

7-2016

# On very large scale test collection for landmark image search benchmarking

CHENG ZHIYONG

*Singapore Management University*, zy.cheng.2011@phdis.smu.edu.sg

Jialie SHEN

*Singapore Management University*, jlshen@smu.edu.sg

**DOI:** <https://doi.org/10.1016/j.sigpro.2015.10.037>

Follow this and additional works at: [https://ink.library.smu.edu.sg/sis\\_research](https://ink.library.smu.edu.sg/sis_research)



Part of the [Databases and Information Systems Commons](#)

---

## Citation

CHENG ZHIYONG and SHEN, Jialie. On very large scale test collection for landmark image search benchmarking. (2016). *Signal Processing*. 124, 13-26. Research Collection School Of Information Systems.

**Available at:** [https://ink.library.smu.edu.sg/sis\\_research/3532](https://ink.library.smu.edu.sg/sis_research/3532)

This Journal Article is brought to you for free and open access by the School of Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email [libIR@smu.edu.sg](mailto:libIR@smu.edu.sg).

# On very large scale test collection for landmark image search benchmarking

Zhiyong Cheng, Jialie Shen\*

School of Information Systems, Singapore Management University, Singapore

---

## A B S T R A C T

High quality test collections have been becoming more and more important for the technological advancement in geo-referenced image retrieval and analytics. In this paper, we present a large scale test collection to support robust performance evaluation of landmark image search and corresponding construction methodology. Using the approach, we develop a very large scale test collection consisting of three key components: (1) 355,141 images of 128 landmarks in five cities across three continents crawled from Flickr; (2) different kinds of textual features for each image, including surrounding text (e.g. tags), contextual data (e.g. geo-location and upload time), and metadata (e.g. uploader and EXIF); and (3) six types of low-level visual features. In order to support robust and effective performance assessment, a series of baseline experimental studies have been conducted on the search performance over both textual and visual queries. The results demonstrate importance and effectiveness of the test collection.

### Keywords:

Large scale landmark image search  
Performance evaluation

---

## 1. Introduction

In general, landmark refers to notable buildings (i.e. Statue of Liberty), architecture with special structure or meaning or purpose (i.e. Beijing National Stadium – “Bird’s Nest”) and famous scenic spots (e.g. Marina Bay in Singapore). Fig. 1 illustrates a few examples. Due to the attractive physical features or/and historical significance, landmarks frequently attract a lot of visitors, who are keen on taking the photos and share them with friends or/and family members via online social communities. Consequently, volume of landmark images increases tremendously in recent years and has accounted for a significant portion of online social images. In recent years, many different algorithms or systems have been developed to support automatic retrieval or visualization of landmark images [1–6]. In

particular, large scale landmark image search emerge as important technical foundation for various real applications [7]. Consequently, numerous efforts have been devoted to improve the corresponding search systems’ performance from different perspectives (e.g. retrieval effectiveness [8–12], visual classification [13], system performance evaluation [14], and result diversification [15,16]).

The technology advancement in landmark image search is largely dependent on studying and analyzing system performance. However, very limited work has been carried out on benchmarking dataset development for the purpose of comparing and evaluating relative algorithms and systems comprehensively. While the importance of the issue has been recognized in the multimedia retrieval and other related communities (e.g. computer vision and signal processing) and a few test collections have been published recently, they generally suffer from one or multiple weaknesses as follows: small scale, unclear definition about search task, lack of diversified landmarks views and

---

\* Corresponding author.



Fig. 1. Examples of landmark images.

limited availability. The issues could be particularly severe when the researchers try to do robust cross-method comparisons. Due to the lack of quality collections, a popular solution taken by scholars is to construct their own datasets by leveraging online public resources, such as Flickr<sup>1</sup> and Google image [8,15–20]. This can easily lead to very expensive and tedious dataset development process. More importantly, the use of self-constructed datasets makes it hard for other scholars to repeat the experimental studies and compare different methods to assess (1) the precise impacts of various systems and (2) identify the state-of-the-art.

In principle, the standard procedure for the performance evaluation of landmark image retrieval systems can include five basic steps: (1) construct a test collection; (2) define specific search tasks; (3) select search queries (text or/and visual queries) and generate associated

ground truth; (4) run each test query through a particular landmark search system; and (5) assess the performance of the system via an empirical distribution of particular measurement metric (e.g. precision, recall and MAP ratio). All five steps are critical for the quality of performance evaluation. In this paper, our main focus is on how to develop very large scale of test collection. To achieve reliable, robust and effective system performance assessment, test collection construction needs to satisfy three key guidelines:

- Given that the size of image collections in many real photo sharing Websites have scaled to billions over the last few years, test collection's scale is required to be sufficiently big to generate statistical meaningful results.
- Real geographic locations in different countries and regions might have very diverse visual appearance and thus test collection should own comprehensive visual coverage of different geographic locations.

<sup>1</sup> <https://www.flickr.com/>

- When more partial views about the same landmark are involved, evaluation process will be more reliable and robust.
- Search task and corresponding image queries for evaluation and associated ground-truth need be clearly defined.

In this research, we make two main contributions to technical advancement of landmark image search:

- a methodology and procedure to construct very large landmark search test collection, and;
- based on the methodology, a very large scale benchmarking test collection is developed to support effective, reliable and robust comparison and assessment of landmark image search system performance.

Totally, the test collection consists of 355,141 images about 128 landmarks in five cities over three continents from Flickr. Besides, six different visual features are extracted as image visual signature. For each landmark, a wide range of visual views have been considered to gain comprehensive coverage. Moreover, a clear definition is given to different search tasks and ground truth information. Using them, a set of empirical studies have been carried out to investigate the search and compare accuracy and efficiency of content-based search methods and text-based search methods. The test results show promising of the test collection.

This paper is organized as follows: [Section 2](#) presents the details about test collection construction. We introduce how to harvest raw dataset, key statistics about dataset and its main structure. [Section 3](#) presents detail empirical study configuration, evaluation system framework and key results on two main retrieval tasks: content-based image retrieval and text-based image retrieval. Finally, we conclude the paper in [Section 4](#) with summary of key research findings and future work.

The results have been partially published in The 19th International Conference on MultiMedia Modeling [14].

## 2. Test collection construction

The construction of the landmark image dataset starts from selecting a set of international cities with various popular landmarks. At this stage, five cities from 3 continents are considered and they include *Beijing*, *Hong Kong*, *Singapore*, *London*, and *New York*. Well-known landmarks of each city are selected from the landmark lists published in Wikipedia<sup>2</sup> and Wikitravel<sup>3</sup> (we also refer to their online tour guide web pages). Altogether, 128 landmarks are identified in the five cities. The number of landmarks in each city can be found in [Table 2](#). [Tables A1–A3](#) in [Appendix A](#) show the details of landmarks in the dataset.

**Table 1**  
Related information crawled for each image.

Surrounding text	Context information	Metadata
Tag	Taken time	Photo ID
Title	Upload time	Uploader ID
Comment	Geo-location	Source page
Description	Contextual URL	EXIF/TIFF

### 2.1. Dataset downloaded

The images of each landmark are collected from Flickr. For the landmark with an unique name, its name is used as the keywords to search images. While for the landmark whose name also corresponds to other landmarks in different cities, the name of corresponding city is also included in the search keywords. For example, “*city hall, singapore*” and “*city hall, new york*” are used to search the images of City Hall in Singapore and New York, respectively. The most relevant images are retrieved using the tag-based method provided by Flickr’s public API.<sup>4</sup> This method requires the returned images must contain the query terms in their tags, and the returned images are sorted in descending order based on relevance. The top 4000 images in the returned list are taken. Notice that not all images in the list can be successfully downloaded. Besides, some landmarks have less than 4000 images tagged with their names in Flickr. Thus, the number of downloaded images for each landmark is 3301 on average before processing. Different kinds of related data associated with each image are also collected. We categorize the associated information into three types: *surrounding text*, *context information* and *metadata*. Details can be found in [Table 1](#). The surrounding text includes title, tags, description and comments, which directly represent the semantic features of the image. We consider four different kinds of context information: *taken time* is about when the image was captured; *upload time* refers to the time when the image was uploaded to Flickr; *geo-location* generally is about the location where the image was taken; *contextual URLs* contains the URLs of photostream, sets and pools to which the image belongs. Each image in Flickr has an unique *photo ID*; *uploader ID* refers the ID of the user who contributed the image; *EXIF/TIFF* contains the image metadata, such as the device used to capture the image and parameter-setting of the device at the time of taking the image. The *source page* of the image in Flickr is kept as the backup reference.

It is worth mentioning that the created dataset will contain some data noise because the tag-based search method is used to collect images. Since social tags are known to be noisy [21,22], there could be some images which are labeled with a landmark but do not contain any related visual contents about the landmark. Besides, some landmark images are labeled with tags but are not about the corresponding landmarks. Because we aim at developing a dataset to facilitate the development of landmark

<sup>2</sup> <http://www.wikipedia.org/>

<sup>3</sup> [http://wikitravel.org/en/Main\\_Page](http://wikitravel.org/en/Main_Page)

<sup>4</sup> <https://www.flickr.com/services/api/>

**Table 2**

The number of landmarks and the distribution of image number across landmarks in each city.

City	Number of landmarks	Distribution of number of images		
		Average	Max	Min
Beijing	25	2874.64	3680	728
London	25	2994.56	3252	1386
Singapore	28	2562.11	3401	589
New York	28	2898.68	3692	741
Hong Kong	22	2523.14	3882	556

**Table 3**

Statistics of surrounding text in the dataset.

Length of Surrounding Text	Average	Max	Min
Number of tags	11.24	160	1
Number of keywords in title	4.07	41	0
Number of keywords in description	65.34	14,303	0
Number of comments	13.74	2055	0
Number of keywords in comment	4.95	593	1

image retrieval systems for real applications, it is essential that the dataset has similar content distributions with the real environment that are faced by users in real image search scenarios from various aspects, such as the distribution of diversified landmark visual contents and the distributions of tag noise and ambiguities. Thus, we have not cleaned the data noise. In the evaluation, we create the ground truth set for the targeted search landmarks, in which the positive images are guaranteed to be contain related visual contents of the landmarks based on the defined judgement criterion and procedures (refer to Sections 3.1.1 and 3.2.2).

## 2.2. Dataset statistics

Using the method described above, totally 419,346 images are collected at Flickr’s medium-scale image resolution, which is  $500 \times 500$  pixels maximum. In the collected images, the most common sizes are “ $500 \times 375$ ”, “ $500 \times 333$ ”, “ $375 \times 500$ ”, and “ $333 \times 500$ ”, accounting for 59.04% of the dataset. We remove the image whose length or width is less than 300. Also, if any piece of the related information listed in Table 1 fails to be downloaded, the corresponding image would not be included to our test collection. Finally, there are 355,141 images left for 128 landmarks.

The distribution of image number for landmarks in each city is shown in Table 2. Meanwhile, Table 3 shows the statistics information of surrounding text of images. Note that more than half of the images do not have any comment, and the minimal number of tags is 1 is because of the used tag-based search method. Fig. 2 shows the distribution of the number of tags per image. In this figure, the number of images has been normalized. To assess the quality of the downloaded images, 20,000 images are randomly drawn from the whole dataset and manually evaluated. 735 images are labeled as low-quality,

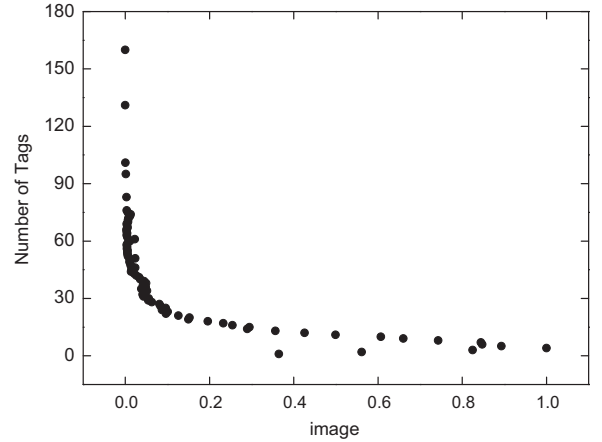


Fig. 2. The number of tags per image.

representing 3.675% of the subset. This ratio can be regarded as an indicator of the proportion of low-quality images in the whole dataset.

## 2.3. Visual features

For the convenience of utilizing the dataset on the performance evaluations of various applications, we extract and provide a set of effective and widely used visual features for each image. They include,

*Color histogram (64D)* [23]: The HSV color space is divided into 64 partitions, and the number of pixels within each partition is then counted for computing the histogram bin of the corresponding color.

*Color auto-correlogram (144D)* [24]: The color auto-correlogram describes the global distribution and the spatial correlation of pairs of colors together. We consider the HSV color space with color quantized into 36 bins, and use 4 distance metrics as [24] to compute the auto-correlogram.

*Gabor texture (72D)* [25]: Wavelet features are extracted at multiple scales and directions from the images using a Gabor wavelet decomposition. The mean and standard deviation of the filter responses are calculated. We extract Gabor features in six different orientations and six different scales.

*Block-wise color moments (225D)*: Each image is divided into  $5 \times 5$  grid partitions. For each grid, the first three color moments (*mean, variance, skewness*) are calculated for each color channel in HSV color space. Each grid region is then characterized by 9 moments, resulting in a 225-dimensional vector for an image.

*Edge histogram (80D)* [26]: 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  $4 \times 4$  grid, and each grid is further divided into small square blocks. Five directional edges are extracted from the small blocks. Then the number of five edge types in each grid is counted to define five histogram bins for the corresponding grid.

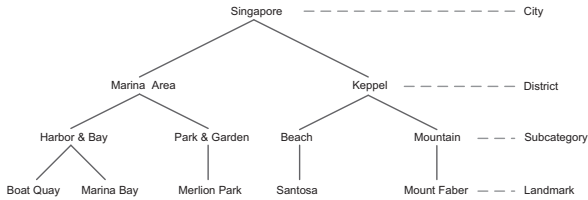


Fig. 3. Hierarchical Structure of the Dataset.

*Bag-of-visual-words* [27]: 500-D bag-of-visual-words (BoVW) is generated for each landmark. For each image, key-points are detected using difference of Gaussian. Then each key-point is described by a 128 dimensional SIFT descriptor [27]. Finally, the descriptors of each image are vector quantized into a vocabulary of visual words, which are generated by *k*-means clustering method.

#### 2.4. Dataset structure

In order to support fast browsing and exploration of the collection, the dataset is organized in hierarchical structure based on geo-location and landmark categories. Under two general categories – *Natural Attractions* and *Man-made Attractions*, 13 subcategories are further defined and the details are as below:

- *A. Natural attractions*: (A1) beach, (A2) island, (A3) mountain, (A4) nature reserve, (A5) wildlife attractions and (A6) park and garden;
- *B. Man-made attractions*: (B1) buildings and monuments, (B2) distinct small town, (B3) harbor and bay, (B4) historic resort, (B5) museum and gallery, (B6) religious architecture, (B7) shopping and commercial center.

Particularly, the landmarks are first divided based on cities and spatial districts; then landmarks in the same district are classified into different subcategories. As an example, Fig. 3 illustrates the hierarchical structure using several landmarks in Singapore. The structure of landmarks in the whole dataset is shown in Appendix A Tables A1–A3.

### 3. Experimental study

In this section, we report two empirical studies on the dataset. The two sets of studies aim to investigate the landmark image search accuracy with the use of basic Content-based Image Retrieval (CBIR) and Text-based Image Retrieval (TBIR) methods. Specifically, in Section 3.1, we study the search accuracy and efficiency of CBIR methods on landmark image search by using different types of landmark views and different visual features. A landmark can have many different representative views (e.g. exterior and interior views, or views of different parts), and users may be interesting in search different aspects of a landmark in real applications. However, different views could have different search difficulties, due to the distinct visual appearances, it is necessary to develop the best search strategies for different types of

landmark views. In this experiments, we provide the baseline performance on five types of landmark views. In Section 3.2, the landmark image search accuracy based on social tags with two popular TBIR methods are studied. Besides, we also explore the performance improvement with the combination of textual and visual features. Besides, by simply categorizing landmarks into two types, we demonstrate that for different types of landmarks, better search performance can be achieved by using different search methods, which implies the necessity of developing different search strategies for different types of landmarks. All the experiments are conducted on a desktop computer with Intel Core i5 2.80 GHz CPU and 4GB memory. In the following, we detail the experimental configuration and report the experimental results for the two set of experiments.

#### 3.1. Content-based landmark search

This section reports the study on the performance of CBIR methods on landmark image search. In particular, we study (1) the search performance of different visual queries which represent different views of a landmark, and (2) the search effectiveness and efficiency of different visual features. insights into the content-based landmark search.

##### 3.1.1. Experimental setup

*Query set*: We select eight landmarks<sup>5</sup> in Singapore and take pictures of various views of each landmark. From the taken photos, images which represent the following types of views are selected as queries, including (1) views from different angles (i.e. front views and side views), (2) partial views (i.e. different parts of the landmark), (3) interior views, (4) close-up exterior views, and (5) far-away exterior views. Fig. 4 shows the examples of different views. For each landmark, we select four queries for each type of view. Finally, total 156 queries are selected.<sup>6</sup>

*Test collection*: The targeted landmarks belong to the subcategories of *park and garden*, *buildings and monuments*, *harbor and bay* and *museum and gallery*. We select images of landmarks in those subcategories to construct a challenging distractor subset, as images of landmarks in the same subcategory are more likely to be similar. Images of landmarks in *Singapore*, *New York* and *London* are used. Altogether, 59 landmarks with 164,690 images are included in the subset.

*Visual features*: We use color histogram (CH), color moments (CM), bag-of-visual-words (VW), and two combinations of them (CH + CM and CH + CM + VW) as visual features in the experiments. A vocabulary with 1000 visual words is generated for the subset using the method described in Section 2.3. Euclidean distance is used for calculating the similarity score. In the combination of different features, the similarity scores are separately computed and normalized, and then uniformly summed together to obtain the final score.

<sup>5</sup> The selected landmarks include *Armenian Church*, *Cathedral of the Good Shepherd*, *Church of Our Lady of Lourdes*, *Church of Saints Peter and Paul*, *Marina Bay*, *Merlion Park*, *National Museum* and *Raffles City*.

<sup>6</sup> Because *Merlion Park* is an open area, it does not have queries representing the interior views.



Fig. 4. Different types of views using images of the Cathedral of the Good Shepherd as examples.

Table 4

Average precision of visual queries in different view types. The representations of the acronyms in the table: AV – views from different angles; PV – partial views; IV – interior views; CUV – close-up exterior views; FAV – far-away exterior views; CH – color histogram; CM – color moments; VW – bag-of-visual-words. The values in the table are the average  $P@10 \pm \text{std}$ .

Visual Feature	AV	PV	IV	CUV	FAV
CH	$0.068 \pm 0.136$	$0.052 \pm 0.079$	$0.017 \pm 0.038$	$0.042 \pm 0.088$	$0.029 \pm 0.076$
CM	$0.096 \pm 0.179$	$0.039 \pm 0.047$	$0.013 \pm 0.045$	$0.031 \pm 0.072$	$0.032 \pm 0.061$
VW	$0.105 \pm 0.182$	$0.068 \pm 0.125$	$0.046 \pm 0.081$	$0.071 \pm 0.139$	$0.117 \pm 0.186$
CH+CM	$0.111 \pm 0.211$	$0.076 \pm 0.148$	$0.034 \pm 0.073$	$0.046 \pm 0.069$	$0.034 \pm 0.075$
CH+CM+VW	$0.242 \pm 0.235$	$0.131 \pm 0.216$	$0.087 \pm 0.165$	$0.165 \pm 0.236$	$0.121 \pm 0.132$

*Ground truth and evaluation:* The task of landmark image retrieval is to search visual views of the desired landmark, which means that positive results must be or at least contain visual views of the targeted landmark. According to this, the judgement criterion is defined as: if an image contains views of the targeted landmark that are recognizable to viewers, then the image is regarded as positive; otherwise, the image is marked as negative. The top 10 results of each query are assessed by human evaluators. As the search results may contain partial or interior views of landmarks, evaluators are required to be familiar with the selected landmarks. Five evaluators in Singapore are recruited for evaluating the search results. There are 3 females and 2 males aged from 20 to 30 years old. They have been to the selected landmarks multiple times, and thus are considered to be familiar with these landmarks. The final judgment on relevance are made based on majority voting. The precision ( $P@10$ ) is used as

the evaluation metric. same landmark are pooled together, and then assessed by human evaluators.

### 3.1.2. Experimental results and analysis

In the following, we present the core results and provide detail analysis on system performance in terms of effectiveness and efficiency.

*On effectiveness:* Table 4 shows the search accuracy of different visual queries in each view type. The values in the table are the average  $P@10$  ( $\pm$  standard deviation) over all queries in the same type. From the table, we can see that the performances of color histogram and color moments are comparable, and the feature of visual words performs slightly better. Although the combination of global and local features improves the search accuracy, it is still pretty low. Content-based landmark search is more challenging than general content-based image retrieval tasks. The mechanism of content-based retrieval method decides it can only return visually similar images (with

respect to the query image) in machine's view. While in landmark image retrieval, the positive results should represent or contain *visual appearance of the targeted landmark*, which implies that a result could be negative even it has similar visual content with the query image.

Queries with the whole view of a landmark are expected to get better results than partial views (*PV*) and interior views (*IV*). Queries of three types (views of different angles (*AV*), close-up views (*CUV*) and far-away views (*FV*) contain the whole exterior view of a landmark. In general, *CUV* contains more details and *FV* contains foreground and background objects (e.g. pedestrians and vehicles). These extra details and objects increase the difficulty of retrieval based on visual features. As a result, *AV* obtains the best results among them. Surprisingly, *PV* gets similar performance as *CUV*. We find that it is because that the captured partial views are usually representative scenes or objects which tend to attract more attentions, resulting in more images about them. The search performance of *IV* is the worst, which is in our expectation. Because interior views typically contain more objects and complicate structures, and the lighting conditions vary greatly at different time or from different angles. In comparison to the performance of different landmarks, we found that the queries of *Marina Bay* and *Merlion Park* obtain much better results than queries of other landmarks, which are all buildings. It implies that buildings are more difficult to search based on visual features. And it can be explained by the viewpoint in [28]: “*buildings tend to have few discriminative visual features and many repetitive structures*”.

*On efficiency:* As a baseline study, we have not used any indexing method in the experiments. For single visual feature, the computation time is spent on computing and sorting the similarity scores of all images in the subset; for the combination of different features, there are two additional time-consuming steps – the normalization and summation of computed scores of individual features. Obviously, higher dimensionality of visual features needs longer processing time. The results are shown in Fig. 5. From the results, we can see that the slight improvement on search performance by combining visual features pays high time cost. With the 1289-dimensional combined features, each query takes 4.14 s, which is much longer than the computation time of 1000-dimensional visual words (1.19 s).

### 3.2. Text-based landmark search

In this section, we investigate the search performance of using tags as textual information sources for landmark search, and also study the effects of combining textual and visual features on search accuracy. In experiments, we use two popular text-based retrieval methods and a basic linearly combination of textual feature and visual features to study the retrieval performance. The results can provide some evidences about the search performance of landmark image search with social tags.

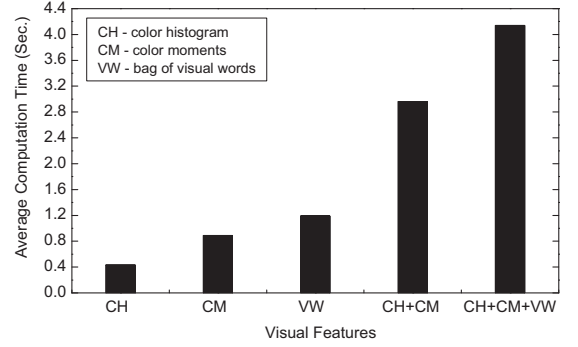


Fig. 5. Computation time per query using different visual features.

#### 3.2.1. Retrieval methods

Social images are usually accompanied with user-contributed annotations, such as title, tags, descriptions and comments. Although social tags are considered to be noisy, they are the most commonly used textual information for social media retrieval [29–31]. To study the landmark image search performance based on social tags, we use two basic text-based retrieval methods OKAPI-BM25 model [32] and Vector space model (VSM) with standard TF-IDF term weighted method [33], which are widely used as baselines in information retrieval. Besides, we also study the performance of the popular late fusion method of text-based retrieval methods (TBIR) and content-based retrieval methods (CBIR) in landmark image search [29,34]. In our implementation, for each image, tags are first tokenized with a standard stop-list, and then concatenated to form the document-term matrix for the image. Terms with occurrence times less than 10 in the corpus are filtered. In the next, we introduce the details of each retrieval method.

*OKAPI-BM25 model:* In this method, the relevant score of an image  $I$  with respect to a query  $q$  is computed as:

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

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  $\log((|D| - |n_t| + 0.5)/(|n_t| + 0.5))$ , where  $|n_t|$  is the number of images whose tags contain  $t$ .  $l_I$  is the number of terms in the image  $I$ , and  $l_{avg}$  is the average number of terms of all images in  $D$ . In our study,  $k_1$  is set to 0.2 [32] and  $b$  is set to 0.75 [35].

*Vector space model (VSM):* The classical TF-IDF weighting scheme is used in this study. In this model, the relevant score is computed as:

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

where  $w_{t,I}$  is the term weight of term  $t$  in the image  $I$ , computed as  $w_{t,I} = tf(t) \cdot \log((|D| - |n_t| + 0.5)/(|n_t| + 0.5))$ . The computation of  $w_{t,q}$  is analogous to  $w_{t,I}$ .



**Weighted linear combination (WLC):** WLC combines the search results of CBIR and TBIR methods using a late fusion method. Specifically, in CBIR, each image is represented by the concatenation of the six types of visual features (described in Section 2.3), and the Euclidean distance is used as similarity measurement. Both text-based retrieval methods are used in WLC: (1) WLC\_BM25 denotes the combination of Okapi-BM25 with the CBIR method, and (2) WLC\_VSM denotes the combination of VSM with the CBIR method. The similarity scores of an image computed by the CBIR method and TBIR method are separately normalized using MinMax [36], and then linearly combined using CombSUM [36,29] with pre-defined weights to compute the final similarity score for the image. The similarity score of CBIR method 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

$$S(q, I) = w_* \cdot S_t + (1 - w) \cdot S_v \quad (3)$$

$S_p$  and  $S_t$  are the similarity scores returned by the CBIR method and TBIR method, respectively.  $w$  is empirically tuned in experiments.
















### 3.2.2. Experimental setup

This section introduces the construction of the query set, test collection, the assessment of the ground truth and the evaluation metrics.

**Query set:** For each city in the dataset, we select five landmarks as targeted landmarks for retrieval, and formulate three text queries for each landmark. Among the three text queries of a landmark, one is the landmark name with the city name (e.g. *national museum, singapore* and *disneyland resort, hong kong*), which is the most direct search method for a landmark. The other two text queries are comprised by the landmark name, city name and *another popular term* to describe the landmark in tags. The term in each query is selected by counting the occurrence times of all the terms in the social tags of the landmark images, which are the landmark subset crawled from Flickr (described in Section 2.1). For each text query, an image query is selected for CBIR. The image query of an image landmark are manually selected to contain the overview or a representative view of the landmark. The selected query landmarks for each city and some examples of the text queries (with additional term) and image queries for each landmark are shown in Table 5. In the figure, the text query examples does not include the city name for simplicity. In total, there are 75 text queries and 75 corresponding textual and visual features (for WLC search methods) in the query set. Notice that in the table, landmarks are labeled with superscript “I” or “II”, which are used to identify two types of landmarks: (1) *Type I* – this type of landmarks is a single building or complex building, such as “London Eye”, and “Statue of Liberty”, and (2) *Type II* – this type of landmarks usually contain a large area and

**Table 5**

Query examples of targeted landmarks in experiments. Landmarks labeled with superscript “I” are Type I landmarks and landmarks with label “II” are Type II landmarks.

City	Landmarks	Query Examples
Beijing	Forbidden City <sup>II</sup> Great Wall <sup>II</sup> Old Summer Palace <sup>II</sup> Temple of Heaven <sup>I</sup> Tiananmen Square <sup>II</sup>	    
Hong Kong	Avenue of Stars <sup>II</sup> Disneyland Resort <sup>II</sup> Peninsula Hotel <sup>I</sup> Tian Tan Buddha <sup>I</sup> Victoria Harbour <sup>II</sup>	    
London	Big Ben <sup>I</sup> Buckingham palace <sup>I</sup> Palace of Westminster <sup>I</sup> Westminster Abbey <sup>I</sup> The London Eye <sup>I</sup>	    
New York	Natural History Museum <sup>I</sup> Central Park <sup>II</sup> Rochefeller Center <sup>I</sup> Saint Patrick's Cathedral <sup>I</sup> Statue of Liberty <sup>I</sup>	    
Singapore	Marina Bay <sup>II</sup> Merilon Park <sup>II</sup> National Museum <sup>I</sup> Santosa <sup>II</sup> Universal Studios <sup>II</sup>	    

**Table 6**Landmark search performance (mean  $\pm$  std.) based on textual and visual features.

Method	P@1	P@5	P@10	MAP@10	MRR
CBIR	0.000 $\pm$ 0.000	0.118 $\pm$ 0.166	0.107 $\pm$ 0.119	0.107 $\pm$ 0.135	0.206 $\pm$ 0.202
Okapi-BM25	<b>0.767 <math>\pm</math> 0.426</b>	0.743 $\pm$ 0.255	0.745 $\pm$ 0.211	0.746 $\pm$ 0.250	0.853 $\pm$ 0.274
VSM	0.493 $\pm$ 0.503	0.636 $\pm$ 0.276	0.647 $\pm$ 0.217	0.613 $\pm$ 0.254	0.695 $\pm$ 0.313
WLC_BM25	0.753 $\pm$ 0.434	<b>0.816 <math>\pm</math> 0.228</b>	<b>0.784 <math>\pm</math> 0.186</b>	<b>0.798 <math>\pm</math> 0.217</b>	<b>0.865 <math>\pm</math> 0.242</b>
WLC_VSM	0.507 $\pm$ 0.503	0.600 $\pm$ 0.296	0.541 $\pm$ 0.186	0.584 $\pm$ 0.296	0.732 $\pm$ 0.281

more points of interests, such as “Universal Studios, Singapore” and “Central Park, New York”. There are 13 Type I landmarks and 12 Type II landmarks in the queries. The underlying intuition of the classification is the big difference of the two types of landmarks. For the Type I, the view of the landmark should always contain the building; and for the Type II, the images of the landmarks could contain different partial views or scenes of the landmark. The difference may lead to different characteristics of the search performance by TBIR and WLC methods.

*Test collection:* With the targeted landmarks, we use the same methodology in Section 3.1 to construct a challenging distractor subset. The targeted landmarks are in seven subcategories: (1) beach, (2) park and garden, (3) buildings and monuments, (4) harbor and bay, (5) historic resort, (6) museum and gallery, and (7) religious architecture. Thus, images of landmarks in these subcategories of the five cities are used to construct the test collection. In total, there are 266,398 images from 98 landmarks are used.

*Ground truth:* Similar to the experiments in Section 3.1, for a targeted landmark, the positive images should contain representative views of the landmark such that the landmark can be easily identified. For images with partial views (such as a scene of a landmark in Type II) which are not discriminatively enough to identify the landmark, they are labeled as negative. Based on the criterion, the positive results are guaranteed to be correct, while some positive results might be mistakenly labeled as negative. In the assessment, the top 10 search results of each method for a landmark are pooled together and assessed by human subjects. As the landmarks are across several countries, evaluators are recruited from China (8 subjects), Singapore (5 subjects) and America (4 subjects).<sup>7</sup> For a landmark, its search results are assessed by three subjects who have been to the landmarks.<sup>8</sup> The majority voting is used to obtain the final annotations, which are then used to evaluate the search performance of each method.

*Evaluation metrics:* In evaluation, we focus on the search accuracy on the top of the list, because it is the most interesting part for users of information retrieval.

*Precision at  $k$  ( $P@k$ ):* It evaluates the proportion of relevant instances in the top  $k$  retrieved results. The values of  $k$  used in experiments are 1, 5, and 10.

*Mean reciprocal rank (MRR):* It averages the inverse of the rank of the first correct answer for each query. It

measures the level of the ranking list at which the information need of the user is first fulfilled.

*Mean average precision (MAP):* It averages the precision at each point of a relevant instance in the ranking list. In experiments, we report the results of MAP@10.

### 3.3. Experimental results

Table 6 shows the average search accuracy over all the queries by different methods. Similar to the results reported in Section 3.1, the accuracy of CBIR method is very low. In contrast, TBIR methods based on social tags can obtain fairly good performance, especially for Okapi-BM25, which can achieve 76.7% for the first search results (P@1) and 74.5% precision in the top 10 results (P@10). VSM does not work as good as Okapi-BM25, but it is still acceptable on the top 10 results. Although the search performance of CBIR is very poor (only 10.7% for P@10), the weighted linear combination of the TBIR and CBIR can improve the search performance in the top search results in general, except that P@1 of Okapi-BM25 is slightly better than WLC\_BM25. The optimal weight of Okapi-BM25 in WLC\_BM25 is 0.3 (for P@10) and the optimal weight of VSM in WLC\_VSM is 0.2 (P@10). The best performance is obtained by setting larger value to CBIR method in the WLC methods, which is consistent with the conclusion in [29].

In the following, we present and analyze the search performance differences of landmarks in Type I and Type II. As Okapi-BM25 is the better TBIR search method, we use its search results and search performances with different weight settings ( $w$ ) in WLC\_BM25 in the following discussion. The average search performance of the two types of landmarks based on Okapi-BM25 and WLC\_BM25 ( $w=0.5$  for Type I and  $w=0.1$  for Type II) are shown in Table 7. With Okapi-BM25 method, the average search performance of landmarks in Type I are much better than that of landmarks in Type II. It is interesting that, the performance of Type I landmarks slightly decreases with the combination of CBIR and TBIR method, while the performance of Type II landmarks increases a lot after combination. Besides, for Type I landmarks, there is no optimal weight  $w$  for all evaluation metrics in WLC\_BM25 method, while for Type II landmarks, the performance with weight  $w=0.1$  is consistently better than other weight settings over all metrics. A potential reason of landmarks in Type I obtaining much better search results than landmarks in Type II is that the visual appearances of a landmark in Type I is much simpler than that of a landmark in Type II. Accordingly, the tags of landmarks in Type I is generally less than that of landmarks in Type II, and thus contain less confusing information and less noise, leading to better text-based search results. For Type I landmarks, CBIR is to search

<sup>7</sup> The landmarks of London are evaluated by subjects from China and Singapore, who have been to these landmarks.

<sup>8</sup> In this experiments, the selected landmarks are the most famous ones in each city, and most search results are exterior views. The judgment is relative easier than the previous experiments.

buildings with similar appearance; and for Type II landmarks, CBIR is to search similar scenes. Although the visual appearance of Type II landmarks is more complex, while in landmark search, it is more difficult to search the same building by CBIR, because different buildings usually have similar visual appearance. For the landmark search task, it is more like a building identification task for the Type I landmarks. Thus, in WLC\_BM25 method, the incorporation of CBIR search results may introduce noisy for Type I landmark search, resulting in performance decrease. Fig. 6 shows the performance varying curves with different weight settings in WLC\_BM25 for two types of landmarks. With the increasing of  $w$ , the performances display different trends for two types of landmarks. For Type I, the overall better performance of WLC\_BM25 are obtained when  $w \in [0.5, 0.8]$  for all evaluation metrics (when  $w=1$ , it is Okapi-BM25 method), while for Type II, the better performance are obtained when smaller weight is set to TBIR method.

From the analysis of the search performance differences of Type I and Type II landmarks, we can find that for different landmarks, different search strategies should be applied to obtain better search results. In the baseline studies, we only analyze the search performance difference on accuracy with a simple categorization of the landmarks. It should be more interesting to analyze the search performance differences for finer landmark categories, which can provide fundamental knowledge for the development of optimal search strategies of different

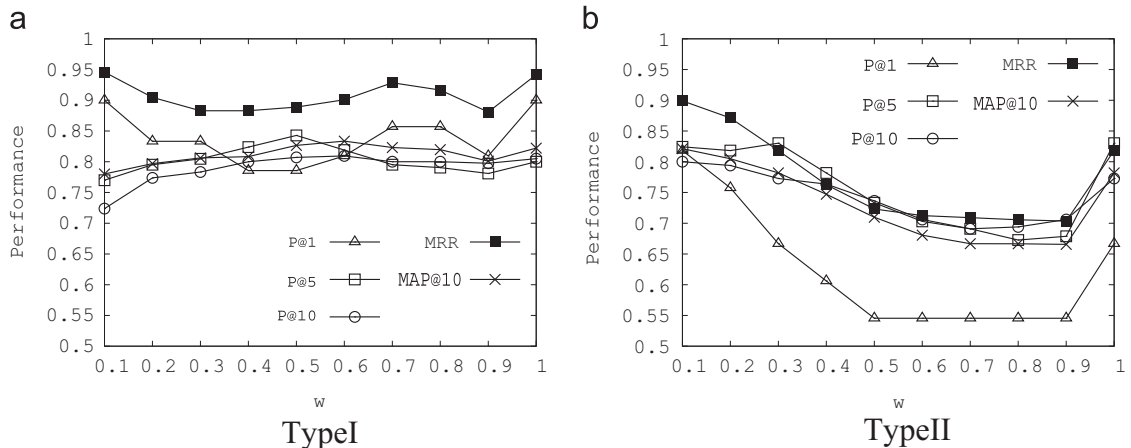
landmarks. Besides, in landmark search, the result diversification is a very practical and interesting research direction, as users usually want to see different aspects of a landmark. In text-based landmark image search, the search results may contain different views of a landmark. However, with the use of image query (e.g. to improve the search performance for the Type II landmarks in this study) will lead to similar search results. How to generate the most representative and diverse views in the top results are worthy of more studies.

#### 4. Conclusion and future work

In this paper, we introduced the construction of a large scale landmark image dataset. The dataset contains various kinds of textual features and provides six types of visual features. Based on the dataset, we identified and discussed several closely related research issues. We also conducted a set of experimental studies on landmark search using visual and textual features. The search accuracy of different visual queries and the search efficiency of different visual features were reported and comprehensively analyzed. The results disclosed the weakness of content-based landmark search. Both search accuracy and efficiency need to be improved. Text-based retrieval methods for landmark image search using social tags can achieve good search performance. And the combination of CBIR and TBIR can improve the search accuracy in general, however, it does not works for all types of landmarks. From the comparisons between search performances of two types of landmarks, we demonstrated that better search performances can be obtained by using different search strategies for different types of landmarks. The results can be used as baselines to facilitate the future related research. The test collection is designed to facilitate the development of large scale landmark image retrieval systems. It can be potentially applied in the study and comparisons of mobile landmark search systems, which have been recently attracting more and more attentions. The specific research

**Table 7**  
Search performance of two types of landmarks.

Metric	Okapi-BM25		WLC_BM25	
	Type I	Type II	Type I	Type II
P@1	0.900	0.786	0.667	0.818
P@5	0.800	0.843	0.830	0.824
P@10	0.805	0.807	0.773	0.800
MAP	0.822	0.827	0.782	0.821
MRR	0.942	0.889	0.818	0.899



**Fig. 6.** The influence of  $w$  in weighted linear combination method (WLC\_BM25) on two types of landmarks. (a) Type I. (b) Type II.

**Table A1**

Hierarchy structure and all landmarks in the collection (*Part I*). Types of landmarks – *A. Natural attractions*: (A1) beach, (A2) island, (A3) mountain, (A4) nature reserve, (A5) wildlife attractions and (A6) park and garden; *B. Man-made attractions*: (B1) buildings and monuments, (B2) distinct small town, (B3) harbor and bay, (B4) historic resort, (B5) museum and gallery, (B6) religious architecture, (B7) shopping and commercial center.

City	Districts	Types	Landmarks	Image number	
Singapore	Marina Area	B3	Boat Quay	2887	
			Marina Bay	1108	
	CBD	A6	Merlion Park	3070	
		B6	Armenian Church	3232	
		B1	City Hall	2805	
	Beach Road		Raffles City	3401	
			National Museum	3128	
		B6	Cathedral of the Good Shepherd	677	
			Church of Our Lady of Lourdes	2887	
			Church of Saints Peter and Paul	3279	
		Tanglin	A6	Singapore Botanic Gardens	3116
		Newton	A4	Bukit Timah Nature Reserve	2861
	Orchard	A6	Istana Park	728	
		B7	Orchard Road	2888	
	Changi	B7	Changi Airport	2762	
		A2	Pulau Ubin	2363	
	Far North	A5	Night Safari Singapore	2834	
			Singapore Zoo	2755	
			Jurong Bird Park	3333	
	Jurong	A4	Sungei Buloh Wetland Reserve	2946	
		Hougang	B6	Church of the Nativity of the Blessed Virgin Mary	589
				Butterfly Park & Insect Kingdom	2088
			Harbourfront center	2863	
	Keppel	B7	Kusu Island	2986	
		A2	Mount Faber	2766	
		A3	Santosa	2137	
		A1	Universal Studios Singapore	2740	
		B1	Haw Par Villa	2510	
	Sourth West	A6			

**Table A2**  
Hierarchy structure and all landmarks in the collection (Part II).

City	Districts	Types	Landmarks	Image number	
New York	Manhattan and Brooklyn	B1	Brooklyn Bridge	2846	
		B5	Ellis Island Immigration Museum	2722	
	Liberty Island	B1	Statue of Liberty	2904	
		B1	Carnegie Hall	3287	
	Manhattan			Chrysler Building	3158
				City Hall	3259
				Dakota Apartments	2259
				Empire State Building	2637
				Federal Hall National Memorial	1022
				Flatiron Building	3140
				Grand Central Terminal	3199
				Lincoln Center	3369
				Macy's Department Store	3198
				New York Public Library	3692
				New York Stock Exchange	3319
				Plaza Hotel	3397
				Radio City Music Hall	3214
				Rockefeller Center	3195
				National September 11 Memorial	2068
			B5	American Museum of Natural History	3447
				Metropolitan Museum of Art	3545
				Solomon R. Guggenheim Museum	3519
				The Morgan Library and Museum	741
				The Museum of Modern Art (MoMA)	2218
		A6	Central Park	2927	
		B6	Saint Patrick's Cathedral	2858	
	B7	Times Square	3086		
	B1	Staten Island Ferry	2937		
Beijing	Changping	B4	Ming Dynasty Tombs	1997	
		B7	Beijing CBD	2781	
	Chaoyang	B1	Beijing National Stadium	2918	
		A6	Chaoyang Park	3435	
				Olympic Green	3202
			B7	Silk Street	2935
			B6	Temple of Heaven	3231
			B4	Bell Tower and Drum Tower	3528
			B5	National Art Museum of China	3050
				National Museum of China	2170
			B1	Tiananmen Square	3179
			B7	Wangfujing	3278
				Xidan	2262
			B4	Yonghegong	3680
	Haidian		A6	Botanical Garden	3312
				Old Summer Palace	3485
			A3	Fragrant Hills	3628
			B7	Zhongguancun	1795
	Shunyi		A6	Olympic Water Park	3382
		Xicheng	A6	Beihai Park	3044
			B7	Beijing Financial Street	728
			B4	Forbidden City	2975
			A6	Shichahai	1925
		Yanqing Xian	B4	Great Wall	2927

**Table A3**  
Hierarchy structure and all landmarks in the collection (*Part III*).

City	Districts	Types	Landmarks	Image number		
London	Between Tower Hamlets and SouthWark	B1	Tower Bridge	2907		
		B5	British Museum	3120		
	Camden	A6	Hampstead Heath	2996		
		B6	St Paul's Cathedral	3146		
	Camden and Barnet	B1	The Gherkin	3208		
		B1	Big Ben	2990		
	City of Westminster	City of Westminster	B1	Marble Arch	3252	
			B4	Trafalgar Square	3122	
			B4	Buckingham palace	2764	
		City of Westminster and Camden	B4	Palace of Westminster	2962	
			B5	National Gallery	3119	
		Kensington and Chelsea	B5	National Portrait Gallery	3111	
			A6	St Jame's Park	1386	
			B5	Tate Britain	3166	
			B6	Westminster Abbey	2951	
			A6	Regent's Park	3202	
	B5		Victoria and Albert Museum	3132		
	City of Westminister and Kensigton and Chelsea	City of Westminister and Kensigton and Chelsea	A4	Natural History Museum	3057	
			A4	Hyde Park	3007	
		Lambeth	B1	Kensington Gardens	3103	
			B1	The London Eye	2768	
	Richmond upon Thames	A6	Bushy Park	3072		
		A4	Richmond Park	3191		
	Southwark	Southwark	B5	Tate Modern	2947	
			B1	The Shard	3185	
		B3	Victoria Harbour	2652		
Hong Kong	Hong Kong Island and Kowloon Peninsula	B3	Victoria Harbour	2652		
		B1	Star Ferry Pier Central	1962		
	Hong Kong Island	A3	Victoria Peak	1660		
		B2	Tai O	2420		
		B6	Tian Tan Buddha	3146		
	The New Territories	B1	Sai Kung Pier	779		
		B1	Sai Kung Pier	779		
	Hong Kong Island	A6	Ocean Park	2699		
		B3	Repulse Bay	3072		
		B2	Stanley Village	556		
		B2	Stanley Village	556		
	Kowloon	Kowloon	B6	Chi Lin Nunnery	3466	
			B6	Chi Lin Nunnery	3466	
		Kowloon	B1	Wong Tai Sin Temple	3358	
			B1	Wong Tai Sin Temple	3358	
		Kowloon	Kowloon	B1	Avenue of Stars	3308
				B1	Clock Tower	1971
			Kowloon	B1	Ocean Terminal	1322
				B1	Peninsula Hotel	2984
			Kowloon	B1	Science Museum	1653
				A6	Kowloon Park	3295
	Kowloon		B7	Kowloon Park	3295	
B7			Nathan Road	3331		
The New Territories	The New Territories	B7	Temple Street	3340		
		A6	Wetland Park	1482		
	The New Territories	A6	Disneyland Resort	3882		
		A2	Lamma Island	3171		

issues include (1) how noise and low quality of mobile captured images can affect the search performance and (2) under mobile environment, how to improve efficiency of landmark image research algorithms. In the near future, we plan to extend the test collection and related ground truth/performance metric to facilitate effective evaluation of region based or object based landmark image search systems.

## Appendix A. Dataset structure

See [Tables A1–A3](#).

## References

- [1] Y. Gao, M. Wang, Z. Zha, J. Shen, X. Li, X. Wu, Visual-textual joint relevance learning for tag-based social image search, *IEEE Trans. Image Process.* 22 (1) (2013) 363–376.
- [2] R. Datta, D. Joshi, J. Li, J.Z. Wang, Image retrieval: ideas, influences, and trends of the new age, *ACM Comput. Surv.* 40 (2) (2008) 1–60.
- [3] M. Wang, K. Yang, X.-S. Hua, H. Zhang, Towards a relevant and diverse search of social images, *IEEE Trans. Multimed.* 12 (8) (2010) 829–842.
- [4] M. Wang, B. Ni, X.-S. Hua, T.-S. Chua, Assistive tagging: a survey of multimedia tagging with human-computer joint exploration, *ACM Comput. Surv.* 44 (4) (2012) 1–24.
- [5] D. Tao, J. Cheng, M. Song, X. Lin, Manifold ranking-based matrix factorization for saliency detection, *IEEE Trans. Neural Netw. Learn. Syst.* (2015) 1–13.

- [6] D. Tao, X. Lin, L. Jin, X. Li, Principal component 2-d long short-term memory for font recognition on single Chinese characters, *IEEE Trans. Cybern.* (2015) 1–10.
- [7] R. Ji, L.-Y. Duan, J. Chen, H. Yao, J. Yuan, Y. Rui, W. Gao, Location discriminative vocabulary coding for mobile landmark search, *Int. J. Comput. Vis.* 96 (3) (2012) 290–314.
- [8] L. Kennedy, M. Naaman, Generating diverse and representative image search results for landmarks, in: *Proceedings of the International Conference Companion on World Wide Web*, 2008.
- [9] L. Zhu, J. Shen, L. Xie, Topic hypergraph hashing for mobile image retrieval, in: *Proceedings of the ACM International Conference on Multimedia*, 2015.
- [10] L. Zhu, J. Shen, H. Jin, R. Zheng, L. Xie, Content-based visual landmark search via multimodal hypergraph learning, *IEEE Trans. Cybern.* (2015) 1–14.
- [11] C. Hong, J. Zhu, Hypergraph-based multi-example ranking with sparse representation for transductive learning image retrieval, *Neurocomputing* 101 (2013) 94–103.
- [12] C. Hong, N. Li, M. Song, J. Bu, C. Chen, An efficient approach to content-based object retrieval in videos, *Neurocomputing* 74 (17) (2011) 3565–3575.
- [13] L. Zhu, J. Shen, H. Jin, R. Zheng, L. Xie, Landmark classification with hierarchical multi-modal exemplar feature, *IEEE Trans. Multimed.* 17 (7) (2015) 981–993.
- [14] Z. Cheng, J. Ren, J. Shen, H. Miao, Building a large scale test collection for effective benchmarking of mobile landmark search, in: *Proceedings of the International Conference on MultiMedia Modeling*, 2013.
- [15] Y.H. Ren, M. Yu, X.J. Wang, L. Zhang, W.Y. Ma, Diversifying landmark image search results by learning interested views from community photos, in: *Proceedings of the International Conference Companion on World Wide Web*, 2010.
- [16] Y. Avrithis, Y. Kalantidis, G. Toliás, E. Spyrou, Retrieving landmark and non-landmark images from community photo collections, in: *Proceedings of the ACM International Conference on Multimedia*, 2010.
- [17] Y. Li, D.J. Crandall, D.P. Huttenlocher, Landmark classification in large-scale image collections, in: *IEEE International Conference on Computer Vision*, 2009.
- [18] W.-C. Chen, A. Battestini, N. Gelfand, V. Setlur, Visual summaries of popular landmarks from community photo collections, in: *Proceedings of the ACM International Conference on Multimedia*, 2009.
- [19] I. Simon, N. Snavely, S.M. Seitz, Scene summarization for online image collections, in: *IEEE International Conference on Computer Vision*, 2007.
- [20] D. Tao, L. Jin, W. Liu, X. Li, Hessian regularized support vector machines for mobile image annotation on the cloud, *IEEE Trans. Multimed.* 15 (4) (2013) 833–844.
- [21] X. Li, C. Snoek, M. Worring, Learning social tag relevance by neighbor voting, *IEEE Trans. Multimed.* 11 (7) (2009) 1310–1320.
- [22] D. Liu, S. Yan, X. Hua, H. Zhang, Image retagging using collaborative tag propagation, *IEEE Trans. Multimed.* 13 (4) (2011) 702–712.
- [23] L.G. Shapiro, G.C. Stockman, *Computer Vision*, Prentice Hall, New Jersey, 2003.
- [24] J. Huang, S. Kumar, M. Mitra, W.-J. Zhu, R. Zabih, Image indexing using color correlogram, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 1997.
- [25] B.S. Manjunath, W.Y. Ma, Texture features for browsing and retrieval of image data, *IEEE Trans. Pattern Anal. Mach. Intell.* 18 (8) (1996) 837–842.
- [26] D.K. Park, Y.S. Jeon, C.S. Won, Efficient use of local edge histogram descriptor, in: *Proceedings of the ACM International Conference on Multimedia*, 2000.
- [27] D. Lowe, Distinctive image features from scale-invariant keypoints, *Int. J. Comput. Vis.* 2 (60) (2004) 91–110.
- [28] D. Chen, G. Baatz, K. Köser, S. Tsai, R. Vedantham, T. Pylvä, K. Roimela, X. Chen, J. Bach, M. Pollefeys, G. Bernd, G. Radek, City-scale landmark identification on mobile devices, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
- [29] Z. Cheng, J. Shen, H. Miao, The effects of multiple query evidences on social image retrieval, *Multimed. Syst.* (2014) 1–15.
- [30] M. Levy, M. Sandler, Music information retrieval using social tags and audio, *IEEE Trans. Multimed.* 11 (3) (2009) 383–395.
- [31] M. Melenhorst, M. Grootveld, M. van Setten, M. Veenstra, Tag-based information retrieval of video content, in: *Proceedings of the International Conference on Designing Interactive User Experiences for TV and Video*, 2008.
- [32] K. Jones, S. Walker, S. Robertson, A probabilistic model of information retrieval: development and comparative experiments—Part 2, *Inf. Process. Manag.* 36 (6) (2000) 809–840.
- [33] J. Benavent, X. Benavent, E. Ves, R. Granados, A.G. Serrano, Experiences at ImageCLEF 2010 using CBIR and TBIR mixing information approaches, in: *Cross-Language Evaluation Forum CLEF 2010*, 2010.
- [34] P. Wilkins, A. Smeaton, P. Ferguson, Properties of optimally weighted data fusion in CBMIR, in: *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2010.
- [35] C. Manning, P. Raghavan, H. Schütze, *An Introduction to Information Retrieval*, Cambridge University Press, Cambridge, 2009.
- [36] J. Shaw, E. Fox, Combination of multiple searches, in: *The Second Text Retrieval Conference (TREC-2)*, 1994.