Effective Multi-Query Expansions: Robust Landmark Retrieval

Yang Wang[†], Xuemin Lin[†], Lin Wu[§♯] and Wenjie Zhang[†]
[†]The University of New South Wales, Kensington, Sydney, Australia
[§]The University of Adelaide, SA, Australia [♯]Australian Centre for Robotic Vision
{wangy, lxue, zhangw}@cse.unsw.edu.au, lin.wu@adelaide.edu.au

ABSTRACT

Given a query photo issued by a user (q-user), the landmark retrieval is to return a set of photos with their landmarks similar to those of the query, while the existing studies on the landmark retrieval focus on exploiting geometries of landmarks for similarity matches between candidate photos and a query photo. We observe that the same landmarks provided by different users may convey different geometry information depending on the viewpoints and/or angles, and may subsequently yield very different results. In fact, dealing with the landmarks with low quality shapes caused by the photography of q-users is often nontrivial and has never been studied.

Motivated by this, in this paper we propose a novel framework, namely multi-query expansions, to retrieve semantically robust landmarks by two steps. Firstly, we identify the top-k photos regarding the latent topics of a query landmark to construct multi-query set so as to remedy its possible low quality shape . For this purpose, we significantly extend the techniques of Latent Dirichlet Allocation. Secondly, we propose a novel technique to generate the robust yet compact pattern set from the multi-query photos. To ensure redundancy-free and enhance the efficiency, we adopt the existing minimum-description-length-principle based pattern mining techniques to remove similar query photos from the $(k + 1)$ selected query photos. Then, a landmark retrieval rule is developed to calculate the ranking scores between mined pattern set and each photo in the database, which are ranked to serve as the final ranking list of landmark retrieval. Extensive experiments are conducted on real-world landmark datasets, validating the significantly higher accuracy of our approach.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

MM'15, October 26–30, 2015, Brisbane, Australia.

ACM 978-1-4503-3459-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2733373.2806233.

Keywords

Landmark Photo Retrieval; Multi-Query Expansions

1. INTRODUCTION

The popularity of personal digital photography has led to an exponential growth of photos with landmarks, *e.g.,* panoramio.com and Picasa Web Album[1]. This highly demands for the research in the area of efficient and effective retrieval of photos based on landmarks (*e.g.,* tower and churches), namely *landmark retrieval.* Given a query photo, the landmark retrieval returns the set of photos with their landmarks highly similar to that of the query photo.

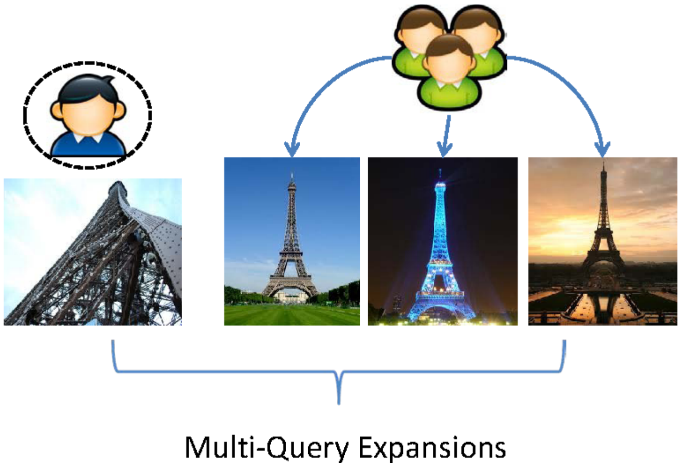


Figure 1: A q-user has issued a biased landmark photo of Eiffel Tower in Paris in the left most photo. Multiple users with the same latent topic as the q-user are selected to recommend three more landmark photos taken at the same place, that complement the given query landmark to construct a multi-query set.

Unlike the conventional image retrieval that performs within the low-level feature spaces (*e.g.,* color and texture), the landmark retrieval is conducted based on geometry information of landmarks. A number of paradigms [8, 10, 14] have been proposed to perform the landmark retrieval under the heterogeneous feature spaces, including the paradigms based on patch level region features [8], mid-level attributes [10], and the combination of low-level features [14].

[1] picasa.google.com

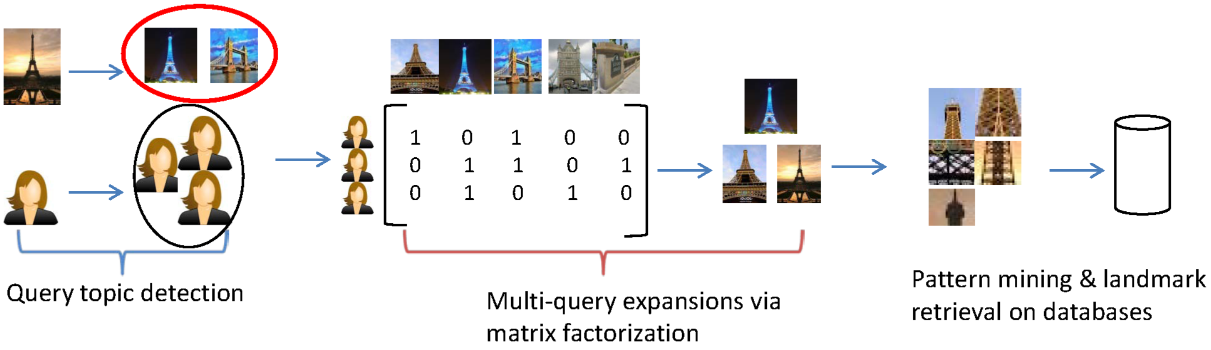

1 0 1 0 0
0 1 1 0 1
0 1 0 1 0
Query topic detection
Multi-query expansions via matrix factorization
Pattern mining & landmark retrieval on databases

Figure 2: The flow chart of our framework. Our framework comprises three main phases: 1) query landmark based topic and user community discovery; 2) select multiple landmark photos from user-query photo matrix to form a multi-query set; and 3) Mining robust and compact pattern set to represent the multi-query set, and calculate the ranking score between such pattern set and each photo in the database, which further lead to the ranking retrieval result for landmark retrieval.

Among the existing techniques, there is one critical assumption: a high quality of query photo is always provided; that is, the landmark captured from a query photo always provides a shape with high quality. Nevertheless, such an assumption is not always true in practice. Indeed, the landmark of a query photo may provide shape with low quality due to various reasons, such as personal preference, photography, etc. For example, as illustrated in Fig. 1, the left most photo taken by a q-user gives the landmark Eiffel Tower with a very low quality shape . Consequently, the existing methods may not be able to return the set of photos that contain the Eiffel Tower since the geometry quality of the landmark in the query photo is too low.

Motivated by this, in this paper we propose a novel method, depicted in Fig. 2, for a robust landmark retrieval through a novel paradigm based on multi-query expansions. Firstly, we propose to identify a set of photos that share the similar latent topics with the query landmark by adopting the Latent Dirichlet Allocation (LDA) techniques [4]. Then, we propose a localized matrix factorization technique over the user-photo matrix that encodes the information from the selected photos. Based on that, a ranking strategy is proposed to generate the top-k photos; this gives the multi-query set with the size $k + 1$ by including the original given query.

We observe that if k is too small then the selected query photos may not be representative, while if k is too large then it will be too expensive to learn robust pattern set (multi-query). Moreover, the multi-query set may contain many similar photos; that is, their landmarks have very similar geometric shapes. To resolve these, we propose to setup a large k, then adopt the existing minimum-description-length-principle [12] based pattern mining techniques to remove similar query photos from the $(k + 1)$ selected query photos. Afterwards, we develop an effective retrieval rule to calculate the ranking score, which represents the similarity between each photo in the database and mined pattern set for multi-query. Finally, the landmark retrieval is performed based on the ranking score of each photo to return the final ranking list.

To sum up, our major contributions are below:

- We propose to tackle the problem of landmarks in query photos with low quality shapes . We present a novel method based on an effective multi-query expansion paradigm.
- A novel strategy is proposed to generate an effective multi-query set. To remove similar photos in the multi query set, we adopt the existing minimum-description-length-principle based pattern mining technique to generate a robust patten set to represent the multi-query. Then, we propose a novel strategy to calculate the ranking score between each photo and pattern set. The final landmark retrieval result is the ranking list of photos as per their ranking score.
- We conducted extensive experiments over real-world landmark photo datasets, validating the higher accuracy of our approach.

The rest of this paper is structured as follows. We review the related work in Section 2. Then, we describe our proposed technique in Section 3, for user communities discovery and multiple queries expansion in Sections 3.1 and 3.2, respectively. A minimum length based pattern mining technique is presented in Section 3.3. We experimentally validate the performance of our approach in Section 4, and conclude this paper in Section 5.

2. RELATED WORK

In this section, we mainly review the related work on landmark retrieval and related query argumentation technique.

2.1 Feature Learning for Landmark Retrieval

The increasing amount of landmark photos has resulted in numerous methods for landmark retrieval [14, 36, 8, 10]. Hays *et al.* [14] presented a feature matching approach to return the K nearest neighbors with respect to the query landmark photo where a query photo and photos in database are represented by aggregating a set of low-level features to perform landmark retrieval. Zhu *et al.* [36] proposed to learn the landmark feature by combining low-level features while assisted with Support Vector Machine (SVM). In [8], a region based recognition method is proposed to detect discriminative landmark regions at patch level, which is seen as the feature for landmark retrieval. To augment semantic interpretation on landmark representation, Fang *et al.* [10] presented an effective approach, namely GIANT, to discover both discriminative and representative mid-level attributes for landmark retrieval. However, these approaches are still using a single query photo for landmark retrieval whilst our approach is focusing on mining robust patterns of landmark photos from an expanded multi-query set.

2.2 Query Argumentation Technique

In [7], a query expansion technique is brought into the visual domain in which a strong spatial constraint between a query image and each result allows for an accurate verification of each return, improving image retrieval performance. This simple method of well-performing query expansion is referred to as Average Query Expansion [7] (AQE) where given a query image, images are ranked using tf-idf scores and corresponding visual words in these images are averaged together with the query visual words, and this resulting query expanded visual word vector is recast to be a new query to re-query the database. Observing that AQE is lack of discrimination, Arandjelovic *et al.* [2] enhanced it using a linear SVM to discriminatively learn a weight vector for re-querying yields a significant improvement over the standard average query expansion method, called DQE. The most related work to ours is [11] (PQE), where a query expansion approach for a particular object retrieval is presented. Our method is different from them in terms of multi-query construction. Zhu *et al.* [37] propose to perform landmark classification with a hierarchical multi-modal exemplar features. There are also research [17] aiming at developing the feature representations for diverse landmark search.

These methods are commonly based on the idea that those particular multiple queries are manually selected or simply retrieved from top-k similar items whilst we automatically determine helpful queries by exploring the latent topics of query landmark as well as the informative user communities. Besides, previous query expansion pipelines are not applicable in the context of social media networks. In other words, this problem cannot be addressed by simple variations of methods in literature.

2.3 Geo-tagging by Exploring User Community

Geo-tagging refers to adding geographical identification metadata into various multimedia data such as images [13] and videos [6] in websites, blogs, and photo-sharing webservices [1, 21, 35]. Associating time and location information (latitude/longitude) with pictures can facilitate geotagging-enabled information services in terms of finding location-based news, photos, or other resources [3, 32, 33, 16]. There are also a number of approaches [15] trying to learn the location based visual codebook for landmark search. Similar to our method, such research area also explore the user communities. However, this kind of research is apparently different from landmark retrieval studied in this paper.

3. PROPOSED TECHNIQUE

In this section, we formally present our techniques with three steps as depicted in Fig. 1.

3.1 Query Topics and User Group Discovery

We aim to discover the latent topics of the query landmark by modeling the relationship between the query landmark and latent topics inherent over entire photo set uploaded by users. To this end, we use Latent Dirichlet Allocation (LDA) [4] to discover a set of latent topics where each topic contains a set of photos describing the similar landmarks. As a byproduct, we construct the user group that consists of the users, who upload the landmark photos sharing the similar topics as the query landmark, which is further utilized to perform multi-query expansions on Section 3.2.

Before shedding lights on how to transfer LDA to detect the latent topics of query landmark photo over the entire photo database, we firstly introduce some preliminaries on some notations and latent topics.

3.1.1 Preliminaries

Formally, given a set of users U, each of which is associated with their landmark photos. We assume that these observed photos can be grouped into clusters as per their latent topics, where each latent topic consists of the set of landmarks; for example, *e.g.,* the photo containing the landmark "Eiffel tower" may belong to the latent topic "architecture"; the landmark "Himalayas" may belong to the topic "mountain". Suppose a set of topics for query landmark have been detected, each photo can be then modeled as a probabilistic distribution over these topics. Based on that, it can also be seen as a generative model where each user album, denoted as d, is composed of a number of topics, and the generation for each photo is probabilistically determined by the topics of the album.

Now, we are ready to reformulate the problem of query photo topic discovery as a LDA model. Denote α, β as the parameter of the Dirichlet prior on the per-album topic distributions (per-topic photo distribution), θ_d as the topic distribution on album d, and ϕ_z as the photo distribution on topic z. LDA assumes the following generative process for a corpus D consisting of $|D|$ user albums:

1. Choosing the number of topics: $|Z|$
2. Choosing $\theta_d \sim Dir(\alpha)$, $d \in \{1, \ldots, |D|\}$
3. Choosing $\phi_z \sim Dir(\beta)$, $z \in \{1, \ldots, |Z|\}$
4. For each photo ω in album d
   - Choosing a topic $z_{\omega d} \sim Multinomial(\theta_d)$
   - Choosing a photo $\omega \sim Multinomial(\phi_{z_{\omega d}})$.

Fig.3 depicts a graphical model for this representation.

As such, the output of LDA over D, denoted as LDA(D), can be seen as a partition of D into multi-groups, each of which consists of a photo set characterizing the same latent topic.

3.1.2 Detecting the candidate latent topics for query landmark

We propose to detect the possible latent topics for query landmark from the entire photo database, and group photos as per each topic z in album d, where each d is described as a mixture of a fixed number of topics Z with topic z having the probability of $P(z|d)$. Each topic z has the probability of generating various photos containing the query photo q in z, denoted as $P(q|z)$. Commonly, we set a probability threshold λ to decide the candidate latent topics for query landmark q. That is, $P(q|z) \geq \lambda$ indicates that z could be one candidate topic for q. Therefore, we further have the following cases regarding probability threshold λ:

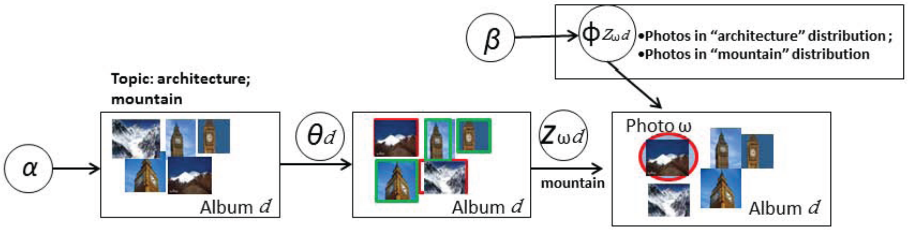

β
$\phi_{z_{\omega d}}$
•Photos in "architecture" distribution;
•Photos in "mountain" distribution
Topic: architecture; mountain
α
Album d
θ_d
Album d
$z_{\omega d}$
mountain
Photo ω
Album d

Figure 3: The LDA model of graphical representation on landmark photo distributions. A topic distribution θ_d (*e.g.*, "architecture" and "mountain") on album d is chosen by Dirichlet prior $Dir(\alpha)$, and photo distribution ϕ_z on topic z is chosen by $Dir(\beta)$. Then each image ω from d can be modeled as a distribution on its topic $z_{\omega d}$ ($\omega \sim Multinomial(\phi_{z_{\omega d}})$) with topic $z_{\omega d}$ characterized by $z_{\omega d} \sim Multinomial(\theta_d)$.

1. if λ is small, we may generate a lot of candidate latent topics, resulting into non-relevant latent topics regarding the query landmark q.

2. if λ is large, the size of candidate topic set may be quite small, which may be not inclusive to involve the true topic for q.

We learn the trade-off value for λ from our empirical studies, which is critical to decide the topic set regarding the query landmark q.

Once the query topic set is determined, it is plausible to recommend a few photos sharing the similar topics to the query photo. To this end, we propose to utilize user communities to assist multi-query set construction. Specifically, a user is seen to have a similar landmark of interest as q-user, if she upload the photos sharing at least one of the similar topics with that of the q-user. Based on that, a user-photo matrix is naturally constructed. Let $M \in \mathbb{R}^{|U|\times|J|}$ be a user-photo matrix that contains $|U|$ users and $|J|$ photos, in which U is the user set, J is the photo set, and $u_q \in U$ where u_q represent the q-user. In M, we have $M(u, j) = 1$ if a user $u \in U$ uploaded a photo j into J, and $M(u, j) = 0$, otherwise.

In next section, we will elaborate the details on selecting the Top-K photos to complement the query landmark to be multi-query set.

3.2 Multi-Query Expansions

To find the appropriate photo set for query expansions, we propose to conduct a non-negative matrix factorization [18] on the user-photo matrix M, where the users (include q-user) and relevant photos are detected by LDA mentioned above. By factorizing the M into a lower-dimensional latent space, each photo's possibility of being recommended to the query user u_q is computed based on the factorized result. Then, photos with top-K highest relevant scores (possibility to share the similar topic with query landmark) are calculated and selected, together with the original single query, to form a robust multi-query set.

Specifically, we perform the non-negative matrix factorization on M by minimizing the objective function below.

$$\arg\min_{P,V} ||M - PV||_F, P \geq 0, V \geq 0, \tag{1}$$

where $||\cdot||_F$ denotes the Frobenius norm. Two low-rank factor matrices $P \in R^{|U|\times L}$ and $V \in R^{L\times|J|}$ are obtained by solving Eq.(1), where P models the mapping of users in the low-dimensional latent space with L dimensions, and V defines the mappings of photos to the same latent space. That is, each user u is represented as the u-th row vector of P denoted by p_u, while each photo j is encoded by the j-th column vector of V denoted by v_j. For a q-user $u_q \in U$, we compute a score as the possibility that photo j will be recommended to u_q from M. The possibility is calculated by the inner product of $P(u_q, .)$ and $V(., j)$ as follows

$$S(u_q, j) =< P(u_q, \cdot), V(\cdot, j) >, \tag{2}$$

where $P(u_q, \cdot)$ and $V(\cdot, j)$ denote the row vector of P corresponding to u_q and j-th column vector of V corresponding to photo j, respectively.

We determine a number of photos to form a multi-query set by ranking all $S(u_q, j)$ scores in descending order and select up to K photos with highest scores, together with q to expand to a new multi-query set. The advantage of using non-negative matrix factorization over a user-photo matrix is: it avoids a brute-force evaluation on each photo's similarity to the query photo; instead, this factorization strategy is more efficient in determining multiple queries.

The landmark retrieval is then performed against a landmark database with respect to this multi-query set. Given multiple queries, a naive way of generating a retrieval list is to issue each query in the set individually, and then combine their retrieved results. However, such a fusion fashion is highly computational demanding when the size of ranking list is large. By contrast, we present a pattern mining method that mines a landmark-specific pattern set as the mid-level representation to represent multi-query set. After that, we aim at calculating the similarity score between such pattern set and each photo in the database. The final ranking retrieval result is formed by ranking all these scores of corresponding photos.

3.3 Multi-Query based Pattern Mining

3.3.1 Why Pattern Mining on Multi-Query?

Although an expanded multi-query set can be obtained by the aforementioned steps, it may fail to form the robust feature representation for multi-query landmark photo set due to the following facts:

1. The multi-query set may cause rather redundancy due to the fact that multiple photos might describe the same landmark from the same perspective but different angles and lighting conditions. Thus, they are not helpful to better represent the landmark.

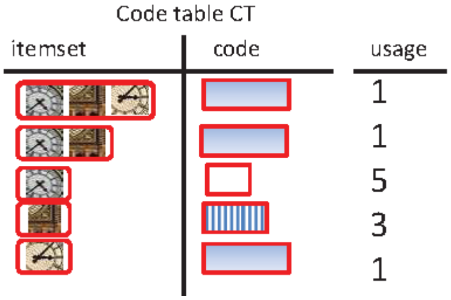


Figure 4: Example on a code table. The lengths of codes represent their lengths in encoding. Note that the usage column is not part of code table, but shown here as illustration: in optimal compression, more shorter the codes should be the more often they are used.

2. The multi-query set may contain the photos characterizing the inconsistent latent topics with the query landmark.

To address the above challenge, we propose to employ a pattern mining strategy to discover the complementary but compact feature representation from a multi-query set, which can serve to be landmark specific mid-level representation to conduct landmark retrieval effectively.

3.3.2 Learning robust pattern set for multi-query set

To faithfully represent a landmark of interest, we mine the frequent patterns among multi-query. A frequent pattern only describes a part of the query set Q, thus, we aim to find a set of patterns that can best describe Q. To this end, we employ an effective pattern mining technique of minimal description length (MDL) [12], which is able to find set of frequent patterns that yield the best compression of Q. Nonetheless, MDL still suffers from the high computational cost in all possible combinations of permutations on patterns. To address this issue, we further deploy the existing KRIMP algorithm [23] which introduces a few heuristic strategies to reduce the complexity of MDL algorithm.

In the following, we will present the technical details based on the principle of minimum description length (MDL), followed by KRIMP algorithm to yield the final robust pattern set to be mid-level representation for multi-query set.

3.3.3 The MDL algorithm

A pattern, denoted as h, consists of a set of items, and each item denotes a patch from a photo. In the rest of paper, we would interchangeably use patch and item. In our case, the multi-query set Q can be regarded as a transaction collection where each photo is treated as a transaction t containing a set of items (patches). The size of a patch is set to be 16×16 and we use root-SIFT descriptors extracted from Hessian-affine regions [22]. A single pattern h can describe only part of the query landmark and a pattern h is considered to be matched/mapped into transaction t if $h \subseteq t$. A set of patterns that together describe a query landmark is referred to as a model H. That is, we want to find the best set of (frequent) itemsets, H^*, to describe Q the best. Essentially, the best model is featured to be able to compress Q the best. To this end, we use the Minimal Description Length Principle (MDL) to seek a set of patterns which are the characteristic of Q.

The key idea of compression is the code table CT, which is a simple two-column translation table that has itemsets on the left-hand side and a code for each itemset on its right-hand side. With such as a code table we find through MDL, the set of itemsets that together optimally describe the multi-query set Q.

Definition 1. *Let I be a set of items and C a set of code words. A code table CT over I and C is a two-column table such that: The first column contains itemsets,* i.e., *subsets over I. The second column contains code words from C such that each code word occurs at most once.*

An itemset X drawn from the powerset of I, *i.e.,* $X \in P(I)$ occurs in CT, denoted by $X \in CT$ if X appears in the first column of CT. For $X \in CT$, $code_{CT}(X)$ denotes its code, *i.e.,* the corresponding element in the second column.

Example 1. *We show an example on code table in Fig4. The first and second columns in the code table present the itemsets and corresponding codes, respectively. Each bar represents a code and its width shows the code length. The usage column is actually not part of code table, however, it is shown for illustrative purposes and will be used later to calculate optimal code length.*

To encode a transaction t from Q over I with code table CT, we require a cover function $cover(CT, t)$ that identifies which elements of CT are used to encode t. The result of a cover function is a disjoint set of elements of CT that cover t. We omit the details on how to compute a cover function, which can be found in [23] [2]. To encode the query set Q using CT, we simply replace each transaction $t \in Q$ by the codes of the itemsets in the cover of t, that is,

$$t \rightarrow \{code_{CT}(X) | X \in cover(CT, t)\}. \quad (3)$$

Since MDL is concerned with the best compression, the codes in CT should be chosen such that the most often used code has the shortest length. Due to the correspondence between code lengths and probability distributions [19], we can calculate the code length $L(d)$ of code d through the following equation:

$$L(d) = -\log(P(d)), \quad (4)$$

where the probability distribution P on Q induced by a cover function is given by the relative usage frequency of each of the itemsets in the code table. To achieve this, we define the usage count of an itemset $X \in CT$ as the number of transactions t from Q where X is used to cover. This normalized frequency represents the probability that the code d is used in the encoding of a transaction $t \in Q$. Then, the optimal code length is $-\log(P(d))$, denoted by $L(code_{CT}(X))$ and a code table is optimal if all its codes have their optimal length.

Example 2. *As illustrated in Fig.4, for the third itemset X, its probability distribution over Q is $P(X|Q) = \frac{5}{11}$, and $L(code_{CT}(X)) = -\log(\frac{5}{11}) = 0.788$, thus X is assigned a code of length 0.788 bits.*

[2] Let h, m be patterns ($h, m \in H$) and t a transaction ($t \in B$). A cover function has three important properties: 1) $h \in cover(H, t) \implies h \in H$; 2) $h, m \in cover(H, t) \implies (h = m)$ or $(h \cap m = \oslash)$; 3) $t = \cup_{h \in cover(H,t)} h$.

For any $t \in Q$, its encoded length, in bits, denoted by $L(t|CT)$ is $L(t|CT) = \sum_{X \in cover(CT,t)} L(code_{CT}(X))$, and the encoded size of Q in bits, is

$$L(Q|CT) = \sum_{t \in Q} L(t|CT). \qquad (5)$$

To use MDL, we need to know $L(CT)$, *i.e.,* the encoded size of a code table. The size of CT in bits, denoted by $L(CT|Q)$ is given by

$$L(CT|Q) = \sum_{X \in CT} L(code_{ST}(X)) + L(code_{CT}(X)), \qquad (6)$$

where ST denotes the standard code table for Q which contains only the singleton itemsets, and we do not take itemsets with zero usage into account, that is, $usage_Q(X) \neq 0$.

Let multi-query set Q be a transaction collection over I, let CT be a code table that is code-optimal for Q and *cover* a cover function. The total compressed size of the encoded Q and CT, in bits, denoted by $L(Q, CT)$ is computed as $L(Q, CT) = L(Q|CT) + L(CT|Q)$. Now we can state our problem using MDL.

Problem statement from MDL principle. The problem of minimal coding set on landmark queries: Let I be a set of items (patches) and Q be transaction set over I, $cover(\cdot)$ a cover function, and F be a candidate set which contains at least all singleton itemsets of I, our goal is to find the a subset of F, denoted by H^* such that for the corresponding code table CT, the total compressed size $L(Q, CT)$ is minimal.

The solution to the above problem allows us to find the best pattern set H^* to describe Q.

3.3.4 The KRIMP Algorithm

To find a set of patterns H^* that best describe the multi-query set Q, we need to examine all $|CT|!$ possible permutations for one transaction (one photo in Q), which is practically impossible. Thus, we employ the KRIMP algorithm that can reduce the computational burden by

- using a heuristic to consider the code table in a fixed order, referred as Standard Cover Order;
- ordering the candidate itemsets such that long, frequently occurring itemsets are given priority, referred as Standard Candidate Order.

We summarize applying KRIMP algorithm to seek the compact patterns for a multi-landmark query in Algorithm 1.

Algorithm 1: The KRIMP algorithm to discover compact set of patterns from a multi-query set Q.

```
Input: A multi-query set $Q$ (transaction collection), a
       candidate set $F$, both cover a set of items $I$.
Output: The best set of patterns $H^*$
Initialize $CT \leftarrow StandardCodeTable(Q)$;
$F_0 \leftarrow F$ in Standard Candidate Order;
for $F \in F_0 \backslash I$ do
    $CT_c \leftarrow (CT \cup F)$ in Standard Cover Order;
    if $L(Q, CT_c) < L(Q, CT)$ then
        $CT \leftarrow CT_C$;
$H^* \leftarrow CT$;
Return $H^*$;
```

3.3.5 Landmark Retrieval Rule

Once the optimal pattern $H^* = \{I_1, \ldots, I_N\}$ regarding multi-query set is mined where I_i denotes the i-th patch(item) set, we aim at calculating the similarity score for each photo based on H^*. To resolve this, we use term frequency and inverse document frequency to (tf-idf) to generate a weight for each patch set I_i that appears in each photo y from a database. We use the term-frequency $tf_{I_i,y}$ to count the number of I_i occurs in photo y (**As aforementioned in section 3.3.3, each patch is generated with** 16×16**, we therefore pre-segment each photo** y **to be a patch set via such size.**). However, because some patch sets can appear in many photos frequently, without giving enough weight to the more meaningful patch sets in the query photo (*e.g.,* some visual patches of London bridge are common architecture features that can be commonly seen in other bridges). Hence, an inverse document frequency factor, idf_{I_i}, is incorporated which diminishes the weight of patch sets that occur very frequently in the database and increases the weight of patch set that occur rarely. Then its tf-idf weighting scheme that assigns the itemset I_i a weight in photo y is given by

$$w^i = tf \cdot idf_{I_i,y} = tf_{I_i,y} \times idf_{I_i}, \qquad (7)$$

where we define $tf_{I_i,y} = 1$ if I_i occurs in y and 0 otherwise, and the idf_{I_i} is a measure of how much information I_i provides, that is, whether I_i is common or rare across all photos in a database. Specifically, we define $idf_{I_i} = \log \frac{|\Omega|}{|\{y : I_i \in y\}|}$, where, as aforementioned, Ω denotes the size of the whole photo database and $|\{p : I_i \in p\}|$ is the number of photos where I_i appears (*i.e.,* $tf_{I_i,y} \neq 0$), from which we can see the larger $|\{y : I_i \in y\}|$ implies more photos where the patch set I_i occurs, the idf_{I_i} will be smaller.

Thus, we may view each photo y as a vector with i-th entry corresponding I_i in H^*, together with a weight given by Eq.(7), which can be seen as a similarity score between between I_i and y. Based on that, we can calculate the total similarity between y and all patch set $I_i (i = 1, 2, \cdot \cdot \cdot, N)$ below:

$$\mathbf{w} = \{w^i\}_{i=1}^{N}. \qquad (8)$$

It can be seen to be similarity between y and H^*(multi-query set). Based on Eq.(8), we define the L_2 normalized overlap score measure $score(y, Q)$ below.

$$score(y, Q) = \sqrt{\sum_{i=1}^{N} |w^i|^2}. \qquad (9)$$

The final output of landmark retrieval is to return the ranking list according to Eq.(9) based on expanded multi-query set Q.

To sum up, we give the entire framework in Algorithm 2.

3.4 Complexity Analysis

Besides applying the efficient Minimal Description Length Principle Pattern Mining techniques to do feature coding, we will analyze the major cost of Algorithm 2 coming from detecting latent topics of query landmark by LDA and non-negative matrix factorization operation. Firstly, for LDA, assume we have $|D|$ user albums, Ω photos, and Z topics.

Algorithm 2: Multi-query expansion and pattern mining based landmark retrieval.

```
Input: User set $U$.
    Photo set $\Omega$.
    Query photo $q$.
    Query user $u_q$.
    Topic detection threshold $\lambda$.
    Number of photos to construct multi-query set $K$.
    Number of reduced lower dimensions or latent factors
    by matrix factorization $L$.
    Final ranking retrieval list $Re$.
Output: Ranked list $R$ from $\Omega$.
$D = \emptyset$;
for $u \in U$ do
    $d_u = \{\omega | u \in U \wedge u \;\; uploaded \;\; \omega \in \Omega\}$;
    $D = D \cup d_u$;
/*Query landmark latent topic set detected by LDA*/
$Z \longleftarrow LDA(D)$;
$C = \emptyset$;
while $z \in Z$ do
    $T \longleftarrow \emptyset$;
    if $P(q|z) > \lambda$ then
        $T = T \cup z$;
Construct the user-photo matrix $M$;
/*Form robust multi-query set*/
Perform matrix factorization on $M$ using Eq.(1);
for photo $j$ sharing at least one of topic set for $q$ do
    Compute $S(u_q, j)$ using Eq.(2);
Select top-K highest $S(u_q, j)$ to form a multi-query set
    $Q$;
/* Pattern mining and landmark retrieval */
Call Algorithm 1 to obtain the robust compact patterns
    $H^*$ that can best describe $Q$;
for photo $y \in \Omega$ do
    Build patch set on $y$ with size of each one to be
    $16 \times 16$ ;
    Compute $score(y, Q)$ according to Eq.(9);
Ranking $score(y, Q)$ to select top-$Re$ to be $R$;
Return $R$;
```

The computational complexity of LDA is $\mathcal{O}(|D|Z\Omega)$. In non-negative matrix factorization, the overall cost for optimizing P in Eq.(1) is $\mathcal{O}(|U||J|L)$ where $|U|$ and $|J|$ represent the number of users and photos in user-photo matrix M, and L is the dimensions of factorized low-dimensional latent space. Optimizing Q has the same time cost. Thus, matrix factorization over M costs $\mathcal{O}(|U||J|L)$. Selecting top-K items by using Eq.(2) costs $|J|$, where $|J|$ is the number of photos encoded in M. Hence, the complexity of the first step is $\mathcal{O}(|U||J|L + |J|)$.

4. EXPERIMENTS

In this section, we first present our experimental settings, and then report the experimental results to verify the effectiveness and efficiency of our approach.

4.1 Datasets and Features

Two datasets are constructed by collecting photos from websites of **Flickr** and **Picasa Web Album**. They are suitable for landmarks retrieval because they contain both user information and corresponding landmark photos.

Table 1: Statistics of datasets.

Dataset	**Flickr**	**Picasa Web Album**
# Landmark	55	16
# Photo per landmark	nearly 1,000	100 ∼ 300
# Total photo	49,840	4,100
# User	7,332	577

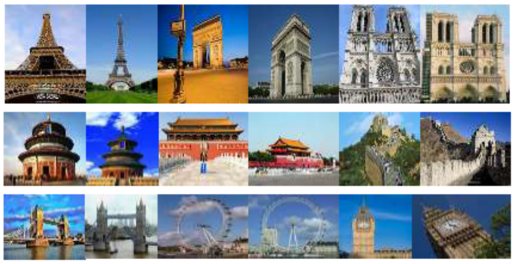


Figure 5: Example landmarks photos from our manually collected Flickr dataset.

Flickr. We use the **Flickr** API to retrieve landmark photos taken at a city posted by a large number of users. We sort out 11 cities: London, Paris, Barcelona, Sydney, Singapore, Beijing, Tokyo, Taipei, Cairo, New York city, and Istanbul. In each city, such as Paris, we obtained photos by querying the associated text tags for famous landmarks such as “Paris Eiffel Tower” or “Paris Triomphe” [3]. Example photos containing a variety of landmarks from this dataset are shown in Fig. 5.

Picasa Web Album. Providing rich information about interesting tourist attractions, this source contains a vast amount of GPS-tagged photos uploaded by users who have visited the landmarks, along with their text tags. We manually download a fraction of photos and their user information on 6 cities: London, Paris, Beijing, Sydney, Chicago, and Barcelona.

The statistics for user information and their uploaded photos over the two datasets are summarized in Table 1. In the training stage, the SIFT descriptor [20] is used as a local descriptor due to its excellent performance in object recognition [5]. Specifically, we adopt a dense sampling strategy to select the interest regions from which SIFT descriptors are extracted. To test the generalization of our framework in which test data differs from the dataset used to generate the quantization, we adopt another dataset. The **Oxford** dataset is used for constructing the vocabulary book [4], upon which local bag-of-words are encoded as representations. The **Oxford Building** dataset contains 5,062 photos, which is a standard set of particular objects for retrieval. The reason for choosing Oxford landmarks is that the photos have the scenes similar to, rather than identical landmarks, those in the two test databases (e.g., buildings, often with some similarities in the architectural style).

[3] In Paris, 12 queries were used to collect the photos from Flickr: La Defense Paris, Eiffel Tower Paris, Hotel des Invalides Paris, Louvre Paris, Moulin Rouge Paris, Musee d’Orsay Paris, Notre Dame Paris, Pantheon Paris, Pompidou Paris, Sacre Coeur Paris, Arc de Triomphe Paris, Paris

[4] http://www.robots.ox.ac.uk/ vgg/data/oxbuildings/

4.2 Settings

Evaluation Metric. We choose precision-recall as the metrics to evaluate both our model and state-of-the-art methods over top-100 retrieval ranked list based on each query landmark. In each landmark, we issue 20 queries and an average precision score is computed for all queries, and these are averaged to obtain a mean Average Precision (mAP) for the each landmark category.

Competitors. In our experiment, we consider the following competitors:

- K-NN [14]: A feature matching approach to return the K nearest neighbors with respect to the query landmark photo where a query photo and photos in database are represented by aggregating a set of low-level features to perform landmark retrieval.
- LF+SVM: Low-level features [36] combined with SVM;
- DRLR [8]: A region based location recognition method that detects discriminative regions at the patch-level.
- GIANT [10]: A method to discover geo-informative attributes that are discriminative and useful for location recognition.
- AQE [7]: Average Query Expansion method that proceeds as follows: given a query region, it ranks a list of photos using tf-idf scores. Bag-of-Word vectors corresponding to these regions are averaged with BoW vectors of the query, resulting in an expanded vector used to re-query the database.
- DQE [2]: Discriminative Query Expansion that enriches a query in the exactly same way as AQE. It considers photos with lower tf-idf scores as negative data to train a linear SVM for further rankings and retrievals.
- PQE [11]: A Pattern based Query Expansion algorithm that combines top-K retrieved photos with a query to find a set of patterns.

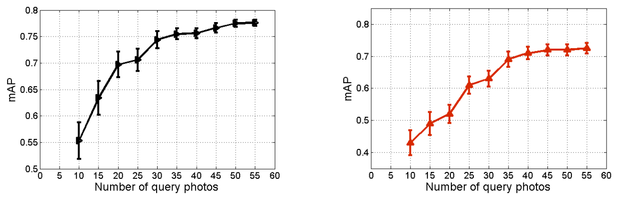
4.3 Learning Parameters

In this section, we learn three important parameters for our technique,

- K: the number for queries after matrix factorization in multi-query set.
- λ: threshold to determine the latent topic set for query landmark.
- L: reduced dimension or latent factor number for matrix factorization.

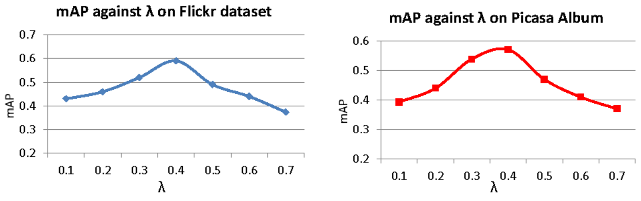
We tuning the parameters over two dataset via the following fashion: that is, we vary one parameter at a time while fixing the others. For instance, when we test the sensitivity of K, we try different values of K within a range of [10,55] while the other two parameters λ and L are fixed to be 0.4 and 64. We adopt mAP as the evaluation metric. For each landmark category, we randomly select 20 distinct queries to yield the mAP value.

The following observations are made:



(a) (b)

Figure 6: Results on mAP scores by varying the number of photos recommended from each user community. (a) mAP scores vs. varied K values over Flickr dataset. (b) mAP scores vs. varied K values over Picasa dataset.



(a) (b)

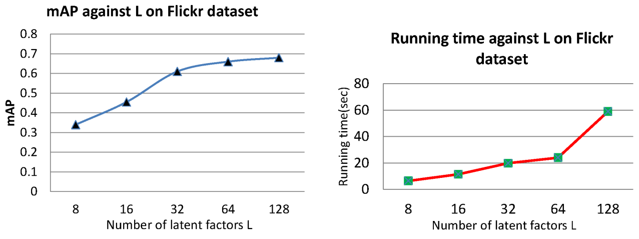
Figure 7: Results on mAP scores by varying the λ. (a) mAP scores vs. varied λ values over Flickr dataset. (b) mAP scores vs. varied λ values over Picasa dataset.

- Fig. 6 depicts the mAP scores against a variety of K values, which indicates that as the number of query photos increases, the values of mAP by our method become higher. This is because a larger query set may be more representative for the queried landmark. As such, our approach is more robust against occlusion or bias-viewpoint captures. However, it may also result in a higher running time. Therefore, we set the value of K to 40 as a trade-off option.
- Fig. 7 reports the mAP scores against varied λ value. Interestingly, as the value of λ increases within a relative small beginning point, the mAP value will be increased as well, since it will prune some topics with a very small probability yielded by LDA. However, when keeping enhanced, the performance will be downgrade, since the larger λ is more likely to prune more topics, which may result into the missing of true topics with a large probability. Within our expectation, the above results over two dataset are consistent with our discussions in section 3.1.2. From our results, we finalize λ to be 0.4.
- Figs. 8 and 9 show the mAP score and running time against varied L value over two datasets, from which we can see the larger L is, the **pros** is that the larger mAP score can be achieved as more original information can be preserved; the **cons** comes from the larger running time for matrix factorization. To achieve a trade-off in terms of effectiveness and efficiency, we set L to be 64 in our experiments.

Based on the above, we just select K to be 40, λ to be 0.4 and L to be 64 for our method, which leads to the corresponding results in the remaining of experimental results for our multi-query expansions.

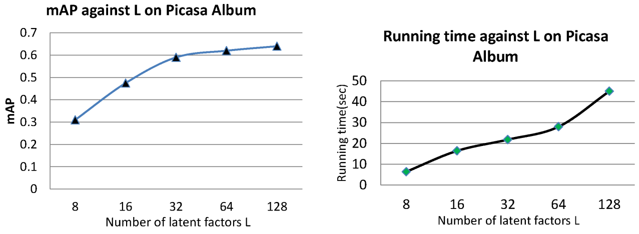
Table 2: The mAP values of selected landmarks by different approaches against Flickr dataset.

Method	London bridge	Triumphal arch	Sagrada familia	Sydney bridge	Singapore city	Temple heaven	Tokyo skytree	Taipei 101	Cairo tower	Brooklyn bridge	Maiden's tower	mAP
K-NN	46.75	32.03	33.55	38.93	29.83	54.33	17.21	32.39	44.37	22.56	64.58	38.78
LF+SVM	40.52	52.35	45.78	36.09	29.17	61.23	38.96	15.99	34.55	33.58	62.33	40.96
DRLR	52.45	35.65	51.48	34.58	40.33	62.45	42.33	40.24	44.58	38.34	57.35	45.52
GIANT	53.15	42.44	50.89	39.15	42.36	64.37	46.73	43.59	47.21	37.82	64.58	49.29
Ours	**57.83**	**64.76**	**57.33**	**45.59**	**63.48**	**68.38**	**60.48**	**59.32**	**70.04**	**54.58**	**72.38**	**61.29**



(a) (b)

Figure 8: (a) The mAP value for landmark retrieval over Flickr dataset against latent factor number L (b) The running time for landmark retrieval over Flicker dataset against latent factor number L.



(a) (b)

Figure 9: (a) The mAP value for landmark retrieval over Picasa Web Album dataset against latent factor number L (b) The running time for landmark retrieval over Flicker Picasa Web Album against latent factor number L.

4.4 Landmark Retrieval and Ranking

We evaluate precision-recall and mAP for our approach and competitors. For each landmark in **Flickr** and **Picasa Web Album**, we randomly sample 20 query photos for each landmark and perform landmark retrieval to return the ranking list, where we obtain the precision-recall value and plot a precision-recall curve, plotting precision $p(r)$ versus the recall r. In our experiment, the average precision-recall is calculated by averaging their values on 20 query photos of each landmarks, and the average values of $p(r)$ over the internal from $r = 0$ to $r = 1$ are shown in Fig.10.

For each dataset, we also use mAP as the evaluation metric to examine the performance of each approach. For a query q, the average precision (AP) is defined as $AP(q) = \frac{1}{L_q}\sum_{r=1}^{n} P_q(r)\theta_q(r)$, where L_q is the ground-truth neighbors of query q in database, n is the number of photos in ranked list, we set $n = 100$, $P_q(r)$ denotes the precision of the top r retrieved photos, and $\theta_q(r) = 1$ if the r-th retrieved photo is a ground-truth within this landmark category and $\theta_q(r) = 0$, otherwise. In our case, given 20 query photos for each landmark, the mAP is defined over all 20 queries for each landmark as: $mAP = \frac{1}{20}\sum_{i=1}^{20} AP(q_i)$. In the step of multi-query expansions, by default, the multi-query set Q are composed of 40 recommended photos.

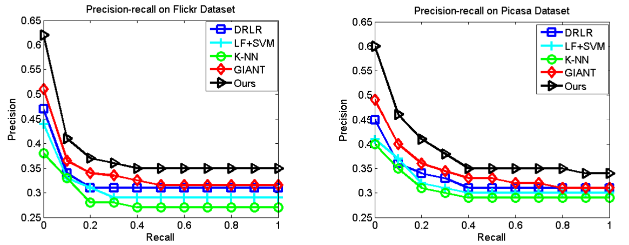


Figure 10: Comparison on average precision-recall curve on top-10 returned photo list over two databases.

Fig.10 shows the average precision-recall curve over two benchmarks. We can see that our method consistently performs better than all competitors. The compared landmark retrieval results with mAP values are shown in Table 2 and Table 3. We observe that: (1) Our method outperforms all competitors, due to the effectiveness of exploiting the complementation of multiple queries as well as the use of the robust yet compact pattern mining technique; (2) The large intra-class variance limits the performance of LF+SVM and K-NN, especially for the **Flickr** dataset; and (3) DRLR detects discriminative regions from a single query photo, which degrades its performance when a query photo was shot from a bad viewpoint. Although GIANT also performs better but second to us, it only utilizes user-generated content to obtain visual attributes. This also demonstrates the need of exploiting user information to assist query understanding.

4.5 Comparing Query Expansion Approaches

Unlike AQE and DQE where plausible queries are selected via a low-level feature matching procedure, our strategy of selecting a multi-query set exploits latent topics and assistance from users to complement a q-user. While PQE uses a similar approach to expanding a single query into a multiple query set with pattern mining being used subsequently, its multi-query set is determined by manual selection. To demonstrate the superiority of our method over existing multiple query methods that can be adopted for landmark retrieval, we compare our method against the above three state-of-the-art approaches with query expansion, which are AQE, DQE, and PQE. Results are shown in Fig.11. We conclude that our multi-query based pattern mining approach outperforms AQE, DQE, and PQE for landmark retrieval by a large margin in terms of mAP values in two databases.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we motivate the problem of landmark retrieval with possibly "ill-shaped" query landmark to be is-

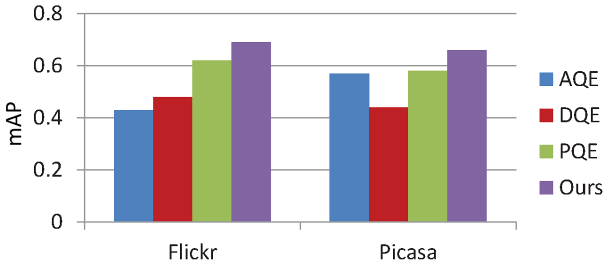


Figure 11: Comparison of query expansion methods.

Table 3: The mAP values of selected landmarks by different approaches against Picasa Web Album dataset.

Landmark	K-NN	LF+SVM	DRLR	GIANT	Ours
London clock	58.37	50.03	60.65	62.20	**71.89**
Eiffel tower	51.24	51.05	46.77	52.13	**61.71**
Forbidden city	44.04	55.28	58.57	58.94	**70.68**
Opera house	49.26	33.48	46.88	48.48	**61.74**
Catalunya plaza	24.87	12.37	21.67	23.23	**42.83**
Chicago plaza	35.48	38.38	33.76	38.74	**50.84**
mAP	43.87	40.09	44.73	47.28	**59.94**

sued. For this purpose, we propose an effective multi-query expansions paradigm to form a multi-photo query set, which sharing the highly similar latent topics with query landmark. Based on such multi-query set, we propose a novel technique to generate the robust pattern set (photo patch set regarding multi-query set). To ensure bias-free and enhance the efficiency, we adopt the existing minimum-description-length-principle based pattern mining techniques to remove similar query photos from the selected multi-query photos, leading to a compact set with smaller size. Then an effective retrieval rule is developed to calculate the similarity score between each photo in the database and mined pattern set (multi-query set). All these similarity scores are ranked to be the final ranking list of landmark retrieval based on query landmark. Experimental results on real-world datasets validate the significantly higher accuracy of our proposed method.

Our future work will investigate learning the deep features for landmark retrieval, such as DeCAF [9], which has been demonstrated to be powerful in many visual recognition applications. We would like to improve the current performance by addressing such gap. Another directions may go to leveraging the multi-view features [29, 30, 28, 27, 24, 34]. Zhu *et al.* [38] propose to learn the hypergraph structure of landmark photos for content based landmark retrieval, which inspires us to explore the (hyper) graph structure [31, 25, 26] of landmarks to improve the performance of landmark retrieval upon our proposed feature learning technique in this paper.

Acknowledgement. Xuemin Lin's research is supported by ARC DP120104168, ARC DP140103578, ARC DP150102728 and NSFC 61232006. Wenjie Zhang's research is supported by ARC DP150103071 and ARC DP150102728.

6. REFERENCES

[1] E. Amitay, N. Harel, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *ACM SIGIR*, 2004.

[2] R. Arandjelovic and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, 2012.

[3] S. Arslan, S. Kim, M. He, and R. Zimmermann. Relevance ranking in georeferenced video search. *Multimedia Systems*, 2010.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet distribution. *J. Mach. Learn. Res.*, 2003.

[5] Y. Boureau, F. Bach, Y. Cun, and J. Ponce. Learning mid-level features for recognition. In *CVPR*, 2010.

[6] X. Cao, L. Wu, J. Rasheed, H. Liu, T. Choe, F. Guo, and N. Haering. Automatic geo-registration for port surveillance. *International Journal of Pattern Recognition and Artificial Intelligence*, 2010.

[7] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: automatic query expansion with a generative feature model for object retrieval. In *ICCV*, 2007.

[8] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros. What makes paris look like paris? *ACM Trans. Graph.*, 2012.

[9] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *arXiv:1310.1531*, 2013.

[10] Q. Fang, J. Sang, and C. Xu. Giant: Geo-informative attributes for location recognition and exploration. In *ACM Multimedia*, 2013.

[11] B. Fernando and T. Tuytelaars. Mining multiple queries for image retrieval: on-the-fly learning of an object-specific mid-level representation. In *ICCV*, 2013.

[12] P. D. Grunwald. *The minimum description length principle*. The MIT press, Cambridge, Massachusetts, USA, 2007.

[13] C. Hauff and G.-J. Houben. Geo-location estimation of flickr images: Social web based enrichment. In *Proceedings of ECIR 12*, 2012.

[14] J. Hays and A. A. Efros. im2gps: estimating geographic information from a single image. In *CVPR*, 2008.

[15] R. Ji, L. Duan, J. Chen, H. Yao, J. Yuan, and W. Gao. Location discriminative vocabulary coding for mobile landmark search. *International Journal of Computer Vision*, 96:290–314, 2012.

[16] E. Kalogerakis, O. Vesselova, J. Hays, A. A. Efros, and A. Hertzmann. Image sequence geolocation with human travel priors. In *Proceedings of ICCV 09*, pages 253–260, 2009.

[17] L. Kennedy and M. Naaman. Generating diverse and representative image search results for landmarks. In *Proceedings of WWW 08*, 2008.

[18] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 2007.

[19] M. Li and P. Vitanyi. *An introduction to Kolmogorov complexity and its applications*. Springer, 1993.

[20] D. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[21] J. Luo, D. Joshi, J. Yu, and A. Gallagher. Geotagging in multimedia and computer vision-a survey. *Multimedia Tools and Applications*, 2011.

[22] O. C. M. Perdoch and J. Matas. Efficient representation of local geometry for large scale object retrieval. In *CVPR 09*, 2009.

[23] J. Vreeken, M. Leeuwen, and A. Siebes. Krimp: mining itemsets that compress. *Data Min. Knowl. Discov*, 2011.

[24] Y. Wang, M. A. Cheema, X. Lin, and Q. Zhang. Multi-manifold ranking: Using multiple features for better image retrieval. In *PAKDD*, 2013.

[25] Y. Wang, X. Lin, L. Wu, and Q. Zhang. Shifting hypergraphs by probabilistic voting. In *PAKDD*, 2014.

[26] Y. Wang, X. Lin, L. Wu, Q. Zhang, and W. Zhang. Shifting multi-hypergraphs via collaborative probabilistic voting. *Knowledge and Information Systems, DOI 10.1007/s10115-015-0833-8*, 2015.

[27] Y. Wang, X. Lin, L. Wu, W. Zhang, and Q. Zhang. Exploiting correlation consensus: Towards subspace clustering for multi-modal data. In *ACM Multimedia*, 2014.

[28] Y. Wang, X. Lin, L. Wu, W. Zhang, and Q. Zhang. Lbmch: Learning bridging mapping for cross-modal hashing. In *ACM SIGIR*, 2015.

[29] Y. Wang, X. Lin, L. Wu, W. Zhang, Q. Zhang, and X. Huang. Robust subspace clustering for multi-view data by exploiting correlation consensus. *IEEE Transactions on Image Processing, DOI 10.1109/TIP.2015.2457339*, 2015.

[30] Y. Wang, X. Lin, and Q. Zhang. Towards metric fusion on multi-view data: a cross-view based graph random walk approach. In *ACM CIKM*, 2013.

[31] Y. Wang, J. Pei, X. Lin, Q. Zhang, and W. Zhang. An iterative fusion approach to graph-based semi-supervised learning from multiple views. In *PAKDD*, 2014.

[32] L. Wu and X. Cao. Geo-location estimation from two shadow trajectories. In *CVPR*, 2010.

[33] L. Wu, X. Cao, and H. Foroosh. Camera calibration and geo-location estimation from two shadow trajectories. *Computer Vision and Image Understanding*, 2010.

[34] L. Wu, Y. Wang, and J. Shepherd. Efficient image and tag co-ranking: a bregman divergence optimization method. In *ACM Multimedia*, 2013.

[35] Y. Zheng, L. Zhang, X. Xie, and W. Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of World Wide Web*, 2009.

[36] J. Zhu, S. C. H. Hoi, M. R. Lyu, and S. Yan. Near-duplicate keyframe retrieval by nonrigid image matching. In *ACM Multimedia*, 2008.

[37] L. Zhu, J. Shen, H. Jin, L. Xie, and R. Zheng. Landmark classification with hierarchical multi-modal exemplar feature. *IEEE Transactions on Multimedia*, 2015.

[38] L. Zhu, J. Shen, H. Jin, R. Zheng, and L. Xie. Content-based visual landmark search via multimodal hypergraph learning. *IEEE Transactions on Cybernetics*, 2015.