. Burges, R. Ragno, and Q. V. Le. Learning to rank with nonsmooth cost functions. In *Advances in Neural Information Processing Systems*, pages 193–200, 2006.

[5] C. J. C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. N. Hullender. Learning to rank using gradient descent. In *Proceedings of the International Conference on Machine Learning*, pages 89–96, 2005.

[6] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen. Hierarchical clustering of www image search results using visual, textual and link information. In *Proc. 12th ACM international conference on Multimedia*, pages 952–959, New York, NY, USA, 2004.

[7] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the International Conference on Machine Learning*, pages 129–136, 2007.

[8] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. S. Kandola. On kernel-target alignment. In *Advances in Neural Information Processing Systems*, pages 367–373, 2001.

[9] R. Datta, W. Ge, J. Li, and J. Z. Wang. Toward bridging the annotation-retrieval gap in image search by a generative modeling approach. In *ACM Multimedia*, pages 977–986, 2006.

[10] D. Grangier and S. Bengio. A discriminative kernel-based approach to rank images from text queries. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(8):1371–1384, 2008.

[11] A. Gretton, O. Bousquet, A. J. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *ALT*, pages 63–77, 2005.

[12] J. He, M. Li, H. Zhang, H. Tong, and C. Zhang. Manifold-ranking based image retrieval. In *ACM Multimedia*, pages 9–16, 2004.

[13] S. C. H. Hoi, R. Jin, and M. R. Lyu. Learning nonparametric kernel matrices from pairwise constraints. In *Proceedings of the International Conference on Machine Learning*, pages 361–368, 2007.

[14] S. C. H. Hoi and M. R. Lyu. A novel log-based relevance feedback technique in content-based image retrieval. In *Proceedings of ACM International Conference on Multimedia (MM2004)*, New York, US, Oct. 10–16 2004.

[15] S. C. H. Hoi and M. R. Lyu. Web image learning for searching semantic concepts in image databases. In *Proc. 13th International World Wide Web conference (WWW2004)*, New York, May 17–22 2004.

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

[17] T. Joachims. Optimizing search engines using clickthrough data. In *KDD*, pages 133–142, 2002.

[18] B. Kulis, M. Sustik, and I. S. Dhillon. Learning low-rank kernel matrices. In *Proceedings of the International Conference on Machine Learning*, pages 505–512, 2006.

[19] M. Lades, J. C. Vorbrüggen, J. M. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. Computers*, 42(3):300–311, 1993.

[20] G. R. G. Lanckriet, N. Cristianini, P. L. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[21] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):1–19, 2006.

[22] W.-H. Lin, R. Jin, and A. Hauptmann. Web image retrieval re-ranking with relevance model. In *WI '03: Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence*, page 242, Washington, DC, USA, 2003.

[23] X. Liu, R. Ji, H. Yao, P. Xu, X. Sun, and T. Liu. Cross-media manifold learning for image retrieval & annotation. In *Proc. 1st ACM International Conference on Multimedia Information Retrieval*, pages 141–148, 2008.

[24] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[25] R.A.Horn and C.A.Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[26] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool in interactive content-based image retrieval. *IEEE Trans. CSVT*, 8(5):644–655, Sept. 1998.

[27] V. Sindhwani, P. Niyogi, and M. Belkin. Beyond the point cloud: from transductive to semi-supervised learning. In *Proceedings of the International Conference on Machine Learning*, pages 824–831, 2005.

[28] 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. PAMI*, 22(12):1349–1380, 2000.

[29] L. Song, A. J. Smola, K. M. Borgwardt, and A. Gretton. Colored maximum variance unfolding. In *Advances in Neural Information Processing Systems*, 2007.

[30] C. Wang, L. Zhang, and H.-J. Zhang. Learning to reduce the semantic gap in web image retrieval and annotation. In *Proceedings of ACM Special Interest Group on Information Retrieval*, pages 355–362, 2008.

[31] Y. M. Wong, S. C. H. Hoi, and M. R. Lyu. An empirical study on large-scale content-based image retrieval. In *Proc. IEEE International Conference on Multimedia & Expo (ICME2007)*, Beijing, P.R. China, July 2–5 2007.

[32] L. Wu, S. C. H. Hoi, J. Zhu, R. Jin, and N. Yu. Distance metric learning from uncertain side information with application to automated photo tagging. In *Proceedings of ACM International Conference on Multimedia (MM2009)*, Beijing, China, Oct. 19–24 2009.

[33] J. Zhu, S. C. H. Hoi, M. R. Lyu, and S. Yan. Near-duplicate keyframe retrieval by nonrigid image matching. In *ACM Multimedia*, pages 41–50, 2008.

[34] J. Zhuang, I. W. Tsang, and S. C. H. Hoi. Simplenpkl: simple non-parametric kernel learning. In *Proceedings of the International Conference on Machine Learning*, page 160, Montreal, Canada, 2009.