Multimedia Systems (2016) 22:509–523
DOI 10.1007/s00530-014-0432-7

SPECIAL ISSUE PAPER

The effects of multiple query evidences on social image retrieval

Zhiyong Cheng · Jialie Shen · Haiyan Miao

Published online: 19 December 2014

Abstract System performance assessment and comparison are fundamental for large-scale image search engine development. This article documents a set of comprehensive empirical studies to explore the effects of multiple query evidences on large-scale social image search. The search performance based on the social tags, different kinds of visual features and their combinations are systematically studied and analyzed. To quantify the visual query complexity, a novel quantitative metric is proposed and applied to assess the influences of different visual queries based on their complexity levels. Besides, we also study the effects of automatic text query expansion with social tags using a pseudo relevance feedback method on the retrieval performance. Our analysis of experimental results shows a few key research findings: (1) social tag-based retrieval methods can achieve much better results than content-based retrieval methods; (2) a combination of textual and visual features can significantly and consistently improve the search performance; (3) the complexity of image queries has a strong correlation with retrieval results' quality—more complex queries lead to poorer search effectiveness; and (4) query expansion based on social tags frequently causes search topic drift and consequently leads to performance degradation.

Z. Cheng · J. Shen (✉)
School of Information Systems, Singapore Management University, 80 Stamford Road, Singapore, Singapore
e-mail: jlshen@smu.edu.sg

Z. Cheng
e-mail: zy.cheng.2011@phdis.smu.edu.sg

H. Miao
A*STAR, Institute of High Performance Computing, Singapore, Singapore
e-mail: miaohy@ihpc.a-star.edu.sg

Keywords Query evidence · Social image retrieval · Performance · Evaluation · Experimentation

1 Introduction

With advancement of mobile computing and Web 2.0 technology, recent years have witnessed explosive growth of social images in scale, variety and complexity. Millions of images are uploaded onto different social networks on a daily basis, exemplified by Geo-coded photo services and online visual discovery tools. This provides people a great convenience to archive, share or exchange information about their daily life. However, developing retrieval techniques to facilitate effective management and exploration of such huge image collections has been a big research challenge for a long time [8, 14, 15, 17, 27, 28]. In general, traditional image retrieval methods can be categorized into two independent paradigms: *content-based image retrieval* (CBIR) and *text-based image retrieval* (TBIR) [11, 56]. In CBIR, low-level visual features extracted from the images are used as content representation for different purposes, such as indexing and searching [54]. In the last two decades, CBIR has been a very active research domain. Various kinds of visual features have been proposed to model image contents [11]. In fact, to gain comprehensive image signature, it is impossible to only consider a single type of visual feature. Thus, multiple visual feature combination becomes a natural solution for the problem [30, 63]. While the method is effective in some specialized image search applications, many issues still remain open [11].

The basic idea of the TBIR system is to leverage text annotations to describe the contents of images. Then, text retrieval techniques can be directly applied to support image search [11, 56]. TBIR achieves better search performance over CBIR. However, the performance of TBIR relies on the quality of the text labels attached to each image. In many real applications, text annotations can be incomplete and biased. To improve the effectiveness and efficiency of image annotation, many automated approaches have been developed [20, 25, 71, 72] to label images with one or several keywords. However, the keywords used are often limited to a small size of corpus derived from a well-annotated image collection.

Compared to traditional image retrieval systems, social image search engines need to face entirely different raw visual information, which contains many heterogeneous components (e.g., low-quality text content + images). Specifically, social images enjoy three unique characteristics:

- *Mass scale*: social images sharing websites contain large-scale images and the scale is continuously growing. For example, the number of images hosted by Flickr reached 6 billion in August 2011 [45], and 3.5 million new images were uploaded daily in 2013 [1].
- *Low-quality annotation*: tag is one of the most important components of social images. Generally, the user-contributed text information is known to be ambiguous, incomplete and overly personalized [29, 34, 36].
- *Lossy and highly duplicated visual contents*: for the same concept, we can observe a great variety of visual appearances in the relevant social images [54]. Besides, the visual quality of images could be also uneven, due to the variant photographic skills of different users and the physical conditions for acquiring the images.

- All the characteristics make the development of efficient and effective social image search techniques very challenging. In real life, the notion of similarity is based on high-level semantic concepts. However, due to semantic gap, low-level feature is unable to represent the concepts accurately and comprehensively. Consequently, when using traditional CBIR schemes to search social images, the performance will degrade greatly. Besides, CBIR systems could return many duplicate or near-duplicate images in social image search, which significantly decreases users' satisfaction on the search results. Efficiency is another big concern of CBIR methods due to the large-scale data size [19, 41]. On the other hand, the user-contributed tags provide a convenient way to index and search social images. Tag-based search methods have been widely used in existing social media platforms, such as Flickr and YouTube. However, the user-provided tags are usually noisy [29, 36] and lack importance or relevance level indications [33]. Besides, the numbers and quality of tags assigned greatly vary. These characteristics impair the performance of traditional TBIR methods when applied to social image search. It is critical to carefully study the performance of classical text retrieval models (e.g., OKAPI-BM25 [26] and vector space model [49]) on tag-based social image search.

On the other hand, how to construct query (either visual-based or text-based query) to effectively describe users' search intents is another important research question. It is well known that the performances of different queries could vary over a large range in different image search engines [57]. Even for the same search intent, very different search results can be obtained based on different queries, especially for social image search, as various kinds of visual appearances can be associated with the same concept. Particularly, in CBIR systems, images with different complexity levels cannot be represented consistently. Generally, it is difficult to represent the image query with complex visual content. Consequently, the search performance for complex visual queries (e.g., *crowd street view*) will be poorer than relatively simple visual queries (e.g., *blue sky*). Therefore, studying the effects of visual complexity on the search performance is very crucial for developing high-performance social image search engine. In tag-based social image search, users often use one or two tags to represent their search intents. However, in most cases, one or two tags cannot completely represent users' desired image contents. For example, when a user searches for a picture of *seagull flying above a sea at sunset*, he/she may use "*sunset*" or "*sunset, sea*" as query and wish that the search engine would return what he/she exactly wants. Furthermore, users sometimes only have a rough idea about what they want. As a result, they cannot precisely describe their search intents. In such cases, *interactive search* becomes an effective tool to help users refine their queries to get satisfying search results. Automatic query expansion based on pseudo relevance feedback is a simple and effective method for interactive query refinement in information retrieval [4]. However, the performance of this method over social image search based on social tags has not been studied in deep previously.

Extensive research efforts have been invested into social image retrieval. Most of those efforts focus on developing new approaches or techniques to improve users' satisfaction on search accuracy [13, 16, 36, 61] or other aspects (e.g., result diversification or attractiveness [18, 62]). In contrast, little attention has been devoted to the fundamental research questions. The study of these fundamental research questions is crucial for the development of effective social image search systems. This study mainly focuses on the fundamental research issues related to the retrieval effectiveness of large-scale social image retrieval. The research questions we address include:

1. How do the CBIR systems perform on large-scale social image search by using a single type of visual feature? How much search accuracy can be improved by combining multiple types of visual features?
2. How does the performance of the classical TBIR methods, such as OKAPI-BM25 [26] and TF-IDF based vector space model (VSM) [49], on social image search by using user-provided tags? How much search accuracy can be improved by leveraging visual features with textual features?
3. How does the visual complexity influence the search performance of CBIR systems? And how does the automatic tag-based query expansion affect the search performance of social image retrieval?

The answers to these questions can help us understand the effectiveness of different information evidences in social images search and gain deep insights about the core research problems related to the development of effective and efficient social image search engines. A social image test collection with carefully selected large set of queries is constructed to facilitate the experimental studies. The effectiveness of textual features of social tags, different types of visual features and their combinations on social image search has been carefully studied and analyzed. To explore the influence of visual query complexity on social image retrieval, a novel framework is proposed to quantitatively measure the image query complexity. Besides, to investigate the effects of automatic text query expansion on search performance, a pseudo relevance feedback method is used to automatically expand the initial query terms with social tags.

While parts of this work have been published in [7] and [52], this paper extends previous studies from several aspects and presents a much more comprehensive study of the effects of multiple information evidences on social image search. Firstly, a more comprehensive study is conducted to analyze the performance of several types of visual features and their combination in social image retrieval. Besides, the performance of query expansion with social tags using pseudo relevance feedback has not been studied in previous studies. In summary, the main contributions of this paper are as follows.

- We study the effects of multiple individual information evidences and their combinations on large-scale social image search, including six popular low-level visual features and textual features based on social tags. To the best of our knowledge, this is the first extensive empirical study on the performance of traditional CBIR and TBIR techniques on social image search. The results not only provide baseline, but also obtain fundamental knowledge about social image search to promote the development of advanced social image retrieval algorithms.
- We propose a quantitative metric to measure the complexity of image query and investigate the relationship between image query complexity and search performance. To the best of our knowledge, this is the first work on modeling visual query complexity quantitatively. The visual query complexity measure can be used to select the query for fair and robust evaluation of CBIR systems. It can help users to select visual queries for better search results.
- We study the influence of visual query complexity on CBIR systems and the effects of automatic tag-based query expansion on the search performance of social image retrieval.
- We construct a large-scale social image collection with associated social tags and carefully select a query set to facilitate the empirical study.
- The rest of the paper is organized as follows. Section 2 reviews the related works and Sect. 3 gives an overview of the whole empirical study. The retrieval systems and visual query complexity measurement are presented in Sects. 4 and 5, respectively. Section 6 introduces the experimental configurations and Sect. 7 describes and analyzes the experimental results. Finally, the paper is concluded in Sect. 8 with detail summary about key research findings and future work.

2 Related work

As an emerging technology to manage large digital image collections, image retrieval has been an active research area since the mid 70s of the last century. Extensive efforts have been devoted to various related research topics [11, 48, 54, 74]. Comprehensive reviews on these topics are beyond the scope of this paper. Thus, this section briefly reviews the existing work on research directions and trends closely related to our study.

2.1 Social image retrieval

Most research efforts in social image search focus on tag-based search methods. The quality of user-provided tags play important role in the performances of tag-based search methods [29, 34, 61]. Therefore, to improve the search performance of tag-based image retrieval methods, various techniques have been proposed and widely studied. They include *automatic image tagging* [70, 71], *retagging/tag refinement* [36, 66, 75], and *tag relevance learning* [29, 33].

Tagging is a mechanism to assign text labels to the images for various purposes [44, 53, 61]. Traditionally, the tagging methods are proposed to assign labels to images at the global level [20, 25, 58, 73]. In recent years, to cope with the diversity and arbitrariness of social images, increasing attention has been given to fine-grained tagging schemes [37, 71, 72], which aim at assigning tags to image at the local level (i.e., image regions). For example, Yang et al. [71] propose a method to encode individual regions by using the group sparse coding with spatial correlation among training regions. *Retagging* and *tag refinement* methods aim at assigning images with the tags which can give better description about the image content, because the user-provided tags are imprecise, incomplete and irrelevant. For example, Liu et al. [36] refine the tag quality using a collaborative tag propagation method, which propagates the information of tags along a particular tag-specific similarity graph. Xu et al. [66] refine the tags by using a regularized LDA methods to estimate tag similarity and tag relevance jointly. *Tag relevance learning* methods estimate the relevant levels of different tags with respect to the image content. For instance, Li et al. [29] learn the relevance scores of tags by a visual neighbor's voting method. Liu et al. [33] rank the tags based on their relevance levels with respect to the targeted image. Zhu et al. [75] estimate the relevance score of tags by a proposed matrix decomposition method. Besides, there are also other methods proposed to improve the tag-based search performance. For example, in [69], a tag tagging method has been proposed to supplement semantic image descriptions to the existing tags by associating a group of property tags, such as color, size and position.

Besides the efforts in improving the performance of tag-based image search, another direction for improving the social image search performance is to exploit the visual content together with text information. A popular method is to use the visual content to rerank the results returned by tag-based search methods [13, 35, 62]. The visual contents of images are used to refine the scores obtained by using tag information in [35]. Fan et al. [13] first estimate the relevance scores of image clusters grouped for a targeted tag and then refine the scores via a random walk process. Wang et al. [62] present a diverse relevance ranking scheme by exploiting both visual contents and tags to rerank social images. More recently, Gao et al. [16] propose a visual-text joint hypergraph learning method to simultaneously utilize the textual and visual information of social images for relevance estimation.

There are also research efforts devoted to other directions, such as query formulation [59, 67], result summarization [60, 68] and diversification [62]. [59] presents an interactive image search system, which enables users to indicate the spatial distribution of colors in the desired images to enhance text-based image search. Similarly, a 2D semantic map is proposed in [67]. The semantic map enables the users to clearly describe the spatial distribution of the targeted semantic concepts in the desired images. For the result summarization and diversification, Wang et al. [60] design a result browsing system, which summarizes the search results with diversified representatives and organizes them in a tree structure to facilitate the browsing of the search results. In [62], a diverse relevance ranking scheme is proposed to consider both the relevance and diversity in social image retrieval.

All the above researches focus on developing new algorithms or techniques in social image retrieval. However, little attention has been given to the study and analysis of the fundamental problems in social image retrieval, such as the performance of traditional content-based and text-based retrieval methods in large-scale social image datasets. There is a lack of systematic investigation on these fundamental research problems in large-scale social image retrieval. Comprehensive analysis can unveil what is important in social image retrieval and provide the insights about right directions to the research community. Therefore, instead of proposing new techniques in social image retrieval, this article presents a comprehensive empirical study to analyze the effects of multiple information evidences on large-scale social image retrieval based on traditional CBIR and TBIR methods.

2.2 Evaluation of social image retrieval

A reliable and robust evaluation is a major force to push forward the research and development of information retrieval-related techniques. In general, there are four essential components for a standard evaluation framework: (1) test collection; (2) test query set; (3) ground truth for the query set; and (4) evaluation metrics. The test collections should be representative for the targeted image retrieval datasets. In social image retrieval, test collections should reflect the characteristics of social images. Thus, the general requirements for the test collections of social image retrieval should at least reflect the properties of large-scale, broad visual concepts, diverse visual appearances and noisy social tags. There have been several social image datasets public for social image retrieval evaluation, such as NUS-WIDE [9] and MIRFLICKR [24].

The query set should be general enough to cover various types of queries that could be issued by users in real life. Thus, the query set should contain queries in different difficulty levels (the difficulty should be independent of retrieval methods). Besides, the query set cannot be favorable to certain search algorithms [52]. Therefore, to select a query set for comprehensive and fair evaluations, it is necessary to estimate the search difficulty of each query in the selection of test queries. Many query difficulty estimation studies have been reported in the text document retrieval domain, while few image query difficulty estimation

Table 1 Summary of symbols and definitions

Symbols	Definitions
D	Social image collection
C	Codebook of visual words of collection D
I	Image in collection D
$\boldsymbol{w_n}$	Word sequence $\{w_1, w_2, ..., w_n\}$ of an image
q	Query in the query set
n_t	Number of images tagged with tag t
l_I	Number of tags of image I
l_{avg}	Average number of tags of images in D
s_{v_i}	Visual similarity score of a SCBIR
w_i	Weight coefficient of s_{v_i} in MCBIR
$S_v(q, I)$	Visual similarity score obtained by MCBIR
$S_t(q, I)$	Textual similarity score obtained by TBIR
w_v	Weight coefficient of $S_v(q, I)$ in the multimodal based retrieval system
$S(q, I)$	Similarity score of multimodal based retrieval system
$C(I)$	Complexity of an image I
$H(I)$	Entropy of an image I
$P(\boldsymbol{w_n})$	Probability of $\{w_1, w_2, ..., w_n\}$ over D
$P(w_k \mid \boldsymbol{w_{k-1}})$	Conditional probability of code w_k given previous word sequence $\boldsymbol{w_{k-1}}$

methods have been proposed. Moreover, the proposed image query difficulty estimation methods are not suitable for the test query selection (as discussed in Sect. 2.3). As a result, the consideration in query set construction for comprehensively and fairly evaluating different retrieval systems is generally ignored in existing image retrieval evaluations.

As it is not necessary and impossible to assess the relevance of all images in the test collection with respect to each query in the query set, *pooling method* is usually used to generate the ground truth for the query set. In the *pooling method*, the top results of all retrieval methods are aggregated and assessed. Although the evaluation metrics and pooling method are standardized, the assessment of the relevant image content with respect to a query is usually subjective. Thus in the assessment of image relevance, several human annotators are required to assess the relevant level of an image with respect to a query. Moreover, to reduce the bias caused by subjective assessments, the queries are usually required to be visible concepts so that they can be clearly observed in the images [7, 29].

Typically, users only pay attentions to the top search results and there are often a large number of relevant images in social image datasets for a query. In such circumstances, *recall* becomes meaningless in the evaluation of social image search performance. $P@N$ and $MAP@N$ are the commonly used evaluation metrics in social image retrieval evaluation, $P@N$ the precision for top N results and $MAP@N$ the mean over the precision values after each correct result in the top N results. Another popular evaluation metric for evaluating social image search performance on accuracy is discounted cumulative gain ($DCG@N$), which considers the relevance levels of the top results.

2.3 Image query complexity

The measure of image query complexity is to estimate the difficulty or predict the performance of an image query. In text document retrieval, query difficulty estimation (QDE) is an important research topic and has been extensively studied for many years [3, 10, 22]. QDE methods can be categorized into two approaches: *pre-retrieval* and *post-retrieval*. *Pre-retrieval* methods predict the query performance by analyzing the properties of the search query and the data collection. On the other hand, *post-retrieval* methods analyze the retrieved results to estimate the performance. Many approaches have been proposed for both methods [3, 10, 21, 22] in text information retrieval. However, few studies have been conducted on estimating image query difficulty.

Li et al. [31, 32] estimate the query image difficulty by analyzing the clarity score, spatial consistency of local features and the appeared consistency of global features between the query image and top results. Tian et al. [57] model the query difficulty prediction as a regression problem by using several characteristics of the search results. Nie et al. [43] predict the search performance of image query by estimating the relevance probability of each image in the results. All the above approaches are post-retrieval methods, relying on analyzing the characteristics of returned results. Xing et al. [65] propose a pre-retrieval method by using textual features associated with images. All of these methods are not suitable for the task of selecting visual queries for fair comparisons between different CBIR systems. The post-retrieval methods can only estimate the query difficulty based on the returned results of a retrieval system, while the performances of a query obtained by different retrieval systems could be very different. Different from the previous method, our method can estimate the retrieval difficulty of a query by only analyzing the complexity of the query's visual content with respect to the dataset (Table 1).

3 Overview

This study empirically investigates the effects of multiple information evidences on large-scale social image retrieval. The considered information evidences include six types of widely used visual features, textual features based on social tags and the combinations of multiple visual features and

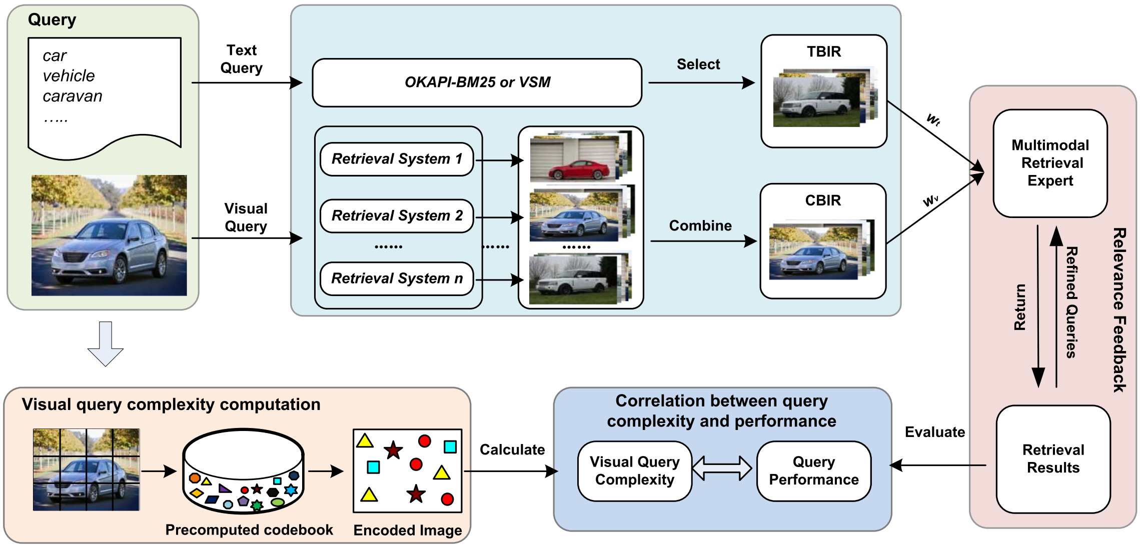

Fig. 1 Overview of our empirical study framework

multimodal features. Besides evaluating the search performance of those information evidences in traditional CBIR and TBIR methods, we also study the influence of text query expansion and visual query complexity on the search performance in terms of accuracy.

Figure 1 gives an overview of this experimental study. A set of retrieval systems have been developed based on different information evidences, including (1) seven content-based image retrieval systems—six of them use a single type of visual feature (called SCBIR for short hereafter) and one combines the six single types of visual features together using a late fusion method (called MCBIR for short hereafter); (2) two text-based retrieval systems based on raw tags—OKAPI-BM25 [26] and vector space model [49]; (3) a multimodal-based retrieval system utilizing both the visual and textual features. As these retrieval systems rely on different visual and textual features, they can be used to study the effects of different information evidences on social image retrieval. Specifically, the SCBIR and TBIR systems are applied to study the effects of individual information evidences on social image search. MCBIR and the multimodal-based retrieval system are used to examine the potentials of feature fusion on search performance improvement. Besides, to gain a good understanding of the effects of visual query complexity, a quantitative metric for the measurement of visual query complexity is proposed based on "bag-of-visual-words" representation of images. On the other hand, the effects of automatic query expansion by social tags are studied via examining the performance changes by adding different numbers of social tags to the text query in the multimodal-based retrieval system. We use a pseudo relevance feedback method to expand the text query with social tags.

4 Retrieval systems

This section gives a detailed introduction about each module on the retrieval system used for our empirical study.

4.1 Content-based retrieval systems

In CBIR systems, visual features and similarity measurements are the two most important components, which can directly affect the retrieval accuracy. Different visual features enjoy different capabilities in representing different characteristics of visual contents in an image, and how to select suitable similarity measure for certain search task is still an open research question. As discussed, social images contain a great variety of visual contents, and the performances of different types of visual features on social image search have not been systematically studied. In this study, the performances of six types of widely used visual features on social image search have been investigated. As a single type of visual feature hardly represents well all the visual contents,[1] the combination of multiple visual features is also studied. Next, we first introduce the retrieval systems based on individual visual feature (SCBIR) and then describe the retrieval system utilizing multiple visual features (MCBIR). Particularly, we give a very comprehensive introduction about each individual visual feature in SCBIR and the late fusion method used in MCBIR.

[1] Note that a particular type of visual feature may work very well for certain types of visual queries. In this study, we focus on the search performance of different visual features on general queries.

4.1.1 *Retrieval systems based on individual visual feature*

In each SCBIR system, an individual visual feature is used. The six individual visual features are color histogram (CH), color auto-correlogram (CA) [23], Gabor texture (GT) [38], Tamura texture (TT) [55], edge histogram (EH) [46] and 64D Global feature (GF) [29]. These features cover the commonly used features in global and local scale. In the following, we introduce these visual features and their corresponding distance measurements.

64D Color histogram The color histogram is an effective and commonly used representation of the color content, usually serving as the baseline feature in CBIR systems. In our implementation, standard RGB color space is used and the color space is divided into 64 partitions. The number of pixels within each partition is counted to compute the histogram bin of the corresponding color. As suggested in [12], Jeffrey divergence or Jensen–Shannon divergence (JSD) is used to compare the color histograms.

144D Color auto-correlogram [23] The color auto-correlogram is the visual feature to characterize the global distribution and the spatial correlation of pairs of colors together. Suppose S is the entire set of image pixels in an image I. By quantizing S into m colors $\{c_1, ..., c_m\}$, the auto-correlogram of I is defined as

$$r^k_{c_i,c_j}(S) = P_{(p_1 \in S_{c_i}, p_2 \in S)}[p_2 \in S_{c_j} | |p_1 - p_2| = k], \qquad (1)$$

where $i, j \in \{1, 2, ..., m\}$ and $P[p \in S_{c_j}]$ are the probability that the color of the pixel is c_j. $|p_1 - p_2|$ is the distance between pixels p_1 and p_2. $k \in d$ and d is fixed as a priori. Auto-correlogram only considers the spatial correlation between identical colors. Colors in the HSV color space are quantized into 36 bins. The distance metric $d = \{1, 3, 5, 7\}$ is used to compute the auto-correlogram. Accordingly, the dimension of the color auto-correlogram is 144. Manhattan distance measure is used for this feature [23].

72D Gabor texture [38] We use a Gabor wavelet to extract Gabor features at multiple scales and directions. The mean and standard deviation of the filter responses are calculated [12]. We extract Gabor features in six different orientations and six different scales, resulting in a 72-dimensional vector.

Tamura texture [55] This descriptor contains six texture features: *coarseness*, *contrast*, *directionality*, *line-likeness*, *regularity* and *roughness*. It has been shown that the first three features are very important [12]. Thus, we create a histogram for *coarseness*, *contrast* and *directionality* to describe the texture and compare the histograms (of different images) using JSD as described in [12].

150D Edge histogram [46] This descriptor extends 80D local edge distribution defined in MPEG-7 to include 5D global and 65D semi-global edge histograms. The edge histogram represents the spatial distribution of five types of directional edges, namely four directional edges and one non-directional edge. Each image is partitioned into 4 × 4 sub-images. Each sub-image is then divided into small square image blocks. Five directional edges are extracted from the image blocks. Then the numbers of the five edge types in a sub-image are defined as the histogram bins for this sub-image. The bins for global and semi-global histograms can be directly obtained from local histograms [46].

64D Global feature [29] It includes three different kinds of visual features: 44D color correlogram, 14D color texture moment and 6D RGB color moment. Three features are extracted from each image and separately normalized into unit length. Then, we form the final 64D feature via line concatenation. Euclidean distance is used as the similarity measurement [29].

4.1.2 *Retrieval system based on multiple visual features*

MCBIR combines the six types of visual features described above together using a late fusion method. Specifically, for a visual query, each SCBIR obtains a similarity score for an image in the database. The similarity scores of an image computed by the SCBIR systems are separately normalized using MinMax [51] and then linearly combined using CombSUM [51, 63] with pre-defined weights to compute the final similarity score for the image. In our study, the similarity score is converted from distance by $s = 1 - d$, where s and d denote the similarity score and the corresponding distance, respectively. Formally, for an image I, its final similarity score with respect to the query q is

$$S_v(q, I) = \sum_{i=1}^{6} w_i \cdot s_{v_i}, \qquad (2)$$

where s_{v_i} is the score returned by the ith SCBIR system. w_i is the weight assigned to the system and $\sum w_i = 1$. The weights are optimized by using Coordinate Ascent method, which has been proven to be effective [40, 63]. An advantage of this method is that it can be optimized directly on average precision (AP) with randomly initialized weights to find a local optimum. Global optimal values can be gained via multiple randomized initialization. Detailed description of this method can be found in [40].

4.2 Text-based retrieval systems

Social images are usually accompanied with annotations, such as title, tags, descriptions and comments. As tag-based search has been widely used in existing commercial systems such as Flickr and Youtube, tags are used as the textual information source in text retrieval systems in this study. To compute the similarity of images, two popular TBIR models are applied:

OKAPI-BM25 model [26] It is a classical text retrieval framework and has been widely used as a baseline method [29]. The relevant score of an image I with respect to a query q is computed as:

$$S_{bm25}(q, I) = \sum_{t \in q} qtf(t) idf(t) \frac{tf(t) \cdot (k_1 + 1)}{tf(t) + k_1 \cdot \left(1 - b + b \cdot \frac{l_I}{l_{avg}}\right)}, \tag{3}$$

where $qtf(t)$ is the frequency of term t in the query q and $tf(t)$ is the frequency of term t in the tag set of image I. $idf(t)$ is the inverse document frequency of t calculated as $\log(\frac{|D|-|n_t|+0.5}{|n_t|+0.5})$, where $|n_t|$ is the number of images labeled with t. l_I is the number of tags in the image I, and l_{avg} is the average number of tags of all images in D. In our study, k_1 is set to 0.2 [26] and b is set to 0.75 [39].

Vector space model (VSM) VSM is the other text-based approach used in this study. The classical *TF-IDF* weighting scheme is used in this study. This method has been widely used as text-based image retrieval baseline in literature [2, 6]. In this model, the relevant score is computed as:

$$S_{vsm}(q, I) = \frac{\sum_{t \in q} w_{t,I} * w_{t,q}}{\sqrt{\sum_{t \in I} w_{t,I}^2} * \sqrt{\sum_{t \in q} w_{t,q}^2}}, \tag{4}$$

where $w_{t,I}$ is the term weight of term t in the image I, computed as $w_{t,I} = tf(t) \cdot \log(\frac{|D|}{|n_t|})$. The computation of $w_{t,q}$ is analogous to $w_{t,I}$.

4.3 Multimodal-based retrieval system

In the multimodal retrieval system, the same method described in Sect. 4.1.2 is used to exploit both the textual and visual features in retrieval. In our implementation, the MCBIR system and OKAPI-BM25 text-based method are used, because MCBIR obtains much better results than all SBCIR systems (as shown in Sect. 7.1.1) and OKAPI-BM25 offers slightly better performance than VSM (as shown in Sect. 7.1.2). The similarity score of an image I with respect to query q is computed as

$$S(q, I) = w_v \cdot S_v + (1 - w_v) \cdot S_t, \tag{5}$$

where S_v and S_t are the similarity scores returned by the MCBIR system and OKAPI-BM25, respectively. w_v is optimized with the Coordinate Ascent method.

4.4 Automatic text query expansion

To investigate the effects of text query expansion by social tags, an automatic query expansion method based on pseudo relevance feedback is embedded into the multimodal-based retrieval system. Particularly, a voting scheme is used to expand the text query with the most frequent social tags in the top search results. Specifically, based on the search results of each search round, the most frequent tags in the top 50 results are integrated to the original text query. Then the updated text query is combined with the initial visual query to form a new query for the next search round. The initial text query is always kept for reducing the risk of topic drift. The effects of expanding different numbers of social tags are explored in our experiments.

5 Visual query complexity estimation

5.1 Complexity definition

Image complexity metric is proposed to measure the complexity level of image visual contents [52]. In this model, images are represented by a list of *visual words* or *keyblocks*, which are similar to the words in text documents. A visual word corpus called *codebook*, which is similar to the word dictionary, needs to be precomputed. The construction of codebook includes three steps: (1) every image in the database is evenly partitioned into blocks with regular geometric configuration, such as cutting an image into nine equal rectangle blocks with three rows and three columns (represented as 3 × 3 for short); (2) each block is represented by a low-level visual feature vector; and (3) a clustering algorithm, K-means, is then used to classify the blocks into clusters. The cluster centers are regarded as *keyblocks* in the *codebook*.

With the generated codebook, an image can be represented by a matrix of indexes by replacing each block in the image with the index of the best match keyblock in the codebook. By transforming the matrix in a certain order to one-dimensional vector, the image is syntactically analogous to a text document, which is essentially a list of terms. The *image complexity* is defined based on the concept of *perplexity* in information theory. The measurement of *image complexity* is expressed as

$$C(I) = 2^{H(I)}, \tag{6}$$

where $H(I)$ is the entropy of an image I, which is estimated based on the Shannon theorem [50]. Suppose that $\boldsymbol{w_n} = \{w_1, w_2, ..., w_n\}$ is a sequence of visual words in an image I in a database D, then the entropy is calculated as follows:

$$H(I) = -\frac{1}{n} \log(P(\boldsymbol{w_n})), \tag{7}$$

where $P(\boldsymbol{w_n})$ is the probability of $\boldsymbol{w_n}$ over D. The key issue of image complexity computation is how to estimate the probability of $\boldsymbol{w_n}$. Next, we introduce a statistical approach to approximate the probability.

5.2 Complexity computation

In the text documents, terms are associated with their neighbor terms to express semantic information. Images, as another important medium for people to represent and express information, and their visual contents should be also correlated and distributed in certain spatial configurations to form the semantic information. Therefore, it is reasonable to assume that a block of an image is correlated with all the other blocks in the image. However, when the size of the codebook is large, it becomes prohibitive to model so many relations. A common solution is to assume that the blocks are connected in the order from left to right and top to bottom [47, 64]. Under this assumption, each block is only conditionally dependent on its previous blocks. More general models can be developed by removing this constraint. This assumption is made for its simplicity.

Given an image I encoded with a codebook C, let $\boldsymbol{w_n}$ denote the code string of I. Based on the chain rule, the probability is written as

$$P(\boldsymbol{w_n}) = \prod_{k=1}^{n} P(w_k|\boldsymbol{w_{k-1}}), \tag{8}$$

where $P(w_k|\boldsymbol{w_{k-1}})$ is the conditional probability of code w_k given the previous code sequence $\boldsymbol{w_{k-1}}$.

It is still very difficult to accurately compute the conditional probabilities in reality. A reasonable approximation is *N-block model* [47, 64], which is analogous to the *n-gram language model* in text processing. The assumption of this model is that each keyblock only depends on its proximate vertical and horizontal neighbors. According to the degree of dependency on remote keyblocks, three popular n-block models are defined: *uni-block*, *bi-block* and *tri-block*. In our previous study [52], *bi-block* model provides the most desirable properties on modeling image complexity. Thus, we use *bi-block* model in this study. In *bi-block* model, the probability of a given code depends only on the preceding code. It can be expressed as:

$$P(\boldsymbol{w_n}) = \prod_{k=1}^{n} P(w_k|w_{k-1}). \tag{9}$$

The conditional probability of a bi-block model is computed as follows. For $\forall w_k, w_{k-1} \in C$, if $N(w_{k-1}w_k) > 0$,

$$P(w_k|w_{k-1}) = \frac{N(w_{k-1}w_k)}{N(w_{k-1})}\left(1 - \frac{N(w_{k-1})}{\sum_{w\in C} N(w)} \cdot \frac{1}{|C|}\right); \tag{10}$$

otherwise,

$$P(w_k|w_{k-1}) = \frac{N(w_{k-1})}{\sum_{w\in C} N(w)} \cdot \frac{1}{|C|} \tag{11}$$

where $N(\boldsymbol{w_k})$ denotes the frequency of a code string $\boldsymbol{w_k}$ in the image database D. $|C|$ is the number of keyblocks in the codebook. Since bi-blocks are sparsely distributed in the images, a small prior probability is assigned to an unseen bi-block $w_{k-1}w_k$ for avoiding zero probability. Accordingly, the amount of prior probability should be discounted from occurring in bi-blocks to satisfy the condition that the sum of probability is around 1 [64]. Note that other more sophisticated smoothing methods [5, 64] can be also applied here.

6 Experimental setup

In this section, we describe the test collections and the experimental setups.

6.1 Configurations for retrieval system evaluation

Test collection To conduct the experiments, we construct a test collection with 100,000 images with the associated tags. The test collection is constructed in three steps: (1) more than one million labeled images are downloaded from Flickr by using random terms as queries; (2) 500 most popular tags in the 1 million social image dataset are selected; and (3) each selected tag is used to randomly choose 200 images, which have to be labeled with the tag from the dataset to construct the test collection.

Query set In our experiments, a query consists of a query tag and an example image to facilitate different retrieval strategies. To reduce the influence of query set on the retrieval performance, two criteria are held in query selection. First, the query concept should be clearly visible in image so that people with common knowledge can easily relate the concept to the visual content [29], such as *tiger, beach, sunset*. In contrast, concepts such as *autumn, london, beautiful* cannot be used as query topics. This criterion is to reduce the influence of subjectivity during the relevance assessment process. Second, the search accuracy of a query could be affected by the number of relevant images with respect to it in the collection, especially when relevant images are sparse. Since it is hard to obtain the accurate number of images related to each concept in such a large collection, social tags are used to estimate the number of images related to a query concept. For each query topic, there should be more than 200 images labeled with its query tag in the collection. Based on the two principles, 100 query concepts are selected. Some query examples are shown in Fig. 2.

Evaluation criterion Because users are generally only interested in the top results returned by search engines,

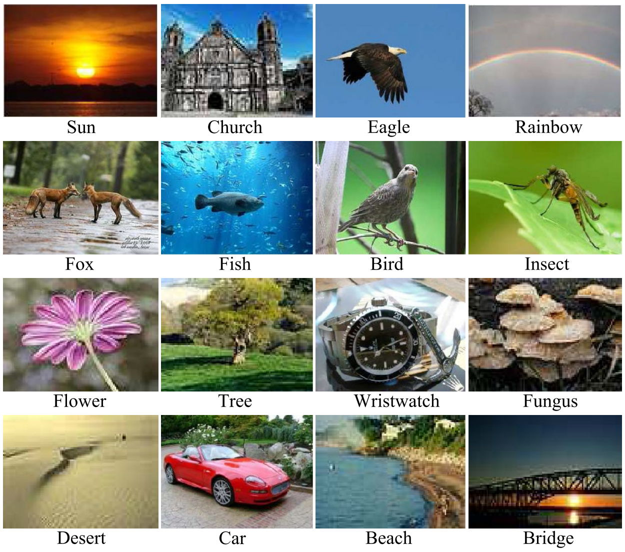

Fig. 2 Query examples: each query consists of a query tag and a query image

relevant images with respect to a query should be ranked as high as possible. *Precision at n* (P@n) is used as performance evaluation metrics. P@n is the proportion of relevant instances in the top n retrieved results, computed as:

$$\mathrm{P@n} = \frac{\text{No. of relevant images in top n. results}}{n} \tag{12}$$

$n = \{10, 20, \ldots, 50\}$ is evaluated. In the presentation of result in Sect. 7, P@n represents the mean precision over all queries in the query set.

6.2 Configurations for visual complexity

The goal of this experiment is to study the effects of the visual query complexity on the search performance of the CBIR systems. In our experiments, images in the test collection are segmented into 6 × 6 blocks. A codebook with 3600 key-blocks is generated[2]. The query images are also cut into 6 × 6 blocks. Then the proposed image complexity metric is used to calculate the complexities of these visual queries. As MCBIR provides the best performance in all CBIR systems (as shown in Sect. 7.1.1), its results are used to analyze the relationship between visual query complexity and search performance.

7 Experimental results

This section presents and analyzes the experimental results. Firstly, the search performances of the retrieval systems are presented in Sect. 7.1. In the next, the effects of visual query complexity in the MCBIR system are reported in Sect. 7.2. Finally, the effectiveness of automatic tag-based query expansion using pseudo relevance feedback on social image search is examined and analyzed in Sect. 7.3.

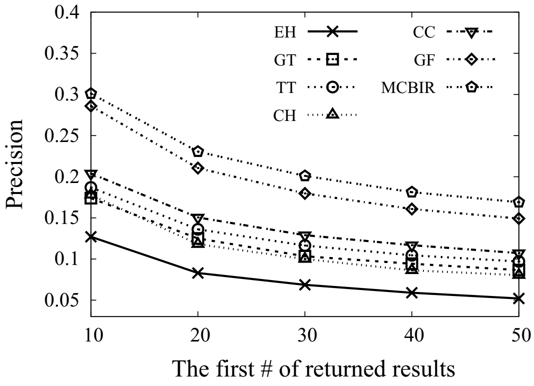

Fig. 3 Search performances of different CBIR systems

7.1 Performance of retrieval systems

The search performances of different CBIR systems are presented and compared in Sect. 7.1.1, and then the performances of TBIR systems and the multimodal-based retrieval system are analyzed in Sects. 7.1.2 and 7.1.3, respectively.

7.1.1 Performances of the CBIR systems

The search performances based on different types of visual features on social image search are presented in Fig. 3 and Table 2. The capital letter (e.g., EH) in the figure and table denotes the corresponding SCBIR system using this type of visual feature. From the results, it is easy to find that the SCBIR system using global feature (GF) obtains the highest search accuracy, while the global feature itself is a combination of three types of color features (color correlogram, color texture moment and color moment). Among the descriptors of single color, texture and edge features, color auto-correlogram (CA) gets the best search performance, and edge histogram (EH) leads to the lowest search accuracy. As discussed in Sect. 1, the visual content of social image is extraordinarily diverse and the quality of social images is spread on various levels. Thus, it is hard for a single type of visual feature to capture such heterogeneous information, resulting in poor performance based on a single type of visual feature. MCBIR, which linearly combines six SCBIR systems with optimized weights, achieves 13.27 %, relatively, improvement on

[2] The configurations of 6 × 6 blocks and the size of codebook are empirically set based on the results in [52].

Table 2 Search performances of different retrieval systems on the test collection

Precision	CBIR							TBIR		Multimodal
	EH	GT	TT	CH	CA	GF	MCBIR	BM25	VSM	
P@10	0.127	0.174	0.187	0.178	0.204	0.286	0.301	0.672	0.672	0.750
P@20	0.083	0.125	0.136	0.118	0.151	0.211	0.231	0.659	0.660	0.717
P@30	0.069	0.104	0.117	0.100	0.129	0.180	0.201	0.640	0.635	0.693
P@40	0.059	0.094	0.105	0.086	0.117	0.161	0.181	0.634	0.621	0.673
P@50	0.052	0.087	0.097	0.080	0.107	0.149	0.169	0.624	0.606	0.660

search accuracy compared to the SCBIR with the best performance (namely, the SCBIR with GF). Although a combination of multiple visual features can provide much better search accuracy than all single types of visual features, the performance is still far from satisfactory. Obviously, we have not covered all visual features that have been proposed so far. However, the chosen visual features with the corresponding distance functions are widely used in the literature. To some extent, the results can demonstrate that CBIR systems cannot get promising search results in large-scale social image searches.

7.1.2 Performances of TBIR systems

The search performances of TBIR systems and the multimodal-based retrieval system are shown in Fig. 4 and Table 2. Compared to the poor performances of CBIR systems, retrieval systems based on textual information achieve much better search results. OKAPI-BM25 and VSM exhibit comparable capability on social image search by using social tags. When the number of top results examined increases, the search accuracies of CBIR systems decrease more quickly than that of TBIR systems. For example, the precision of MCBIR drops from 30.1 to 16.9 % when the number of results taken into consideration increases from top 10 to top 50, while the performances of BM25 and VSM are relatively stable. This demonstrates that textual information is much more effective and reliable in social image search than visual information.

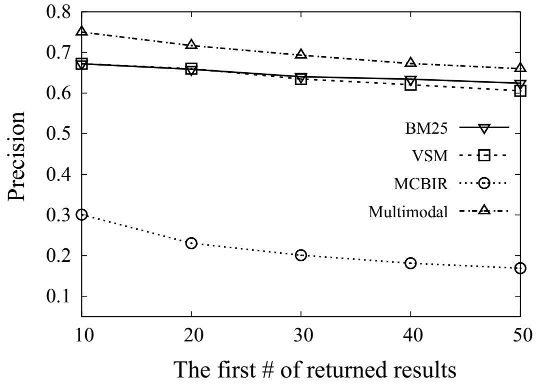

Fig. 4 Search performances of retrieval systems based on features from different modalities

7.1.3 Performance of multimodal-based retrieval system

The multimodal-based retrieval system linearly combines the results of OKAPI-BM25 and the MCBIR system. OKAPI-BM25 is used in the multimodal-based retrieval system because it provides slightly better results than VSM. As shown in Fig. 4, the combination of textual and visual information can consistently improve the search performance. An interesting observation gained is that although the performance of MCBIR is much poorer than TBIR, the best performance is obtained by assigning a higher weight to MCBIR (0.75), based on the optimized weight using coordinate ascent method. This is largely due to the unique characteristics of the social image. Specifically, most social images only have a few tags and each tag only appears once per image. As a result, for a certain query, many images share the same textual similarity score obtained by OKAPI-BM25. The visual similarity scores of images provided by MCBIR are all different and can be used to differentiate those images with the same textual similarity score. Thus, the working mechanism of multimodal-based retrieval system in social images can be regarded as that the TBIR system retrieves the most related images, and then the CBIR system reranks these images. More detailed analysis about the weight is given in [7].

7.2 Effects of visual query complexity

Figure 5 shows the search performances (P@20) of MCBIR with respect to different image query complexities. The values of P@20 and query complexity in this figure are normalized using MinMax [51]. The curve is processed using adjacent averaging smoothing method. Although there are some fluctuations in the curve, we can still observe a clear relation between the query complexity and search performance: with the increase of query complexity, the corresponding search accuracy decreases. In other words, the image query with higher complexity results in poorer search performance. The correlation presented can

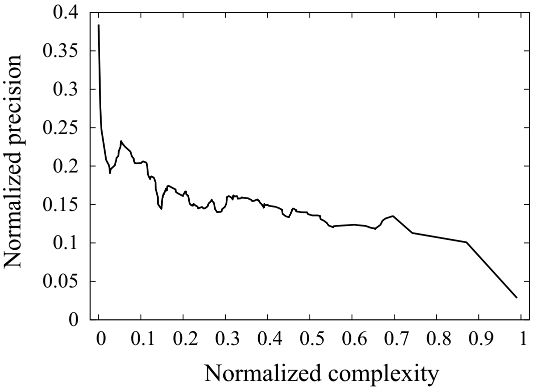

Fig. 5 The variations of search performances with visual query complexity in MCBIR system

Table 3 Search precision at P@50 of tag-based relevance feedback with different numbers of extended terms. The column of "Iteration" represents different search rounds in an interactive search session. For example, the second row ("0") shows the initial search results. Columns of "Number of extended tags" represent different expansion strategies

Iteration	Number of extended tags				
	1	2	3	4	5
0	0.660	0.660	0.660	0.660	0.660
1	0.597	0.559	0.527	0.527	0.526
2	0.614	0.552	0.519	0.515	0.498
3	0.613	0.552	0.517	0.509	0.491

be seen as statistical results of the complexities of queries with their corresponding search performance. The fluctuations in the curve might be due to the relatively small statistical samples (the used test queries) used in the experiment.

7.3 Effects of query expansion based on social tags

An important factor in relevance feedback is the information used to refine the query. In pseudo relevance feedback methods, the results returned by the previous search round becomes determinant. Because the multimodal-based retrieval system provides the best results, it is used in this experiment. Another key issue in text query expansion is how many tags to be added to the initial text query, because adding new terms to the original query bears the risk of topic drift. This is hard to predict by theoretical analysis. We examine the number of extended tags from 1 to 5 in experiments. The top 50 results are used to generate the extended tags. The results (P@50) obtained by extending different numbers of tags are given in Table 3. Rows from second to fifth present the search performance of the first four search rounds, respectively. Note that the second row is the initial results given by the multimodal-based retrieval system. Columns from second to sixth denote different expansion strategies. For example, column "1" refers to the strategy of adding one tag to the initial query. From the table, we can see that the performances of query expansion based on pseudo term-based relevance feedback decrease in all cases. Moreover, the search accuracy becomes lower either with more iterations or when more tags are added to the initial query.

The results are beyond our expectation. Initially, we expected that the performance should be improved with the expansion of one or two tags. The reasons for our expectation are as follows. Firstly, the expansion tags are the most frequently appearing tags in returned results, so the probability of the extended tags in relevant images should be larger than in irrelevant images, because the distribution of tags in irrelevant images, which generally contain diverse visual contents, should be much more spread than the distribution of tags in relevant images. Secondly, these frequent co-occurring tags should be relate to the original query since they are used to describe similar visual contents. To find out the reasons for the performance decreasing, we study the search accuracy variations between the first search round and the second search round of all queries in the results of adding one tag to the original query. The number of queries with increased search accuracy (from the first search round to the second search round) is 39, and the number of queries with decreased search accuracy is 50. The performances of the 11 other queries were unchanged.[3] It shows that with query expansion, more queries tend to get worse search results. We then classify the queries into two sets—performance-increased set (*PIS*) and performance-decreased set (*PDS*). The average increase precision is 7.08 % (based on the first and second search rounds) in *PIS*, while the average decrease precision is 18.04 % (based on the first and second search rounds) in *PDS*. It shows that the search accuracy decrease of a query in *PDS* is much larger than the search accuracy increase of a query in *PIS* on average.

We further study the extended terms for each query in the two sets. Table 4 gives some examples of the extended tags for initial text queries in the two sets. In the table, the column of "Initial query" shows the original queries at the beginning and the column of "Extended term" shows the extended term in the first iteration for the corresponding initial query in the same row. The column of "0" presents the search accuracy of the first search round and the column of "1" gives the search accuracy of the second search round (the search performance with the extended query). It is easy to find that queries in *PIS* generally get a related tag (with respect to the original query concept) extended,

[3] Notice that there are 100 queries in total as described in Sect. 6

Table 4 Some examples of query performance changes before and after query expansion in performance-increased set (*PIS*) and performance-decreased set (*PDS*)

Examples of performance-increased queries				Examples of performance-decreased queries			
Initial query	0	Extended term	1	Initial query	0	Extended term	1
Church	0.82	*Architecture*	0.96	*Snow*	0.96	*Deer*	0.84
Ocean	0.80	*Beach*	0.88	*Beach*	0.64	*Palm*	0.14
Hamster	0.88	*Roden*t	0.96	*Hat*	0.94	*Christmas*	0.56
Rose	0.64	*Flower*	0.86	*Basketball*	0.88	*Raw*	0.54
Temple	0.84	*Church*	0.96	*Flag*	0.84	*Beijing*	0.30

The column of "0" is the search precision of initial queries in the first search round, and the column of "1" is the search precision of extended queries in the second search round

while the extended tags for queries in *PDS* are somewhat unexpected. This observation explains the large decrease of search accuracy in *PDS*. It is known that the text-based query expansion works well only if the refined query is consistent with the original query concept [42]. It is obvious that the topics of queries in *PDS* can be different after query expansion. One possible reason of this phenomenon is the low quality of raw tags. The tags contributed by common users are known to be noisy. The users may only tag objects they are interested in while ignoring other useful information in the image. Consequently, some highly correlated semantic concepts related to the theme of an image are absent. Thus, using raw tags to expand query terms often results in topic drift. In other words, the extended tag has the corresponding visual content appearing in some images in the top results, while its concept is not consistent with the original query concept in general, such as "*deer*" for "*snow*" and "*palm*" for "*beach*" as shown in Table 4. The visual content of these concepts may appear in the top results, but they are not generally related to the initial query concept, resulting in the decrease of performance.

8 Conclusion

This paper reports a set of empirical studies on the effects of multiple information evidences on large-scale social images search. An extensive set of experiments have been conducted on a large-scale social image test collection. Search performances based on different types of visual features and textual information are compared and analyzed. Besides, a novel metric is proposed to quantitatively measure visual query complexity. Based on the image complexity measurement, the effects of visual query complexity on search performance are studied. We also explore the influence of automatic tag-based query expansion based on a pseudo relevance feedback method on the retrieval performance. Experimental results demonstrate that:

- The search accuracy based on individual visual feature is poor. Although the combination of multiple visual features can greatly improve the search results, the performance is still rather poor. On the other hand, retrieval systems based on textual features of social tags can achieve much better search performance.
- The combination of textual and visual features can enhance the search accuracy significantly and consistently. An interesting observation is that although the overall search performance based on visual features is poor, better results are obtained by assigning a higher weight to the CBIR module than the TBIR model in the multimodal based retrieval system.
- The retrieval accuracy of visual query demonstrates an inverse correlation with the complexity level of visual contents.
- Automatic text query expansion using social tags often leads to performance degradation. The main reason is that social tags' quality is low and query expansion based on the text labels usually causes topic drift.
- The empirical results indicate that although social tags are noisy, they are still very important for accurate social image search. Due to the great variety of visual content, the visual features cannot provide satisfactory results. On the other hand, the visual features are useful to refine the search results returned by TBIR systems. We hope the results can provide some guidelines for the development of more advanced social image search engines in the future. In this study, we have not fully explored the features of social tags in TBIR methods. For example, we only use the TF-IDF weighting scheme in VSM and have not studied the effects of different parameter settings in OKAPI-BM25. Besides, it is also worth studying the effectiveness of different types of visual features on different types of visual queries in CBIR systems. Another important direction for future work is to explore the other aspects of search performance (besides accuracy) in large-scale social image

retrieval, such as the efficiency and result diversification of retrieval methods based on different information evidences.

Acknowledgments Jialie Shen and Zhiyong Cheng are partially supported by Mobile plus Cloud Computing Theme Research Program Award—Microsoft Research. Portions of this work appeared in [7, 52].

References

1. Jeffries, A.: The man behind flickr on making the service 'awesome again'. The Verge. 2013–03–20. Retrieved 2013–08–29
2. Benavent, J., Benavent, X., Ves, E., Granados, R., Serrano, A.G.: Experiences at ImageCLEF 2010 using CBIR and TBIR mixing information approaches. In: Cross-Language Evaluation Forum CLEF 2010 (2010)
3. Carmel, D., Yom-Tov, E., Darlow, A., Pelleg, D.: What makes a query difficult? In: Proceedings of ACM SIGIR (2006)
4. Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval. ACM Comput. Surv. **44**(1), 1–50 (2012)
5. Chen, S., Goodman, J.: An empirical study of smoothing techniques for language modeling. Technical report (1998)
6. Cheng, P., Yeh, J., Ke, H., Chien, B., Yang, W.: NCTU-ISU's evaluation for the user-centered search task at ImageCLEF 2004. In: Cross-Language Evaluation Forum CLEF 2004 (2004)
7. Cheng, Z., Ren, J., Shen, J., Miao, H.: The effects of heterogeneous information combination on large scale social image search. In: Proceedings of ACM ICIMCS (2011)
8. Cheng, Z., Ren, J., Shen, J., Miao, H.: Building a large scale test collection for effective benchmarking of mobile landmark search. In: Proceedings of MMM (2013)
9. Chua, T., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: NUS-WIDE: a real-world web image database from National University of Singapore. In: Proceedings of ACM CIVR (2009)
10. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Predicting query performance. In: Proceedings of ACM SIGIR (2002)
11. Datta, R., Joshi, D., Li, J., Wang, J.: Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. **40**(2), Article 5 (2008)
12. Deselaers, T., Keysers, D., Ney, H.: Features for image retrieval: an experimental comparison. Inf. Retr. **11**(2), 77–107 (2008)
13. Fan, J., Shen, Y., Zhou, N., Gao, Y.: Harvesting large-scale weakly-tagged image databases from the web. In: Proceedings of IEEE CVPR (2010)
14. Gao, Y., Wang, F., Luan, H., Chua, T.: Brand data gathering from live social media streams. In: Proceedings of ACM ICMR (2014)
15. Gao, Y., Wang, M., Tao, D., Ji, R., Dai, Q.: 3-d object retrieval and recognition with hypergraph analysis. IEEE Trans. Multimed. **21**(9), 4290–4303 (2012)
16. Gao, Y., Wang, M., Zha, Z., Shen, J., Li, X., Wu, X.: Visual-textual joint relevance learning for tag-based social image search. IEEE Trans. Image Process **22**(1), 363–376 (2013)
17. Gao, Y., Wang, M., Zha, Z., Tian, Q., Dai, Q., Zhang, N.: Less is more: efficient 3-d object retrieval with query view selection. IEEE Trans. Image Process **13**(5), 1007–1018 (2011)
18. Geng, B., Yang, L., Xu, C., Hua, X., Li, S.: The role of attractiveness in web image search. In: Proceedings of ACM MM (2011)
19. Grauman, K., Fergus, R.: Learning binary hash codes for large-scale image search. In: Machine learning for computer vision, pp. 49–87. Springer (2013)
20. Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C.: TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation. In: Proceedings of IEEE ICCV (2009)
21. He, B., Ounis, I.: Inferring query performance using pre-retrieval predictors. In: String processing and information retrieval (2004)
22. He, B., Ounis, I.: Query performance prediction. Inf. Syst. **31**(7), 585–594 (2006)
23. Huang, J., Kumar, S., Mitra, M., Zhu, W.J., Zabih, R.: Image indexing using color correlogram. In: Proceedings of IEEE CVPR (1997)
24. Huiskes, M., Lew, M.: The MIR flickr retrieval evaluation. In: Proceedings of ACM MIR (2008)
25. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using crossmedia relevance models. In: Proceedings of ACM SIGIR (2003)
26. Jones, K., Walker, S., Robertson, S.: A probabilistic model of information retrieval: development and comparative experiments—part 2. Information Processing and Management **36**(6), 809–840 (2000)
27. Kennedy, L., Naaman, M.: Generating diverse and representative image search results for landmarks. In: Proceedings of WWW (2008)
28. Kurashima, T., Iwata, T., Irie, G., Fujimura, K.: Travel route recommendation using geotags in photo sharing sites. In: Proceedings of CIKM (2010)
29. Li, X., Snoek, C., Worring, M.: Learning social tag relevance by neighbor voting. IEEE Trans. Multimed. **11**(7), 1310–1320 (2009)
30. Li, X., Snoek, C., Worring, M.: Unsupervised multi-feature tag relevance learning for social image retrieval. In: Proceedings of ACM CIVR (2010)
31. Li, Y., Geng, B., Yang, L., Xu, C., Bian, W.: Query difficulty estimation for image retrieval. Neurocomputing **95**, 48–53 (2012)
32. Li, Y., Geng, B., Zha, Z., Tao, D., Yang, L., Xu, C.: Difficulty guided image retrieval using linear multiview embedding. In: Proceedings of ACM MM (2011)
33. Liu, D., Hua, X., Yang, L., Wang, M., Zhang, H.: Tag ranking. In: Proceedings of WWW (2009)
34. Liu, D., Hua, X., Zhang, H.: Content-based tag processing for internet social images. Multimed. Tools Appl. **51**(2), 723–728 (2011)
35. Liu, D., Wang, M., Yang, Y., Hua, X., Zhang, H.: Tag quality improvement for social images. In: Proceedings of IEEE ICME (2009)
36. Liu, D., Yan, S., Hua, X., Zhang, H.: Image retagging using collaborative tag propagation. IEEE Trans. Multimed. **13**(4), 702–712 (2011)
37. Liu, X., Cheng, B., Yan, S., Tang, J., Chua, T., Jin, H.: Label to region by bi-layer sparsity priors. In: Proceedings of ACM MM (2009)
38. Manjunath, B., Ma, W.: Texture features for browsing and retrieval of image data. IEEE Trans. Pattern Anal. Mach. Intell. **18**(8), 837–842 (1996)
39. Manning, C., Raghavan, P., Schutze, H.: An introduction to information retrieval. Cambridge University Press, (2009)
40. Metzler, D., Croft, W.: Linear feature-based models for information retrieval. Inf Retr **10**(3), 257–274 (2007)
41. Mu, Y., Shen, J., Yan, S.: Weakly-supervised hashing in kernel space. In: IEEE Proceedings of CVPR (2010)
42. Natsev, A., Haubold, A., Tesic, J., Xie, L., Yan, R.: Semantic concept-based query expansion and re-ranking for multimedia retrieval. In: Proceedings of ACM MM (2007)
43. Nie, L., Wang, M., Zha, Z., Chua, T.: Oracle in image search: A content-based approach to performance prediction. ACM Trans. Inf. Syst. **30**(2), 13 (2012)

44. Nov, O., Ye, C.: Why do people tag?: Motivations for photo tagging. Commun. ACM **53**(7), 128–131 (2010)
45. Parfeni, L.: Flickr boasts 6 billion photo uploads. Softpedia. Retrieved 2012–03-01
46. Park, D., Jeon, Y., Won, C.: Efficient use of local edge histogram descriptor. In: Proceedings of ACM MM (2000)
47. Rao, A., Srihari, R., Zhu, L., Zhang, A.: A method for mearsuring the complexity of image databases. IEEE Trans. Multimed. **4**(2), 160–173 (2002)
48. Rui, Y., Huang, T., Chang, S.: Image retrieval: Current techniques, promising directions, and open issues. J. Vis. Commun. Image Represent **10**(1), 39–62 (1999)
49. Salton, G., Wong, A., Yang, C.: A vector space model for automatic indexing. Commun ACM **18**(11), 613–620 (1975)
50. Shannon, C.: Prediction and entropy of printed english. Bell Syst. Tech. J. **30**(1), 50–64 (1951)
51. Shaw, J., Fox, E.: Combination of multiple searches. In: The Second Text REtrieval Conference (TREC-2), pp. 243–252 (1994)
52. Shen, J., Cheng, Z.: On effects of visual query complexity. In: The Era of Interactive Media, pp. 531–541. Springer (2013)
53. Shen, J., Wang, M., Yan, S., Hua, X.S.: Multimedia tagging: past, present and future. In: Proceedings of ACM MM Conference (2011)
54. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. **22**(2), 1349–1380 (2000)
55. Tamura, H., Mori, S., Yamawaki, T.: Textural features corresponding to visual perception. IEEE Trans. Syst., Man, Cybern. **8**(6), 460–473 (1978)
56. Tamura, H., Yokoya, N.: Image database systems: a survey. Pattern Recognit **17**(1), 29–43 (1984)
57. Tian, X., Lu, Y., Yang, L.: Query difficulty prediction for web image search. IEEE Trans. Multimed. **14**(4), 951–962 (2012)
58. Tsai, D., Jing, Y., Liu, Y., Rowley, H., Ioffe, S., Rehg, J.: Large-scale image annotation using visual synset. In: Proc. of IEEE ICCV (2011)
59. Wang, J., Hua, X.: Interactive image search by color map. ACM Trans. Intell. Syst. Technol. 3(1), Article 12 (2011)
60. Wang, J., Jia, L., Hua, X.: Interactive browsing via diversified visual summarization for image search results. Multimed. Syst. **17**(5), 379–391 (2011)
61. Wang, M., Ni, B., Hua, X., Chua, T.: Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. ACM Comput. Surv. **44**(4), Article 25 (2012)
62. Wang, M., Yang, K., Hua, X., Zhang, H.: Towards a relevant and diverse search of social images. IEEE Trans. Multimed. **12**(8), 829–842 (2010)
63. Wilkins, P., Smeaton, A., Ferguson, P.: Properties of optimally weighted data fusion in CBMIR. In: Proceedings of ACM SIGIR (2010)
64. Wu, L., Li, M., Li, Z., Ma, W., Yu, N.: Visual language modeling for image classification. In: Proceedings of ACM MIR (2007)
65. Xing, X., Zhang, Y., Han, M.: Query difficulty prediction for contextual image retrieval. In: Advances in Information Retrieval, pp. 581–585. Springer (2010)
66. Xu, H., Wang, J., Hua, X., Li, S.: Tag refinement by regularized LDA. In: Proceedings of ACM MM (2009)
67. Xu, H., Wang, J., Hua, X., Li, S.: Interactive image search by 2D semantic map. In: Proceedings of WWW (2010)
68. Xu, H., Wang, J., Hua, X., Li, S.: Hybrid image summarization. In: Proceedings of ACM MM (2011)
69. Yang, K., Hua, X., Wang, M., Zhang, H.: Tag tagging: Towards more descriptive keywords of image content. IEEE Trans. Multimed. **13**(4), 662–673 (2011)
70. Yang, Y., Huang, Z., Shen, H., Zhou, X.: Mining multi-tag association for image tagging. World Wide Web **14**(2), 133–156 (2011)
71. Yang, Y., Huang, Z., Yang, Y., Liu, J., Shen, H., Luo, J.: Local image tagging via graph regularized joint group sparsity. Pattern Recognit **46**(5), 1358–1368 (2013)
72. Yang, Y., Yang, Y., Huang, Z., Shen, H., Nie, F.: Tag localization with spatial correlations and joint group sparsity. In: Proceedings of IEEE CVPR (2011)
73. Yang, Y., Yang, Y., Shen, H.: Effective transfer tagging from image to video. ACM Trans. Multimed. Comput. Commun. Appl. **9**(2), 14 (2013)
74. Zhou, X., Huang, T.: Relevance feedback in image retrieval: a comprehensive review. Multimed. Syst. **8**(6), 536–544 (2003)
75. Zhu, G., Yan, S., Ma, Y.: Image tag refinement towards low-rank, content-tag prior and error sparsity. In: Proceeding of ACM MM (2010)