Signal Processing 93 (2013) 1426–1434

Contents lists available at SciVerse ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

A comprehensive study on learning to rank for content-based image retrieval

Yangxi Li*, Chao Zhou, Bo Geng, Chao Xu, Hong Liu

Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing 100871, PR China

ARTICLE INFO

Article history:
Received 25 January 2012
Received in revised form
5 June 2012
Accepted 9 June 2012
Available online 17 June 2012

Keywords:
Learning to rank
Content-based image retrieval

ABSTRACT

Recently, various learning to rank approaches have been proposed in the information retrieval realm, with their promising performance in general document and web page retrieval applications. Based on these achievements, in this paper, we investigate and discuss whether learning to rank approaches can be adapted to content-based image retrieval (CBIR). Given the complex structure of image representation, it is also challenging how to design visual features for learning to rank algorithms that not only scale up well, but also model various visual modalities and the spatial distributions of local features. We answer this question by introducing some scalable visual-based ranking features for learning to rank. Specifically, we firstly adopt several well performed ad hoc ranking models to generate the bag-of-visual-words-based ranking features. Besides, images are divided into different salient regions and spatial blocks, respectively, and ranking features are extracted from each region and block. Finally, image global features-based similarities are also concatenated with the existing ranking features. Extensive experiments with three state-of-the-art learning to rank algorithms are performed over four popular image retrieval databases, together with some insightful conclusions to facilitate the adaptation of learning to rank approaches to CBIR.

1. Introduction

In recent years, various learning to rank algorithms have been proposed in information retrieval field and achieved promising performance in document and web page retrieval applications. The objective of this work is to present an extensive study of learning to rank approaches to the content-based image retrieval (CBIR) problem [1–3], i.e., searching the images that are perceptually similar to the query one based on the visual information. Inspired by the success in text-based document retrieval techniques, where an inverted indexing structure is built to obtain a high-efficiency searching, bag-of-visual-words (BOV) representation is designed to represent an image that is analogous to the text-based document structure and widely used in CBIR applications [3–8]. Therefore, various text-based document information retrieval techniques [9], such as vector space model, language model and Okapi-BM25 can be readily available to CBIR.

There are two key issues which affect the performance of BOV-based CBIR applications. The first is how to select a robust ranking model from so many mature text information retrieval techniques for CBIR [10,11]. Certain ranking models may be more effective for some image databases, and less effective for others. There is no single ranking model that is consistently superior to the others. Besides, since many kinds of visual features are designed and could be utilized to boost the performance, visual features selection is another key issue in CBIR. It is

* Corresponding author.
E-mail addresses: liyangxi@gmail.com (Y. Li), chaozhou88@gmail.com (C. Zhou), bogeng1985@gmail.com (B. Geng), xuchao@cis.pku.edu.cn (C. Xu), liuhong@pku.edu.cn (H. Liu).

0165-1684/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.sigpro.2012.06.012

desirable to conduct image retrieval with different features under different ranking configurations and combine their results together to obtain an optimal ranking list. However, since these retrieval approaches provide heterogeneous types of outputs, it is not easy to conduct such combination directly.

Learning to rank approaches are suitable to address this problem and can effectively combine the results from different features and ranking models [12]. Firstly, a set of queries and their associated images, including the relevance degrees are provided to the learning algorithm. These data are used to train a ranking function that could be used in the prediction. When a new query image is given, one database image's relevance degree to the query image is computed with its ranking feature, as well as the learned ranking function.

In this paper, by reviewing the progress of learning to rank approaches in text information retrieval, we investigate the ranking abilities of existing learning to rank algorithms in CBIR applications. Based on the existing ranking features in text information retrieval, we carefully design scalable ranking features for images from four different perspectives. The first set features are *BOV-based ranking features* which are ranking scores derived from several ad hoc ranking models. Since ranking features in text information retrieval are always derived from different fields of documents, such as title, abstract, and body, we divide images into several regions according to their salient degrees to generate image "fields", and BOV-based ranking features are extracted from each image "field". Besides, inspired by works in spatial pyramid matching [13], we design ranking features with a *spatial pyramid* manner to best preserves the image's spatial information, meanwhile brings little computational cost to the system. Specifically, we split the image into blocks with different sizes from coarse to fine. For each block, BOV-based ranking features are computed and all the blocks' features are concatenated. Finally, image similarities of global features are adapted as a supplement to the BOV-based features. To keep the system's scalability, the global features are converted into binary codes with LSH [14] for efficient similarity computation. Global features are also embedded into low dimensional features with linear multiview embedding (LME) [15] to compute the visual PageRank [16] of database images. With these ranking features, extensive experiments are performed with three state-of-the-art learning to rank approaches on four different image retrieval datasets, where the results provide some useful conclusions for further applications. In addition, the effectiveness of feature selection and feature dimension reduction of visual ranking features are also evaluated.

The rest of the paper is organized as follows. In Section 2, we discuss the learning to rank problems and introduce three representative algorithms evaluated in our experiments. In Section 3, we elaborate the designed visual ranking features in detail. A comprehensive evaluation of learning to rank algorithms for CBIR is conducted in Section 4. Section 5 finishes the paper with a conclusion.

2. Learning to rank problems

In recent years, more and more machine learning technologies have been used to train the ranking model, and a new research area named "learning to rank" has gradually emerged [17]. In general, most of the state-of-the-art learning to rank algorithms learn the optimal way of combining features extracted from document and query through discriminative training. According to the hypotheses and data used in model learning, [17] classifies the learning to rank algorithms into three categories: the pointwise approach, the pairwise approach, and the listwise approach.

2.1. The pointwise approach

The pointwise approach utilizes existing machine learning technologies, such as regression, into ranking problem and predict each document's relevance degree directly, although this may not be correct when the target is to generate a ranking list of a set of documents.

Perceptron-based Ranking (PRank) is one famous algorithm in pointwise approach [18]. The goal of PRank is to find a direction defined by a parameter vector $\mathbf{w}$, after projecting the documents onto which one can easily use thresholds to distinguish the documents into different relevant degrees. This goal is achieved by an iterative learning process. On iteration t, the learning algorithm gets an instance $\mathbf{x}_j$ associated with query q and predicts $\hat{y}_j = \arg\min_k\{\mathbf{w}^T\mathbf{x}_j - b_k < 0\}$. Given the ground truth label y_j, if the algorithm makes a mistake by predicting the category of $\mathbf{x}_j$ as $\hat{y}_j$ instead of y_j, then there is at least one threshold, indexed by k, for which the value of $\mathbf{w}^T\mathbf{x}_j$ is on the wrong side of b_k. To correct the mistake, PRank moves the value of $\mathbf{w}^T\mathbf{x}_j$ and b_k toward each other.

Since the input data in the pointwise algorithm is a single document, the relative order between documents cannot be naturally considered in the learning process. Given this problem, the pointwise approach can always be a sub-optimal solution to ranking.

2.2. The pairwise approach

The pairwise approach does not focus on accurately predicting the relevance degree of each document, it cares about the relative order between two documents. In these algorithms, the ranking problem is reduced to a classification problem on document pairs. That is to say, the goal of learning is to minimize the number of miss-classified pairs. One of the most famous pairwise approach is RankSVM [19].

For a set of training documents as $\{\mathbf{x}_i\}_{i=1}^m \in \mathcal{R}^n$, assume that the preference relation that $\mathbf{x}_i$ is preferable to $\mathbf{x}_j$ is denoted by $\mathbf{x}_i \succ \mathbf{x}_j$, the goal of ranking learning is to induce a ranking function $f : \mathcal{R}^n \to \mathcal{R}$ under a set of constraints

$$\forall \mathbf{x}_i \succ \mathbf{x}_j : f(\mathbf{x}_i) > f(\mathbf{x}_j). \tag{1}$$

The value of $f(\mathbf{x}_i)$ is known as the ranking score of $\mathbf{x}_i$. We assume f to be a linear function. In that case, we have $f(\mathbf{x}_i) = \mathbf{w}^T\mathbf{x}_i$. Just like in classification SVM, an approximate solution can be obtained by introducing a nonnegative slack variables ξ_{ij} and minimizing the upper bound of the ranking loss $\sum \xi_{ij}$. Formally, the optimization problem

of RankSVM is defined as follows

$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}+\gamma\sum_{i,j}\xi_{ij},$$
$$\text{s.t.} \quad \forall \mathbf{x}_i \succ \mathbf{x}_j : \mathbf{w}^T\mathbf{x}_i \geq \mathbf{w}^T\mathbf{x}_j+1-\xi_{ij},$$
$$\forall i,j : \xi_{ij} \geq 0, \tag{2}$$

where γ is a parameter that allows trading-off margin size against training error. We could rearrange the constraints in Eq. (2) as

$$\forall \mathbf{x}_i \succ \mathbf{x}_j : \mathbf{w}^T(\mathbf{x}_i-\mathbf{x}_j) \geq 1-\xi_{ij}, \tag{3}$$

the optimization problem becomes equivalent to a classification SVM on the feature differences of pairwise examples. Therefore, it can be solved using algorithms similar to those used for SVM classification. Assume that $\mathbf{w}^*$ is the solution that optimizes Eq. (2). For a set of new documents, we can obtain the ranking score for each by

$$f(\mathbf{x}) = \mathbf{w}^{*T}\mathbf{x}, \tag{4}$$

and then rank them according to their scores.

2.3. The listwise approach

In the listwise approach, the input space contains the entire group of documents associated with query q, e.g., $\mathbf{x}=\{\mathbf{x}_j\}_{j=1}^m$, and the output space contains the ranked list of the documents. Thus, the loss function for this category methods measures the inconsistency between the ranking list derived from the ranking model and the ground truth ranking list π_y.

ListNet [20] is one representative algorithm of this category. In ListNet, it first defines the permutation (i.e., ranking list) probability distribution based on the scores of ranking function f with Luce model [21]. Given the ranking scores of the documents outputted by scoring function f, i.e., $\mathbf{s}=\{s_j\}_{j=1}^m$, where $s_j=f(\mathbf{x}_j)$, the Luce model defines a probability for each possible permutation π of the documents based on the chain rule as follows:

$$P(\pi|\mathbf{s}) = \prod_{j=1}^{m}\frac{\varphi(s_{\pi^{-1}(j)})}{\sum_{u=j}^{m}\varphi(s_{\pi^{-1}(u)})}, \tag{5}$$

where $\pi^{-1}(j)$ denotes the document ranked at the jth position of permutation π, and φ is a transformation function. With the Luce model, ListNet also defines another permutation probability distribution $P_y(\pi)$ based on the ground truth label. Then the ListNet ranking loss is defined as the K–L divergence between these two distributions

$$L(f;\mathbf{x},\pi_y) = D(P(\pi|\varphi(f(w,\mathbf{x})))||P_y(\pi)). \tag{6}$$

A neural network model is employed in ListNet, and the gradient descent approach is used to minimize the K–L divergence loss.

2.4. Training complexity

For pointwise approach, the training complexity is roughly proportional to the number of documents. For example, for n documents with k labels, training complexity of each iteration in PRank is $O(nk)$. For pairwise approach, the training complexity is proportional to the number of pairs, i.e., $O(n^2)$. For listwise approach, such as ListNet, the training complexity is in the exponential order of n, because the complexity of evaluation of the K–L divergence loss for each query is $O(n!)$, which is intractable for practical applications. To resolve this problem, a top-k version of the K–L divergence loss is introduced in ListNet [20], which reduces the training complexity from n-factorial to the polynomial order of n.

Intuitively, pairwise approach and listwise approach model the ranking problem in a more natural way than the pointwise approach, and thus can address some problems that pointwise approach has encountered.

3. Visual ranking feature construction

To rank images for CBIR with learning to rank approaches, one crucial issue is the construction of feature vector $\mathbf{x}$. In this paper, we generate the image ranking features from four perspectives: ranking scores from ad hoc ranking models (*BOV-based features*); salient region-based ranking features which are designed to mimic the features extracted from different fields in document (*salient region-based features*); *spatial pyramid features* extracted from blocks of images to best capture the spatial information; and image similarities from *global features* which try to capture complementary properties among images.

3.1. BOV-based features

Since images with BOV representation are analogous to documents, we directly transfer a set of ranking features from learning to document rank into image ranking features. The first part of 12-dimensional BOV-based features is five ranking scores from ad hoc ranking models including vector space model with cosine distance, Okapi-BM25 probabilistic model, and language models with three different smooth strategies. The second part is six statistical measurements of query image pairs, including the number of common words, sum of term frequency (TF)/inverse document frequency (IDF)/ TF $\times$ IDF, and maximum of TF/IDF. Word number of the image is also one dimension of BOV-based features. This set features are designed to mimic the most frequently used features in learning to document rank.

3.2. Salient region-based features

Document field information plays an important role in text information retrieval. Features from different fields, i.e., body, anchor, and title, have different contributions to the final ranking performance. In image of CBIR, there is not a clear definition of title or body, and other fields. Thus, we resort to saliency detection method to describe the importance degree of different image regions, and define image "fields" according to salient degrees. For example, the most salient region of one image may correspond to the title and abstract of a document, while those un-salient regions may correspond to document body.

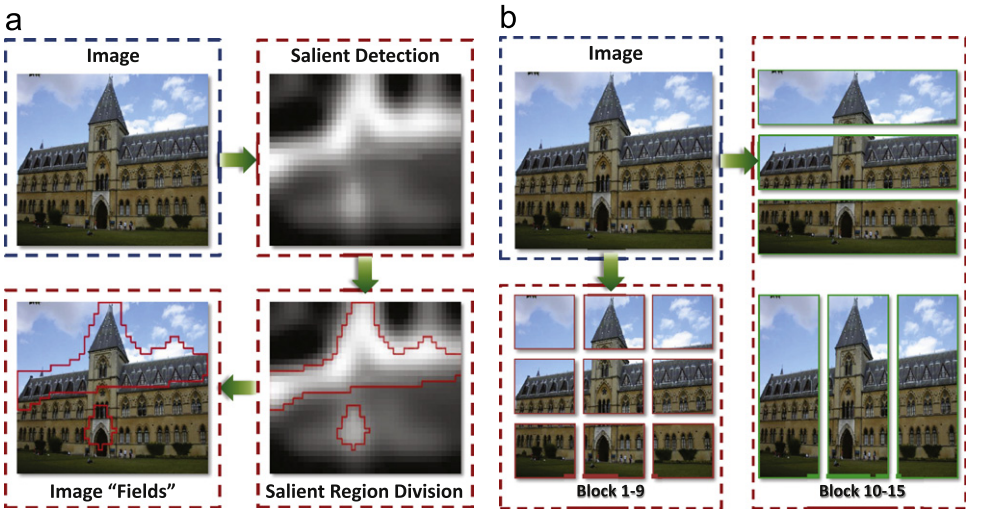

Fig. 1. Illustration of visual ranking feature's extraction. (a) Salient region-based features. (b) Spatial pyramid-based features.

Specifically, for each image, its saliency map is derived with a mature and efficient saliency detection method [22]. Empirically, we divide each image's visual words into two fields: visual words contained in the top 50% salient regions, and visual words contained in other unsalient regions. For each field, the 12-dimensional BOV-based features is calculated (as shown in Fig. 1(a)), the dimensionality of salient region-based features is 2 fields $\times$ 12 = 24.

3.3. Spatial pyramid features

Despite its simplicity and efficiency, BOV representation discards too much spatial structure information of images. Previous works in CBIR prove that spatial information is crucial for image retrieval and brings much performance improvement [3,5]. Inspired by Lazebnik's work spatial pyramid matching [13], we propose to extract ranking features with a spatial pyramid manner to best preserve image's spatial information, meanwhile brings little computational burden to the system.

In this paper, each image is divided in two layers. The whole image is evenly split into three rows and three columns in the first layer. In the second layer the image is divided into 3×3 grids. Thus, each image is converted into 15 blocks (three rows, three columns and nine grids, as shown in Fig. 1(b)). For each block we, respectively, calculate the 12-dimensional BOV-based features described above, the dimensionality of spatial pyramid features equals to 15 blocks $\times$ 12 = 180.

3.4. Global features-based ranking features

Besides mining the saliency and spatial information of images, we also construct ranking features with the help of image's global features which provide complementary information of images global appearance [23–26].

To improve the image matching efficiency of the high-dimensional global features, locality-sensitive hashing (LSH) [14] is employed in this paper to approximate the near neighbor search in the high dimensional space. The LSH algorithm is summarized as follows:

1. Given a group of hashing functions, $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$, where the atomic hashing function $f_i(v)$ accepts a D dimensional feature v. The hashing function of LSH is designed to approximate the cosine distance between vectors. The basic idea of this technique is to choose a random hyperplane (defined by a normal unit vector r) and use the hyperplane to hash input vector v. For example, given an input vector v and a hyperplane defined by r, the hashing function could be $f(v) = \mathrm{sgn}(r \cdot v)$. That is $f(v) = \pm 1$, depending on which side of the hyperplane v lies. Each possible choice of r defines a single hashing function.
2. Randomly choose B hashing functions from $\mathcal{F}$ to construct a hashing function set $\mathcal{H} = \{h_1, h_2, \ldots, h_B\}$, which projects the D dimensional input vector to a binary code with length of B.
3. Repeat step 2 L times to obtain L different hashing function sets $\mathcal{G} = \{\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_L\}$.

Given a query image Q, which is also represented as a D dimensional feature vector, LSH maps Q to B dimensional space L times by hashing function sets obtained in step 3. Then K nearest neighbors of Q are found by searching the L hashing tables and comparing the hamming distance between Q and the feature vectors that have the same hashing values of Q. In our case, we set K=200 and assign these 200 nearest neighbors a normalized score from 1 to 0 according to their similarity.

Besides computing query image similarities, we also compute visual PageRank [16] for database images with global features. PageRank measures dataset images' own importance without considering queries. Suppose S^* is the column normalized, symmetrical adjacency matrix where $S_{u,v}$ measures the visual similarity between image u and v, the visual PageRank v PR of database images is iteratively defined as $v\mathrm{PR} = S^* \times v\mathrm{PR}$, which is similar to

web PageRank [27]. A damping factor α is added for easy of computation. Formally, one image's visual PageRank is defined as follows:

$$\nu\mathrm{PR} = \alpha S^* \times \nu\mathrm{PR} + (1-\alpha)\left[\frac{1}{n}\right]_{n\times 1}, \quad (7)$$

we set $\alpha = 0.85$ following [16]. Two kinds of similarities are calculated in visual PageRank. The first similarity is computed with Euclidean distance of image's low dimensional embedding, which is derived with a multiple feature embedding algorithm LME [15] from images' multiple global features. Another similarity is computed as the histogram intersection of image's BOV representation.

Table 1 lists all the components of our proposed ranking features for learning to rank image scheme.

3.5. Dimension reduction and feature selection

Feature dimension reduction tries to reduce the size of features by transforming or combining the original features. It has been shown very effective in many applications [28–30], such as face detection, but little work has been done in ranking.

Differently, feature selection [31] is often used in learning to rank to reduce the feature size, which chooses a part of features from raw features, and discards others. In feature selection, two properties of each feature are defined firstly, i.e., *importance* and *similarity*. In the computation of feature importance, each dimension of feature is used to rank database images and the ranking performance such mAP or NDCG is taken as this feature's importance. In the computation of feature similarity between two features, they are firstly used to rank database images individually, the correlation of these two ranking list is then taken as the feature similarity. With these two properties, the loss function of feature selection is defined as to select the most important (high importance) and representative (low similarity) features. A greedy algorithm is utilized in [31] to select features one by one from raw features.

Feature selection has been proved useful for ranking problem. We study the effectiveness of both dimension reduction algorithm (PCA [32] in this paper) and feature selection for image ranking applications in experiments section.

4. Experiments

We perform extensive experiments on four image retrieval datasets with three ranking algorithms to give a comprehensive analysis of learning to rank approaches for CBIR.

4.1. Datasets and experimental setup

Oxford5K. The *Oxford5K* dataset consists of overall 5062 high resolution images, which are taken at 11 different Oxford landmarks together with some distracters. Five queries are, respectively, selected for each landmark.

Oxford505K. We also build an *Oxford505K* dataset which is composed of *Oxford5K* and 500 K randomly selected unlabeled images from *ImageNet* dataset [33], to test the scalability of ranking models. These 500 K images are assumed not to contain images of the ground truth of *Oxford5K* and only used as "distracters" for the experiments.

Paris. The *Paris* dataset [34] contains 6390 high resolution images of Paris landmarks. It also contains 55 queries.

NUS-WIDE. We use NUS-WIDE-OBJECT dataset [35] which contains 31 object categories and 30,000 images in total. We randomly select 50 images from all 31 categories as the queries for image retrieval task.

Experimental setup. All the datasets are evenly split into five folds. We choose three folds for learning the ranking model, one to validate the parameters, and the rest one to evaluate the performance. We switch the fold configuration

Table 1
The components of proposed ranking features.

ID	Feature description	Category
1	Vector space model score	BOV-based feature
2	Okapi-BM25 score	
3	Language model with JM smooth	
4	Language model with DIR smooth	
5	Language model with ABS smooth	
6	The number of common words	
7	Sum of TF	
8	Sum of IDF	
9	Sum of TF × IDF	
10	Max of TF	
11	Max of IDF	
12	$\|D\|$: visual words number	
13–36	For both salient and un-salient regions, the BOV features are extracted, respectively. Dimensionality: 2 × 12=24	Salient based feature
37–216	For each 15 blocks, the 12 BOV features are extracted, respectively. Dimensionality: 15 × 12=180	Spatial pyramid feature
217–222	Global features similarity. Dimensionality: 6	Global feature
223–224	Visual PageRank features. Dimensionality: 2	Visual PageRank

in a round-robin fashion so that each fold is ensured to be taken as test and validation set for once, respectively, resulting into learning under five different configurations. The average performance over five configurations is reported. For the first three datasets, the image's relevance level are labeled as "Good" (perfect match), "Ok" (good match), "Junk" (partial match), and "Absent" (total irrelevance). For NUS-WIDE dataset, images are labeled as "Relevant" and "Irrelevant".

Global features. For the first three datasets, six kinds of global features are utilized, which are 225-D block-wise color moments, 75-D edge direction histogram, 128-D wavelet texture, 144-D Color Correlogram, 64-D Color Histogram, and 554-D biologically inspired feature [36]. For NUS-WIDE dataset, only the first five kinds of global features provided by their website are utilized.

Visual word generation. In the visual word generation, all the images' SIFT [37] features are abstracted firstly. These features are then clustered into K classes with clustering algorithm such as K-means or approximate K-means. Afterwards, a unique number is assigned to each clustering centroid, and these numbers are taken as visual word. All the clustering centroid-visual word pairs are recorded into dictionary, which is also known as codebook. After deriving dictionary, each SIFT features is replaced with the visual word of the nearest clustering centroid in the dictionary, then each image is represented as a set of visual words. In our experiments, the dictionary size of two Oxford related datasets are 1 M, the dictionary size of Paris dataset is 500 K, and 500 for NUS-WIDE dataset.

Evaluation metric. The performance evaluations of all the experiments are based on two measurements, mean average precision (mAP) and normalized discounted cumulative gain at different rank truncation levels (NDCG@n). mAP is one of the most frequently used measurements to evaluate the average performance of a ranking algorithm. mAP denotes the mean of the average precision (AP), where the AP computes the area under precision/recall curve with noninterpolated manner and prefers relevant samples with higher rank. Since AP is evaluated only for binary judgment, we define relevance level 0 as irrelevant and all the other relevance degrees as relevant for all the datasets. To measure the ranking performance for multiple degree relevance, NDCG [38] is proposed as a cumulative, multilevel measurement of ranking quality, which is usually truncated at a particular rank level. For a given query q_i, the NDCG is calculated as

$$\mathcal{N}_i = N_i \sum_{j=1}^{L} \frac{2^{r(j)}-1}{\log(1+j)}, \tag{8}$$

where $r(j)$ is the relevance degree of the jth document, N_i is the normalization coefficient to make the perfect order list with $\mathcal{N}_i = 1$, and L is the ranking truncation level at which NDCG is computed. In this paper, we evaluate NDCG@n by setting the truncation level n at 10, 50 and 100.

4.2. Experimental results and analysis

Retrieval performance on four datasets is listed in Figs. 2–4 and Table 2. For the first three datasets, performance of proposed features is compared with the best performed single model (VSM), BOV features (BOV), BOV features with saliency detection (BOV+Sal), BOV features with saliency detection as well as spatial pyramid (BOV+Sal+SP). NDCG with different truncation levels and mAP are reported. For the NUS-WIDE dataset, the proposed features are compared with the best performed model (VSM) and BOV features (BOV).

Ranking model comparison. On all the datasets, both pairwise method RankSVM and listwise method ListNet can steadily given a promising performance for CBIR and significantly improve the ranking performance over VSM ($>10\%$ over all the metrics on average), while pointwise method PRank outperforms VSM only on *Oxford505K* dataset. This is because both pairwise and listwise methods are designed particularly for ranking problem. They are based on reasonable and suitable hypotheses and loss functions compared to pointwise method. Besides, the

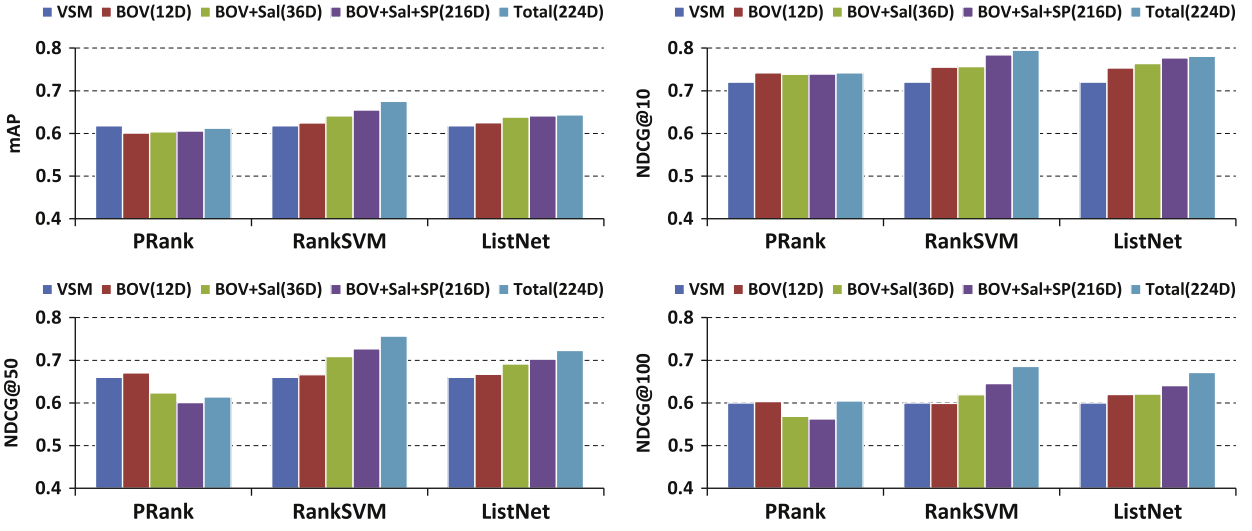

Fig. 2. Performance comparison on *Oxford5K* dataset.

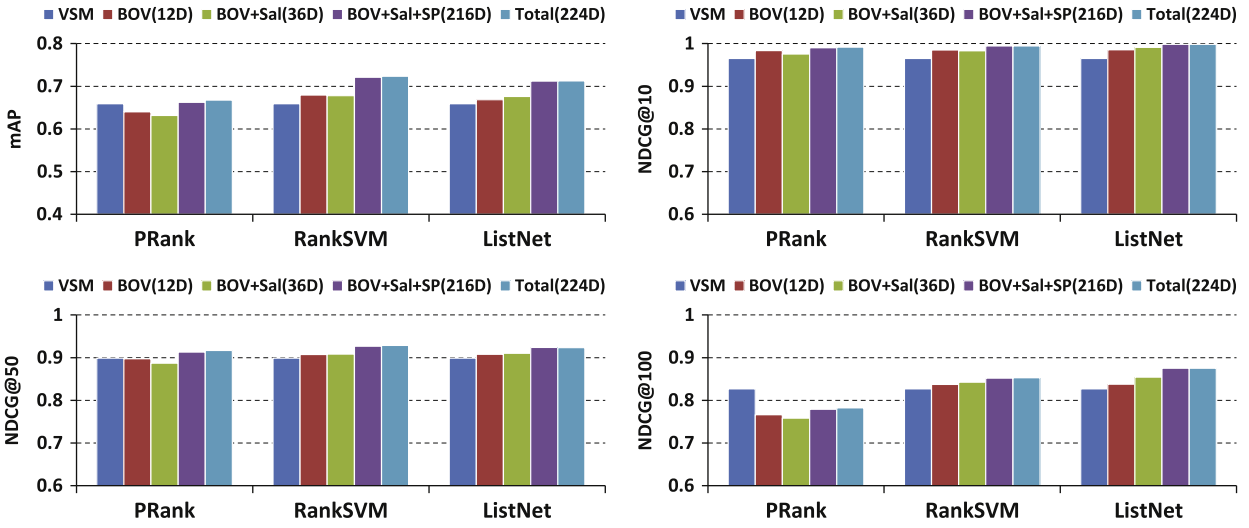

Fig. 3. Performance comparison on *Paris* dataset.

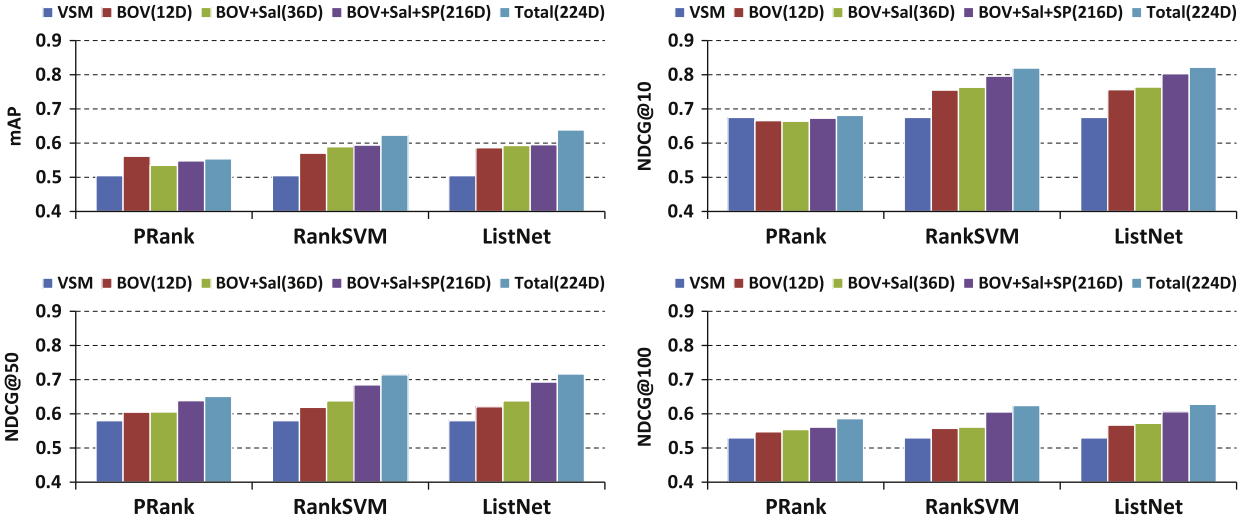

Fig. 4. Performance comparison on *Oxford505K* dataset.

Table 2
Performance comparison on *NUS-WIDE* dataset, mAP reported.

	PRank	RankSVM	ListNet
VSM	0.161	0.161	0.161
BOV (12D)	0.147	0.173	0.173
Total (19D)	0.147	0.182	0.182

performance of RankSVM and ListNet is comparable over each dataset, although ListNet is proved superior to RankSVM in text retrieval. Thus, we conclude that both pairwise and listwise methods are suitable for CBIR applications and could obtain steadily performance improvement over single ranking model, such as VSM.

Visual ranking feature analysis. Experimental results on all the datasets demonstrate that each proposed feature component has some contribution to the final performance improvement, and the concatenated features obtain best performance under all the metrics. These results not only demonstrate the correctness of proposed visual ranking features, but also provide useful clues for further research in feature design for learning to image rank.

4.3. Feature selection and dimension reduction results and analysis

The performance of feature selection (FS) and dimension reduction (DR) at different dimensions over the first three datasets is listed in Tables 3–5. Due to the limitation of page length, only mAP is reported. To facilitate the comparison, we also list the performance of raw features. On all the three datasets, best FS results outperform raw features more than 5% on average. Compared to the best performed ranking model VSM, pairwise and listwise methods with FS obtain over 17% performance gain,

Table 3
Performance comparison on *Oxford5K* dataset, mAP reported.

# dimension	12	20	36	50	100	150	224
FS							
PRank	0.628	0.630	0.639	0.647	**0.648**	0.643	*0.612*
RankSVM	0.672	0.673	0.682	**0.695**	0.681	0.673	*0.675*
ListNet	0.643	0.643	0.652	0.664	**0.686**	0.660	*0.643*
DR							
PRank	0.589	0.602	0.608	**0.616**	0.607	0.615	*0.612*
RankSVM	0.624	0.630	0.635	0.634	**0.636**	0.634	*0.675*
ListNet	0.642	0.633	0.633	0.635	**0.638**	0.638	*0.643*

Table 4
Performance comparison on *Paris* dataset, mAP reported.

# dimension	12	20	36	50	100	150	224
FS							
PRank	**0.709**	0.705	0.702	0.702	0.678	0.670	*0.667*
RankSVM	0.703	0.708	0.722	0.722	0.723	**0.735**	*0.723*
ListNet	0.703	**0.725**	0.700	0.701	0.674	0.667	*0.713*
DR							
PRank	0.623	0.625	0.637	**0.649**	0.645	0.621	*0.667*
RankSVM	0.678	0.677	0.688	0.692	0.704	**0.715**	*0.723*
ListNet	0.642	0.645	0.655	0.654	0.660	**0.659**	*0.713*

Table 5
Performance comparison on *Oxford505K* dataset, mAP reported.

# dimension	12	20	36	50	100	150	224
FS							
PRank	0.522	0.530	0.549	0.557	**0.558**	0.553	*0.554*
RankSVM	0.620	0.623	0.634	0.635	**0.641**	0.633	*0.623*
ListNet	0.623	0.623	0.632	**0.646**	0.636	0.631	*0.638*
DR							
PRank	0.519	0.522	0.528	0.531	**0.537**	0.533	*0.554*
RankSVM	0.611	0.610	0.615	0.614	**0.617**	0.614	*0.623*
ListNet	0.612	0.612	0.621	0.621	**0.625**	0.622	*0.638*

which is quite significant. In most experiments, FS features with 100-D obtain comparable performance with raw 224-D features, which indicates the effectiveness of feature selection for feature redundancy remove. However, DR results underperform the raw features in most cases. This is because in FS the computation of feature's importance and similarity are performance related, i.e., the computation needs ground truth information. Thus, FS could be seen as a supervised dimension reduction method, which guarantees that the bad performed features are eliminated from raw features. Conversely, PCA conducts DR only from the feature's distribution in feature space, which is separated from the further learning to rank jobs. To sum up, feature selection is a helpful technology for learning to image rank.

5. Conclusions

In this paper, by reviewing the progress of learning to rank approaches in text information retrieval, we investigate whether existing learning to rank algorithms can be adapted to CBIR applications. Based on the existing ranking features in text IR, we carefully design scalable ranking features for learning to rank images from four different perspectives. With designed ranking features, extensive experiments are performed with three state-of-the-art learning to rank approaches on four different image retrieval datasets. Based on the results, we conclude that: (1) pairwise and listwise learning to rank methods are suitable for CBIR applications, both of which obtains more than 10% performance gain over single ranking model and 17% gain with the help of feature selection; (2) the proposed visual ranking features are both effective and efficient for learning to rank applications; and (3) feature selection is a helpful tool for further performance improvement in CBIR applications.

Acknowledgments

This Paper is supported by Chinese 973 Program (2011CB302400), NSFC (60975014) and NSFB (4102024).

References

[1] D. Li, J. Peng, Z. Li, Q. Bu, LSA-based multi-instance learning algorithm for image retrieval, Signal Processing 91 (8) (2011) 1993–2000.
[2] J. Zhang, L. Ye, Local aggregation function learning based on support vector machines, Signal Processing 89 (11) (2009) 2291–2295.
[3] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Object retrieval with large vocabularies and fast spatial matching, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007 (CVPR'07), IEEE, 2007, pp. 1–8.
[4] J. Yu, D. Liu, D. Tao, H. Seah, Complex object correspondence construction in two-dimensional animation, IEEE Transactions on Image Processing 11 (2012) 3257–3269.
[5] Z. Wu, Q. Ke, M. Isard, J. Sun, Bundling features for large scale partial-duplicate web image search, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009 (CVPR 2009), IEEE, 2009, pp. 25–32.
[6] M. Subrahmanyam, R. Maheshwari, R. Balasubramanian, Local maximum edge binary patterns: a new descriptor for image retrieval and object tracking, Signal Processing 92 (6) (2012) 1467–1479.
[7] X. Tian, D. Tao, X. Hua, X. Wu, Active reranking for web image search, IEEE Transactions on Image Processing 19 (3) (2010) 805–820.
[8] X. Tian, D. Tao, Y. Rui, Sparse transfer learning for interactive video search reranking, Arxiv preprint arxiv:1103.2756.
[9] C. Manning, P. Raghavan, H. Schutze, Introduction to Information Retrieval, Cambridge University Press, Cambridge, 2008.
[10] B. Geng, L. Yang, C. Xu, A study of language model for image retrieval, in: IEEE International Conference on Data Mining Workshops, 2009 (ICDMW'09), IEEE, 2009, pp. 158–163.
[11] B. Geng, L. Yang, C. Xu, X. Hua, Ranking model adaptation for domain-specific search, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, ACM, 2009, pp. 197–206.
[12] B. Geng, L. Yang, C. Xu, X. Hua, S. Li, The role of attractiveness in web image search, in: Proceedings of the 19th ACM International Conference on Multimedia, ACM, 2011, pp. 63–72.
[13] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, vol. 2, IEEE, 2006, pp. 2169–2178.
[14] M. Datar, N. Immorlica, P. Indyk, V. Mirrokni, Locality-sensitive hashing scheme based on p-stable distributions, in: Proceedings of the 20th Annual Symposium on Computational Geometry, ACM, 2004, pp. 253–262.
[15] Y. Li, B. Geng, Z. Zha, D. Tao, L. Yang, C. Xu, Difficulty guided image retrieval using linear multiview embedding, in: Proceedings of the 19th ACM International Conference on Multimedia, ACM, 2011, pp. 1169–1172.

[16] Y. Jing, S. Baluja, Pagerank for product image search, in: Proceeding of the 17th International Conference on World Wide Web, ACM, 2008, pp. 307–316.
[17] T. Liu, Learning to rank for information retrieval, Foundations and Trends in Information Retrieval 3 (3) (2009) 225–331.
[18] K. Crammer, Y. Singer, Pranking with ranking, Advances in Neural Information Processing Systems 14 (2001) 641–647.
[19] T. Joachims, Optimizing search engines using clickthrough data, in: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2002, pp. 133–142.
[20] Z. Cao, T. Qin, T. Liu, M. Tsai, H. Li, Learning to rank: from pairwise approach to listwise approach, in: Proceedings of the 24th International Conference on Machine Learning, ACM, 2007, pp. 129–136.
[21] R. Luce, Individual Choice Behavior, John Wiley, 1959.
[22] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (11) (1998) 1254–1259.
[23] M. Song, D. Tao, C. Chen, J. Bu, J. Luo, C. Zhang, Probabilistic exposure fusion, IEEE Transactions on Image Processing 1 (2012) 341–357.
[24] M. Song, D. Tao, C. Chen, X. Li, C. Chen, Color to gray: visual cue preservation, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9) (2010) 1537–1552.
[25] X. Tian, D. Tao, Y. Rui, Sparse Transfer Learning for Interactive Video Search Reranking, TOMCCAP.
[26] D. Tao, NeNMF: an optimal gradient method for non-negative matrix factorization, IEEE Transactions on Signal Processing 99 (2012) 1.
[27] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Computer Networks and ISDN Systems 30 (1–7) (1998) 107–117.
[28] X. Wang, D. Tao, Z. Li, Subspaces indexing model on Grassmann manifold for image search, IEEE Transactions on Image Processing 99 (2011) 1.
[29] N. Guan, D. Tao, Z. Luo, B. Yuan, Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent, IEEE Transactions on Image Processing 20 (7) (2011) 2030–2048.
[30] N. Guan, D. Tao, Z. Luo, B. Yuan, Non-negative patch alignment framework, IEEE Transactions on Neural Networks 99 (2011) 1.
[31] X. Geng, T. Liu, T. Qin, H. Li, Feature selection for ranking, in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2007, pp. 407–414.
[32] H. Hotelling, Analysis of a complex of statistical variables into principal components, Journal of Educational Psychology 24 (6) (1933) 417.
[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-scale Hierarchical Image Database, in: IEEE International Conference on Computer Vision, IEEE Computer Society, 2009.
[34] 〈http://www.robots.ox.ac.uk/∼vgg/data/parisbuildings/index.html〉.
[35] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, NUS-WIDE: A Real-world Web Image Database from National University of Singapore, 2009, p. 48.
[36] D. Song, D. Tao, Biologically inspired feature manifold for scene classification, IEEE Transactions on Image Processing 19 (1) (2010) 174–184.
[37] D. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.
[38] K. Järvelin, J. Kekäläinen, IR evaluation methods for retrieving highly relevant documents, in: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2000, pp. 41–48.