Image Search Reranking With Hierarchical Topic Awareness

Xinmei Tian, *Member, IEEE*, Linjun Yang, *Member, IEEE*, Yijuan Lu, *Member, IEEE*,
Qi Tian, *Senior Member, IEEE*, and Dacheng Tao, *Fellow, IEEE*

***Abstract*—With much attention from both academia and industrial communities, visual search reranking has recently been proposed to refine image search results obtained from text-based image search engines. Most of the traditional reranking methods cannot capture both relevance and diversity of the search results at the same time. Or they ignore the hierarchical topic structure of search result. Each topic is treated equally and independently. However, in real applications, images returned for certain queries are naturally in hierarchical organization, rather than simple parallel relation. In this paper, a new reranking method "topic-aware reranking (TARerank)" is proposed. TARerank describes the hierarchical topic structure of search results in one model, and seamlessly captures both relevance and diversity of the image search results simultaneously. Through a structured learning framework, relevance and diversity are modeled in TARerank by a set of carefully designed features, and then the model is learned from human-labeled training samples. The learned model is expected to predict reranking results with high relevance and diversity for testing queries. To verify the effectiveness of the proposed method, we collect an image search dataset and conduct comparison experiments on it. The experimental results demonstrate that the proposed TARerank outperforms the existing relevance-based and diversified reranking methods.**

***Index Terms*—Image search reranking, relevance, topic coverage (TC), topic-aware reranking (TARerank).**

I. Introduction

Most of the frequently-used commercial Web image search engines, e.g., Bing, Google, and Yahoo!, are implemented by indexing and searching the textual information associated with images, such as image file names, surrounding texts, universal resource locator, and so on. Although text-based image search is effective for large-scale image collections, it suffers from the drawback that textual information cannot comprehensively and substantially describe the rich content of images. As a consequence, some irrelevant images are observed in the search results.

To tackle the difficulties in text-based image search, visual reranking has been proposed. It incorporates visual information of images to refine the text-based search results. Generally, text-based search is first applied to obtain a coarse result from a large text-indexed image database. Then the top returned images are reordered via various reranking approaches by mining their visual patterns. Many reranking methods have been proposed in recent years. According to their reranking objectives, the existing methods can be categorized into two classes, i.e., relevance-based reranking [1]–[7] and diversified reranking [8]–[11].

The objective of relevance-based reranking is to maximize the relevance of the returned image list through reordering. However, since maximizing the relevance of each item in the list is the only objective, the resulting ranking list tends to return a large number of redundant images that convey repetitive information. For example, duplicate, near duplicate, and visually similar images tend to appear in the top of the list. As discussed in [12], users usually prefer search results consisting of images that are not only highly relevant but also covering broad topics. Therefore, diversified reranking is proposed to allow the search results to convey more information by maximizing the topic coverage (TC).

Although the existing diversified reranking methods improve the diversity in some cases, they suffer from two challenges. First, although both relevance and diversity are considered, optimizations are performed in a two-step manner [9], [11], i.e., firstly conducting relevance-based reranking to maximize the relevance, and then enriching the TC by diversifying the relevance-based reranking result. The two-step optimization that maximizes the relevance and

Manuscript received December 11, 2012; revised February 19, 2014 and October 22, 2014; accepted October 26, 2014. Date of publication January 20, 2015; date of current version September 14, 2015. The work of X. Tian was supported in part by the National Science Foundation of China (NSFC) under Grant 61390514 and Grant 61201413, in part by the Fundamental Research Funds for the Central Universities under Grant WK2100060007, Grant WK2100060011, and Grant WK2100100021, and in part by the Specialized Research Fund for the Doctoral Program of Higher Education under Grant WJ2100060003. The work of Y. Lu was supported in part by the Texas State University Research Enhancement Program, in part by Army Research Office under Grant W911NF-12-1-0057, and in part by the National Science Foundation under Grant CRI 1305302. The work of Q. Tian was supported in part by Army Research Office (ARO) under Grant W911NF-12-1-0057, in part by the Faculty Research Awards by NEC Laboratories of America, and in part by the NSFC under Grant 61429201. The work of D. Tao was supported by the Australian Research Council Projects under Grant DP-140102164, Grant FT-130101457, and Grant LP-140100569. This paper was recommended by Associate Editor X. Jiang.

X. Tian is with the CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System, University of Science and Technology of China, Hefei 230027, China (e-mail: xinmei@ustc.edu.cn).

L. Yang is with Microsoft Corporation, Redmond, WA 98052-7329 USA (e-mail: linjuny@microsoft.com).

Y. Lu is with the Department of Computer Science, Texas State University, San Marcos, TX 78666 USA (e-mail: yl12@txstate.edu).

Q. Tian is with the Department of Computer Science, University of Texas at San Antonio, San Antonio, TX 78249 USA (e-mail: qitian@cs.utsa.edu).

D. Tao is with the Centre for Quantum Computation & Intelligent Systems and the Faculty of Engineering and Information Technology, University of Technology, Sydney, Ultimo, NSW 2007, Australia (e-mail: dacheng.tao@uts.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2014.2366740

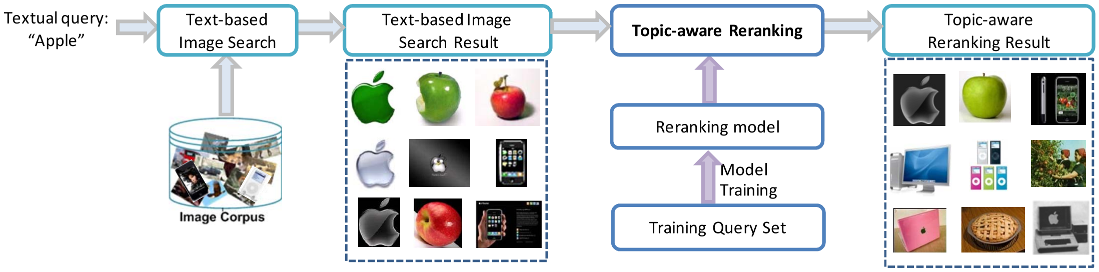


Fig. 1. Framework of the proposed topic-aware reranking (TARerank) method illustrated with the query "apple." When the textual query is submitted to a text-based image search engine, an initial search result is returned which may contain some irrelevant or duplicate images. Our proposed TARerank method reorders those images to obtain a more satisfactory result which consists of relevant and diverse images.

diversity separately can hardly achieve the joint optimum. Second, the diversified reranking usually models topic diversity through low-level visual features [9], which may not reflect users' perception on the semantic diversity due to the semantic gap. Although Song *et al.* [8] tried to use automatic annotation to bridge this gap, it is restricted by the scalability and accuracy of the automatic annotation in practical large-scale databases.

In addition, both relevance-based reranking and diversified reranking do not capture the hierarchical topic structure of search results very well. They usually treat topics equally and independently. However, different topics have different levels of importance. Generally, covering a more popular/important topic is preferred to covering a rare topic. Moreover, it can only deal with the simplest situation where all topics are independent to each other. In real applications, images returned for a certain query are naturally in hierarchical organization, rather than simple parallel relation. For example, the query apple includes two main categories, "fruit apple" and "products of Apple company." In the topic fruit apple, it further includes several sub-topics, e.g., apple trees, red apple, apple pie, etc.

To address the above problems, this paper proposes a new reranking method, termed TARerank. The framework of TARerank is presented in Fig. 1. When a textual query is submitted to a text-based image search engine, an initial search result is returned which may contain some irrelevant or duplicate images. Our proposed TARerank method reorders those images to obtain a more satisfactory result which consists of relevant and diverse images. TARerank can describe the hierarchical topic structure of search results in one model, and seamlessly captures both relevance and diversity in image search results simultaneously.

TARerank learns a reranking model from a training set by jointly optimizing relevance and diversity. A set of features is first extracted to describe the relevance and diversity properties of an arbitrary ranking result. Then, a reranking model is learned to capture the dependency between the low-level features and the semantic-level TC and relevance. Once the model is learned, we can use it to predict a reranking result which consists of highly relevant images covering broad topics for a new query. This method is built in the framework of structured learning and can be efficiently solved by using the cutting plane method.

In order to capture the hierarchical topic structure, a new criterion, called normalized cumulated topic coverage (NCTC), is also proposed. This measurement takes topic importance into consideration, and is well-suited for dealing with hierarchical topics. Since irrelevant images have no contribution to TC, NCTC also captures the relevance character.

In short, the main contributions introduced in this paper are summarized as follows.

1) Topic aware reranking is proposed as a learning-based reranking method. It directly learns a model from a training set by jointly optimizing relevance and diversity.
2) We propose a new criterion, NCTC, to seamlessly quantify relevance and diversity simultaneously. NCTC is a highly general measurement. It can handle the hierarchical TC and also take topic importance into consideration. The commonly used criterion topic recall (TRecall) [13] is a special case of NCTC.
3) To learn the TARerank model, we design a set of features to describe the relevance and diversity properties of a ranking result. By introducing a learning procedure, the gap between low-level visual feature diversity and high-level semantic topic diversity is bridged to some extent.

The rest of this paper is organized as follows. Firstly, we briefly review the related work in Section II and then present the proposed NCTC measurement in Section III. In Section IV, we introduce the proposed TARerank problem, as well as its learning and prediction. By analyzing the properties of most wanted diverse search results, a set of corresponding features is defined in Section V. The experimental results are presented and analyzed in Section VI, followed by the conclusion in Section VII.

II. Related Work

Image search plays an important role in our daily life. Considerable research efforts have been made to improve image search performance from various aspects, e.g., novel visual feature design [14]–[17], feature generation [18]–[21], semantic annotation [22]–[26], machine learning tools [27]–[30], and ranking and reranking algorithms [2], [9], [31]–[33]. Among them, visual reranking draws increasing attention since it leverages the advantages of both content-based [34] and text-based image retrieval. As aforementioned, existing reranking methods can be classified into two categories, i.e., relevance-based reranking and diversified reranking.

Relevance-based reranking focuses on improving the quality of search results from the relevant aspects, boosting the rank of relevant images. Most visual reranking work in earlier years belongs to this category. Yan *et al.* [5] proposed to rerank the image search results in classification way. It introduces the pseudo-relevance feedback assumption in document retrieval to obtain pseudo-positive and pseudo-negative training samples for relevance classifier training. Hsu *et al.* [3] modeled the reranking process as a random walk over a graph that is constructed by using images as the nodes and the edges between them being weighted by visual similarities. Jing and Baluja [2] applied the well-known PageRank algorithm to image search reranking by directly treating images as documents and their visual similarities as probabilistic hyperlinks. Tian *et al.* [4] proposed a general graph-based reranking framework and formulated visual reranking as an optimization problem from the Bayesian perspective. The problem in relevance-based reranking is that they mainly rely on visual consistency to perform reranking, therefore visually similar images are often ranked nearby. Near-duplicate images present less information to users, especially in response to queries that are ambiguous, such as apple. Many researchers have found that users are not very clear on what they want when performing such searches. Thus, a diverse result covering rich topics may meet the various needs of users more effectively and could help them reach their search targets more quickly.

Since search results with rich TC are preferred by users, various methods have been proposed to achieve the diversity objective at the reranking stage. In [10], a retrieval model is designed to return diversified image search results by utilizing the textual information associated with the images, i.e., tags, titles, and descriptions. In [8], TC relations between an image pair are measured via their associated words that are annotated automatically. By taking TC relations as probabilistic linkage between images, a method similar to PageRank is adopted to deduce the topic richness score for each image, and a diversified result is sequentially derived by choosing images which have high topic richness and cover new topics. Cao *et al.* [35] extended VisualRank [2] to cluster the images into several groups. In [9], the images are first clustered via clustering algorithms based on the maximal marginal relevance (MMR) rule and then the diverse result is formed by picking up one representative image from each cluster. Yang *et al.* [11] conducted a relevance-based reranking first to obtain the relevance score of each image, then sequentially selected images which were both relevant and different from images already selected.

Although promising improvements have been made, existing reranking methods have problems in optimizing relevance and diversity simultaneously. The separate two-step optimization of relevance and diversity can hardly achieve joint optimum [9], [11]. Besides, criterion which can measure relevance and diversity seamlessly is highly desired. To solve those problems, we propose a new reranking method and a new criterion to achieve the joint optimum.

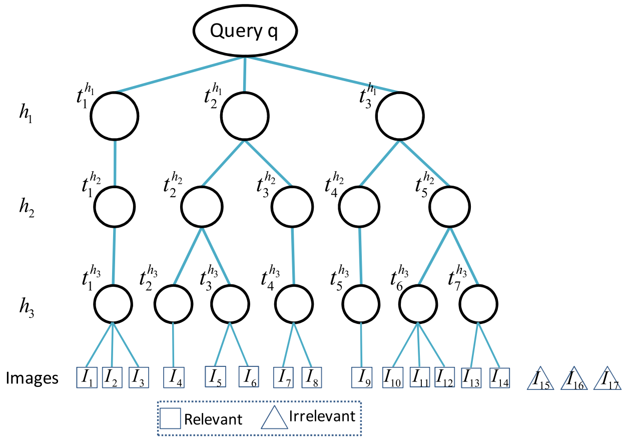


Fig. 2. Illustration of hierarchical topic structure for query q. The text-based image search engine returns 17 images for it, 14 relevant ones and three irrelevant ones. The 14 relevant images are organized into hierarchical topics.

III. NCTC

As discussed in Section I, the performance of a ranking result should be measured from two aspects, relevance and diversity. It is expected to use one criterion to measure both aspects at the same time and take topic importance into consideration. This paper proposes such a criterion called NCTC to capture the relevance, diversity, and topic structure. We will detail the proposed NCTC as follows.

A. TC

For a query q, suppose there are N images $\mathcal{I} = \{I_1, \ldots, I_N\}$ returned in the text-based search stage. A ranking vector $\mathbf{y} = [y_1, \ldots, y_N]^T$ is adopted to represent the ranks of these N images, where y_i denotes the rank of I_i. For example, if we have four images $\{I_1, I_2, I_3, I_4\}$, $\mathbf{y} = [3, 2, 1, 4]^T$ means the order of these four images are $<I_3, I_2, I_1, I_4>$.

For a ranking vector $\mathbf{y}$, we use TC@k to denote the TC of the top-k ranked images in it. In this paper, hierarchical topics are adopted to capture the real Web image data distribution. For each query, all relevant images are organized into different topics and subtopics, as shown in Fig. 2. Irrelevant images do not belong to any topic. The root node denotes the query itself. TC@k should consider the TC in each topic layer. Therefore, we can define TC@k as the weighted sum of TC tc_{h_i} in each layer h_i

$$\mathrm{TC}@k = \frac{1}{z}\sum_{i=1}^{N_h} wh_{h_i} * \mathrm{tc}_{h_i} \tag{1}$$

where N_h is the number of the topic layer. For example in Fig. 2, $N_h = 3$. The wh_{h_i} is the weighting for layer h_i and $z = \sum_{i=1}^{N_h} wh_{h_i}$ is a normalization constant. We use τ_{h_i} to denote the set of topics in layer h_i, for example, $\tau_{h_1} = \{t_1^{h_1}, t_2^{h_1}, t_3^{h_1}\}$ in Fig. 2. Then wh_{h_i} is defined as

$$wh_{h_i} = \frac{1}{\log_2(1 + |\tau_{h_i}|)} \tag{2}$$

which means larger TC in top layers is preferred.

tc_{h_i} measures to what degree the topics in layer h_i are covered. The most direct way for calculating tc_{h_i} is to define it as the ratio of covered topic numbers to the total topic numbers in h_i

$$\text{tc}_{h_i} = \frac{\sum\limits_{t \in \tau_{h_i}} \delta(t)}{|\tau_{h_i}|}. \tag{3}$$

$\delta(t)$ is a binary function to denote whether topic $t \in \tau_{h_i}$ is covered by the top-k images in ranking vector $\mathbf{y}$ or not, i.e., $\delta(t) = 1$ if t is covered and otherwise $\delta(t) = 0$.

A problem existing in (3) is that, it does not consider the importance of different topics. Therefore, we propose to use a topic importance weighted ratio to calculate tc_{h_i}

$$\text{tc}_{h_i} = \frac{\sum\limits_{t \in \tau_{h_i}} wt_t * \delta(t)}{\sum\limits_{t \in \tau_{h_i}} wt_t}. \tag{4}$$

wt_t is the weighting for topic t and is defined as

$$wt_t = \log_2(1 + n_t) \tag{5}$$

where n_t denotes the number of images belonging to topic t. Equation (5) means that covering a topic containing more images provides more information than covering a topic containing fewer images. However, in some applications rare topics might be more important than popular topics. In this case, we can adjust wt_t and assign larger weighting to rare topics.

TC is a general measurement which considers hierarchical topic coverage and topic importance. If we only consider TC in a certain topic layer and set equal wt_t for each topic, then TC degenerates to TRecall used in [13] and [35].

B. NCTC

The TC@k can accurately measure the TC of the top-k ranked images in $\mathbf{y}$. However, it does not differentiate the order of these top-k ranked images. For example, given two ranking vectors $\mathbf{y}_1 = [1, 2, 3, 4, 5, 6]^T$ and $\mathbf{y}_2 = [4, 3, 2, 1, 5, 6]^T$, their TC@4 are the same. To measure the overall quality of a ranking vector, we propose a single value measurement, NCTC. NCTC@k is defined as the weighted sum of TC@1 to TC@k

$$\text{NCTC}_{\mathbf{y}}@k = \frac{1}{z}\sum_{i=1}^{k}(1 - \rho_i)\text{TC}@i \tag{6}$$

where $\rho_i = (k - i)/k$ is the forgetting factor. A larger ρ_i is assigned to a smaller i since TC@i has already incorporated TC@1 to TC@$(i-1)$ to some extent. The normalization constant z is chosen to guarantee a perfect ranking vector's NCTC@$k = 1$.

C. Discussion

The proposed NCTC measures both the relevant and hierarchical TC of a ranking result. For a query q and the N images returned for it, the ideal ranking result should be the one which has the highest NCTC. To illustrate the advantage of NCTC measurement, we use the toy data in Fig. 2 as an example. There are 17 images returned in total, 14 relevant and three irrelevant. Supposing we can only return three images to users, which three should be selected? Here we discuss three different ranking results which are constructed via different criteria. Result 1: three images are selected by maximizing relevance, i.e., they are all relevant but may belong to duplicate topics, $< I_{10}$, I_{11}, $I_{12} >$. Result 2: three images are selected by maximizing the TC in layer h_3 without considering the hierarchical topic structure, $< I_1, I_4, I_6 >$. Result 3: three images are selected by maximizing NCTC, $< I_{10}, I_5, I_1 >$.

TABLE I
NUMBER OF TOPICS COVERED BY THREE RANKING RESULTS IN TOPIC LAYERS h_1, h_2, AND h_3, RESPECTIVELY

Top 3 ranked images	Number of topics covered		
	h_1	h_2	h_3
Result 1 $< I_{10}, I_{11}, I_{12} >$	1	1	1
Result 2 $< I_1, I_4, I_6 >$	2	2	3
Result 3 $< I_{10}, I_5, I_1 >$	3	3	3

Table I lists the number of topics covered by those three results in topic layers h_1, h_2, and h_3, respectively. The best ranking result should maximize the TC in different layers. Table I shows that Result 3, the ideal ranking result defined by NCTC, achieves the maximum TC in all topic layers. This highly diverse result efficiently shows more information about the query, thus it can satisfy different kinds of users with broad search interests and help them reach their search targets more quickly.

IV. TARerank

A. Problem Formulation

For a query q, the text-based image search engine returns a list of images by processing textual information. We denote the top-N ranked image set as $\mathcal{I} = \{I_1, \ldots, I_N\}$. A ranking vector $\bar{\mathbf{y}} = [\bar{y}_1, \ldots, \bar{y}_N]^T$ is adopted to represent the ranks of these images in text-based search results, where $\bar{y}_i$ denotes the rank of I_i in text search results. The aim of TARerank is to reorder the N images to obtain a new ranking vector $\mathbf{y} = [y_1, \ldots, y_N]^T$ in which the top-ranked images are not only relevant to the query but also cover broad topics.

In this paper, a supervised learning-based reranking method, called TARerank, is proposed. It directly learns a reranking model by optimizing the NCTC on a training set. The training set comprises m queries $\{q^{(i)}\}_{i=1}^{m}$. For each query q in the training set, we already know the relevance degree and hierarchical topic labels of all the images. Then an optimal ranking vector $\mathbf{y}^*$ can be derived via straightforward greedy selection by maximizing criterion NCTC($\mathbf{y}$), or minimizing a loss $\Delta(\mathbf{y})$ equivalently. Here we define $\Delta(\mathbf{y})$ as

$$\Delta(\mathbf{y}) = 1 - \text{NCTC}_{\mathbf{y}}@k. \tag{7}$$

Minimizing $\Delta(\mathbf{y})$ ensures high relevance and high TC in the top-k ranked images in $\mathbf{y}$.

Algorithm 1 Greedy Selection For $\mathbf{y}^* = \arg\min_{\mathbf{y}\in\mathcal{Y}} \Delta\mathbf{y}$

```
Input: $\mathcal{I}$, $\bar{\mathbf{y}}$, $Dep$
Initialization: $\mathcal{S} = \emptyset$, $y_i^* = N$ for $i = 1, \ldots, N$
for $k = 1, \ldots, Dep$ do
    $I_i \leftarrow \arg\min_{I_j : I_j \in \mathcal{I}, I_j \notin \mathcal{S}} \Delta(\mathbf{y})$, where $\mathbf{y}$ is defined as: $\mathbf{y} = \mathbf{y}^*$ and $y_j = k$;
    $y_i^* = k$;
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{I_i\}$;
end for
return $\mathbf{y}^*$
```

Our aim is to learn a model $f(\cdot)$ which should satisfy the following constraints:

$$\forall \mathbf{y} \in \mathcal{Y} \backslash \mathbf{y}^* : \ f(\mathbf{y}^*) > f(\mathbf{y}) \tag{8}$$

where $\mathcal{Y}$ is the set of all possible $\mathbf{y}$ with $|\mathcal{Y}| = O(N!)$. It means that a good model should assign a higher value to optimal ranking vector $\mathbf{y}^*$ than any other nonperfect ones.

In this paper, we consider the simplest linear model $f(\cdot) = \mathbf{w}^T \psi(\mathbf{y})$, where $\mathbf{w}$ is the weighting vector and $\psi(\mathbf{y})$ is a feature vector which describes the relevance and diversity attributes for ranking vector $\mathbf{y}$. We will detail $\psi(\mathbf{y})$ later in Section V. With the linear model, the constraints in (8) translate to

$$\forall \mathbf{y} \in \mathcal{Y} \backslash \mathbf{y}^* : \ \mathbf{w}^T \psi(\mathbf{y}^*) > \mathbf{w}^T \psi(\mathbf{y}). \tag{9}$$

With m training queries $\{q^{(i)}\}_{i=1}^m$, we formulate the learning problem by using the powerful structural SVMs [36]

$$\begin{aligned} \min_{\mathbf{w}, \xi \geq 0} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{m}\sum_{i=1}^{m} \xi_i \\ \text{s.t.} \ & \forall i, \forall \mathbf{y} \in \mathcal{Y}^{(i)} \backslash \mathbf{y}^{(i)*} \\ & \mathbf{w}^T \psi\left(\mathbf{y}^{(i)*}\right) \geq \mathbf{w}^T \psi(\mathbf{y}) + \Delta(\mathbf{y}) - \xi_i \end{aligned} \tag{10}$$

where ξ are the slack variables and $C > 0$ controls the trade-off between model complexity and training errors. $\mathbf{y}^{(i)*}$ is the optimal ranking vector for $q^{(i)}$, which has the minimum loss $\Delta(\mathbf{y})$. $\Delta(\mathbf{y})$ on the right hand side of the constraints is utilized to give a more severe penalty to $\mathbf{y}$ which violates far from $\mathbf{y}^{(i)*}$.

The greedy selection algorithm for deriving $\mathbf{y}^{(i)*}$ is given in Algorithm 1. Due to the computation cost and the need in real applications (users often only examine the images returned in the top 1 to 2 pages, about 20–40 images), we only need to select the top subset, for example top-*Dep* images. The parameter *Dep* is utilized to denote how many top-ranked images we evaluated in $\mathbf{y}^*$.

B. Learning TARerank Model

Now we discuss how to solve the learning problem (10) and how to use the learned model to predict rich topic-covering ranking vectors for new incoming test queries. As will be introduced in Section V, the proposed feature $\psi(\mathbf{y})$ consists of three sub-feature vectors, i.e., $\psi(\mathbf{y}) = (\psi_1^T, \psi_2^T, \psi_3^T)^T$, where ψ_j is the jth sub-feature vector. The three sub-feature vectors have different dimensions. To avoid the influence of imbalanced feature dimensions, we modify (10) by introducing balance parameters $\{\gamma_j\}_{j=1}^3$ for $\mathbf{w} = [\mathbf{w}_1^T, \mathbf{w}_2^T, \mathbf{w}_3^T]^T$ where $\mathbf{w}_j$ is the sub-weighting vector corresponding to feature ψ_j. Then we get the new learning problem

Algorithm 2 Cutting Plane Algorithm to Solve (11)

```
Input: $(\mathcal{I}^{(1)}, \bar{\mathbf{y}}^{(1)}, \bar{\mathbf{y}}^{(1)*}), \ldots, (\mathcal{I}^{(m)}, \bar{\mathbf{y}}^{(m)}, \bar{\mathbf{y}}^{(m)*})$, $C$,$\epsilon$
Initialization: $\mathcal{W}^{(i)} \leftarrow \emptyset$ for all query $i = 1, \ldots, m$
repeat
    for $i = 1, \ldots, m$ do
        $H(\mathbf{y}; \mathbf{w}) \equiv \Delta(\mathbf{y}) + \mathbf{w}^T \psi(\mathbf{y}) - \mathbf{w}^T \psi(\mathbf{y}^{(i)*})$
        Compute $\hat{\mathbf{y}} = \arg\max_{\mathbf{y}\in\mathcal{Y}^{(i)}} H(\mathbf{y}; \mathbf{w})$
        Compute $\xi_i = \max\{0, \max_{\mathbf{y}\in\mathcal{W}_i} H(\mathbf{y}; \mathbf{w})\}$
        if $H(\hat{\mathbf{y}}; \mathbf{w}) > \xi_i + \epsilon$ then
            $\mathcal{W}^{(i)} \leftarrow \mathcal{W}^{(i)} \cup \{\hat{\mathbf{y}}\}$
            $\mathbf{w} \leftarrow$ optimize (11) over $\mathcal{W} = \cup_i \mathcal{W}^{(i)}$
        end if
    end for
until no $\mathcal{W}^{(i)}$ has changed during iteration.
return $\mathbf{w}$
```

$$\begin{aligned} \min_{\mathbf{w}, \xi \geq 0} \quad & \frac{1}{2}\sum_{j=1}^{3} \gamma_j \|\mathbf{w}_j\|^2 + \frac{C}{m}\sum_{i=1}^{m} \xi_i \\ \text{s.t.} \ & \forall i, \forall \mathbf{y} \in \mathcal{Y}^{(i)} \backslash \mathbf{y}^{(i)*} \\ & \mathbf{w}^T \psi\left(\mathbf{y}^{(i)*}\right) \geq \mathbf{w}^T \psi(\mathbf{y}) + \Delta(\mathbf{y}) - \xi_i \end{aligned} \tag{11}$$

$\gamma_j > 0$ is the weighting coefficient for $\|\mathbf{w}_j\|^2$. For balance, the $\mathbf{w}_j$ corresponding to features with lower dimension should be slacked by a smaller γ_j. We empirically set $\gamma_j = |\psi_j|$.

For solving this structural learning problem (11), the cutting plane algorithm [36] is utilized, as given in Algorithm 2. To deal with a large amount of constraints, Algorithm 2 iteratively adds constraints into a working set $\mathcal{W}$. For each query $q^{(i)}$, it starts with an empty working set $\mathcal{W}^{(i)}$ and then the most violated constraint is selected and added into active constraint set $\mathcal{W}^{(i)}$ if its violation is larger than a tolerance constant ϵ. With the updated working set $\mathcal{W} = \cup_i \mathcal{W}^{(i)}$, we resolve (11) until the active constraint set $\mathcal{W}^{(i)}$ does not change for all training queries. As proven in [36], the learning procedure is guaranteed to converge in polynomial time.

A key step in Algorithm 2 is to find out the most violated constraints $\hat{\mathbf{y}} = \arg\max_{\mathbf{y}\in\mathcal{Y}} H(\mathbf{y}; \mathbf{w})$ for each query. Finding the exact maximum validated ranking vector $\hat{\mathbf{y}}$ is intractable since there are $N!$ possible candidates in $\mathcal{Y}$. Therefore, we resort to the following greedy selection method to complete it, as given in Algorithm 3.

C. Prediction on Test Query

After learning the optimal parameter vector $\mathbf{w}$, we use the learned model to predict the rich topic-covering ranking result for new incoming queries. The optimal ranking vector $\hat{\mathbf{y}}$ should be selected according to

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}\in\mathcal{Y}} \mathbf{w}^T \psi(\mathbf{y}). \tag{12}$$

However, it is intractable to find out $\hat{\mathbf{y}}$ by examining all $N!$ possible permutation in $\mathcal{Y}$. Therefore, we also resort to

Algorithm 3 Greedy Selection For $\hat{\mathbf{y}} = \arg\max_{\mathbf{y}\in\mathcal{Y}} \Delta(\mathbf{y}) + \mathbf{w}^T\psi(\mathbf{y})$

```
**Input:** $\mathcal{I}$, $\bar{\mathbf{y}}$, $Dep$, $\mathbf{w}$
**Initialization:** $\mathcal{S} = \emptyset$, $y_i^* = N$ for $i = 1, \dots, N$
**for** $k = 1, \dots, Dep$ **do**
    $I_i \leftarrow \arg\max_{I_j: I_j\in\mathcal{I}, I_j\notin\mathcal{S}} \Delta(\mathbf{y}) + \mathbf{w}^T\psi(\mathbf{y})$, where $\mathbf{y}$ is defined as: $\mathbf{y} = \mathbf{y}^*$ and $y_j = k$;
    $y_i^* = k$;
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{I_i\}$;
**end for**
**return** $\mathbf{y}^*$
```

the greedy selection method to complete this procedure. The greedy selection algorithm is similar to Algorithm 3, except we must replace the objective $\Delta(\mathbf{y}) + \mathbf{w}^T\psi(\mathbf{y})$ with $\mathbf{w}^T\psi(\mathbf{y})$.

V. Feature Construction

In this section, we will detail how to derive a set of proper features $\psi(\mathbf{y})$ to describe the properties of a ranking vector $\mathbf{y}$. We investigate three important properties that a perceptual good ranking result should have: relevance, TC, and representativeness. For each of those criteria, we define related features to measure them. The feature vector can be defined as $\psi(\mathbf{y}) = (\psi_1^T, \psi_2^T, \psi_3^T)^T$, where ψ_j is the sub-feature vector corresponding to the jth criterion. In the following subsections, we will detail how to derive sophisticated ψ by addressing the above three criteria respectively.

A. Relevance

All top-ranked images should be relevant. Irrelevant images in the top list affect user experience. We define relevance related features to measure the relevance quality of $\mathbf{y}$.

The relevance feature ψ_1 should measure how relevant the top-*Dep* ranked images in $\mathbf{y}$ are. For each query, a relevance score vector $\bar{\mathbf{r}} = [\bar{r}_1, \dots, \bar{r}_N]^T$ expresses the relevance of images to this query with $\bar{r}_i$ corresponding to image I_i. The $\bar{\mathbf{r}}$ can be obtained through any existing relevance-based reranking method, or directly obtained from a text-based search.

We define the relevance feature as the weighted sum of the relevance scores of the top-*Dep* ranked images in $\mathbf{y}$, that is

$$\psi_1 = \frac{1}{z}\sum_{y_i \le Dep} \beta_i * \bar{r}_i \tag{13}$$

where β_i is the weight for $\bar{r}_i$ and $z = \sum_{y_i \le Dep} \beta_i$ is the normalization constant.

Since we desire more relevant samples to have higher ranks, a larger β_i should be assigned to an image with a higher rank. In this paper, we empirically set β_i as

$$\beta_i = \frac{1}{\log_2(1 + y_i)}. \tag{14}$$

The relevance feature is used to maintain the relevance information obtained from any cues. The text-based search results essentially provide a way for deriving $\bar{\mathbf{r}}$, i.e., setting $\bar{r}_i$ according to the rank of I_i in text-based search results. Besides, we can also resort to relevance-based reranking methods to obtain refined relevance score vectors. Through various text-based search technologies and relevance-based reranking methods, we can derive a set of relevance score vectors $\{\bar{\mathbf{r}}_d\}$, $d = 1, \dots, d_1$. Then ψ_1 can be extended to a d_1-dimensional vector $\psi_1 = [\psi_{1_1}, \dots, \psi_{1_{d_1}}]^T$ with ψ_{1_d} defined on $\bar{\mathbf{r}}_d$ according to (13).

B. TC

Images with duplicate topics, although relevant, cannot provide rich information. Therefore diverse topics among top-ranked images are highly preferred. Besides, due to the ambiguity of the text query terms, a diverse ranking result can satisfy various users. Features relating to TC will be utilized to measure the topic richness of the top-ranked images.

To ensure the top-*Dep* ranked images in $\mathbf{y}$ cover rich topics, we require these images to be visually dissimilar to each other. Therefore, we define the TC feature ψ_2 as the minimum visual dissimilarity among the top-*Dep* ranked images, that is

$$\psi_2 = \min_{y_i \le Dep, y_j \le Dep, i \ne j} (1 - s_{ij}) \tag{15}$$

where s_{ij} is the visual similarity between images I_i and I_j. Maximizing the minimal dissimilarity ensures that, in top-*Dep* ranked image set each image is highly dissimilar to others.

The similarity s_{ij} between images I_i and I_j is calculated from their visual features $\mathbf{x}_i$ and $\mathbf{x}_j$ as

$$s_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{2\sigma^2}\right)^2. \tag{16}$$

As we can see from (16), s_{ij} is influenced by the scaling parameter σ and the utilized visual feature $\mathbf{x}$. Since there is no good solution to determine which kind of visual feature or which σ should be used, we can utilize a set of visual features $\{\mathbf{x}_p\}_{p=1,\dots,m}$ and a set of scaling parameters $\{\sigma_q\}_{q=1,\dots,n}$. By calculating a set of ψ_2 via (15) with each visual feature and variance scale, we can augment ψ_2 to a long feature vector $\psi_2 = [\psi_{2_1}, \dots, \psi_{2_{d_2}}]^T$ with dimensionality $d_2 = m \times n$.

C. Representativeness

Besides the above two criteria, there is the third that should be considered—representativeness. We define an image as representative if it is located in a dense area with many similar images. Representativeness has dual connections to both relevance and TC. On one hand, it is widely assumed in relevance-based reranking that frequently occurring images are more likely to be relevant [1], [2]. From this point of view, representativeness is part of relevance-related feature. On the other hand, in TARerank, we require top-ranked images to cover rich topics. However, there are usually a set of relevant images that belong to the same topic, therefore determining which should be used to represent the topic is problematic. Generally, more representative images are often preferred. Due to the importance of representativeness, we also define features for measuring the representativeness of the top-*Dep* ranked images in $\mathbf{y}$.

Intuitively, an image is more representative if it is located in a dense area with many images around it. Therefore, we can measure the representativeness of image I_i with its probability density p_i. p_i can be estimated through kernel density estimation (KDE) [37], [38]

$$p_i = \frac{1}{|\mathcal{N}_i|} \sum_{I_j \in \mathcal{N}_i} k(\mathbf{x}_i - \mathbf{x}_j) \tag{17}$$

where $\mathcal{N}_i$ is the set of neighbors of image I_i and $k(\mathbf{x})$ is a kernel function that satisfies both $k(x) > 0$ and $\int k(\mathbf{x})d(\mathbf{x}) = 1$. The Gaussian kernel is adopted in this paper.

With the representativeness p_i for each image, we can define the representativeness feature ψ_3 for ranking vector $\mathbf{y}$ as the weighted sum of p_i of the top-*Dep* ranked images

$$\psi_3 = \frac{1}{z} \sum_{y_i \le Dep} \beta_i * p_i. \tag{18}$$

The weighting β_i and normalization constant z are defined in the same way as that in (13).

The estimation of p_i via KDE is also influenced by the scaling parameter σ and the utilized visual feature $\mathbf{x}$. Similar to the TC feature, we also augment ψ_3 to a d_3-dimensional feature vector $\psi_3 = [\psi_{3_1}, \ldots, \psi_{3_{d_3}}]^T$ with each ψ_{3_i} estimated via (18) with different variance scales and visual features.

VI. Experiments

In order to demonstrate the effectiveness of the proposed TARerank method, we conduct several experiments on a Web image search dataset.

A. Experimental Setting

1) Dataset Collection: There is no publicly available benchmark dataset which has been labeled with hierarchical topics. Therefore, we collected a dataset from Web image search engines. Due to the laborious nature of labeling hierarchical topics for training queries, this preliminary dataset currently consists of 23 948 images and 26 queries. (The topic label is not required for a test query.) For each query, we have retrieved the images (at most, the top 1000 ranked) returned by a text-based image search engine.

2) Relevance and Topic Labeling: For each image, its relevance degree with respect to the corresponding query is judged by human labelers on two levels, i.e., "relevant" and "irrelevant." For each query, the human labelers are also required to group all relevant images into different topics. The images belonging to the same topic are further divided into several subtopics if necessary, until the labelers think there is no need to continue this operation. The numbers of topic layers in these queries vary from 1 to 6.

3) Visual Features: We extract several low-level visual features to describe the images' content and use them for calculating similarity and density. These features include: 1) attention-guided color signature [39]; 2) color spatialet [40]; 3) scale-invariant feature transform [41]; 4) multilayer rotation invariant edge orientation histogram [42]; 5) histogram of gradient [43]; 6) the combination of the above five features and daubechies wavelet [44] as well as facial feature [45], as described in [40]; and 7) color moment in lightness color-opponent dimensions space [46]. More details of these extractions of visual features can be found in [40]. For fair comparison, in our experiments all other methods also utilize these features for calculating the similarity between images. In calculating the TC and representativeness features in (16) and (18), seven different σs are adopted for each kind of visual feature, resulting $|\psi_2| = |\psi_3| = 49$. A set of scaling parameters $\{\sigma_1, \ldots, \sigma_7\}$ are empirically defined as

$$\sigma_i = \text{scale}_i * \text{MeanDist} \tag{19}$$

where MeanDist is the average distance of K nearest neighbors over all N images and scale $= \{1/4, 1/2, 1/\sqrt{2}, 1, \sqrt{2}, 2, 4\}$. K is set as 15 in this paper.

4) Dataset Split for Fourfold Cross Validation: We split the 26 queries into fourfolds with each fold comprising 7, 7, 6, and 6 queries, respectively. Each time, we use twofolds queries for training, onefold queries for validation and the remaining fold queries for testing. We repeat the experiments four times and let each fold be used once for testing.

5) Evaluated Methods: We compared TARerank with several methods, including the text search baseline (Text), one typical relevance-based reranking method—Bayesian reranking (BR) [4], one typical diversified reranking method—MMR [9] based on text search results (MMR-Text), as well as the two-step combination of applying MMR to the post-process BR result, denoted as MMR-BR. BR, MMR-Text, and MMR-BR are all unsupervised methods. For fair comparison, their optimal parameters are also selected on the validation set and then applied on the test set to get the fourfold cross validation results. Here we do not evaluate the method proposed in [10] and [11] due to the lack of tags, which are essentially required in those methods but often unavailable for general Web images.

6) Evaluation Measures: The measurements used for performance evaluation in this paper include: 1) the aforementioned NCTC; 2) existing relevance measurement averaged precision (AP) [47] and normalized discounted cumulated gain (NDCG) [48]; and 3) existing diversity measurement TRecall [13]. AP is the mean of the precision values obtained when each relevant image occurs. The AP of top-k ranked images is defined as

$$\text{AP@}k = \frac{1}{Z_k} \sum_{i=1}^{k} [\text{precision}(i) \times \text{rel}(i)] \tag{20}$$

where rel(i) is a binary function denoting the relevance of the ith ranked image with "1" for relevant and "0" for irrelevant. precision(i) is the precision of top-i ranked images

$$\text{precision}(i) = \frac{1}{i} \sum_{j=1}^{i} \text{rel}(j). \tag{21}$$

Z_k is a normalization constant that is chosen to guarantee AP@$k = 1$ for a perfect ranking result list. The perfect ranking result list is derived by ordering images according to their ground-truth relevance labels. The TRecall is calculated in a similar way, and is also normalized by a constant to guarantee a perfect ranking result list's TRecall@k = 1. The perfect ranking result list is derived by ordering images according to their ground-truth topic labels.

TABLE II

RERANKING COMPARISON OF DIFFERENT METHODS. CROSS-VALIDATION IS CONDUCTED ACCORDING TO NCTC, FOR FAIR TC COMPARISON. TARERANK MARKED BY "*" MEANS IT OUTPERFORMS ALL OTHER FOUR METHODS SIGNIFICANTLY

Method	*Dep*-5				*Dep*-10				*Dep*-20			
	NCTC	TRecall	AP	NDCG	**NCTC**	TRecall	AP	NDCG	**NCTC**	TRecall	AP	NDCG
Text	65.3	71.5	78.3	85.4	59.4	60.4	70.9	81.4	59.1	**51.0**	65.9	79.3
MMR-Text	64.6	*72.3*	76.9	84.6	59.4	60.0	67.0	78.4	*59.4*	50.3	58.7	74.2
BR	61.5	70.0	*78.4*	85.2	58.8	57.7	*72.3*	*82.7*	58.8	49.6	65.6	78.8
MMR-BR	52.1	57.7	61.9	72.2	51.9	48.8	57.0	70.2	51.2	42.1	54.0	68.6
TARerank	**67.3***	**76.2***	**83.1***	**88.8***	**60.1***	**61.2***	**73.9***	**83.7***	**61.0**	49.8	**66.7***	**79.6**

TABLE III

RERANKING COMPARISON OF DIFFERENT METHODS. CROSS-VALIDATION IS CONDUCTED ACCORDING TO NDCG, FOR FAIR RELEVANCE COMPARISON

Method	*Dep*-20			
	NCTC	TRecall	AP	NDCG
Text	59.1	51.0	65.9	79.3
MMR-Text	59.1	51.0	65.9	79.3
BR	51.1	45.8	*67.2*	*79.4*
MMR-BR	53.6	44.0	*66.7*	78.1
TARerank	**60.5**	**51.6**	**69.7**	**81.7**

B. Experimental Results and Analysis

In this section, the results of experiments with various settings are presented and analyzed. We have tested a set of $Dep = \{5, 10, 20\}$. Table II presents the experimental results of the proposed TARerank and the four baseline methods. For fair comparison, in all methods their optimal parameters are selected via fourfold cross-validation by optimizing their performance in terms of NCTC on the validation set.

1) Comparison of NCTC: We first analyze their performance in terms of NCTC. Table II shows that the proposed TARerank presents the best performance among the five methods, and achieves consistent improvements over three *Dep* (5, 10, 20) settings (compared with Text baseline). The NCTC in relevance-based reranking method BR decreases because BR has the only objective of improving the relevance and neglects the diversity. For diversified reranking method MMR-Text, its performances on *Dep*-5, *Dep*-10, and *Dep*-20 slightly decrease, keep stable, and then slightly increase, respectively. The reason is that MMR-Text post-processes the top-ranked images in Text result by selecting a visually diverse image set. The gap between visual diversity and semantic topic diversity causes limited improvements (sometimes even deterioration). For relevance-diversified two-step method MMR-BR, it accumulates the TC reduction in the BR step. This error accumulation, coupled with the limited power of MMR, makes it hard for MMR-BR to improve the TC.

2) Correlation With TRecall: TRecall is a diversity measurement which has been used in some diversified reranking work for evaluation [13], [35]. The main difference between NCTC and TRecall is that NCTC is much more general and takes the hierarchical topic structure and the topic importance into consideration. By comparing NCTC and TRecall of the five methods in Table II, we can find that they are roughly consistent, i.e., methods achieving high NCTC generally also have high TRecall. Specifically, their correlation coefficients measured via Kendall τ ($\in [-1, 1]$) [49] are 0.875, 1.0, and 0.5 on *Dep*-5, *Dep*-10, and *Dep*-20, respectively. Since both TRecall and NCTC are used for TC measuring, the positive correlation between them partially verifies the capacity of NCTC in measuring reranking performance. Since TRecall is just a special case of NCTC, they are not perfectly correlated.

3) Comparison of Relevance: We have analyzed the performance of TARerank in terms of NCTC above. Now we examine whether it improves relevance and diversity simultaneously. The performance in terms of relevance corresponds to the AP and NDCG columns in Table II. We find that TARerank also achieves excellent performance in improving relevance, even better than the relevance-based reranking method BR. However, since the results in Table II are obtained via cross-validation according to NCTC, the relevance comparison between TARerank and BR here may be unfair since they have different ranking objectives. Considering this, we further conduct another cross-validation where optimal parameters are selected for all methods according to NDCG. The results are presented in Table III. Here we take only *Dep*-20 for illustration. This table shows that TARerank also outperforms BR. This phenomenon demonstrates the power of TARerank in improving relevance and diversity simultaneously.

Overall, MMR-Text can only slightly improve the diversity of Text, while sacrificing relevance. BR improves the relevance of Text, while sacrificing diversity. Two-step method MMR-BR improves diversity and relevance in two separate steps and the errors are easily accumulated. As a consequence, MMR-BR can hardly achieve satisfactory results. Our proposed TARerank directly optimizes the relevance and diversity simultaneously in one objective and achieves the best performance.

To verify whether the improvement of TARerank is statistically significant, we further perform a statistical significance test. Here we conduct a paired T-test with a 5% level of significance between TARerank and the other four methods. The results are reported in Table II. A mark of "*" is given if TARerank significantly outperforms all other methods. It shows that the differences are significant in most cases, especially when $Dep \leq 10$.

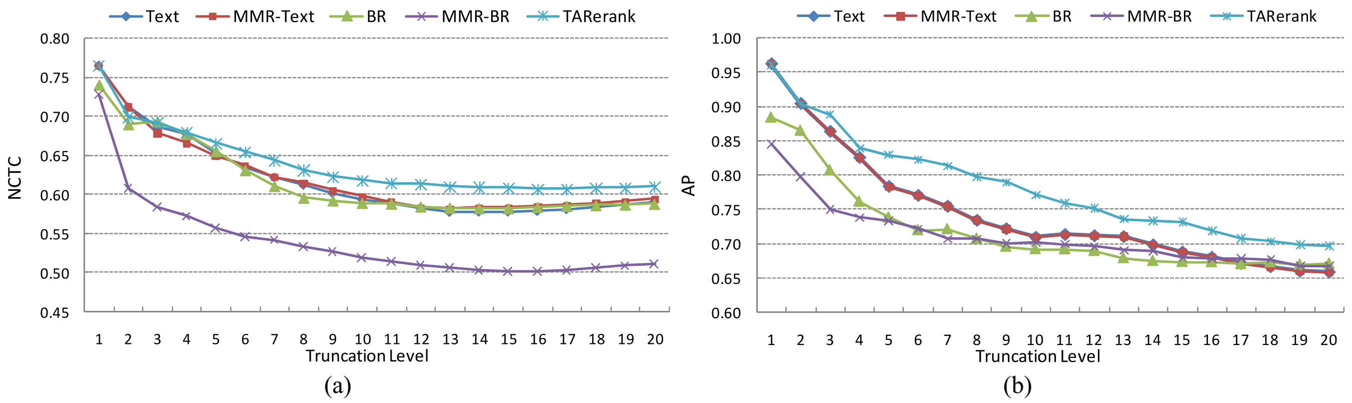


Fig. 3. Experimental results of TARerank, the text search baseline, and other reranking methods. (a) and (b) Measure the NCTC and AP at different truncation levels respectively. Since the result of MMR-Text is close to Text, the Text curve is almost covered by that of MMR-Text (best viewed in color).

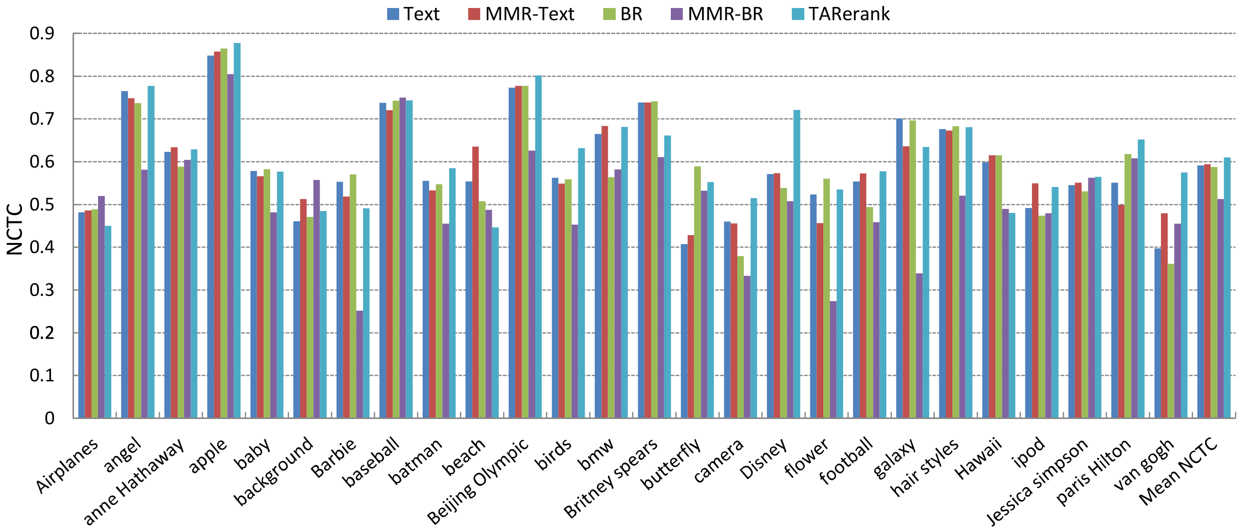


Fig. 4. Performance in terms of NCTC@20 of TARerank, as well as the other four methods on each query. TARerank outperforms Text on 19/26 queries, and obtains the highest performance over all five methods on 11/26 queries.

4) Comparison of Performance (NCTC, AP) at Different Truncation Levels: In Tables II and III, only the performances at truncation level *Dep* are given. To further examine their effectiveness at truncation levels from 1 to *Dep*, we also illustrate the curves of NCTC@1-20 and AP@1-20, as shown in Fig. 3. From Fig. 3(a), we find that TARerank gets stable improvements at different truncation levels with the only exception of NCTC@2, which is slightly degraded. Fig. 3(b) shows that the text search baseline is consistently improved by TARerank at different truncation levels, while BR and MMR-BR improve the Text only at levels 17–20.

5) TARerank on Each Query: Besides the overall performance on the whole dataset, we also analyze the performance of TARerank on each query. Here we take the experiments with *Dep*-20 for illustration and present the results in terms of NCTC@20 and AP@20 for each query in Figs. 4 and 5, respectively. From Fig. 4, we can find that for most queries, NCTC is improved after reranking via TARerank. Specifically, TARerank outperforms Text on 19 out of 26 queries and obtains the highest performance over all five methods on 11 out of 26 queries. As for AP@20, Fig. 5 shows that BR and MMR-BR improve the AP of Text on some queries, for example "baby" and "batman" for BR, and "camera" and "Paris Hilton" for MMR-BR. However, they also suffer from sudden decreases on many queries, for example "angle," "Disney," and "football." Compared with BR and MMR-BR, TARerank improves the Text much steadier and rarely shows large decreases on queries.

MMR-BR performs the reranking in a two-step manner, i.e., first using BR to improve relevance and then utilizing MMR to improve the diversity of the BR result. This two-step process creates the problem of error accumulation, which is the reason why MMR-BR is not as stable as TARerank. The performance of MMR-BR highly depends on the BR result. As shown in Fig. 5, for those queries BR fails, the MMR-BR shows either a sudden increase ("airplanes," camera) or a sudden decrease (angel). As we know, BR tends to return near-duplicate images in the top of the reranking result. MMR-BR increases the diversity by eliminating the visually duplicate images from BR result sequentially. Those near-duplicate images may be relevant, but they can also be noisy. As a consequence, if the eliminated near-duplicate images are

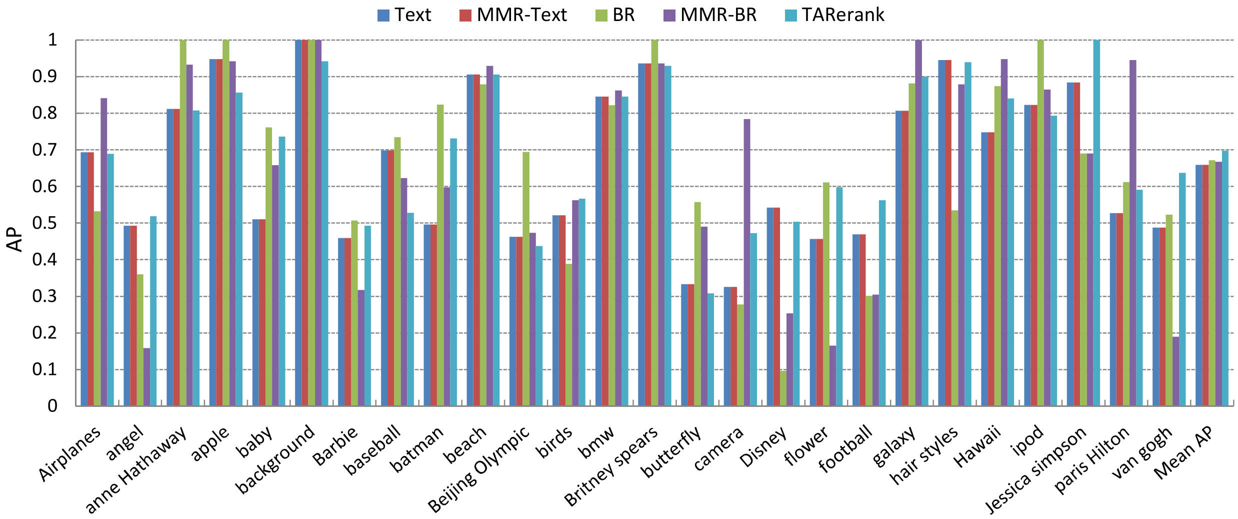


Fig. 5. Performance in terms of AP@20 of TARerank, as well as the other four methods on each query. TARerank outperforms Text on 15/26 queries. Compared with BR and MMR-BR, TARerank improves the Text more steadily and rarely shows large decreases on queries.

noisy, MMR-BR can improve the performance of BR, leading to a sudden increase. Otherwise, a sudden decrease will be observed if the eliminated near-duplicate images are relevant.

Fig. 6 gives the top-10 images returned on query "Van Gogh" by Text, MMR-Text, BR, MMR-BR, and TARerank. MMR-Text improves the diversity of Text, but introduces some irrelevant images at the same time. BR improves the relevance but returns some near-duplicate images (for example, the "sunflower" paintings). MMR-BR accumulates the errors in BR and MMR, therefore it performs the worst. Our proposed TARerank achieves the best performance and returns the paintings of Van Gogh without duplication.

6) Individual Feature Evaluation: As introduced in Section V, our proposed feature $\psi(\mathbf{y})$ consists of three sub-feature vectors which correspond to relevance (Fea1), TC (Fea2), and representativeness (Fea3) respectively. Here we further investigate the effectiveness of each of those three features and their late fusion. The experimental results are presented in Table IV. Fea1 is a 1-D feature vector defined according to the relevance information provided by the text-based search result. Since there is no other information utilized, the performance of TARerank with only Fea1 is almost the same as Text. For TARerank with only Fea2, it improves the TC of Text to some extent, but AP and NDCG decrease. This is because Fea2 only focuses on selecting visually diverse images and neglects the relevance property. As a consequence, some visually different, but irrelevant, images are returned. For TARerank with only Fea3, it outperforms Text in terms of AP and NDCG, but underperforms Text in terms of TRecall and NCTC since representative images may be visually duplicated. Overall, compared to TARerank with all features combined ("AllCombined"), the individual features do not perform well. This is because those three features characterize very different but highly complementary properties of a good search result. All of them are essentially required to learn a satisfactory reranking model. "LateFusion" denotes the performance that we combine the reranking results of "Fea1," "Fea2," and "Fea3." This late fusion is performed as follows. We assign three scores $\{S_1^I = 1/(r_{\text{Fea1}}), S_2^I = 1/(r_{\text{Fea2}}), S_1^I = 3/(r_{\text{Fea3}})\}$ for each image I, where $r_{\text{Fea}i}$ is the rank of image I in the ranking result of "Fea$i$," $i = 1, 2, 3$. The final score of image I is the average of those three scores. The late fusion is obtained by ranking all images according to their final score in descending order. We can see that LateFusion performs better than the individual features, but achieves much lower performance than AllCombined (early fusion).

TABLE IV
RERANKING COMPARISON OF TARERANK WITH THREE INDIVIDUAL FEATURES

Method	*Dep*-5			
	NCTC	TRecall	AP	NDCG
Fea1	65.3	71.5	78.3	85.4
Fea2	65.5	71.8	76.5	84.4
Fea3	64.1	70.2	79.1	86.0
LateFusion	65.8	72.1	80.5	86.7
AllCombined	**67.3**	**76.2**	**83.1**	**88.8**

7) Sensitivity of TARerank to Parameter C: Our proposed TARerank has only one free parameter C in structural support vector machine (11). In the experiments, we test a set of Cs ∈ [1000, 100, 10, 1, 0.1, 0.01]. The results presented above are obtained via cross-validation over all Cs. To investigate the sensitivity of TARerank to this parameter, here we examine its performance with each C, as presented in Table V. From this table, we find that TARerank outperforms Text with various Cs stably for *Dep*-10 and *Dep*-20. For *Dep*-5, TARerank is more sensitive to C and the NCTC decreases slightly when $C \leq 10$. By comparing their best C (1000 for *Dep*-5 and *Dep*-10, 0.1 for *Dep*-20), we find that a lower *Dep* usually prefers a larger C, and vice versa. This provides a rough guideline for setting proper C empirically in practical applications. An intuitive

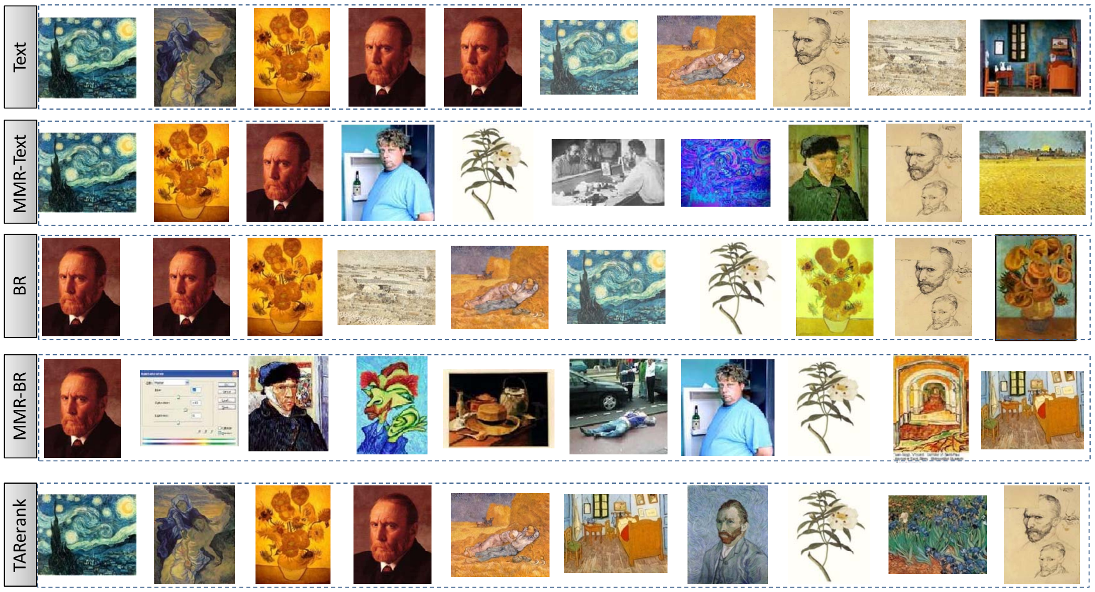


Fig. 6. Top-ten images returned on query Van Gogh by Text, MMR-Text, BR, MMR-BR, and TARerank.

TABLE V
TARerank With Different Cs. TARerank Outperforms Text With Various Cs Stably For *Dep*-10 and *Dep*-20. By Comparing Their Best C (1000 for *Dep*-5 and *Dep*-10, 0.1 for *Dep*-20), We Find That a Lower *Dep* Usually Prefers a Larger C, and Vice Versa

	Text	Validation over all Cs	C					
			1000	100	10	1	0.1	0.01
Dep-5	65.3	67.3	**67.6**	66.1	64.3	64.7	64.3	64.3
Dep-10	59.4	60.1	**61.2**	60.7	60.8	60.0	60.9	60.9
Dep-20	59.1	61.0	59.1	59.9	60.1	60.5	**62.3**	62.3

explanation for this phenomenon is that the trade-off parameter C balances the effects of the two terms: model complexity and training error. A larger C indicates that a smaller training error is required. For a lower *Dep*, the learning problem is much easier with fewer constraints that can easily be satisfied, therefore a smaller training error can be ensured, leading to a larger C. When *Dep* increases, the learning problem becomes more challenging and the training error will be bigger, therefore a smaller C is preferred.

8) Complexity Analysis and Comparison: The time complexity for MMR-Text is $O(DepMN)$, where M is the dimension of the low-level visual features and N is the number of images for reranking. The time complexity for BR is $O(MN^2 + N^3)$ approximately. Therefore, the time cost for MMR-BR is $O(DepMN + MN^2 + N^3)$. In TARerank, the time complexity for extracting feature $\psi(\mathbf{y})$ for a given $\mathbf{y}$ is $O(DepMN)$. For the training of TARerank, it is guaranteed to converge in polynomial time [36]. Besides, the model only needs to be trained once offline. Therefore, we mainly analyze the time complexity during the online testing stage for TARerank, which is $O((DepMN + d)DepN)$, where d is the dimension of $\psi(\mathbf{y})$. Since d is usually much smaller than $DepMN$, the online testing time cost for TARerank can be approximated by $O(Dep^2MN^2)$. In summary, among the four methods MMR-Text has the lowest time complexity, and the time cost for TARerank in the testing stage is comparable to that of BR and MMR-BR when *Dep* is small.

Besides theoretical analysis, we also test the time cost experimentally. They are implemented using C++ and run on a server with 2.67-GHz Intel Xeon CPU and 16 GB memory in single thread, $N = 200$, $Dep = 20$. MMR-Text takes less than 0.01 s. For BR and MMR-BR, they take about 0.1 s for reranking. For TARerank, it takes about 2 min for training the model from 13 queries, and takes less than 0.4 s for testing. It is worth emphasizing that in the testing stage, TARerank can be processed in parallel for efficiency and then its time cost is further reduced to $O(Dep^2MN)$. From the theoretical analysis and the statistical numbers discussed above, we can see that TARerank achieves the best reranking performance with acceptable time complexity.

VII. Conclusion

In this paper, we propose a new diversified reranking method, TARerank, to refine text-based image search results. This method not only takes topic importance into consideration, but also directly learns a reranking model by optimizing a criterion related to reranking performance in terms of both relevance and diversity in one stage simultaneously. To better model the hierarchical topic structure of search results and describe the relevance and diversity in one criterion seamlessly, NCTC is proposed to quantify the hierarchical TC. Compared with the two-step optimization in other diversified reranking methods, TARerank can achieve the joint optimum of improving relevance and diversity. Besides, the learning procedure can bridge the gap between low-level visual feature diversity and high-level semantic topic diversity to some extent. These two advantages ensure the superiority of TARerank. By conducting extensive experiments on a Web image dataset, we have demonstrated the effectiveness of the proposed method. Furthermore, we find that both the relevance and TC are improved in our proposed TARerank. We believe that this method is a promising new paradigm for Web image search reranking.

Our future work will explore some additional objectives. One is to involve semantic information in TC feature construction and further bridge the gap between visual diversity and topic diversity. Currently, the NCTC can only deal with two relevance levels. Thus, generating multilevel relevance in the NCTC and TARerank is a direction for future research.

References

[1] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking via information bottleneck principle," in *Proc. ACM Int. Conf. Multimedia*, Santa Barbara, CA, USA, 2006, pp. 35–44.

[2] Y. Jing and S. Baluja, "VisualRank: Applying pagerank to large-scale image search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1877–1890, Nov. 2008.

[3] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking through random walk over document-level context graph," in *Proc. ACM Int. Conf. Multimedia*, Augsburg, Germany, 2007, pp. 971–980.

[4] X. Tian, L. Yang, J. Wang, X. Wu, and X.-S. Hua, "Bayesian visual reranking," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 639–652, Aug. 2011.

[5] R. Yan, A. G. Hauptmann, and R. Jin, "Multimedia search with pseudo-relevance feedback," in *Proc. ACM Int. Conf. Image Video Retrieval*, Champaign, IL, USA, 2003, pp. 238–247.

[6] X. Tian, D. Tao, X.-S. Hua, and X. Wu, "Active reranking for Web image search," *IEEE Trans. Image Process.*, vol. 19, no. 3, pp. 805–820, Mar. 2010.

[7] X. Tian, D. Tao, and Y. Rui, "Sparse transfer learning for interactive video search reranking," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 8, no. 3, pp. 1–19, 2012.

[8] K. Song, Y. Tian, W. Gao, and T. Huang, "Diversifying the image retrieval results," in *Proc. ACM Int. Conf. Multimedia*, Santa Barbara, CA, USA, 2006, pp. 707–710.

[9] R. H. van Leuken, L. G. Pueyo, X. Olivares, and R. van Zwol, "Visual diversification of image search results," in *Proc. Int. World Wide Web Conf. (WWW)*, Madrid, Spain, 2009, pp. 341–350.

[10] R. van Zwol, V. Murdock, L. G. Pueyo, and G. Ramírez, "Diversifying image search with user generated content," in *Proc. Multimedia Inf. Retrieval*, Vancouver, BC, Canada, 2008, pp. 67–74.

[11] K. Yang, M. Wang, X.-S. Hua, and H.-J. Zhang, "Social image search with diverse relevance ranking," in *Advances in Multimedia Modeling*. Berlin, Germany: Springer, 2010, pp. 174–184.

[12] J. G. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proc. ACM SIGIR Spec. Interest Group Inf. Retrieval*, Melbourne, VIC, Australia, 1998, pp. 335–336.

[13] M. Sanderson, J. Tang, T. Arni, and P. Clough, "What else is there? Search diversity examined," in *Proc. Eur. Conf. IR Res. Adv. Inf. Retrieval (ECIR)*, Toulouse, France, 2009, pp. 562–569.

[14] L. Liu, L. Shao, X. Zhen, and X. Li, "Learning discriminative key poses for action recognition," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1860–1870, Dec. 2013.

[15] W. Fu, M. Johnston, and M. Zhang, "Low-level feature extraction for edge detection using genetic programming," *IEEE Trans. Cybern.*, vol. 44, no. 8, pp. 1459–1472, Aug. 2014.

[16] L. Shao, R. Yan, X. Li, and Y. Liu, "From heuristic optimization to dictionary learning: A review and comprehensive comparison of image denoising algorithms," *IEEE Trans. Cybern.*, vol. 44, no. 7, pp. 1001–1013, Jul. 2014.

[17] B. Liu, Y. Xiao, P. S. Yu, Z. Hao, and L. Cao, "An efficient orientation distance-based discriminative feature extraction method for multi-classification," *Knowl. Inf. Syst.*, vol. 39, no. 2, pp. 409–433, 2014.

[18] J. Yu, D. Liu, D. Tao, and H. S. Seah, "On combining multiple features for cartoon character retrieval and clip synthesis," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 5, pp. 1413–1427, Oct. 2012.

[19] Q. Huang, D. Tao, X. Li, L. Jin, and G. Wei, "Exploiting local coherent patterns for unsupervised feature ranking," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 6, pp. 1471–1482, Dec. 2011.

[20] D. Tao, L. Jin, Y. Wang, and X. Li, "Rank preserving discriminant analysis for human behavior recognition on wireless sensor networks," *IEEE Trans. Ind. Informat.*, vol. 10, no. 1, pp. 813–823, Feb. 2014.

[21] L. Yang and Y. Zhou, "Exploring feature sets for two-phase biomedical named entity recognition using semi-CRFs," *Knowl. Inf. Syst.*, vol. 40, no. 2, pp. 439–453, 2014.

[22] M. Wang *et al.*, "Unified video annotation via multigraph learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 5, pp. 733–746, May 2009.

[23] S. Zhang *et al.*, "Automatic image annotation using group sparsity," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, San Francisco, CA, USA, 2010, pp. 3312–3319.

[24] J. Tang *et al.*, "Correlative linear neighborhood propagation for video annotation," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 2, pp. 409–416, Apr. 2009.

[25] R. Hong *et al.*, "Image annotation by multiple-instance learning with discriminative feature mapping and selection," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 669–680, May 2014.

[26] D. Tao, L. Jin, Y. Yuan, and Y. Xue, "Ensemble manifold rank preserving for acceleration-based human activity recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.

[27] D. Tao, L. Jin, W. Liu, and X. Li, "Hessian regularized support vector machines for mobile image annotation on the cloud," *IEEE Trans. Multimedia*, vol. 15, no. 4, pp. 833–844, Jun. 2013.

[28] B. C. Wallace and I. J. Dahabreh, "Improving class probability estimates for imbalanced data," *Knowl. Inf. Syst.*, vol. 41, no. 1, pp. 33–52, 2014.

[29] E. Eaton, M. desJardins, and S. Jacob, "Multi-view constrained clustering with an incomplete mapping between views," *Knowl. Inf. Syst.*, vol. 38, no. 1, pp. 231–257, 2014.

[30] L. I. Kuncheva and J. J. Rodríguez, "A weighted voting framework for classifiers ensembles," *Knowl. Inf. Syst.*, vol. 38, no. 2, pp. 259–275, 2014.

[31] S. Zhang, M. Yang, T. Cour, K. Yu, and D. N. Metaxas, "Query specific fusion for image retrieval," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Florence, Italy, 2012, pp. 660–673.

[32] Y. Huang, Q. Liu, S. Zhang, and D. N. Metaxas, "Image retrieval via probabilistic hypergraph ranking," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Francisco, CA, USA, 2010, pp. 3376–3383.

[33] S. Bashir and A. Rauber, "Automatic ranking of retrieval models using retrievability measure," *Knowl. Inf. Syst.*, vol. 41, no. 1, pp. 189–221, 2014.

[34] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A power tool for interactive content-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 644–655, Sep. 1998.

[35] L. Cao *et al.*, "RankCompete: Simultaneous ranking and clustering of Web photos," in *Proc. 19th Int. Conf. World Wide Web*, Raleigh, NC, USA, pp. 1071–1072, 2010.

[36] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *J. Mach. Learn. Res.*, vol. 6, pp. 1453–1484, Sep. 2005.

[37] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Stat.*, vol. 33, no. 3, pp. 1065–1076, 1962.

[38] M. Kristan and A. Leonardis, "Online discriminative kernel density estimator with Gaussian kernels," *IEEE Trans. Cybern.*, vol. 44, no. 3, pp. 355–365, Mar. 2014.

[39] Y. Rubner, L. Guibas, and C. Tomasi, "The earth mover's distance, multi-dimensional scaling, and color-based image retrieval," in *Proc. ARPA Image Understanding Workshop*, New Orleans, LA, USA, pp. 661–668, 1997.

[40] J. Cui, F. Wen, and X. Tang, "Real time Google and live image search re-ranking," in *Proc. ACM Int. Conf. Multimedia*, Vancouver, BC, Canada, 2008, pp. 729–732.

[41] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[42] W. T. Freeman and M. Roth, "Orientation histograms for hand gesture recognition," in *Proc. Int. Workshop Autom. Face Gesture Recognit.*, Zurich, Switzerland, pp. 296–301, 1994.

[43] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, 2005, pp. 886–893.

[44] M. Unser, "Texture classification and segmentation using wavelet frames," *IEEE Trans. Image Process.*, vol. 4, no. 11, pp. 1549–1560, Nov. 1995.

[45] R. Xiao, H. Zhu, H. Sun, and X. Tang, "Dynamic cascades for face detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Rio de Janeiro, Brazil, 2007, pp. 1–8.

[46] M. Stricker and M. Orengo, "Similarity of color images," *SPIE Stor. Retrieval Still Image Video Databases*, vol. 2420, pp. 381–392, Feb. 1995.

[47] (Nov. 8, 2010). *Trecvid Video Retrieval Evaluation*. [Online]. Available: http://www-nlpir.nist.gov/projects/trecvid/

[48] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, 2002.

[49] S. M. Kendall and J. D. Gibbons, *Rank Correlation Methods*. London, U.K.: Edward Arnold, 1990.

Xinmei Tian (M'13) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively.

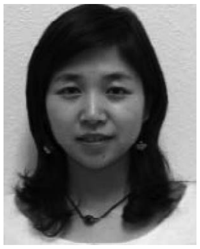
She is an Associate Professor with the CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System, University of Science and Technology of China. Her current research interests include multimedia information retrieval and machine learning.

Prof. Tian was the recipient of the Excellent Doctoral Dissertation of the Chinese Academy of Sciences Award and the Nomination of the National Excellent Doctoral Dissertation Award in 2012 and 2013, respectively.

Linjun Yang (M'08) received the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands, in 2013.

He is currently a Senior Development Lead with Microsoft, Redmond, WA, USA, where he focuses on developing the state-of-the-art image understanding technologies to improve multimedia search experience. He has authored over 50 referred papers.

Dr. Yang was the recipient of the Best Paper Award from ACM Multimedia and ACM Conference on Information and Knowledge Management in 2009.



Yijuan Lu (M'05) received the Ph.D. degree in computer science from the University of Texas at San Antonio, San Antonio, TX, USA, in 2008.

She is an Associate Professor with the Department of Computer Science, Texas State University, San Marcos, TX, USA. Her current research interests include multimedia information retrieval, computer vision, and machine learning. Her research has been funded by the National Science Foundation, Texas Department of Transportation, Department of Defense, Army Research, and Texas State. She has published extensively and has served on Program and Organizing Committee for several international conferences.

Prof. Lu was the recipient of the 2013 International Conference on Multimedia and Expo Best Paper Award, the 2012 International Conference on Internet Multimedia Computing and Service (ICIMCS) Best Paper Award, and is one of the top winners of the 2013 Eurographics Shape Retrieval Contest competitions in Large-Scale Sketch-Based 3-D Retrieval Track, Range Scan Track, and Low-Cost Depth-Sensing Camera Track, the 2014 College Achievement Award, the 2012 Dean nominee for Texas State Presidential Award for Excellence in Scholarly/Creative Activities, and a nominee for 2008 Microsoft Research Faculty Summit.

Qi Tian (M'96–SM'03) received the B.E. degree in electronic engineering from Tsinghua University, Beijing, China, the M.S. and Ph.D. degrees in electrical and computer engineering from Drexel University, Philadelphia, PA, USA, and the University of Illinois at Urbana-Champaign, Urbana, IL, USA, in 1992, 1996, and 2002, respectively.

He is currently a Professor with the Department of Computer Science, University of Texas at San Antonio, San Antonio, TX, USA. He took a one-year faculty leave at Microsoft Research Asia from 2008 to 2009. His current research interests include multimedia information retrieval and computer vision. His research projects were funded by the National Science Foundation, ARO, Department of Homeland Security, San Antonio Life Science Institute, Center for Infrastructure Assurance and Security, and University of Texas at San Antonio. He has published over 260 refereed journal and conference papers.

Dr. Tian was the recipient of the Best Paper Awards in Pacific-Rim Conference on Multimedia (PCM) 2013, Multimedia Modeling 2013, and ICIMCS 2012, the Top 10% Paper Award in International Workshop on Multimedia Signal Processing 2011, the Best Student Paper in International Conference on Acoustics, Speech and Signal Processing 2006, the Best Paper Candidate in PCM 2007, the 2010 ACM Service Award, the Faculty Research Awards from Google, Mountain View, CA, USA, NEC Laboratories of America, Princeton, NJ, USA, FX Palo Alto Laboratory, Akiira Media Systems, and HP Laboratories, Palo Alto, CA, USA. He is a Guest Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, the *Journal of Computer Vision and Image Understanding*, *Pattern Recognition Letter*, the *EURASIP Journal on Advances in Signal Processing*, the *Journal of Visual Communication and Image Representation*, and an Editorial Board Member of the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUIT AND SYSTEMS FOR VIDEO TECHNOLOGY, the *Multimedia Systems Journal*, the *Journal of Multimedia*, and the *Journal of Machine Visions and Applications*.

Dacheng Tao (M'07–SM'12–F'15) is Professor of Computer Science with the Centre for Quantum Computation and Intelligent Systems, and the Faculty of Engineering and Information Technology, University of Technology, Sydney, Ultimo, NSW, Australia. His current research interests include statistics and mathematics to data analytics, computer vision, data science, image processing, machine learning, neural networks, and video surveillance.

Prof. Tao was the recipient of the Best Theory/Algorithm Paper Runner up Award in IEEE ICDM'07, the Best Student Paper Award in IEEE ICDM'13, and the 2014 ICDM 10 Year Highest-Impact Paper Award. His research results have expounded in one monograph and 100+ publications at prestigious journals and prominent conferences, such as IEEE T-PAMI, T-NNLS, T-IP, T-CYB, JMLR, IJCV, Neural Information Processing Systems, International Conference on Machine Learning, Computer Vision and Pattern Recognition, International Conference on Computer Vision, European Conference on Computer Vision, International Conference on Artificial Intelligence and Statistics, ICDM, and ACM SIG Knowledge Discovery and Data Mining,