Query Difficulty Estimation for Image Search With Query Reconstruction Error

Xinmei Tian, *Member, IEEE*, Qianghuai Jia, and Tao Mei, *Senior Member, IEEE*

***Abstract*—Current image search engines suffer from a radical variance in retrieval performance over different queries. It is therefore desirable to identify those "difficult" queries in order to handle them properly. Query difficulty estimation is an attempt to predict the performance of the search results returned by an image search system. Most existing methods for query difficulty estimation focus on investigating statistical characteristics of the returned images only, while neglecting very important information, i.e., the query and its relationship with returned images. This relationship plays a crucial role in query difficulty estimation and should be explored further. In this paper we propose a novel query difficulty estimation method with query reconstruction error. This method is proposed based on the observation that, given the images returned for an unknown query, we can easily deduce what the query is from those images if the search results are high quality (i.e., lots of relevant images returned); otherwise, it is difficult to deduce the original query. Therefore, we propose to predict the query difficulty by measuring to what extent the original query can be recovered from the image search results. Specifically, we first reconstruct a visual query from the returned images to summarize their visual theme, and then use the reconstruction error, i.e., the distance between the original textual query and the reconstructed visual query, to estimate the query difficulty. We conduct extensive experiments on two real-world Web image datasets and demonstrate the effectiveness of the proposed method.**

***Index Terms*—Image retrieval, image search quality, query difficulty estimation, query reconstruction.**

I. Introduction

WITH the exponential growth of Web images, many image search systems have emerged over the past decades. However, image search systems suffer from a radical variance in search performance over different queries. Even for the image search systems that perform well over various queries on average, their search performance for some "difficult" queries might not be satisfying. This is because the quality of image search results depends on many factors: chosen search algorithms, ranking functions, indexing features, the base image database, etc. Previous research has shown that no setting can always perform optimally for all queries. In other words, different queries have different search difficulty levels. For some popular queries with lots of relevant images in the database, (e.g.,, "dog" and "white house"), they are easy to retrieve, and the search engine can return adequate results for them. While for other rare or long queries, it may be difficult to yield satisfactory results. Automatically identifying "difficult" queries will allow users or image retrieval systems to provide a better search experience. For example, for users, they can rephrase the "difficult" queries to improve retrieval effectiveness if an instant evaluation of query difficulty is provided. For the image retrieval engines, they can adopt alternative retrieval strategies for different queries according to their difficulty levels. Therefore, it is desirable to identify those difficult queries in order to handle them properly.

Query difficulty estimation, or query performance prediction, refers to an estimation of the search difficulty level for a given query by estimating the performance of search results returned for this query. Query difficulty estimation has a wide variety of applications in the information retrieval area, such as query refinement, query suggestion, and distributed retrieval. It has been widely explored in text retrieval for many years [1]–[10]. In [1], the Clarity Score query difficulty estimation method was proposed to measure the ambiguity of a query through the difference between the language models created from the top retrieved documents and all documents in the collection. Yom-Tov *et al.* estimated the performance of the search results by measuring the agreement between the top results returned by the full query and its sub-queries [2]. Carmel *et al.* proposed to use the distances among the query, the relevant documents, and the whole collection to estimate the query difficulty [4]. Imran and Sharan proposed two query difficulty predictors based on the co-occurrence information among query terms [8]. They assumed that the higher co-occurrence of query terms indicated more information conveyed, which led to an easier query, or a lower query difficulty level.

However, little research has been conducted on query difficulty estimation for image search. Compared with query difficulty estimation in text retrieval, query difficulty estimation in image search is more challenging. For text document search, both the input and output are in the textual domain. However, for text-based image retrieval, the input (textual queries) and the output (visual images) are in two different domains, i.e., textual and visual domains. Due to this domain gap, current query

Manuscript received April 15, 2014; revised August 15, 2014 and October 23, 2014; accepted October 24, 2014. Date of publication November 24, 2014; date of current version December 23, 2014. This work was supported by the 973 project under Contract 2015CB351803, the NSFC under Contract 61390514 and Contract 61201413, the Specialized Research Fund for the Doctoral Program of Higher Education No. WJ2100060003, and the Fundamental Research Funds for the Central Universities No. WK2100060007 and No. WK2100060011. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Enrico Magli.

X. Tian and Q. Jia are 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; jqh218@mail.ustc.edu.cn).

T. Mei is with Microsoft Research, Beijing 100190, China (e-mail: tmei@microsoft.com).

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/TMM.2014.2368714

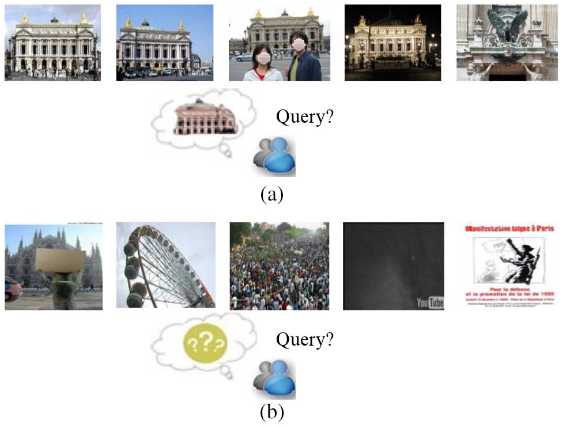


Fig. 1. (a) and (b) list the top five ranked images returned for two unknown queries respectively. We can easily deduce that the query for (a) may be "opera garnier". However, we have no idea what the query is for (b) since the search results are too poor.

difficulty estimation methods in image search mainly focus on investigating some statistical characteristics of the returned images, such as coherence, prominence, and consistency. For example, Rudianc *et al.* exploited the coherence of the top-ranked images to predict the query performance [11]. Li *et al.* measured the query difficulty by further considering the prominence character of image search results [12]. The prominence is defined as the distance between a pseudo-relevant image set and the whole image collection. Tian *et al.* measured the tightness of the top-ranked images to estimate the query difficulty [13].

All of the above methods predict the query performance by analyzing the distribution characteristics of returned images, while neglecting the most important information–the query. The query describes the user's search intent and also decides the relevance labels of the returned images together with the ground truth query performance. As a consequence, the query plays a crucial role in query difficulty estimation and should be explored in greater detail. In this paper, we propose a novel query difficulty estimation method by exploring the relationship between the query and the image search results.

Our proposed query difficulty estimation method is built based on the observation that: given the images returned for an unknown query, if this query is "easy" with a lot of relevant images returned, we can easily deduce the unknown query from only the returned images; otherwise, if the query is "difficult" and the search results contain many irrelevant images, it is difficult to recover the original query from the search results. As an example shown in Fig. 1, the top five ranked images returned for two unknown queries are given in Fig. 1(a) and (b), respectively. We can easily deduce that the query for (a) is "opera garnier", but have no idea about the query for (b) since the search results are too noisy.

The explanation behind this phenomena is that the image search results are actually the visual representation of the query. It interprets the textual query from a visual perspective. A good search result, corresponding to an "easy" query, conveys enough information to recover the original query. Inspired by this, we propose to predict the query difficulty by measuring to what extent the original query can be recovered from the image search results, *i.e.*, the query reconstruction error. Specifically, for a textual query Q and its top ranked images returned by the search engine, we first deduce a visual query Q' from those returned images by analyzing their visual content. This reconstructed visual query is the summary of the visual theme of the search results. Then, we compare Q and Q' for query performance prediction. For an "easy" query Q, the search results consist of images relevant to Q. Therefore, the visual theme of those images can sufficiently represent the query Q. In other words, the reconstructed Q' must be consistent with the original query Q with a small reconstruction error. On the contrary, if Q is a "difficult" query, then the search results are highly noisy with a lot of irrelevant images returned. As a consequence, the visual query Q' recovered from the search result is very different from the original query Q. In other words, the reconstruction error is large. This indicates that the difference between the orignal query Q and the reconstructed query Q'- the reconstruction error - correlates with the query difficulty.

Motivated by the above observation, there are two key problems to be solved in our proposed query difficulty estimation method based on query reconstruction error. One is how to reconstruct the visual query Q', while the other is how to measure the reconstruction error, *i.e.*, the consistent degree between the textual query and its reconstructed visual query. Regarding the first problem, we use the popular BOVWs (bag-of-visual-words) [14] model for image visual representation. Each image in L can be viewed as a visual document consisting of a set of visual words (More detail about BOVWs will be introduced in Section III-A). Then, we formulate a visual query reconstruction as an optimization problem and select a set of the most discriminative visual words from the images in L for reconstructing the visual query. For the second, instead of directly comparing the textual query and the visual query, we formalize this task by measuring the distribution difference between their search results.

The main contributions of this paper can be summarized as follows:

- We propose a novel query reconstruction error-based query difficulty estimation method, which explores the relationship between the textual query and the returned images. To the best of our knowledge, very few works exist that investigate the relationship between the textual query and the returned images for query difficulty estimation in image search.
- We propose a visual query reconstruction method for summarizing the visual theme of the returned image search results. We formulate the query reconstruction as an optimization problem by selecting a set of visual words that are strongly associated with the search results.
- We propose a method for measuring the query reconstruction error which tackles the problem of the domain gap between the textual query and the reconstructed visual query.

The rest of this paper is organized as follows. Section II briefly introduces the related work. In Section III, the proposed query reconstruction error-based query difficulty estimation method is presented in detail. Experimental results are given and analyzed in Section IV, followed by the conclusion in Section V.

II. Related Work

Query difficulty estimation has been explored in the information retrieval field for many years, especially for text document retrieval. Many methods investigate the statistical characteristics (robustness, coherence, etc.) of the returned documents with respect to the whole collection in order to estimate query difficulty [1], [3], [7]. For example, Cronen-Townsend *et al.* first proposed the Clarity Score query difficulty estimation method [1]. It measures the ambiguity of a query through the difference between the language models created from top retrieved documents and all documents in the collection. Ranking Robustness, proposed by Zhou and Croft, defined the similarity between the ranked lists returned from the original collection and the corrupted collection for query difficulty prediction [3]. He *et al.* proposed the coherence score predictor to measure the portion of coherent document pairs in the top of the returned document set [7]. A pair of documents is supposed to be coherent if their similarity exceeds one certain threshold.

On the other hand, some methods investigate the relationships between the textual query and the returned documents to estimate query difficulty [2], [8], [4], [15]. Yom-Tov *et al.* estimated the query performance by measuring the agreement between the top returned results of a full query and the top returned results of each of the query terms [2]. Imran and Sharan proposed two predictors based on the co-occurring information among query terms. They assumed that higher co-occurrence of query terms in the documents meant more information conveyed, which led to an easier query or lower query difficulty level [8]. Carmel *et al.* demonstrated that in text document search, a query's difficulty was mainly related to three components: the query itself, the set of returned documents, and the whole collection of documents [4]. Therefore, they proposed to use the relationships between them for query difficulty estimation. Rudinac *et al.* proposed two simplified coherence indicators to select query expansions for spoken content retrieval [15]. The indicators they proposed capture the tightness of the topical consistency in the top results of the initial results lists returned by the unexpanded query and several query expansion alternatives.

All the above work in query difficulty estimation is designed for text document retrieval. There is little research regarding query difficulty estimation for image retrieval. As aforementioned, due to the domain gap and the semantic gap, existing query difficulty estimation methods for image retrieval mainly focus on investigating the statistical characteristics of the returned images. Li *et al.* estimated the image retrieval performance by analyzing the Clarity Score [1] between the query image and the returned images [16]. Xing *et al.* adopted the textual information (surrounding text, image URL, etc.) to predict whether a query is difficult or easy [17]. This method leverages the noisy textual information and neglects the rich contents of the returned images. Rudinac *et al.* exploited the coherence of the top-ranked video to predict the query performance for selecting the best video search result [11]. Li *et al.* estimated the query difficulty by measuring the prominent character of image search results, i.e., the distance between a pseudo-relevant image set and the whole image collection [12]. Kofler *et al.* explored the visual variance of video search results, the search log, and the click-through data of a user to predict a failure of a video search query [18], [19]. Tian *et al.* proposed a model to automatically predict the query performance based on several features which were designed by exploring the visual characteristics of the returned images [20], [13]. Different from those methods, our proposed approach attempts to investigate the relationship between the query and the visual theme of the returned images to estimate the query difficulty.

Our visual query reconstruction method is similar to query expansion methods, therefore we also review related works in this topic. Query expansion exploits the top-ranked results to enrich the original query representation, and then re-submits the expanded query into the search engine. It has a long history in information retrieval and was introduced to the image search domain by Chum *et al.* [21]. In [21], spatial verification was first applied to the top 1000 ranked images returned for the visual query to get a reliable set of images for query expansion. Then, several methods were investigated to build a new query from the spatially verified images. The average query expansion (AQE) method shows good performance and is usually considered as a baseline. Joly and Buisson implemented a query expansion method for logo retrieval [22]. Their query expansion method is very similar to [21] except that a contrario adaptive thresholding strategy is applied for the spatial verification. Chum *et al.* further extended their previous work [21] from three aspects, including tf-idf failure recovery, improved spatial verification and reranking, as well as relevant spatial context learning [23]. Recently Arandjelovic and Zisserman proposed a discriminative query expansion approach [24]. This method trained a linear SVM classifier on both positive and negative data for image ranking. Xie *et al.* proposed a contextual query expansion that was built on common visual patterns [25]. They found contextual synonymous visual words and expanded a word in the query image with its CSVWs to boost retrieval accuracy.

In this paper we mainly consider the case that the original input query is textual, since searching by key words is still one of the most popular ways for image search. Though we mainly take textual query input for illustration, it is worth noticing that our proposed method is also suitable for visual query input [26]–[28]. Another important feature is that our proposed method only relies on the image's visual features for query difficulty estimation. We do not require any auxiliary information, for example the tag and surrounding text information. This is because some images on the Web may not have rich additional information. However, in the cases where additional information is available, we believe that incorporating them would be very helpful, as demonstrated in [26]–[28].

III. Query Difficulty Estimation with Query Reconstruction Error

The proposed query difficulty estimation with query reconstruction error investigates the consistency between the original textual query and the reconstructed visual query to predict the query performance. The visual query is reconstructed from the image search results and it represents the visual theme of the returned images. It assumes that the images returned for an "easy" query have a clear visual theme which can well interpret the original textual query from a visual perspective. On the contrary, the images returned for a "difficult" query are too noisy, therefore the visual theme deduced from them is very different from the original textual query.

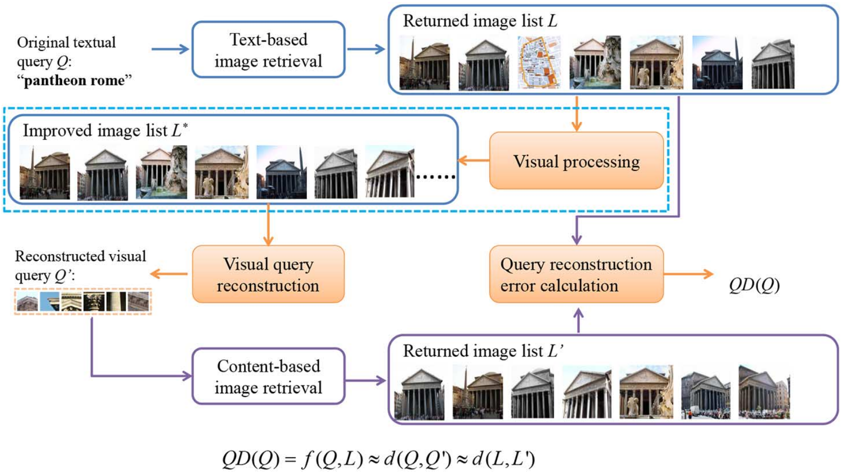

Original textual query Q: "pantheon rome"
Text-based image retrieval
Returned image list L
Improved image list L^*
Visual processing
Reconstructed visual query Q':
Visual query reconstruction
Query reconstruction error calculation
$QD(Q)$
Content-based image retrieval
Returned image list L'
$QD(Q) = f(Q,L) \approx d(Q,Q') \approx d(L,L')$

Fig. 2. Framework of the proposed query reconstruction error-based query difficulty estimation model. In our proposed method, the query difficulty is formulated as a function of both Q and L, i.e., $QD(Q) = f(Q, L)$ and is approximated by the query reconstruction error $d(Q, Q')$. In order to tackle the domain gap problem, we use $d(L, L')$ to approximate $d(Q, Q')$. The L and L' are the ranked image lists returned for Q and Q', respectively, and they can be regarded as the expansion of the simple queries. To reconstruct more accurate visual query Q', an improved image list L^* is obtained from L via visual processing.

Fig. 2 shows the framework of our query reconstruction error-based query difficulty estimation method. Given a textual query Q, a ranked list of images L is returned by the image search system. Then a visual query Q' is reconstructed from L to represent the visual theme of the returned images. To deal with the noise problem in L, we first apply visual processing to get an improved image List L^*, then reconstruct the visual query Q' from L^* instead of L. The query difficulty is defined as the reconstruction error $d(Q, Q')$, *i.e.*, the distance between Q and Q'. Since Q and Q' are in different domains, it is difficult to directly measure their distance. To solve this problem, we propose to approximate $d(Q, Q')$ by $d(L, L')$, *i.e.*, the distance between L and L'. The L' is the image list returned by a content-based image retrieval system for the visual query Q'. For query Q, the larger $d(L, L')$ is, the higher its query difficulty level will be.

There are three key components in our proposed query difficulty estimation method. The first is how to reconstruct the visual query Q' from the returned images list L. The second is how to derive L' for Q'. The third is how to measure the distance between L and L'. We will detail our solutions to those problems in the following sections.

A. *Visual Query Reconstruction*

Given a textual query Q, a ranked image list $L = \{I_1, I_2, \cdots, I_N\}$ is returned for it by the text-based image search engine. The I_i in L denotes the i-th ranked image. The visual query reconstruction summarizes the visual content of images in L via a visual query Q'.

To derive Q', we propose to adopt the popular BOVWs (bag-of-visual-words) [14] model for image visual representation. The BOVWs model has been successfully employed in many applications, such as object recognition [30]–[32], image segmentation [33]–[35], and large-scale image retrieval [14], [36]–[38]. In BOVWs, the local features (for example, SIFT [29]) are extracted from each image. A vocabulary/dictionary V with $|V|$ visual words is constructed via clustering algorithms. With BOVWs representation, each image in L can be viewed as a visual document consisting of a set of visual words. Therefore, we define the reconstructed query Q' as a set of visual words which are strongly associated with the images in L. Here we propose two strategies for query reconstruction.

Query Reconstruction Via Representative and Discriminative Visual Word Selection: The visual words in Q' should be both representative and discriminative in L. Based on this rule, our query reconstruction contains the following two steps.

We first identify the representative visual words in L, *i.e.*, those visual words which frequently appear in L. We build a language model $P(w|L)$ to capture the distribution of visual words over images in L. The language model is

$$P(w|L) = \sum_{I \in L} P(w|I)P(I|L) \tag{1}$$

where $w \in V$ is a visual word, and I is an image in L. $P(w|I)$ is defined as the term frequency of visual word w occurring in image I. $P(I|L)$ indicates the importance of I in language model building. We only use the top-100 ranked images in L for estimating $P(w|L)$ since they tend to be more relevant according to the widely used pseudo relevance feedback assumption [39], [40]. Here we use ranking position weighted prior for the top-100 images and zero for all others in L.

$$P(I|L) = \begin{cases} \frac{1+\sum\limits_{i=r}^{100} \frac{1}{i+1}}{200}, & \text{Rank of } I \text{ in } L \text{ is } r \text{ and } r \leq 100, \\ 0, & \text{else} \end{cases} \tag{2}$$

The $P(w|L)$ describes the distribution of visual words in L. The larger $P(w|L)$ is, the more representative the visual word w will be. To avoid selecting visual words which are generally more common than others (stop words, etc.), we further identify those visual words which are not only representative, but also discriminative. To accomplish this goal, we adopt two different functions to estimate the importance score of each visual word. The two functions are:

- Doszkocs' variant of CHI-squared (CHI) [41]

$$score(w) = \frac{P(w|L) - P(w|C)}{P(w|C)}; \tag{3}$$

- Kullback-Leibler distance (KLD) [42]

$$score(w) = P(w|L) \times \log \frac{P(w|L)}{P(w|C)}. \tag{4}$$

The $P(w|L)$ is the language model for L defined in Eq. (1). $P(w|C)$ is the language model for the whole image collection C. It is defined as the term frequency of visual word w over all images in the database C.

The two discriminative estimation functions can effectively identify those visual words which are both representative and discriminative. The $score(w)$ is proportional to $P(w|L)$ and is inversely proportional to $P(w|C)$. The higher the $score(w)$ is, the more discriminative the visual word w in L is from the collection C.

We rank all visual words according to their importance scores derived via Eq. (3) or Eq. (4) in descending order, and select the top-K ranked visual words to reconstruct the visual query

$$Q' = \{w'_k, k = 1, 2, \cdots, K\} \tag{5}$$

where w'_k is the visual word with the k-th highest $score(w)$. K is a free parameter which decides how many visual words are included in Q'. In general, K should not be too small or too large. Its effect will be discussed later in the experiments. We will also analyze the effect of the two visual word ranking functions (CHI and KLD) in experiments.

Query Reconstruction Via Optimization: The visual query reconstruction can also be formulated as an optimization problem from the probabilistic perspective.

Supposing the visual query Q' is a random variable, the optimal reconstructed query Q'^* is the one which can interpret the image list L with maximum probability. From this perspective, we formulate the visual query reconstruction as the following optimization problem:

$$Q'^* = \arg\max_{Q'} P(L|Q'). \tag{6}$$

$P(L|Q')$ denotes to what extend we can recover L from a given Q'. The optimal Q'^* is the one with which can recover L with the maximum probability. According to Bayes' formula, $P(L|Q')$ can be rewritten as

$$P(L|Q') = \frac{P(Q'|L)P(L)}{P(Q')}. \tag{7}$$

For $P(Q'|L)$ and $P(Q')$, we estimate them according to

$$P(Q'|L) = \prod\nolimits_{w \in Q'} P(w|L) \tag{8}$$

$$P(Q') = \prod\nolimits_{w \in Q'} P(w|C). \tag{9}$$

Substituting Eqs. (7)–(9) into Eq. (6), we have

$$\begin{aligned} Q'^* &= \arg\max_{Q'} P(L|Q') \\ &= \arg\max_{Q'} P(L) \frac{\prod_{w \in Q'} P(w|L)}{\prod_{w \in Q'} P(w|C)} \cdot \end{aligned} \tag{10}$$

For a given L, $P(L)$ is a fixed value, therefore we can neglect it in the optimization problem

$$Q'^* = \arg\max_{Q'} \prod\nolimits_{w \in Q'} \frac{P(w|L)}{P(w|C)}. \tag{11}$$

Eq. (11) indicates that the optimal visual query Q'^* should consist of the top-K visual words which have the highest $\frac{P(w|L)}{P(w|C)}$.

Comparing Eq. (11) and Eq. (3), we can find that the visual queries reconstructed via those two methods are equivalent.

Query Reconstruction Via Visual Processing: It is challenging to reconstruct an accurate visual query from the highly noisy image list L. If we can reduce the noise in L to some extent, a better Q' would be derived. Therefore, instead of directly using L for query reconstruction, we propose to use the de-noised L in Eq. (2). Visual reranking has proven effective to refine the image search results, therefore we adopt VisualRank to reorder the images in L to obtain an improved image list L^* [43]–[48].

VisualRank is proposed by Jing and Baluja [48]. They applied the well-known PageRank algorithm to rerank Google image search by treating images as documents and their visual similarities as probabilistic hyper-links. For query Q and its N ranked images in L, VisualRank constructs a graph $\mathcal{G}$ with the images as the nodes and the edges between them being weighted by visual similarity. We denote the visual similarity matrix as $\mathbf{W} \in \mathbb{R}^{N \times N}$. The (i, j) element w_{ij} in $\mathbf{W}$ denotes the similarity between the i-th ranked image and j-th ranked image in L. In this paper, we use the popular Bag-of-Visual-Words representation with the intersection kernel to calculate the visual similarity. Then, reranking is formulated as a random walk process over the graph with stochastic matrix $\mathbf{P}$ derived by column normalizing the similarity matrix $\mathbf{W}$. The state probability of the nodes $\mathbf{r}$ is iteratively updated as

$$\mathbf{r} = \mu \mathbf{P} \mathbf{r} + (1 - \mu)\mathbf{v} \tag{12}$$

where μ is the trade-off parameter and $\mathbf{v}$ is a damping vector which is set according to the images' ranks in L.

The stationary state probability of the random walk process is regarded as the reranked scores for the images. We derive improved L^* by sorting the N images according to their reranked scores in descending order. We then replace L in Eq. (2) with L^* for query reconstruction.

B. Query Reconstruction Error Calculation

After reconstructing the visual query Q' from the returned list of images L, we attempt to calculate the reconstruction error, i.e., $d(Q, Q')$- the distance between the original textual query Q and the visual query Q' for measuring the query difficulty.

The challenge in calculating $d(Q, Q')$ is that Q and Q' are in two different domains, *i.e.,* textual and visual domains. To tackle this problem, we submit the visual query Q' to a content-based image retrieval system and obtain a list of images L' returned for it. Instead of directly calculating the distance between Q and Q', we use $d(L, L')$- the distance between L and L' to approximate it. This approximation has two advantages. First, it successfully solves the domain gap problem for comparing Q and

Q'. Second, the search result list can be treated as an extension of the queries. Compared with $d(Q, Q')$, which is directly defined on two simple queries, $d(L, L')$ derived by comparing two image lists is more robust since auxiliary knowledge is involved in the search result lists.

Image Retrieval for Reconstructed Query Q' : We propose to use the query likelihood model to conduct the retrieval for reconstructed query Q'. It estimates the probability $P(I|Q')$ that an image I is relevant to a given query Q'. Using Bayes' rule, it can be rewritten as

$$\begin{aligned} P(I|Q') &= \frac{P(I)}{P(Q')} P(Q'|I) \\ &= \frac{P(I)}{P(Q')} \prod\nolimits_{w \in Q'} P(w|I) \end{aligned} \tag{13}$$

where $P(I)$ is the prior probability of image I to be relevant to Q'. $P(w|I)$ is the term frequency of visual word w occurring in image I. Since $P(Q')$ is the same for all images, it can be ignored. However, due to the sparsity of $P(w|I)$, the $P(Q'|I)$ will be zero for many images. To tackle this problem, the Jelinek-Mercer smoothing method is applied [1], [49]

$$P(I|Q') = \frac{P(I)}{P(Q')} \prod\nolimits_{w \in Q'} (\lambda P(w|I) + (1 - \lambda) P(w|C)). \tag{14}$$

The $\lambda \in [0, 1]$ is the trade-off parameter and is empirically set as 0.8 throughout this paper.

For $P(I)$, in general image retrieval, each image in collection C has equal $P(I)$ since there is no prior information available. However, in our case, the initial search result list L has provided us with such prior information. Those images which are in collection C but not in L are regarded as irrelevant to Q by the text-based image search engine. Since Q' and Q is relevant to some extent, those images have high probability to be irrelevant to Q' also. Therefore, we neglect those images which are not in L and only consider the images in L for ranking. This process has two advantages. First, it uses the initial text-based search result for pre-filtering, which can prevent lots of noisy images in the content-based image retrieval. Second, the efficiency of the retrieval process can be largely improved since only N images need to be ranked.

We rank N images in L according to $P(I|Q')$ in descending order to derive the new ranked image list L' for query Q'

$$L' = \{I'_1, I'_2, \cdots, I'_N\}, \tag{15}$$

where I'_i is the image with the i-th highest $P(I|Q')$.

Query Difficulty Estimation: As aforementioned, the query difficulty estimation turns into the task of measuring the distance between L and L'. In this paper, we propose to measure $d(L, L')$ via the *Kullback-Leibler* divergence of the visual word probability distributions of L' from L.

Specifically, to measure the query difficulty (or query performance) of Q at truncation level T, the distance $d(L, L')@T$ is defined as the KL-divergence between the language models of the top-T ranked images in L and in L'

$$\begin{aligned} d(L, L')@T &= D_{KL}\{P(w|L_T)|P(w|L'_T)\} \\ &= \sum\nolimits_{w \in V} P(w|L_T) \log \frac{P(w|L_T)}{P(w|L'_T)} \end{aligned} \tag{16}$$

where L_T and L'_T denote the truncated lists of L and L' respectively. L_T (L'_T) consists of the top-T ranked images in $L(L')$. The reason why we introduce a truncation level T here is that the query performance at different truncation levels is different. The ground truth search performance measurements, such as AP and NDCG, are also defined according to different truncation levels: AP@T , NDCG@T [50].

$P(w|L_T)$ and $P(w|L'_T)$ in Eq. (16) are the language models for L_T and L'_T. They are estimated according to Eq. (17) and Eq. (18), respectively

$$P(w|L_T) = \sum\nolimits_{I \in L_T} P(w|I) P(I|L_T), \tag{17}$$

$$P(w|L'_T) = \sum\nolimits_{I \in L'_T} P(w|I) P(I|L'_T). \tag{18}$$

$P(w|I)$ is the term frequency of visual word w occurring in image I. $P(I|L_T)$ and $P(I|L'_T)$ indicate the importance of I in the language model estimation. Here we utilize two different strategies for it. The first strategy assigns equal weights for the images in L_T and L'_T

$$P(I|L_T) = \begin{cases} \frac{1}{T}, & I \in L_T \\ 0, & \text{others} \end{cases} \tag{19}$$

$$P(I|L'_T) = \begin{cases} \frac{1}{T}, & I \in L'_T \\ 0, & \text{others} \end{cases} \tag{20}$$

The other strategy assigns a weight to the ranking position. This assumes that an image with a higher rank is more important, and therefore should have a larger weight. It defines the weight as a non-linear decreasing function of the rank of I in L_T and L'_T

$$P(I|L_T) = \begin{cases} \frac{1 + \sum\limits_{i=r}^{T} \frac{1}{i+1}}{2T}, & I \in L_T \text{ and its rank in } L_T \text{ is } r \\ 0, & \text{others} \end{cases} \tag{21}$$

$$P(I|L'_T) = \begin{cases} \frac{1 + \sum\limits_{i=r}^{T} \frac{1}{i+1}}{2T}, & I \in L'_T \text{ and its rank in } L'_T \text{ is } r \\ 0, & \text{others} \end{cases} \tag{22}$$

We denote the proposed query reconstruction error (QReCE)-based query difficulty estimation method with those two different weighting strategies as eq-QReCE and w-QReCE respectively. We will analyze their performance later.

The overall procedure of our proposed query reconstruction error-based query difficulty estimation method is summarized in Algorithm 1.

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

Fig. 3. Example images in Web353 dataset.

Algorithm 1 Query Reconstruction Error-Based Query Difficulty Estimation (**QReCE**)

```
Input: the given query $Q$, a returned image list $L = \{I_i\}_{i=1}^{N}$,truncation level $T$, $K$, codebook $V = \{w_i\}_{i=1}^{|V|}$.
Output: Query difficulty level $QD(Q) = d(L, L')@T$
1) Query Reconstruction:
Construct language model $P(w|L)$ by Eq. (1);
for each $w \in V$ do
    Compute $score(w)$ by Eq. (3) or Eq. (4);
end for
Rank all visual words according to $score(w)$ in descending order;
Derive the reconstructed query $Q' = \{w'_k, k = 1, \cdots, K\}$;
2) Image Retrieval for Reconstructed Query:
for each image $I \in L$ do
    Compute $P(I|Q')$ by Eq. (13);
end for
Rank images according to $P(I|Q')$ in descending order to obtain a new list $L' = \{I'_j\}_{j=1}^{N}$;
3) Query Difficulty Estimation:
Construct language models for $L_T$ and $L'_T$ by Eq. (17) and Eq. (18);
Compute $d(L, L')@T$ by Eq. (16);
return $d(L, L')@T$
```

IV. Experiments

In this section, we investigate the effectiveness of our proposed image search query difficulty estimation method by conducting experiments on real Web image search datasets. Several state-of-the-art query difficulty estimation methods [11], [12], [13] for image search are taken as baselines.

A. Experimental Setup

Dataset: In order to demonstrate the capacity of the proposed query difficulty estimation method, we conduct experiments on a large public Web image dataset "Web353"[51]. This dataset consists of the top-ranked images returned for 353 textual queries by a popular image search engine. The 353 queries are diverse in topics, which contain landmarks, people, animal, plant, sports, flag, and instruments, etc. For each query, about top-200 ranked images are collected. Each image is manually labeled as relevant or irrelevant to the corresponding query. The ground truth search performance is calculated based on the labels. Fig. 3 shows some example images in this dataset.

Ground-Truth Performance and Correlation Measurement: For each query, we can derive the ground truth search performance based on the manual relevance labels. Here, it is measured via the commonly used truncated average precision (AP) at different truncation level T. To evaluate the effectiveness of the query difficulty estimation method, the widely used correlation coefficient criterion is adopted. It measures the correlation coefficient between the ground truth performance of all queries and the predicted ones given by query difficulty estimation methods. Since our $d(L, L')$ is inversely proportional to the query performance, we report the correlation coefficient between $\mathbf{r} = [AP^{(Q_1)}@T, \cdots, AP^{(Q_M)}@T]$ and $\mathbf{r}' = [-d^{(Q_1)}(L, L')@T, \cdots, -d^{(Q_M)}(L, L')@T]$, where M is the number of queries in the dataset, $AP^{(Q_i)}@T$ is the ground truth performance of the i-th query, and $d^{(Q_i)}(L, L')@T$ is the predicted query difficulty of the i-th query. The higher the correlation coefficient is, the better the query difficulty estimation method performs.

In query difficulty estimation, the widely used correlation measurements include Pearson's r liner correlation [52], nonparametric rank correlation Kendall's τ [53], and Spearman's ρ [54]. All the above three correlation coefficients vary from -1 to 1, where -1 means perfect reverse and 1 means perfect agreement. In our experiments, all three correlation measurements are adopted.

Image Representation: As aforementioned, we adopt the BOVWs model for image representation. The scale-invariant-feature transform (SIFT) [38] features are extracted from each image with dense sampling. Then we build a codebook with 1000 visual words by clustering all the SIFT features via K-means clustering. Each SIFT feature is quantized into the nearest visual word. Finally, each image is viewed as a visual word based document.

Baselines: In this paper, we compare the proposed query reconstruction error (QReCE)-based query difficulty estimation method with several recently proposed ones, including Visual Clarity Score (VCS) [13], Coherence Score (CoS) [11], Representativeness Score (RS) [13], and Inner Coherence Score (ICS) [12]. VCS is a variant of Clarity Score [1] applied to the image search query difficulty estimation. It measures query difficulty via the difference between the language model of the returned images and the language model of the whole image collection. The CoS measures the portion of coherent image pairs in the top-ranked images. A pair of images is coherent if their similarity exceeds a certain threshold which is empirically set. The RS is defined as the mean of the density of the top ranked images in the returned results. The density is estimated via kernel density estimation. The ICS is defined as the diameter of the pseudo relevant image set R in the returned images. A small ICS reflects a compact set of pseudo relevant images, which stands for a good performance.

B. Experimental Results

In this section, we first analyze the performance of our QReCE-based query difficulty estimation approach. Then we compare our method with other baseline methods. Since users only focus on the performance of images returned in the first several pages, we test different T_s and report the correlation coefficients at truncation levels of $T \in \{20, 40, 60\}$.

TABLE I
THE CORRELATION COEFFICIENTS COMPARISON OF THE PROPOSED QRECE USING CHI AND KL FOR QUERY RECONSTRUCTION

		Kendall's τ (P-value)	Pearson's r (P-value)	Spearman's ρ (P-value)
T=20	KL	0.307 (*8.2e-18*)	0.438 (*5.2e-18*)	0.445 (*1.6e-18*)
	CHI	**0.314** (***1.5e-18***)	**0.458** (***1.1e-19***)	**0.453** (***2.9e-19***)
T=40	KL	0.272 (*2.5e-14*)	0.373 (*4.3e-13*)	0.392 (*2.1e-14*)
	CHI	**0.286** (***1.2e-15***)	**0.384** (***7.9e-14***)	**0.411** (***7.7e-16***)
T=60	KL	0.247 (*4.2e-12*)	0.307 (*4.1e-09*)	0.363 (*2.0e-12*)
	CHI	**0.260** (***3.0e-13***)	**0.317** (***1.1e-09***)	**0.383** (***9.0e-14***)

TABLE II
THE CORRELATION COEFFICIENTS COMPARISON OF THE PROPOSED QRECE USING ORIGINAL L AND IMPROVED L^* FOR QUERY RECONSTRUCTION

		Kendall's τ (P-value)	Pearson's r (P-value)	Spearman's ρ (P-value)
T=20	QReCE with original L	0.314 (*1.5e-18*)	0.458 (*1.1e-19*)	0.453 (*2.9e-19*)
	QReCE with improved L^*	**0.343** (***<1e-20***)	**0.494** (***<1e-20***)	**0.493** (***<1e-20***)
T=40	QReCE with original L	0.286 (*1.2e-15*)	0.384 (*7.9e-14*)	0.411 (*7.7e-16*)
	QReCE with improved L^*	**0.343** (***<1e-20***)	**0.484** (***<1e-20***)	**0.493** (***<1e-20***)
T=60	QReCE with original L	0.260 (*3.0e-13*)	0.317 (*1.1e-09*)	0.383 (*9.0e-14*)
	QReCE with improved L^*	**0.308** (***5.8e-18***)	**0.407** (***1.6e-15***)	**0.446** (***1.3e-18***)

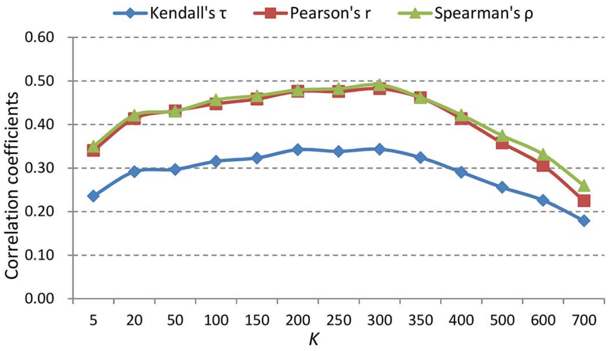


Fig. 4. The correlation coefficients of our QReCE-based query difficulty estimation with different K. It shows that a moderate K is preferred and the performance is steady when K is around 200-300. ($T = 40$).

Effect of Visual Word Ranking Functions in Query Reconstruction: In our proposed QReCE-based query difficulty estimation, the query reconstruction plays the crucial role. As discussed in Section III-A, we apply two functions, CHI and KL, to rank the visual words for reconstructing the visual query. Table I presents the correlation coefficients of our method with the visual query reconstructed using those two functions respectively. It shows that the CHI outperforms KL consistently over different T_s. Comparing Eq. (3) and Eq. (4), we can find that both CHI and KL qualify the importance of visual words according to the ratio $P(w|L)/P(w|C)$. The difference between them is that an additional weight $P(w|L)$ is multiplied to $\log(P(w|L)/P(w|C))$ in KL. It indicates that KL tends to select visual words which have a higher frequency in L. However, visual words in images are not as discriminative as the textual words in documents. There are many noisy visual words which might be extracted from backgrounds but have high frequency. Therefore, the visual words selected via KL may be noisier and lead to a lower performance. In the following experiments, the CHI is applied for visual query reconstruction.

Query Reconstruction With Improved L^* : As discussed in Section III-A, it is challenging to reconstruct an accurate visual query from the highly noisy image list L. Therefore, instead of directly using the initial image list, we replace L in Eq. (2) with the improved image list L^* for query reconstruction. Table II shows the correlation coefficients comparison between the QReCE-based query difficulty estimation using improved L^* and original L. From the results, we can find that the performance is boosted when L^* is adopted for query reconstruction. The underlying reasons are explained as follows. From Eq. (1), we can see that the estimation of $P(I|L)$ plays a crucial rule in visual query reconstruction. $P(I|L)$ indicates the importance of I in language model building. $P(I|L)$ is weighted by the image's ranking position in L in Eq. (2). It is reasonable since an image with a higher rank is more likely to be relevant and consequently, more important. Beyond the ranking position, the importance of an image is also reflected by its popularity or representativeness (termed "authority" in VisualRank) among all images in L. Therefore, we apply VisualRank to measure the "authority" of images. By replacing the L in Eq. (2) with L^*, a better estimation of $P(I|L)$ will be derived, which leads to a better query reconstruction result.

Effect of Free-Parameter K: In our QReCE-based query difficulty estimation method, there is one free-parameter K, the number of visual words in the reconstructed query Q'. We analyze the sensitivity of our method to this parameter by varying K from 5 to 700. The experimental results are shown in Fig. 4. It shows that the performance increases when K grows from 5 to 150, and is quite stable when K is around 200-300. When K is larger than 350, the performance gradually decreases. Since a single visual word contains limited information, the reconstructed query with small K is not informative enough to summarize L. When K is too large, the reconstructed query may involve noisy visual words which are less discriminative. Therefore, a moderate K is preferred. As shown in Fig. 4, our method achieves a steady performance when K is around 200-300. We empirically set $K = 300$ in this paper.

TABLE III
CORRELATION COEFFICIENTS AND P-VALUES OF OUR PROPOSED METHOD AS WELL AS BASELINE METHODS ON WEB353 DATASET

Method		VCS	CoS	RS	ICS	eq-QReCE	w-QReCE
T=20	Kendall's τ	0.120	0.310	0.195	0.263	0.316	**0.343**
	(P-value)	(*7.6e-04*)	(*5.2e-18*)	(*4.3e-08*)	(*1.7e-13*)	(*8.0e-19*)	(<***1e-20***)
	Pearson's r	0.155	0.462	0.287	0.381	0.453	**0.494**
	(P-value)	(*3.5e-03*)	(*5.0e-20*)	(*4.0e-08*)	(*1.2e-13*)	(*2.7e-19*)	(<***1e-20***)
	Spearman's ρ	0.182	0.448	0.286	0.387	0.458	**0.493**
	(P-value)	(*6.1e-04*)	(*8.7e-19*)	(*4.4e-08*)	(*4.5e-14*)	(*1.1e-19*)	(<***1e-20***)
T=40	Kendall's τ	0.132	0.253	0.178	0.229	0.327	**0.343**
	(P-value)	(*4.0e-06*)	(*1.6e-12*)	(*5.9e-07*)	(*1.4e-10*)	(*5.4e-20*)	(<***1e-20***)
	Pearson's r	0.187	0.390	0.228	0.324	0.448	**0.484**
	(P-value)	(*4.1e-04*)	(*3.0e-14*)	(*1.5e-05*)	(*4.5e-10*)	(*8.8e-19*)	(<***1e-20***)
	Spearman's ρ	0.194	0.365	0.262	0.337	0.469	**0.493**
	(P-value)	(*2.4e-04*)	(*1.4e-12*)	(*6.2e-07*)	(*8.1e-11*)	(*1.0e-20*)	(<***1e-20***)
T=60	Kendall's τ	0.138	0.232	0.175	0.222	0.274	**0.308**
	(P-value)	(*1.1e-04*)	(*8.0e-11*)	(*9.1e-07*)	(*4.7e-10*)	(*1.6e-14*)	(***5.8e-18***)
	Pearson's r	0.216	0.353	0.189	0.285	0.359	**0.407**
	(P-value)	(*4.4e-05*)	(*8.2e-12*)	(*3.6e-04*)	(*5.3e-08*)	(*3.7e-12*)	(***1.6e-15***)
	Spearman's ρ	0.202	0.335	0.257	0.325	0.401	**0.446**
	(P-value)	(*1.3e-04*)	(*1.0e-10*)	(*1.0e-06*)	(*4.2e-10*)	(*4.6e-15*)	(***1.3e-18***)

Comparison Between Our Proposed Method and Baselines: We compare our proposed QReCE-based query difficulty estimation method with several state-of-the art baselines. The correlation coefficients and the corresponding P-values are presented in Table III. For CoS, RS and ICS, we have tried various parameter settings and report the highest correlation coefficients. The crucial parameter in CoS is the visual threshold. It is defined as $P\%$ of image pairs in the dataset that have a smaller visual similarity than this value. We varied P from 5 to 95 at intervals of 5, to find the best threshold empirically. For RS, the key parameter is the size of the neighbors in KDE (kernel density estimation). We also varied this parameter from 5 to 30 at intervals of 5, to find the best one. For ICS, we take the top-T ranked images in L as the pseudo relevant set. eq-QReCE and w-QReCE denotes the equal or ranked weighted strategies adopted in Eq. (17) and Eq. (18) for language model building.

Among the four baseline methods, we can see that CoS and ICS yield mediocre performances. The correlation coefficients for CoS and ICS with the retrieval performance become continuously worse with the increase of T. The reason is that, when T increases, more and more irrelevant images occur in the top-ranked image results and thus it is more difficult to precisely measure the tightness of the returned images precisely. For VCS, the correlation coefficients are much worse than others, however, its performance improves with the increase of T. [1] illustrates that the clarity method needs a great number of documents to adequately measure the coherence of the ranked list. Thus, the performance of VCS is poor at a small value of T.

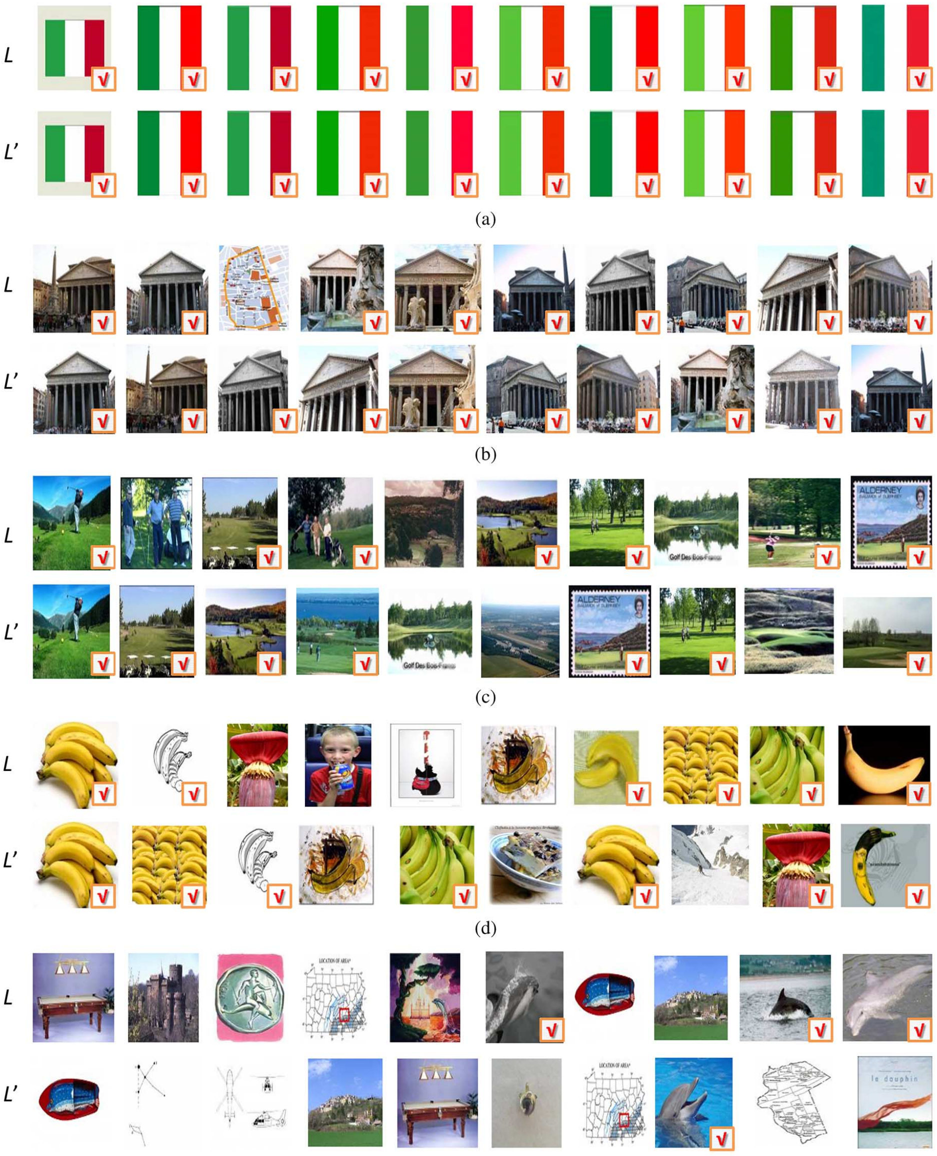
From Table III, we can observe that our approach consistently outperforms other methods over all T_s. The P-value is far less than 0.05, which indicates that the correlation between the estimated query difficulty and the ground truth search performance is statistically significant. w-QReCE achieves better performance than eq-QReCE, demonstrating that a ranking position weighted strategy is more effective than an equal weight strategy for building language models for L and L' [Eq. (17) and Eq. (18)]. We also observe that the performance of our model is better at a smaller T. This is because as T increases, the overlap between L_T and L'_T grows. As a consequence, their difference is smaller according to Equation (16), resulting in lower correlation coefficients.

Fig. 5 further illustrates the top-10 ranked images (ordered left to right) in L and L' for several queries, including “flag italy”, “pantheon rome”, “golf course”, “banana”, and “dolphin”. L is the initial ranked image list returned for the original textual query Q. The L' is the ranked image list returned for the reconstructed visual query Q'. Query relevant images are marked by a red “$\surd$”. It shows that the better L is (higher AP), the L' is more similar with L, resulting a smaller distance between L and L' (and *vice versa*). Those examples show that the proposed QReCE method can accurately indicate the query difficulty level.

Comparison Between QReCE and Supervised Regression Method: In [13], a general supervised query difficulty estimation framework, CombRegression, is proposed. It combines several query difficulty estimators via a regression model. In CombRegression, several predictors including CoS, VCS, and RS, are concatenated into a feature vector to represent a query. Then, a supervised regression model is trained on a training set to derive a query difficulty estimation model. We also compare our unsupervised QReCE method with this supervised one, as shown in Table IV. We find that our simple query difficulty prediction performs even better than this complex, supervised regression model which combines several predictors. As previously discussed, those predictors used in CombRegression only reflect the statistical characteristics of the returned images, while our method further explores the relationship between the query and the returned images. In other words, our QReCE is complimentary to those predictors (CoS, VCS, RS) used in CombRegression. Therefore, incorporating QReCE into CombRegression (combining QReCE and predictors already used in CombRegression) can further improve the query difficulty estimation performance.

C. Experiments on MSRA-MM Dataset

We also test our approach on the dataset MSRA-MM [55]. This dataset collected 60257 images from Microsoft Bing



(e)

Fig. 5. Top 10 ranked images in L and L', ordered left to right. Query relevant images are marked by a red "$\surd$". L is the initial ranked image list returned for the original textual query Q. L' is the ranked image list returned for the reconstructed visual query Q'. It shows that the better L is (higher AP), L' is more similar with L, resulting in a smaller distance between L and L' (and vice versa). Those examples show that the proposed QReCE method can well indicate the query difficulty level. (a) Query "flag italy". Ground truth performance of L is $\mathrm{AP@10} = 1$, the distance between L and L' is 0 (b) Query "pantheon rome". Ground truth performance of L is $\mathrm{AP@10} = 0.79$, the distance between L and L' is 0.06 (c) Query "golf course". Ground truth performance of L is $\mathrm{AP@10} = 0.52$, the distance between L and L' is 0.14 (d) Query "banana". Ground truth performance of L is $\mathrm{AP@10} = 0.41$, the distance between L and L' is 0.24 (e) Query "dolphin". Ground truth performance of L is $\mathrm{AP@10} = 0.07$, the distance between L and L' is 0.45.

TABLE IV
CORRELATION COEFFICIENTS COMPARISON BETWEEN QRECE AND THE SUPERVISED REGRESSION METHOD WHICH COMBINES SEVERAL QUERY DIFFICULTY PREDICTORS. ($T = 40$)

	Kendall's τ (P-value)	Pearson's r (P-value)	Spearman's ρ (P-value)
CombRegression [13]	0.317 (*7.0e-19*)	0.431 (<*1e-20*)	0.465 (<*1e-20*)
w-QReCE	**0.343** (<***1e-20***)	**0.484** (<***1e-20***)	**0.493** (<***1e-20***)

image search for 68 representative queries. For each image, its relevance to the corresponding query is labeled with three levels: very relevant, relevant and irrelevant. These three levels are indicated by scores 2, 1 and 0, respectively. Here we adopt the truncated normalized discounted cumulative gain (NDCG) [50], which is widely used for graded relevance judgments, to measure the ground-truth performance for each query in this dataset. Other experimental settings are the same as in Web353.

Table V shows the correlation coefficients comparison of our approach and four baseline methods with different performance metrics. From the results, we can see that our approach achieves the best performance in almost all cases. We also compare our unsupervised QReCE method with the supervised regression query difficulty estimation method in [13]. The results are shown in Table VI. It also demonstrates that our method outperforms the supervised regression model which combines several query difficulty predictors.

TABLE V
CORRELATION COEFFICIENTS AND P-VALUE OF QUERY DIFFICULTY PREDICTION METHODS ON MSRA-MM DATASET

Method		VCS	CoS	RS	ICS	w-QReCE
T=20	Kendall's τ	0.037	0.200	0.150	0.178	**0.305**
	(P-value)	(*6.6e-01*)	(*1.8e-02*)	(*7.1e-02*)	(*3.2e-02*)	(***2.4e-04***)
	Pearson's r	0.072	0.226	0.164	0.220	**0.299**
	(P-value)	(*5.6e-01*)	(*6.4e-02*)	(*1.8e-01*)	(*7.1e-02*)	(***1.3e-02***)
	Spearman's ρ	0.061	0.292	0.207	0.257	**0.451**
	(P-value)	(*6.2e-01*)	(*1.6e-02*)	(*9.0e-02*)	(*3.5e-02*)	(***1.3e-04***)
T=40	Kendall's τ	0.129	0.183	0.174	0.205	**0.282**
	(P-value)	(*1.2e-01*)	(*2.8e-02*)	(*3.6e-02*)	(*1.3e-02*)	(***6.9e-04***)
	Pearson's r	0.066	0.299	0.199	0.272	**0.335**
	P-value)	(*5.9e-01*)	(*1.3e-02*)	(*1.0e-01*)	(*2.5e-02*)	(***5.2e-03***)
	Spearman's ρ	0.171	0.285	0.261	0.294	**0.448**
	(P-value)	(*1.6e-01*)	(*1.8e-02*)	(*3.2e-02*)	(*1.5e-02*)	(***1.5e-04***)
T=60	Kendall's τ	0.088	0.246	0.204	0.220	**0.300**
	(P-value)	(*2.9e-01*)	(*3.0e-03*)	(*1.4e-02*)	(*8.0e-03*)	(***3.0e-04***)
	Pearson's r	0.091	0.315	0.226	0.249	**0.320**
	(P-value)	(*4.6e-01*)	(*9.0e-03*)	(*6.4e-02*)	(*4.1e-02*)	(***7.8e-03***)
	Spearman's ρ	0.127	0.355	0.284	0.320	**0.455**
	(P-value)	(*3.0e-01*)	(*3.0e-03*)	(*1.9e-02*)	(*8.0e-03*)	(***1.2e-04***)

TABLE VI
CORRELATION COEFFICIENTS COMPARISON BETWEEN QRECE AND THE SUPERVISED REGRESSION METHOD WHICH COMBINES SEVERAL QUERY DIFFICULTY PREDICTORS ON MSRA-MM DATASET. ($T = 20$)

	Kendall's τ (P-value)	Pearson's r (P-value)	Spearman's ρ (P-value)
CombRegression [13]	0.203 (*1.5e-02*)	0.268 (*2.7e-02*)	0.317 (*9.0e-03*)
w-QReCE	**0.305** (***2.4e-04***)	**0.299** (***1.3e-02***)	**0.451** (***1.3e-04***)

TABLE VII
CORRELATION COEFFICIENTS AND P-VALUE OF QUERY DIFFICULTY PREDICTION METHODS ON NUS-WIDE-OBJECT DATASET

Method	VCS	CoS	RS	ICS	w-QReCE
Kendall's τ	0.246	0.041	0.039	0.166	**0.284**
(P-value)	(*1.2e-02*)	(*7.0e-01*)	(*7.1e-01*)	(*9.1e-02*)	(***3.8e-03***)
Pearson's r	0.380	0.087	0.022	0.253	**0.408**
(P-value)	(*6.4e-03*)	(*5.5e-01*)	(*8.8e-01*)	(*7.6e-02*)	(***3.2e-03***)
Spearman's ρ	0.345	0.065	0.035	0.253	**0.425**
(P-value)	(*1.4e-02*)	(*6.6e-01*)	(*8.5e-01*)	(*7.6e-02*)	(***2.1e-03***)

TABLE VIII
CORRELATION COEFFICIENTS COMPARISON BETWEEN QRECE AND THE SUPERVISED REGRESSION METHOD WHICH COMBINES SEVERAL QUERY DIFFICULTY PREDICTORS ON NUS-WIDE-OBJECT DATASET

	Kendall's τ (P-value)	Pearson's r (P-value)	Spearman's ρ (P-value)
CombRegression [13]	0.252 (*1.1e-02*)	0.385 (*5.9e-03*)	0.373 (*7.3e-03*)
w-QReCE	**0.284** (***3.8e-03***)	**0.408** (***3.2e-03***)	**0.425** (***2.1e-03***)

D. Experiments on NUS-WIDE-OBJECT Dataset

We further conduct experiments on the benchmark dataset NUS-WIDE-OBJECT [56]. This dataset consists of 30000 images belonging to 31 object categories. We follow the experimental settings in [12] and randomly select 50 images as queries. The Bag-of-Visual-Words features are used for image representation. Other experimental settings are the same as in Web353 and MSRA-MM.

We first compare our approach with the four baseline methods. The experimental results are presented in Table VII. We can see that our approach achieves the best performance on this dataset too. For the four baseline methods, the VCS outperforms the others while CoS and RS are the worst. The reason why CoS and RS fail on this dataset is that the returned images are all visually similar with the query image. Therefore the coherence score and representativeness score are high for almost all queries, leading to low discriminative ability. Due to the limited space, we only report the experimental results when $T = 20$ in Table VII. For $T = 40$ and $T = 60$, we have similar observations. As in Web353 and MSRA-MM, we also compare our unsupervised QReCE method with the supervised regression query difficulty estimation method CombRegression in [13]. The results are shown in Table VIII. It also demonstrates that our method outperforms the supervised regression model, which combines several query difficulty predictors.

The last experiment conducted on this dataset compares our proposed query reconstruction method with existing query expansion methods. As discussed in Section II, our query reconstruction is somewhat similar to query expansion. Here, we

TABLE IX
COMPARISON OF QUERY EXPANSION METHODS. WE REPLACE THE QUERY RECONSTRUCTION METHOD IN QRECE WITH THE QUERY EXPANSION METHODS PROPOSED IN [23] AND [24] RESPECTIVELY. THE CORRELATION COEFFICIENTS OF QRECE WITH VARYING QUERY EXPANSION/RECONSTRUCTION METHODS ARE REPORTED. (NUS-WIDE-OBJECT DATASET)

	Kendall's τ (P-value)	Pearson's r (P-value)	Spearman's ρ (P-value)
[23]	0.212 (*3.1e-02*)	0.244 (*8.8e-02*)	0.307 (*3.0e-02*)
[24]	0.222 (*2.4e-02*)	0.314 (*2.7e-02*)	0.336 (*1.7e-02*)
Our	**0.284** (***3.8e-03***)	**0.408** (***3.2e-03***)	**0.425** (***2.1e-03***)

compare our proposed method with two representative query expansion methods which are proposed in [23] and [24]. We test QReCE with our proposed query reconstruction method and the two query expansion methods applied respectively. The correlation coefficients of QReCE with varying query expansion/reconstruction methods are reported in Table IX. It shows that our proposed query reconstruction method performs the best, which demonstrates the effectiveness of our method.

E. Application Discussion

For a given query Q and its search result list L, the output of our method is a value which indicates the performance of L. As shown in Fig. 5, a larger $d(L, L')$ indicates that L has a lower search performance (AP), and *vice versa*. In other words, if we have multiple search result lists $\{L_1, L_2, \cdots, L_M\}$ returned for the same query Q via different search engines or search strategies, we can automatically compare the performances of those search result lists. This leads to many attractive applications. For example, for a given query Q, by comparing the performances of its several search result lists returned by different search engines, we can automatically select the best search engine for this query [20], [13]. We can also apply it to guide multiple search result merging, for example distributed search result merging as discussed in [2] and [13]. It can also be used to guide image search reranking as discussed in [12] and query expansion selection as discussed in [15].

In summary, the applications for our query difficulty estimation method are as broad as other methods. Those applications have been thoroughly discussed before [2], [13]–[15], [20]. The only factor that affects their successfulness in those applications is the query difficulty estimation performance. The better the query difficulty estimation performance is, the more successful the applications will be. In this paper, our emphasis is on how to construct a favorable query difficulty estimation method. We have proven that our proposed method achieves the best performance and therefore, it can be successfully applied in various applications.

V. CONCLUSION

In this paper, we propose a novel approach to automatically estimate the query difficulty for Web image search. Our method investigates the relationship between the textual query and the returned images for query difficulty estimation. This is different from existing methods, which only focus on investigating the statistical characteristics of the returned images, and neglect the query itself. We propose a novel method to reconstruct a visual query from the returned images and then adopt the query reconstruction error for query difficulty estimation. Extensive experiments on two real world image datasets demonstrate the effectiveness of our proposed method.

REFERENCES

[1] S. Cronen-Townsend, Y. Zhou, and W. B. Croft, "Predicting query performance," in *Proc. ACM SIGIR Special Interest Group Inf. Retrieval*, 2002, pp. 299–306.

[2] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow, "Learning to estimate query difficulty: Including applications to missing content detection and distributed information retrieval," in *Proc. ACM SIGIR Special Interest Group Inf. Retrieval*, 2005, pp. 512–519.

[3] Y. Zhou and W. B. Croft, "Ranking robustness: A novel framework to predict query performance," in *Proc. 15th ACM Int. Conf. Inf. Knowl. Manag.*, 2006, pp. 567–574.

[4] D. Carmel, E. Yom-Tov, A. Darlow, and D. Pelleg, "What makes a query difficult," in *Proc. ACM SIGIR Special Interest Group Inf. Retrieval*, 2006, pp. 390–397.

[5] Y. Zhou and W. B. Croft, "Query performance prediction in web search environments," in *Proc. ACM SIGIR Special Interest Group Inf. Retrieval*, 2007, pp. 543–550.

[6] F. Diaz, "Performance prediction using spatial autocorrelation," in *Proc. ACM SIGIR Special Interest Group Inf. Retrieval*, 2007, pp. 583–590.

[7] J. He, M. Larson, and M. De Rijke, "Using coherence-based measures to predict query difficulty," in *Proc. Eur. Conf. Inf. Retrieval*, 2008, pp. 689–694.

[8] H. Imran and A. Sharan, "Co-occurrence based predictors for estimating query difficulty," in *Proc. IEEE Int. Conf. Data Mining Workshops*, Dec. 2010, pp. 867–874.

[9] A. Shtok, O. Kurland, D. Carmel, F. Raiber, and G. Markovits, "Predicting query performance by query-drift estimation," *ACM Trans. Inf. Syst.*, vol. 30, no. 2, pp. 1–11, 2012.

[10] O. Kurland, A. Shtok, S. Hummel, F. Raiber, D. Carmel, and O. Rom, "Back to the roots: A probabilistic framework for query-performance prediction," in *Proc. CIKM*, 2012, pp. 823–832.

[11] S. Rudinac, M. Larson, and A. Hanjalic, "Leveraging visual concepts and query performance prediction for semantic-theme-based video retrieval," *IJMIR*, vol. 1, no. 4, pp. 263–280, 2012.

[12] Y. Li, B. Geng, D. Tao, Z.-J. Zha, L. Yang, and C. Xu, "Difficulty guided image retrieval using linear multiple feature embedding," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1618–1630, Dec. 2012.

[13] X. Tian, Y. Lu, and L. Yang, "Query difficulty prediction for web image search," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 951–962, Aug. 2012.

[14] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. ICCV*, 2003, pp. 1470–1477.

[15] S. Rudinac, M. Larson, and A. Hanjalic, "Exploiting result consistency to select query expansions for spoken content retrieval," in *Proc. Eur. Conf. Inf. Retrieval*, 2010, pp. 645–648.

[16] Y. Li, Y. Luo, D. Tao, and C. Xu, "Query difficulty guided image retrieval system," in *Advances in Multimedia Modeling*, ser. Lecture Notes in Comput. Sci. Berlin, Germany: Springer-Verlag, 2011, pp. 479–482.

[17] X. Xing, Y. Zhang, and M. Han, "Query difficulty prediction for contextual image retrieval," in *Proc. Eur. Conf. Inf. Retrieval*, 2010, pp. 581–585.

[18] C. Kofler, L. Yang, M. Larson, T. Mei, A. Hanjaliic, and S. Li, "When video search goes wrong: Predicting query failure using search engine logs and visual search results," in *Proc. ACM Multimedia*, 2012, pp. 319–328.

[19] C. Kofler, L. Yang, M. Larson, T. Mei, A. Hanjalic, and S. Li, "Predicting failing queries in video search," *IEEE Trans. Multimedia*, vol. 16, no. 7, pp. 1973–1985, Nov. 2014.

[20] X. Tian, Y. Lu, L. Yang, and Q. Tian, "Learning to judge image search results," in *Proc. ACM Multimedia*, 2011, pp. 363–372.

[21] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *Proc. ICCV*, 2007, pp. 1–8.

[22] A. Joly and O. Buisson, "Logo retrieval with a contrario visual query expansion," in *Proc. ACM Multimedia*, 2009, pp. 581–584.

[23] O. Chum, A. Mikulk, M. Perdoch, and J. Matas, "Total recall II: Query expansion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 889–896.

[24] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2911–2918.

[25] H. Xie, Y. Zhang, J. Tan, L. Guo, and J. Li, "Contextual query expansion for image retrieval," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1104–1114, Jun. 2014.

[26] X. Li, C. G. Snoek, and M. Worring, "Learning tag relevance by neighbor voting for social image retrieval," in *Proc. MIR*, 2008, pp. 180–187.

[27] G.-J. Qi, C. C. Aggarwal, Q. Tian, H. Ji, and T. S. Huang, "Exploring context and content links in social media: A latent space method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 850–862, May 2012.

[28] D. Rafailidis, A. Axenopoulos, J. Etzold, S. Manolopoulou, and P. Daras, "Content-based tag propagation and tensor factorization for personalized item recommendation based on social tagging," *ACM Trans. Interactive Intell. Syst.*, vol. 3, no. 4, pp. 1–27, 2014.

[29] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2006, vol. 2, pp. 2169–2178.

[30] D. Liu, G. Hua, P. A. Viola, and T. Chen, "Integrated feature selection and higher-order spatial feature extraction for object categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 590–597.

[31] T. Chen, K.-H. Yap, and D. Zhang, "Discriminative soft bag-of-visual phrase for mobile landmark recognition," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 612–622, Apr. 2014.

[32] J. M. Winn, A. Criminisi, and T. P. Minka, "Object categorization by learned universal visual dictionary," in *Proc. ICCV*, 2005, pp. 1800–1807.

[33] L. Yang, P. Meer, and D. J. Foran, "Multiple class segmentation using a unified framework over mean-shift patches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–8.

[34] I. González-Díaz, C. E. Baz-Hormigos, and F. Díaz-de-María, "A generative model for concurrent image retrieval and ROI segmentation," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 169–183, 2014.

[35] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jan. 2006, pp. 2161–2168.

[36] W. Zhou, M. Yang, H. Li, X. Wang, Y. Lin, and Q. Tian, "Towards codebook-free: Scalable cascaded hashing for mobile image search," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 601–611, Apr. 2014.

[37] B. Liu, Z. Li, L. Yang, M. Wang, and X. Tian, "Real-time video copy-location detection in large-scale repositories," *IEEE MultiMedia*, vol. 18, no. 3, pp. 22–31, Mar. 2011.

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

[39] J. G. Carbonell, Y. Yang, R. E. Frederking, R. D. Brown, Y. Geng, and D. Lee, "Translingual information retrieval: A comparative evaluation," in *Proc. 15th Int. Joint Conf. Artificial Intelligence*, 1997, pp. 708–714.

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

[41] C. Carpineto, R. de Mori, G. Romano, and B. Bigi, "An information-theoretic approach to automatic query expansion," *ACM Trans. Inf. Syst.*, vol. 19, no. 1, pp. 1–27, 2001.

[42] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience, 1991.

[43] 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.

[44] 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.

[45] Y. Liu, T. Mei, and X.-S. Hua, "Crowdreranking: Exploring multiple search engines for visual search reranking," in *Proc. ACM SIGIR Special Interest Group Inf. Retrieval*, 2009, pp. 500–507.

[46] T. Mei, Y. Rui, S. Li, and Q. Tian, "Multimedia search reranking: A literature survey," *ACM Comput. Survey*, vol. 46, no. 3, pp. 1–38, 2014.

[47] T. Yao, C.-W. Ngo, and T. Mei, "Circular reranking for visual search," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1644–1655, Apr. 2013.

[48] 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.

[49] C. Zhai and J. D. Lafferty, "A study of smoothing methods for language models applied to information retrieval," *ACM Trans. Inf. Syst.*, vol. 22, no. 2, pp. 179–214, 2004.

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

[51] J. Krapac, M. Allan, J. Verbeek, and F. Juried, "Improving web image search results using query-relative classifiers," *Proc. IEEE Conf. Comput, Vis, Pattern Recog.*, pp. 1094–1101, 2010.

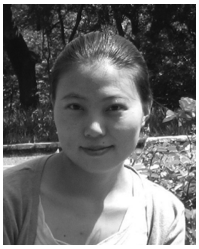
[52] E. Kreyszig, *Advanced Engineering Mathematics*. New York, NY, USA: Wiley, 1997.

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

[54] J. D. Gibbons and S. Chakraborty, *Nonparametric Statistical Inference*. New York, NY, USA: Chapman and Hall, 1992.

[55] M. Wang, L. Yang, and X.-S. Hua, "Msra-mm: Bridging research and industrial societies for multimedia information retrieval," Microsoft Research Asia, Beijing, China, Tech. Rep. MSR-TR-2009-30, 2009.

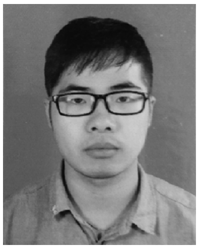
[56] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: A real-world web image database from National University of Singapore," in *Proc. CIVR*, 2009, p. 48.



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 in the CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, University of Science and Technology of China, Hefei, China. She Her current research interests include multimedia information retrieval and machine learning.

Dr. Tian received the Excellent Doctoral Dissertation of Chinese Academy of Sciences award in 2012 and the Nomination of National Excellent Doctoral Dissertation award in 2013.



Qianghuai Jia received the B.S. degree from Hangzhou Dianzi University, Hangzhou, China, in 2012, and is currently working towards the Master's degree from the CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, University of Science and Technology of China, Hefei, China.

He was previously a Research Intern for Netease, Hangzhou, Zhejiang. His research interests lie primarily in multimedia search, information retrieval, and machine learning.

Tao Mei (M'07-SM'11) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively.

He is a Lead Researcher with Microsoft Research, Beijing, China. He has authored or co-authored over 100 papers in journals and conferences and 10 book chapters, and edited three books. He holds 13 U.S. patents and has over 20 pending patents. His current research interests include multimedia information retrieval and computer vision.

Dr. Mei was the recipient of several awards from prestigious multimedia journals and conferences, including the IEEE CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY Best Paper Award in 2014, the IEEE TRANSACTIONS ON MULTIMEDIA Prize Paper Award in 2013, the Best Student Paper Award at the IEEE VCIP in 2012, the Best Paper Award at the ACM ICIMCS in 2012, and the Best Paper Award at ACM Multimedia in 2009 and 2007. He is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, *ACM/Springer Multimedia Systems*, and *Neurocomputing*, and a Guest Editor of six international journals. He was the General Co-Chair of ACM ICIMCS 2013, and the Program Co-Chair of IEEE ICME 2015, IEEE MMSP 2015, and MMM 2013. He is a Senior Member of the ACM.