Location-Based Parallel Tag Completion for Geo-Tagged Social Image Retrieval

JIAMING ZHANG and SHUHUI WANG, Key Lab of Intellectual Information Processing, Institute of Computing Technology, Chinese Academy of Sciences
QINGMING HUANG, University of Chinese Academy of Sciences

Having benefited from tremendous growth of user-generated content, social annotated tags get higher importance in the organization and retrieval of large-scale image databases on Online Sharing Websites (OSW). To obtain high-quality tags from existing community contributed tags with missing information and noise, tag-based annotation or recommendation methods have been proposed for performance promotion of tag prediction. While images from OSW contain rich social attributes, they have not taken full advantage of rich social attributes and auxiliary information associated with social images to construct global information completion models. In this article, beyond the image-tag relation, we take full advantage of the ubiquitous GPS locations and image-user relationship to enhance the accuracy of tag prediction and improve the computational efficiency. For GPS locations, we define the popular geo-locations where people tend to take more images as Points of Interests (POI), which are discovered by mean shift approach. For image-user relationship, we integrate a localized prior constraint, expecting the completed tag sub-matrix in each POI to maintain consistency with users' tagging behaviors. Based on these two key issues, we propose a unified tag matrix completion framework, which learns the image-tag relation within each POI. To solve the optimization problem, an efficient proximal sub-gradient descent algorithm is designed. The model optimization can be easily parallelized and distributed to learn the tag sub-matrix for each POI. Extensive experimental results reveal that the learned tag sub-matrix of each POI reflects the major trend of users' tagging results with respect to different POIs and users, and the parallel learning process provides strong support for processing large-scale online image databases. To fit the response time requirement and storage limitations of Tag-based Image Retrieval (TBIR) on mobile devices, we introduce Asymmetric Locality Sensitive Hashing (ALSH) to reduce the time cost and meanwhile improve the efficiency of retrieval.

CCS Concepts: • **Information systems** → **Social tagging systems**; *Collaborative and social computing systems and tools*; *Image search;*

Additional Key Words and Phrases: Tag matrix completion, geo-location information, social image retrieval, asymmetric locality sensitive hashing

ACM Reference Format:
Jiaming Zhang, Shuhui Wang, and Qingming Huang. 2017. Location-based parallel tag completion for geo-tagged social image retrieval. ACM Trans. Intell. Syst. Technol. 8, 3, Article 38 (April 2017), 21 pages.
DOI: http://dx.doi.org/10.1145/3001593

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400 and 2015CB351802, National Natural Science Foundation of China: 61332016, 61303160, 61572488, 61620106009 and 61672497, 863 program of China: 2014AA015202, Postdoctoral Science Foundation of China: 2014T70126, Basic Research Program of Shenzhen: JCYJ20140610152828686, and in part by Key Research Program of Frontier Sciences, CAS: QYZDJ-SSW-SYS013.
Authors' addresses: J. Zhang and S. Wang, Key Lab of Intellectual Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China; emails: jiaming.zhang@vipl.ict.ac.cn, wangshuhui@ict.ac.cn; Q. Huang, University of Chinese Academy of Sciences, Beijing, China; email: qmhuang@ucas.ac.cn.

DOI: http://dx.doi.org/10.1145/3001593

1. INTRODUCTION

The Online Sharing Websites (OSW), such as Flickr[1] and Panoramio,[2] have experienced vigorous evolution in the Web 2.0 era. Having benefited from tremendous growth of User-Generated Content (UGC) on OSW, the massive number of social tags provide rich information in understanding the content of online images. Therefore, it has become more and more important to discover the true semantic information from the social tags toward the need of efficient large-scale online image retrieval.

However, according to the principle of least effort [Halpin et al. 2007; Kipp and Campbell 2006], the majority of users usually prefer to choose abstract and fuzzy phrases as tags for the images uploaded by themselves in order to save time on tedious tagging jobs. This phenomenon leads to the certain level of incompleteness and noise existed in the manually annotated tags of the ever-growing images on OSW. Therefore, it gives rise to a challenging research problem, especially in the mobile application scenario [Ji et al. 2014, 2015], which demands more compact information representation, that how to achieve a sufficient number of high-quality tags for social images based on existing user-generated tags with massive absence and noise.

There are two possible paradigms to solve this problem. One feasible way is classifier-based models [Goh et al. 2005; Barnard et al. 2003; Jiang et al. 2007], which formulate the problem with a standard multi-class classification [Carneiro et al. 2007] or multi-label classification [Hariharan et al. 2010; Zha et al. 2009], and the missing tags are obtained via image annotation process [Zhou et al. 2011; Carneiro et al. 2007]. However, classifier-based methods are highly dependent on the quantity and quality of manual tags annotated by OSW users. Moreover, the rich information in the social attributes (e.g., location, time, user, and group) of images from OSW may not be easily incorporated by classifier-based models.

Another way to solve this problem is tag refinement and completion, which aims at alleviating the number of noisy tags [Xu et al. 2009; Zhu et al. 2010; Sang et al. 2012; Li et al. 2016] and enhancing the number of informative tags [Chen et al. 2010; Wu et al. 2013] by modeling the relation between visual content and tags. Generally, the tag refinement and completion can be achieved by information averaging [Li et al. 2009; Chen et al. 2010] and latent factor learning [Wu et al. 2013]. For example, neighborhoods [Li et al. 2009] and graphs are usually exploited, and the tagging information from visually similar social images is borrowed to construct probabilistic descriptions on the uncertainty of the social tags. Wu et al. [2013] propose an *image-tag matrix* completion framework based on the matrix factorization method, and automatically fill in the missing tags and correct noisy tags for given images. However, existing works have not taken full advantage of rich social attributes and auxiliary information associated with social images.

In general, we consider three key issues to address the tag completion problem. First, the similarities calculated independently on tag space and visual space are different. Such difference should be minimized in order to achieve more semantically consistent representation on both spaces. Second, the correlation among individual tags reflects the tag co-occurrence in the real world, and thus provides important hints on the true semantics of visual content. Last but not the least, the social attributes from OSW provide rich information in deriving the true semantics of the images. For example, the location of the image may be strongly correlated with the tags with geographical information. The images which are all taken around Big Ben in London have high probability to be attached to “Big Ben” or “London.” Moreover, it can be regarded as

[1]http://www.flickr.com.
[2]http://www.panoramio.com.

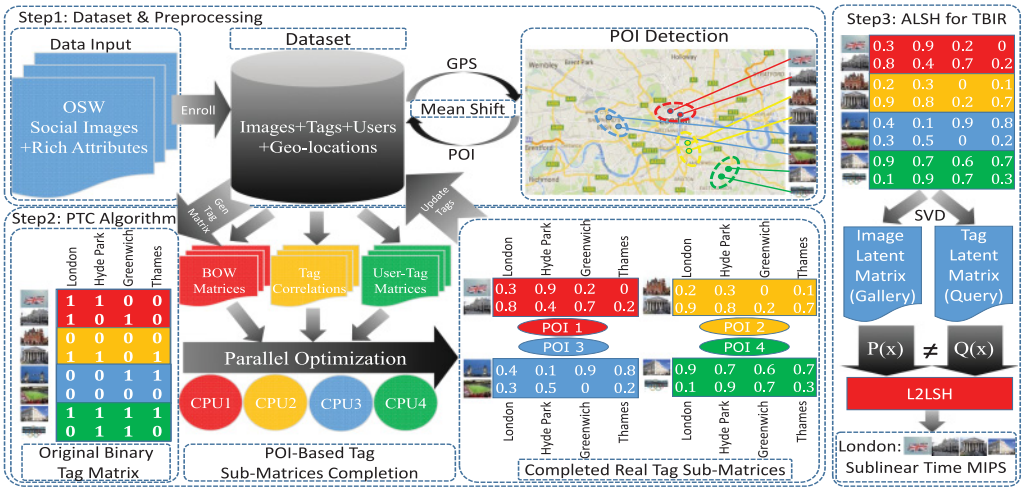


Fig. 1. Framework of Location-Based Parallel Tag Completion (PTC). It consists of two steps. The first step is the data partition preprocessing step dealing with GPS locations (by Mean Shift procedure). The second step is the proposed matrix completion model for image tagging. Besides the main process, we also introduce Asymmetric Locality Sensitive Hashing (ALSH) to conduct more efficient tag-based image retrieval.

noise annotation if the GPS location indicates "Big Ben" while the images are tagged with "Paris" or "Effiel Tower." The diversified backgrounds and preference styles of online users also result in complementary expressions in social tags to the visual content. For example, if the user has taken many images during his or her vacation on Fifth Avenue in New York, he or she might give "Fifth Avenue" and "New York" tag to all of these images. These social attributes have strong correlation with the users' tagging behavior, which may not explicitly be represented in the visual content. Therefore, by jointly considering the contents, tags, and social attributes (e.g., locations and users), a better social tag learning model can be achieved toward the real-world applications.

In this article, beyond the image-tag relation, we take advantage of rich social attributes of images available on OSW, especially the ubiquitous GPS locations and image-user relationship, to enhance the accuracy of tag completion and improve the computational efficiency. For geo-locations, we define the popular places where people tend to take images as Points of Interest (POI). We discover the POIs by mean shift approach to obtain the geographical clustering result on OSW image collections. We then propose a POI-based tag matrix completion framework which processes the images within each POI in parallel. For image-user relationship, we integrate a localized prior constraint into our proposed model, expecting the completed tag sub-matrices to maintain consistency with users' tagging behaviors in single POI. We formulate the tag matrix completion algorithm with a unified matrix factorization framework which combines both serial modeling and parallel learning steps. To solve the proposed model, an efficient proximal sub-gradient descent algorithm is designed. The model optimization can be easily parallelized and distributed to learn the tag sub-matrix for each POI. Experimental results demonstrate that our approach achieves higher performance in social tag completion on real-world social media images. The POI-based parallel tag matrix completion method is formulated into a unified model computation framework, as illustrated in Figure 1.

The contributions of this article are summarized below:

Overall: Social Attributes. We propose a unified framework which considers information in visual content, tags, and other ubiquitous social attributes such as the location information and the associated user behaviors for effectively learning social tags.
POI: Parallel Processing. We decompose the overall tag completion problem into a set of sub-problems with the help of location information. By localizing in geographical coordinate space and matrix partition with respect to the POI, the computational process of the tag matrix completion can be largely accelerated.
User: Preference Prior. We introduce a user-related prior constraint term into the formulation of our framework. It improves the quality of completed tag matrix, which is validated by promising performance in automatic image annotation.

The rest of this article is structured as follows. We overview the related work on several research aspects in Section 2. Section 3 gives notations and definitions of the tag completion problem, and provides a description in detail for our proposed framework and algorithm. We summarize the experiment results on automatic image annotation and tag-based image retrieval in Section 4. Section 5 concludes this study with some suggestions for future work.

2. RELATED WORK

In view of automatic image annotation, researchers have proposed numerous effective approaches, while many other literatures study approaches for tag-based image retrieval. As OSWs provide rich social attributes, some researchers have proposed a lot of methods that embed social attributes into social image tagging.

2.1. Automatic Image Tagging

Many algorithms for automatic image annotation have been proposed in the past decades. Both global [Goh et al. 2005] and local visual features [Barnard et al. 2003; Jiang et al. 2007] are taken into account for feasible solutions of image annotation. Moreover, several recent works [Russell et al. 2008; Wei et al. 2011] focus on spatial structure of visual content for performance promotion. Liu et al. [2009b] proposed a tag ranking method based on probability density estimation and tag similarity graph. Similar to Liu et al. [2009b], Li et al. [2009], Wei et al. [2010], Mishra et al. [2014], and Lee et al. [2014] proposed a visually similar neighbor-oriented tag relevance voting approach for tag ranking. Most content-based algorithms for automatic image annotation require fully annotated image samples for training confidential models. Despite the developments made by these algorithms, the room for performance improvement of existing automatic image annotation techniques is restricted by this limitation.

Meanwhile, several researchers choose other tag-oriented perspectives to solve image annotation problems, such as image retagging, tag refinement, tag propagation, and the like. Li et al. [2009] propose a neighbor voting method for social tagging. Guillaumin et al. [2009] propose a tag propagation (TagProp) method to transfer tags through a weighted nearest neighbor graph. Chen et al. [2010] propose an image retagging approach processing in batch-mode. Liu et al. [2011] propose a graph-based algorithm, including both tag-specific visual similarity graphs and tag semantic similarity graphs to handle the image retagging problem. However high-quality annotated tags are also essential for these approaches mentioned above. Therefore, they do not match the reality of manual tags on the OSW.

Besides general content-based image annotation techniques, many recent works exploit multi-label learning techniques to deal with image annotation as a multi-label classification problem. Desai et al. [2011] introduce a discriminative model in multi-label learning. Hariharan et al. [2010] combine a Support Vector Machine (SVM) with

multi-label learning to manage large-scale data collection. Zha et al. [2009] propose a graph-based multi-label learning approach for image annotation. Gao et al. [2011] propose a hypergraph learning approach, which aims to estimate the relevance of images. Chen et al. [2012] propose an SVM classifier per query to learn relevance scores of its associated photos. These multi-label learning approaches usually need complete and well class assignments in the period of model training. However, manually annotated tags on OSW contain many incorrect and noisy ones, which do not match the requirement of multi-label learning approaches.

In recent years, matrix completion techniques are also introduced to address the poor initial image annotation problem, and good experiment results have been obtained. Goldberg et al. [2010] propose a matrix completion for transductive classification. Lin et al. [2013] propose an linear sparse construction-based tag completion method to refine the tags by using the semantic and visual similarities in the same group of the test images. Wu et al. [2013] build a concise tag matrix completion computational framework. They not only strengthen the consistency between the similarity of tag semantic and visual content, but also restrict the tag correlation consistency between the completed and observed tag matrix.

2.2. Tag-Based Image Retrieval

Rich surrounding texts brought by manual annotation such as tags make it easy to conduct tag-based image retrieval (TBIR) by regarding tags as queries. In many recent works, visual consistency between images and semantic information of tags are both considered [Liu et al. 2009a; Wei et al. 2013; Wu et al. 2014]. They investigate the bag-of-words and bag-of-visual-words representations that are extracted from both the textual information and visual content of images, respectively. Hu et al. [2008] propose an image ranking method based on multiple-instance learning, which utilizes sets of regions as image representation. Liu et al. [2009a] utilize an optimization framework to automatically rank images based on their relevance to a given tag without any intermediate tag processing. Haruechaiyasak and Damrongrat [2010] adopt content-based image retrieval techniques to refine the retrieval results achieved by tag-based image retrieval method. Wu et al. [2014] propose a two-step self-tune manifold similarity ranking scheme that aims at preserving both visual and semantic resemblance in the similarity ranking, including the visual-based similarity ranking and the semantic-oriented similarity re-ranking. Yu et al. [2015] propose a learning-based ranking model, in which both the click and visual feature are adopted simultaneously in the learning process. Most of existing relative algorithms are formulated on the basis of large margin structured output learning by modeling the visual consistency with a hypergraph regularizer term.

2.3. User-Aware Tag Recommendation

There is also a good deal of literature utilizing rich user information as social attributes to enhance the performance of tag recommendation. Sang et al. [2012] propose a user-aware tag refinement approach which utilizes user information as additional prior. Wang et al. [2013] propose a social recommendation method aiming at user-aware tag recommendation. Qian et al. [2014] propose an image retagging approach aiming at a wide range coverage of semantics, in which both the tag-image relevance and the semantic compensations to existing tags are fused to refine the tag list. Yang et al. [2014] propose a tag refinement method which leverages massive social images and their associated tags as the social assistance to learn the classifiers to directly refine noisy tags for given social images. These methods have improved the tag recommendation with social attribute modeling. However, there are still many other attributes which can be utilized to enhance the tag recommendation performance.

ALGORITHM 1: POI-based Tag Matrix Completion Algorithm

Input:
Original Tag Sub-Matrices: $\widehat{T}_k, k \in \{1, \ldots, p\}$
Image-User Sub-Matrices: $\widehat{U}_k, k \in \{1, \ldots, p\}$
Visual Feature Sub-Matrices: $V_k, k \in \{1, \ldots, p\}$
Parameters: $\alpha, \beta, \gamma, \eta, \lambda, \epsilon$
Output:
Completed Tag Sub-Matrices: $T_k, k \in \{1, \ldots, p\}$;
Initialization: $W_1 = I_{d\times m}, T_k^1 = \widehat{T}_k^1, t = 1$;
while $\|\mathcal{L}^{t+1} - \mathcal{L}^t\| \geq \epsilon\mathcal{L}^t$ **do**
  Step size $\delta_t = \delta_0/t$;
  {The loop below is executed in parallel}
  for $k = 1$ to p **do**
    Calculate $\overline{T}_k^{t+1}$: Equation (13)
    Update T_k^{t+1}: Equation (17)
    t = t + 1;
  end for
  Calculate $\overline{W}^{t+1}$: Equation (14)
  Update W^{t+1}: Equation (18)
end while
return $T_k^t, k \in \{1, \ldots, p\}$;

2.4. Location-Based Visual Content Analysis

Besides numerous research works on image annotation, there are many research works focusing on combining geographical attributes and visual content [Moxley et al. 2008; Crandall et al. 2009; Liu et al. 2010, 2014; Zheng et al. 2014]. Moxley et al. [2008] adopt a geographical-based search strategy to provide candidate tags and images which are similar in visual content. To analyze large-scale online image collections with both geographical and visual content information, Crandall et al. [2009] formularize the image location estimation as a classification problem by classifying images into POI categories. Liu et al. [2014] propose a unified framework using sub-space learning in personalized and geo-specific tag recommendation for images on OSW.

3. APPROACH

Our POI-based parallel tag matrix completion framework consists of two steps. The first step is the POI detection step dealing with GPS locations of images by a mean shift procedure, which aims at matrix partition preprocessing. The second step is the proposed POI-based tag matrix completion model. Detailed information is provided as follows.

3.1. Notations and Problem Definitions

The problem that we try to solve is that given a large-scale image collection with rich annotated tags, how can we automatically complement the missing tags and filter noisy tags for tag-related applications? First, we denote n as the number of images uploaded by l users in the dataset and m as the number of unique tags. To address this problem, our goal is to automatically complete a real tag matrix $T \in \mathbb{R}^{n\times m}$ based on an observed binary tag matrix $\widehat{T} \in \{0, 1\}^{n\times m}$, where T_{ij} indicates the probability of assigning tag j to image i. Each element $\widehat{T}_{ij}$ of $\widehat{T}$ is set to 1 if tag j is assigned to image i, and otherwise 0. The ith row of $\widehat{T}$ can be regarded as a term frequency (TF) vector of all tags for image i. Similarly, we can define the corresponding observed user-tag matrix

$\widehat{U} \in \mathbb{R}^{l \times m}$, where $\widehat{U}_{rj} = \sum_i \widehat{T}_{ij}$ if image i belongs to user r. The rth row of $\widehat{U}$ can be considered as a histogram of tags for user r.

Besides tag matrices, the visual content is also involved in our proposed method. We represent the visual content of images by $V \in \mathbb{R}^{n \times d}$, where the ith row corresponds to a d dimension visual feature of image i. Furthermore, to take the relationship between different tags into account, we define the tag correlation matrix $R \in \mathbb{R}^{m \times m}$. We use cosine distance to measure the correlation score between two tags as follows:

$$R_{ij} = \widehat{T}_{\cdot i}^{\top} \cdot \widehat{T}_{\cdot j}, \tag{1}$$

where $\widehat{T}_{\cdot i}$ is the ith column vector of $\widehat{T}$, as in Liu et al. [2014].

3.2. Finding POIs Using Mean Shift

In general, a large value of n makes the original tag matrix T very large. Therefore, the computational burden for directly handling such large matrices is costly and prohibitive. As discussed in the previous section, geographically adjacent images may have similar visual content or semantic information with higher probability. Therefore, we utilize the geo-locations as the auxiliary information to partition the whole image collection into isolated blocks for improving the computation efficiency.

We observe from OSM data that there are lots of images uploaded around certain places. That is to say, the areas with massive uploaded images with similar GPS information are potential POIs. Therefore, we apply a mean shift approach to detect POIs from the GPS location information (latitudes and longitudes) of social images. With the POI detection results, we can partition all relative matrices into a set of sub-matrices according to POIs. The key step of a mean shift procedure is calculated as:

$$m(\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathbf{x}_i g\left(\|\mathbf{x} - \mathbf{x}_i\|^2 / h\right)}{\sum_{i=1}^{n} g\left(\|\mathbf{x} - \mathbf{x}_i\|^2 / h\right)} - \mathbf{x}, \tag{2}$$

where $\mathbf{x} = \langle lat, lon \rangle$ and $\mathbf{x}_i = \langle lat_i, lon_i \rangle$ denote the latitude and longitude of the POI center and the ith image, respectively. The kernel function g is used for density estimation with bandwidth parameter h. The mean shift algorithm is performed in an iterative process, where the update rule is:

$$\mathbf{x}^{(l+1)} = \mathbf{x}^{(l)} + m\left(\mathbf{x}^{(l)}\right) \tag{3}$$

The GPS locations clustering results are gained until the mean shift procedure converges. After that, we achieve our goal for finding POIs. Then, we obtain partitioned sub-matrices of each detected POI for our proposed framework.

3.3. POI-Based Tag Completion Algorithm

Without loss of generality, we take POI k as an example for clearer presentation of the formulation of our proposed algorithm. Correspondingly, the notations with subscript k refer to the k POI. For better description, we permutate images in each sub-matrix of $\widehat{T}$ grouping by the users' order in each sub-matrix $\widehat{U}_k \in \mathbb{R}^{l_k \times n_k}$ of $\widehat{U}$ and denote the new one as $\widehat{T}_k \in \mathbb{R}^{n_k \times m}$ ($k \in \{1, 2, \ldots, p\}$, $\sum_{k=1}^{p} n_k = n$). For example, image 1 and 2 belong to user I, image 3, 4, and 5 belong to user II, and so on. To build the unified framework of POI-based tag matrix completion, we consider three types of significant constraint terms.

The first type is image-wise constraint terms. To address the coherence in visual content and tags, we penalize the difference of similarities in visual feature space and tag space with a Frobenius norm $\|T_k T_k^{\top} - V_k V_k^{\top}\|_F^2$, corresponding to POI k. However, low-level visual features are less capable compared with tags for a semantic representation of the image. To reduce this semantic gap, we introduce a feature mapping

matrix $W \in \mathbb{R}^{d\times m}$, which can directly map the visual feature into textual semantic space. Then, the visual constraint term is defined as $\|F_k\|_F^2$, where

$$F_k = T_k T_k^\top - V_k W W^\top V_k^\top. \tag{4}$$

Besides the visual constraint term, we also introduce a user-related prior constraint term according to the least effort principle. For all of the images uploaded by the same user belonging to the same POI, the user's tagging behavior tends to have no difference among these images. Without loss of generality, suppose user r has r_u image in POI k. Then, we build an auxiliary matrix $A_k \in \mathbb{R}^{l_k\times n_k}$ defined as follows:

$$A_k = \begin{bmatrix} C_{(l_k-r_k)\times(n_k-r_k)} & 0 \\ 0 & I_{r_k} \end{bmatrix}, \tag{5}$$

where $C_{ri} = \frac{1}{r_u}$ if image i belongs to user r, and $C_{ri} = 0$ otherwise. The identity matrix $I_{r_k} \in \mathbb{R}^{r_k\times r_k}$ is attached to the rest of r_k images' lack of user information in POI k by assigning an anonymous user to each image.

Similar as the ith row of T_k depicting the actual tag distribution of image i, the ith row in the product $A_k\widehat{U}_k$ reveals the average tag distribution for all the images related to user r. So, $A_k\widehat{U}_k$ can be regarded as a refined prior estimation for T_k using the group of images for each user. With the assistance of $A_k\widehat{U}_k$, we define the POI-based user-related prior constraint term by calculating the difference of similarities between $A_k\widehat{U}_k$ and T_k in a Frobenius norm. The user-related prior constraint term is denoted as $\|G_k\|_F^2$, where G_k is denoted as follows:

$$G_k = T_k T_k^\top - A_k\widehat{U}_k\widehat{U}_k^\top A_k^\top. \tag{6}$$

Both visual and user-related prior constraint terms compare the differences of similarities among different images.

The second type is tag-wise constraint terms. Tag co-occurrence is proved to be effective in image tagging [Wu et al. 2008]. Its key idea is that the more common tags two images share, the higher semantic similarity they have beyond the tags. To maintain the tag co-occurrence consistency of T_k before and after the optimization, we expect a minor difference between the completed and the original tag correlation matrix. The tag correlation constraint term is denoted as $\|H_k\|_F^2$, where $H_k = T_k^\top T_k - R_k$.

Since we reconstruct the completed tag matrix T_k based on an observation version $\widehat{T}_k$, the completed one should be similar to the observed one. That is, we prefer the solution of T_k with a small value of a tag consistency constraint term denoted as $\|K_k\|_F^2$, where $K_k = T_k - \widehat{T}_k$. These two tag-wise constraint terms focus on the preservation of consistency between the completed and the observed tag sub-matrix.

The last-but-not-least type is the regularization term. To avoid a dense solution of T_k, we require that only a small number of entries of T_k be nonzero, that is, several unique tags are attached to each image. As studied in many sparse coding literatures [Mairal et al. 2010], we consider introducing an ℓ_1-norm regularization term $\|T_k\|_1$ for a sparse solution of T_k. For the shared mapping matrix W, we also add an ℓ_1-norm regularization term $\|W\|_1$ for sparsity.

Finally, with respect to all of these criteria, we formulate our POI-based tag matrix completion framework as follows:

$$\min_{T_{1,2\ldots,p},W} \sum_{k=1}^{p} \mathcal{L}_k + \eta\|W\|_1 \tag{7}$$

$$\mathcal{L}_k = \|F_k\|_F^2 + \alpha\|G_k\|_F^2 + \beta\|H_k\|_F^2 + \gamma\|K_k\|_F^2 + \lambda\|T_k\|_1, \tag{8}$$

where $\alpha, \beta, \gamma, \lambda, \eta > 0$ are parameters whose values are fixed in a cross-validation procedure.

3.4. Optimization in Parallel

As we can see from the formulation above, the ℓ_1-norm regularization terms $\|T_k\|_1$ and $\|W\|_1$ make the whole objective function non-smooth. While the subgradient descent approach is one of the commonly used iterative methods dealing with non-convex optimization problems, its remarkable calculation efficiency in each iteration makes it more practical in processing large-scale image datasets. So, we adopt the subgradient descent approach to solve the non-smooth optimization problem that we proposed above.

However, we may get dense immediate solutions $T_k^t, k \in \{1, \ldots, p\}$ if we directly use the subgradient descent approach to solve the original optimization problem. It will significantly increase the calculation time per iteration. To avoid this potential difficulty, we split the objective function into two parts according to the composite function optimization method [Cartis et al. 2011]. In particular, we construct an auxiliary function as follows:

$$B_k = \|F_k\|_F^2 + \alpha\|G_k\|_F^2 + \beta\|H_k\|_F^2 + \gamma\|K_k\|_F^2 \tag{9}$$

Then, the original loss function in Equation (7) (denoted as $\mathcal{L}$) can be rewritten as:

$$\mathcal{L} = \sum_{k=1}^{p}(B_k + \lambda\|T_k\|_1) + \eta\|W\|_1 \tag{10}$$

We divide the optimization procedure into two steps for each iteration t.

At the first step, we calculate the subgradients of the auxiliary function subject to both T_k^t and W^t as follows:

$$\nabla_{T_k^t} B_k = 2F_kT_k^t + 2\alpha G_kT_k^t + 2\beta T_k^tH_k + \gamma K_k \tag{11}$$

$$\nabla_{W^t} B_k = 2\left(\sum_{k=1}^{p} V_k^\top F_k V_k\right) W^t \tag{12}$$

Then, we update the immediate solutions $\overline{T}_k^{t+1}$, $\overline{W}^{t+1}$ of auxiliary function by:

$$\overline{T}_k^{t+1} = T_k^t - \delta_t \nabla_{T_k^t} B_k \tag{13}$$

$$\overline{W}^{t+1} = W^t - \delta_t \nabla_{W^t} B_k, \tag{14}$$

where δ_t is the step size.

At the second step, we are going to solve another optimization problem:

$$T_k^{t+1} = \arg\min_{T_k} \frac{1}{2}\left\|T_k - \overline{T}_k^{t+1}\right\|_F^2 + \lambda\delta_t\|T_k\|_1 \tag{15}$$

$$W^{t+1} = \arg\min_{W} \frac{1}{2}\left\|W - \overline{W}^{t+1}\right\|_F^2 + \eta\delta_t\|W\|_1 \tag{16}$$

Combining the immediate solutions, we obtain the solution as follows:

$$T_k^{t+1} = \max\left(\mathbf{0}, \overline{T}_k^{t+1} - \lambda\delta_t \mathbf{1}_n\mathbf{1}_m\right) \tag{17}$$

$$W^{t+1} = \max\left(\mathbf{0}, \overline{W}^{t+1} - \eta\delta_t \mathbf{1}_d\mathbf{1}_m\right), \tag{18}$$

where $\mathbf{1}_d$ is a vector with all ones of the d dimensions.

Parallel processing within POIs. After the introduction of our proposed POI-based tag matrix completion algorithm in single POI above, we discuss the whole optimization procedure in all of the POIs. Since the matrices T, V, and U are divided into different POIs in the clustering step 3.2, we conduct the optimization procedure in parallel on different POI-specific sub-matrices.

For the tag sub-matrix T_k of POI k, its calculation process is independent from sub-matrices in other POIs. But for the feature mapping matrix W, it will lead to abnormal synchronization if each W_k differs from one another among POIs in a parallel processing environment. So we make W shared by all of the sub-matrices V_k in parallel computation. Algorithm 1 illustrates the main steps in our solution for the optimization problem.

3.5. Tag-Based Image Retrieval with Asymmetric *Locality Sensitive Hashing* (ALSH)

In order to fit the response time requirement and storage limitations of TBIR on mobile devices, we introduce hashing techniques to reduce the time cost, and meanwhile, improve the efficiency of retrieval. Instead of traditional Nearest Neighbor Search (NNS), we formulate the TBIR into the problem of *Maximum Inner Product Search* (MIPS) [Shrivastava and Li 2014]. Then, we fuse the Asymmetric *Locality Sensitive Hashing* (ALSH) method proposed by Shrivastava and Li [2014] into our framework for image retrieval. First, we expatiate the key issues of the MIPS problem, and then we provide technical details of the ALSH approach.

3.5.1. Locality Sensitive Hashing (LSH) for L2 Distance. *Locality Sensitive Hashing* (LSH) is a widely used hashing techniques in information retrieval. Different from the hashing methods with a single hash function, LSH has a family function which has the property for maintaining locality based on probabilistic bucketing. The property is that more similar input objects have a higher collision probability in the range space than less similar ones. The definition of LSH is described as follows:

Definition (*Locality Sensitive Hashing (LSH)*). A Family $\mathcal{H}$ is called S_0, cS_0, p_1, p_2-sensitive if, for any two points $x, y \in \mathbb{R}_D$, h chosen uniformly from $\mathcal{H}$ satisfies the following:

(1) if $Sim(x, y) \geq S_0$, then $\Pr_{\mathcal{H}}\left(h(x) = h(y)\right) \geq p_1$
(2) if $Sim(q, x) \leq cS_0$, then $\Pr_{\mathcal{H}}\left(h(x) = h(y)\right) \leq p_2$,

where $p_1 > p_2$ and $c < 1$ is needed for efficient approximate nearest neighbor search.

Specific to the TBIR task, we introduce a novel LSH function family for all $L_p (p \in (0, 2])$ distances proposed by Datar et al. [2004]. This scheme provides an LSH family (L2LSH) for L_2 distances when $p = 2$, particularly. The hash function in L2LSH is designed as:

$$h^{L2}_{a,b}(x) = \left\lfloor \frac{a^\top x + b}{r} \right\rfloor, \tag{19}$$

where $\lfloor\rfloor$ is the floor operator and the random scalar b is uniformly generated from $[0, r]$ (r is a fixed parameter which can be tuned). Each element $a_i \in N(0, 1)$ of the random vector a is generated from independent identically distributed gaussian distribution. Then, the collision probability of L2LSH can be calculated as

$$\Pr\left(h^{L2}_{a,b}(x) = h^{L2}_{a,b}(y)\right) = F_r(d), \tag{20}$$

$$F_r(d) = 1 - 2\Phi(-r/d) - \frac{2}{\sqrt{2\pi}(r/d)}\left(1 - e^{-(r/d)^2/2}\right), \tag{21}$$

where $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$ is the cumulative density function (cdf) of the standard normal distribution $N(0, 1)$ and $d = \|x - y\|_2$ is the Euclidean distance between the vectors x and y. This collision probability $F_r(d)$ is a monotonically decreasing function referring to the distance d. Hence, $h_{a,b}^{L2}$ is guaranteed as an LSH for L_2 distances.

3.5.2. Maximum Inner Product Search. Focusing on the MIPS problem of TBIR, we concern ourselves with the optimal retrieval results $\hat{p} \in \mathcal{S}$ achieved by maximizing (or approximately maximizing) the **inner product** $q^\top p$, where $q \in \mathbb{R}^k$ is the latent feature vector of the given query tag. Formally, we are interested in efficiently computing

$$p = \arg\max_{x \in \mathcal{S}} q^\top x. \tag{22}$$

The MIPS problem is related to the problem of NNS, which, instead, requires computing

$$p = \arg\min_{x \in \mathcal{S}} \|q - x\|_2^2 = \arg\min_{x \in \mathcal{S}} \left(\|x\|_2^2 - 2q^\top x \right). \tag{23}$$

These two problems are equivalent if the norm of every element in $x \in \mathcal{S}$ is constant. The value of the norm $\|q\|_2$ has no effect on the solution of the MIPS as it is a constant throughout and does not change the characteristics of arg max or arg min. As studied in Shrivastava and Li [2014], there are many scenarios in which MIPS arises naturally at places where the norms of the elements in $\mathcal{S}$ have very significant variations [Koenigstein et al. 2012] and can not be controlled. As a consequence, existing fast algorithms for the problem of approximate NNS can not be directly used for solving MIPS.

3.5.3. Asymmetric Locality Sensitive Hashing. In the LSH algorithm (Section 3.5.1), we use the same hash function h for both the preprocessing step and the query step. We assign buckets in the hash table to all the candidates $x \in \mathcal{S}$ using h. The theory behind LSH still works if we use hash function h_1 for preprocessing $x \in \mathcal{S}$ and a different hash function h_2 for querying, as long as the probability of the event $\{h_2(q) = h_1(x)\}$ increases with $Sim(q, x)$, and there exist p_1 and p_2 with the required property. The traditional LSH definition does not allow this asymmetry, but it is not a required condition in the proof. For this reason, we can relax the definition of c-NN without losing runtime guarantees.

Definition (*Asymmetric* Locality Sensitive Hashing *(ALSH)*). A Family $\mathcal{H}$, along with the two vector functions $Q : \mathbb{R}^D \mapsto \mathbb{R}^{D'}$ (**Query Transformation**) and $P : \mathbb{R}^D \mapsto \mathbb{R}^{D'}$ (**Preprocessing Transformation**), is called (S_0, cS_0, p_1, p_2)-sensitive if for a given c-NN instance with query q, and the hash function h chosen uniformly from $\mathcal{H}$ satisfies the following:

(1) if $Sim(q, x) \geq S_0$, then $\Pr_{\mathcal{H}}\left(h(Q(q)) = h(P(x))\right) \geq p_1$
(2) if $Sim(q, x) \leq cS_0$, then $\Pr_{\mathcal{H}}\left(h(Q(q)) = h(P(x))\right) \leq p_2$.

Here, x is any point in the collection $\mathcal{S}$.

When $Q(x) = P(x) = x$, we recover the vanilla LSH definition with $h(.)$ as the required hash function. Coming back to the problem of MIPS, if Q and P are different, the event $\{h(Q(x)) = h(P(x))\}$ will not have probability equal to 1 in general. Thus, $Q \neq P$ can counter the fact that self similarity is not highest with inner products. We just need the probability of the collision event $\{h(Q(q)) = h(P(y))\}$ to satisfy the conditions of the definition of c-NN for $Sim(q, y) = q^\top y$. The query transformation Q is only applied on the query and the pre-processing transformation P is applied to $x \in \mathcal{S}$ while creating hash tables. Meanwhile, the source of randomization h for both q and $x \in \mathcal{S}$ is the same. It is this asymmetry which will allow us to solve MIPS efficiently. In Section 3.5.4, we

will explicitly show a construction (and hence the existence) of the ALSH function for solving MIPS.

3.5.4. From MIPS to NNS. As argued previously, $\|x - y\|_2 = \sqrt{(\|x\|_2^2 + \|y\|_2^2 - 2x^\top y)}$ is not monotonic in the inner product $x^\top y$ unless $\|x\|_2^2$ and $\|y\|_2^2$ are constants. Therefore, $h_{a,b}^{L2}$ is not suitable for MIPS. Without loss of generality, we set the norm of the query q normalized for the MIPS problem, that is, $\|q\|_2 = 1$, because the arg max operator is independent of $\|q\|_2$ in computing optimization Equation (22). Then, the norm of x is the main difference between NNS and MIPS . In particular, we can choose to let $U \leq 1$ be a constant such that

$$\|x_i\|_2 \leq U < 1, \forall x_i \in \mathcal{S} \tag{24}$$

If this is not the case, then during the one-time preprocessing, we can always divide all x_is by $\max_{x_i \in \mathcal{S}} \frac{\|x_i\|_2}{U}$. We are sure that scaling all x_i's by the same constant U does not change the optimal solution of $\arg\max_{x \in \mathcal{S}} q^\top x$.

The key step in our ALSH algorithm can be divided into two steps. First, we define two vector transformations $P : \mathbb{R}^D \mapsto \mathbb{R}^{D+M}$ and $Q : \mathbb{R}^D \mapsto \mathbb{R}^{D+M}$ as follows:

$$P(x) = \left[x; \|x\|_2^2; \|x\|_2^4; \cdots; \|x\|_2^{2^M}\right] \tag{25}$$

$$Q(x) = [x; 1/2; 1/2; \cdots; 1/2], \tag{26}$$

where $[;]$ is the concatenation operator. $P(x)$ appends M scalars of the form $\|x\|_2^{2^i}$ at the end of the vector x, while $Q(x)$ simply appends M"1/2" to the end of the vector x.

By observing that

$$\|P(x)\|_2^2 = \|x\|_2^2 + \|x\|_2^4 + \cdots + \|x\|_2^{2^M} + \|x\|_2^{2^{M+1}} \tag{27}$$

$$\|Q(x)\|_2^2 = \|q\|_2^2 + M/4 = 1 + M/4 \tag{28}$$

$$Q(q)^\top P(x_i) = q^\top x_i + \frac{1}{2}\left(\|x\|_2^2 + \|x\|_2^4 + \cdots + \|x\|_2^{2^M}\right), \tag{29}$$

we obtain the following key equality:

$$\|Q(q) - P(x_i)\|_2^2 = (1 + M/4) - 2q^\top x_i + \|x\|_2^{2^{M+1}}. \tag{30}$$

Since $\|x_i\|_2 \leq U < 1$, then we have $\|x\|_2^{2^{M+1}} \to 0$ at the lower rate (exponential to exponential). The term $(1 + M/4)$ is a fixed constant. As long as M is not too small (e.g., $m \geq 3$ would suffice), we have

$$\arg\max_{x \in \mathcal{S}} q^\top x \simeq \arg\min_{x \in \mathcal{S}} \left\|Q(q) - P(x)\right\|_2^2 \tag{31}$$

Then, the solution of approximate NNS is connected to the solution of MIPS after the key "concatenation" procedure. This works only after shrinking the norms, as norms greater than 1 will instead blow the term $\|x\|_2^{2^{M+1}}$. Transformations P and Q, when norms are less than 1, make the L2 distance $\|Q(q) - P(x_i)\|_2$ rank correlate with the (un-normalized) inner product.

4. EXPERIMENTS

We evaluate the performance of our proposed PTC approach on two application tasks: automatic image annotation and tag-based image retrieval.

Table I. Statistics of the Datasets Used in the Experiments

Dataset	Image		User		Tag				
	#Total	#GPS	#User	#Image per User	#Tag	#Tag per Image		#Image per Tag	
						mean	max	mean	max
YFCC100M	99,206,564	48,469,829	581,099	170.72	N/A	N/A	N/A	N/A	N/A
London	1,338,388	771,099	16,225	47.52	1,000	5.2	63	4,016.4	481,957
New York	1,210,094	732,555	15,344	47.74	1,000	5.5	71	4,031.5	299,867

4.1. Dataset and Experiment Settings

According to our application scenario, we use a large-scale social image database published by Yahoo Web Lab[3] called YFCC100M to conduct the experiments. The first row in Table I shows some statistical information about this dataset.

Users' tagging behavior in specific POI may become uncertain as the geographical scope goes larger. According to "*landmark-scale*" POI defined in Crandall et al. [2009], we fix the bandwidth parameter h as 0.005 in the Mean Shift procedure for POI detection. This bandwidth parameter is in correspondence with 500 meters as maximum geographical radius for the POIs detected in our experiments. On the basis of POI setting, we extract two city subsets, *London* and *New York* (also used as the name of the subset), from this large dataset by geographical restriction, for the reason that there are more images in them than other cities.

For each city subset, we first choose a maximum bounding rectangle on the world map and then select images whose latitude and longitude fall in the region. Then, we obtain 1,026,345 and 924,707 images in total for *London* and *New York*, respectively. As studied in Wu et al. [2009], the tag distribution among images is extremely unbalanced and the majority of tags belong to a few images. Then, we rank the tags according to their number of annotated images and select the top 1,000 to serve as the vocabulary in the experiment. After this operation, the size of *London* shrinks to 771,099, and *New York* to 732,555, respectively. The second and the third row in Table I show some statistical information about *London* and *New York*.

For both *London* and *New York*, we extract dense SIFT [Vedaldi and Fulkerson 2010] descriptor as the local visual feature. Then, we cluster randomly chosen 1,000,000 descriptor samples into 1,000 visual words. Each local feature descriptor is quantized to one of these 1,000 visual words for Bag-Of-Visual-Words representation. After cross-validation procedure, we set $\alpha = 100$, $\beta = 10$, $\gamma = 1$, $\lambda = 1$, $\eta = 1$. For the TMC method, we adopt the parameter settings reported in their paper. The initial step size δ_0 is set as 10^{-6} according to experience. For the ALSH, the parameters are determined as $r = 2.5, U = 0.83, M = 3$, being the same with Shrivastava and Li [2014].

4.2. Comparison Methods

We compare the proposed method and its simplified variant with a baseline state-of-the-art approach as follows:

—**Tag Matrix Completion (TMC)** [Wu et al. 2013], which directly completes the tag matrix by exploiting the tag correlation and image examples similarity to ensure the consistency between the observed tag matrix and the completed tag matrix.
—**Tag Propagation (TagProp)** [Guillaumin et al. 2009], which utilizes a weighted nearest neighbor graph to propagate the tag information from the tagged images to the un-tagged images.
—**Tag Completion via Linear Sparse Reconstruction (LSR)** [Lin et al. 2013], which adopts linear sparse construction framework to complete the tag

[3]http://webscope.sandbox.yahoo.com/.

Table II. Performance Comparison about MAP for Automatic Image Annotation

London	MAP@5							MAP@20						
Methods	FastTag	TagRel	TMC	TagProp	LSR	PTC-U	PTC	FastTag	TagRel	TMC	TagProp	LSR	PTC-U	PTC
e=1	77.81	76.45	82.37	78.35	79.57	83.15	85.68	60.90	61.77	66.87	63.78	66.44	67.55	71.18
e=2	76.62	76.86	82.49	80.96	81.82	83.78	87.50	60.17	62.14	67.93	66.79	67.73	68.86	72.47
e=3	76.01	78.12	82.79	81.13	81.97	84.03	88.17	59.78	62.86	67.61	67.35	68.37	69.46	73.04
e=4	77.58	78.39	82.87	81.63	82.39	84.36	88.64	61.34	64.31	67.75	68.01	68.96	69.53	73.38
e=5	78.36	80.07	83.61	82.58	84.16	85.67	**89.01**	63.13	65.05	68.49	68.82	69.25	70.68	**74.78**
New York	MAP@5							MAP@20						
Methods	FastTag	TagRel	TMC	TagProp	LSR	PTC-U	PTC	FastTag	TagRel	TMC	TagProp	LSR	PTC-U	PTC
e=1	68.88	71.26	77.13	73.92	75.86	78.06	83.39	56.10	58.61	62.48	60.07	62.31	64.30	70.36
e=2	68.11	72.41	77.41	74.24	77.21	78.38	85.58	56.06	59.43	62.81	60.90	63.09	64.34	69.97
e=3	67.46	72.39	77.68	75.56	78.43	78.66	87.33	55.65	59.67	62.92	61.77	63.68	64.39	**70.92**
e=4	69.27	74.90	77.60	76.40	78.75	78.70	86.84	57.11	60.64	63.00	62.39	64.40	64.21	70.90
e=5	71.40	75.38	78.24	76.73	79.58	79.35	**87.50**	58.98	61.39	62.98	62.70	65.11	64.27	70.46

representation for given images by keeping the reconstruction coefficients consistent in both image-specific and tag-specific views.

—**PTC**, our method containing both the user-related prior constraint term in loss function and the POI-based matrix partition strategy for parallel processing.

—**PTC-U**, our proposed method without the user-related prior constraint term, corresponding to the case of $\alpha = 0$ in Equation (8).

4.3. Automatic Image Annotation

Given a query image q in automatic the image annotation task, we simply rank all the tags in descending order of their probability scores attached to image q, corresponding to the qth row in T. In particular, to test the robustness of our proposed method to the number of initial tags, we vary the number of initial training tags (denoted as e) for each training image from $\{1, 2, 3, 4, 5\}$. Without loss of generality, suppose image i has m_i manually annotated tags, corresponding to m_i non-zero entries in the ith row of the observed tag matrix in the training set. If $e \leq m_i$, we randomly select e tags as partial annotation for image i. Otherwise, if $e > m_i$, we drop out image i from the training set. We use the Mean Average Precision (MAP) on the top s ($s \in \{5, 10, 20\}$) of completed tags to measure the performance of different algorithms.

As shown in Table II, the annotation accuracy goes up along with the increase of the number of initial tags for all methods in vertical comparison. It is in line with our expectation because more initial tags for each image bring richer information and lead to lower prediction risk. In horizontal comparison, we observe that the proposed PTC approach outperforms its simplified variant PTC-U method and the PTC-U method outperforms the TMC method. This experimental phenomenon demonstrates that both our matrix partition strategy and user-related prior constraint term make contributions to the enhanced performance. The matrix partition strategy makes locality consistency more compact in tag space. And the user-related prior constraint term is in coincidence with common sense. Figure 2 shows several annotation results selected from *London* in condition of $e = 5$.

4.4. Tag-Based Image Retrieval via ALSH

In the tag-based image retrieval task, we consider a simple scenario that the query is a single tag. Given the completed tag matrix T, we follow the PureSVD approach in Cremonesi et al. [2010] to generate image and tag latent vectors. Formally, the SVD of T is calculated as follows:

$$T = X\Sigma Y^{\top}, \tag{32}$$

where $X \in \mathbb{R}^{n\times f}, Y \in \mathbb{R}^{m\times f}$, and $\Sigma \in \mathbb{R}^{f\times f}$. The *latent dimension* f is appropriately chosen as 850 in our experiment. The rows of matrix $Z = X\Sigma$ are regarded as the

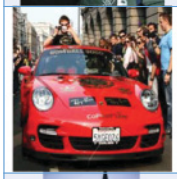
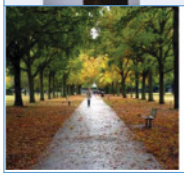
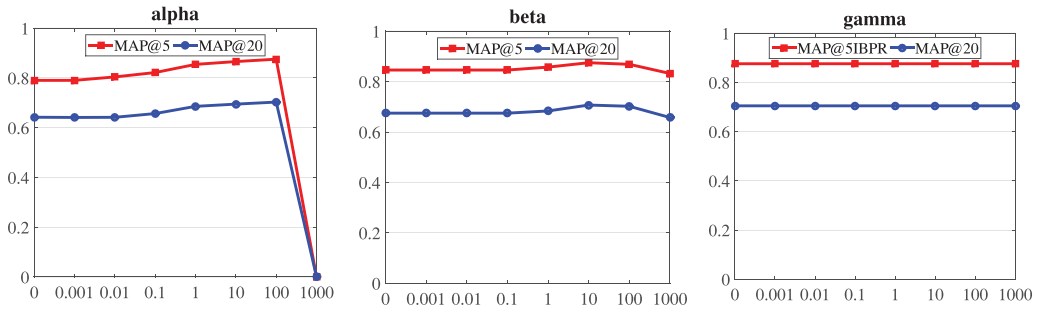
Photos	Ground Truth	TMC	PTC-U	PTC
	london united kingdom westminster big ben parliament palace	london united kingdom biorhythms big ben palace silhouette	united kingdom westminster big ben southbank parliament palace	london united kingdom westminster big ben parliament palace
	london england united kingdom great britain greater london trafalgar square	london england united kingdom great britain greater london city of westminster	london england great britain greater london city of westminster feggy	london england great britain greater london city of westminster rebel
	london england united kingdom great britain greater london river thames	london england united kingdom great britain greater london lambeth	london united kingdom great britain live river thames lambeth	london united kingdom great britain river thames comedian lambeth
	london england united kingdom olympics stratford athletics	london england united kingdom olympics lifelog athletics	london england united kingdom stratford construction athletics	london england united kingdom olympics stratford athletics
	london england big ben architecture night westminster	england big ben architecture lifelog night trafalgar square	london big ben architecture westminster trafalgar square one	london england big ben architecture westminster trafalgar square
	london olympics stadium london 2012 stratford olympic park	london england uk united kingdom 2012 lifelog	london 2012 park olympics olympic park stadium	london 2012 park olympics olympic park stadium

Fig. 2. Examples of image annotation results by different methods.



(a) α (b) β (c) θ

Fig. 3. Sensitivity analysis of α, β, γ in terms of MAP@5, MAP@20 on the New York dataset.

Table III. Performance of Tag-Based Image Retrieval with Single-Tag Queries

London	MAP@5														
	Normal			HashLen = 64			HashLen = 128			HashLen = 256			HashLen = 512		
	TMC	PTC-U	PTC	TMC	PTC-U	PTC	TMC	PTC-U	PTC	TMC	PTC-U	PTC	TMC	PTC-U	PTC
r=1	64.03	64.03	64.03	56.27	56.89	57.20	58.94	60.00	60.14	59.33	60.45	60.57	60.86	61.40	61.48
r=2	72.92	72.93	72.94	64.21	65.79	66.42	69.09	70.34	70.58	69.35	70.08	70.21	69.55	70.12	70.33
r=3	80.86	80.86	80.86	73.23	74.12	74.76	76.15	76.88	77.07	76.29	77.23	77.42	75.51	77.50	77.70
r=4	87.75	87.79	87.79	78.43	79.87	80.27	82.17	81.91	82.06	82.63	84.95	84.11	83.55	84.47	84.90
r=5	90.21	90.24	90.26	80.34	81.70	82.12	84.27	85.05	85.27	84.76	85.20	85.33	85.00	85.40	85.42
New York	MAP@5														
	Normal			HashLen = 64			HashLen = 128			HashLen = 256			HashLen = 512		
	TMC	PTC-U	PTC	TMC	PTC-U	PTC	TMC	PTC-U	PTC	TMC	PTC-U	PTC	TMC	PTC-U	PTC
r=1	88.58	88.58	88.58	76.92	77.09	77.23	81.46	81.58	81.74	82.10	82.09	82.26	82.60	82.59	82.68
r=2	93.49	93.49	93.49	82.18	82.54	82.88	87.30	87.32	87.49	87.86	87.81	87.97	88.04	87.93	88.12
r=3	95.43	95.43	95.43	85.06	85.22	85.37	88.90	88.88	89.04	89.28	89.31	89.37	89.48	89.50	89.54
r=4	97.14	97.15	97.15	87.06	87.31	87.44	90.96	90.94	91.08	91.36	91.37	91.41	91.67	91.68	91.76
r=5	98.74	98.74	98.74	87.15	87.43	87.64	91.48	91.50	91.51	91.81	91.86	91.88	91.89	92.01	92.07

image latent vectors, while rows of Y are regarded as tag latent vectors. Then, we can compare different images by the inner product between the corresponding image and tag latent vectors. Hence, the solution of the MIPS problem gives top-ranked images for the given query tag based on the inner product $y_i^\top z_j, \forall j$.

Since every tag can be used as a query, we exploit all of the 1,000 tags in the vocabulary as queries. We keep the same setting of initial training tags as in the automatic image annotation task. However, we do not distinguish training or testing images. Instead we gather all the images in each dataset to serve as gallery images for retrieval. The rule of the relevance between image and query tag [Wu et al. 2009] we adopt in the experiment is that an image is relevant if its annotation contains the query.

Table III shows the MAP at top 5 results of tag-based image retrieval using single tag queries for *London* and *New York*. We can see that there is almost no significant difference in performance between all of the three methods without ALSH. It means that our method has little performance improvement compared with the TMC method. We attribute this phenomenon to poor original annotation which we used as ground truth.

We also observe another phenomenon that the performance of TBIR increases with the increment of the length of hash codes. The reason is that longer hash codes contain more discriminant information than shorter hash codes, which is a very common property of hashing techniques. Moreover, it is essential to apply ALSH for sublinear time retrieval within acceptable accuracy loss. Figure 4 illustrates some examples of single tag queries and the images returned by different methods in *New York*. Each word on the left side is the tag query. Besides each query, images displayed in three rows are the retrieval results corresponding to our proposed PTC method, the PTC-U method, and the TMC method from top to bottom, respectively.

4.5. Computational Efficiency Analysis

We evaluate the computational efficiency of our proposed PTC method and the TMC method. To create a fair environment for comparison, we use the same hardware and software platform to calculate the running time in each iteration. Both of the two algorithms are implemented on MATLAB R2014a, and executed on the Intel (R) Core (TM) i7-4770K CPU @3.50GHz and 32GB RAM PC. Figure 5 reveals the running time per iteration of both PTC and TMC methods. The shape of the curves demonstrates that our proposed PTC method has much less computational time cost than the TMC method as the increase of scalability. There is no surprising that our parallel computational framework conducted by the matrix partition strategy is the key point of efficiency improvement.

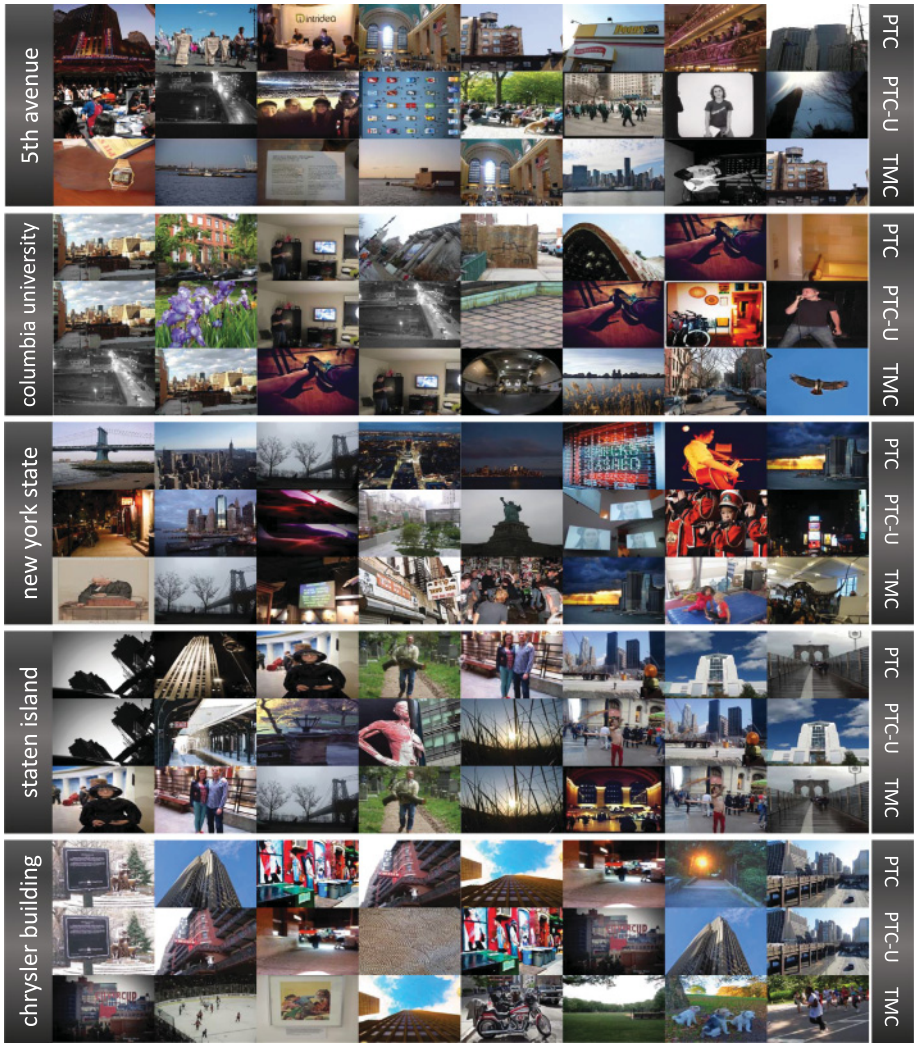


Fig. 4. Illustration of some examples in TBIR with single tag queries.

4.6. Parameter Sensitivity Analysis

We evaluate the sensitivity of the parameters in PTC with the *New York* dataset in the condition of $e = 5$. We select three significant parameters $\{\alpha, \beta, \theta\}$ to illustrate the MAP@5 and MAP@20 scores of AIA. For each parameter to be analyzed, we fix other parameters as the original value ($\alpha = 100, \beta = 10, \gamma = 1$). The parameters α, β, and θ range from 10^{-3} to 10^{3} (also includes 0) for sensitivity test.

Both of the curves in Figure 3(a) keep the ascending trend when $\theta \leq 100$, and they both tend to descend from $\theta = 100$ to $\theta = 1{,}000$ which leads to an ill-posed condition. To balance the performance, we choose 100 as the optimal value for α.

In Figure 3(b), we can see that the MAP scores of AIA first increase and then decrease at the inflection point $\beta = 10$. To obtain a relatively better performance, we choose 10 as the optimal value for β, respectively.

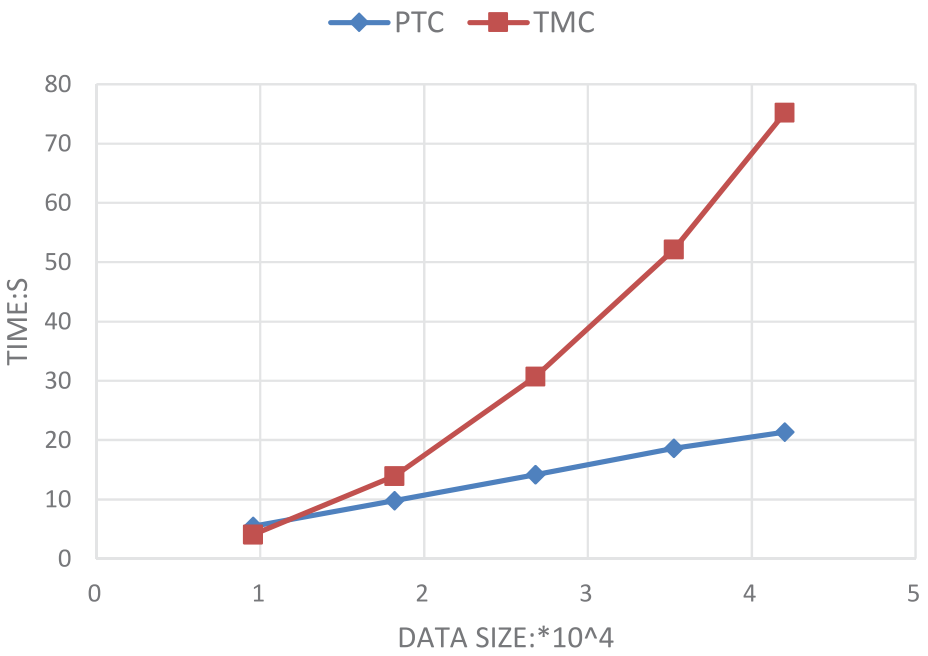


Fig. 5. CPU Execution Time of different methods.

Illustrated in Figure 3(c), we can see that the MAP scores of AIA are almost stable in a wide range of γ. The trend of all scores goes up with the increment of β and all of the metrics level off in the condition of $\beta \geq 1$. We choose 1 instead of other values as the optimal value for γ to make our model more flexible in parameter tuning.

4.7. Space Complexity Analysis

The space requirement for TMC, LSR, and Non-negative Matrix Factorization (NMF) is $O(n \times m)$. Since TagProp utilizes the K-NearestNeighbor (KNN) technique to reduce the size of search space, its space requirement is $O(n \times K)$. By reducing the problem scale with POI-specific relation matrix partition, the space requirement of our method is $O(n_{max} \times m)$, where n_{max} is the number of images in the largest POI. Typically, $m > 5K$ and $n > 10n_{max}$. Therefore, our method achieves the smallest space requirement than other state-of-the-art models.

4.8. Time Complexity Analysis

For time complexity analysis, we first calculate time complexity for each term in our loss function, then we aggregate them into a final result. The time complexity of each term and corresponding subgradients is listed as:

$$F_k. : O(l_k * m * n_k + l_k^2 * m)$$
$$G_k. : O(l_k * m * n_k + n_k^2 * m)$$
$$H_k. : O(l_k * m^2 + l_k * m * n_k)$$
$$K_k. : O(n_k * m^2)$$
$$\nabla_{T_k} B_k. : O(l_k * m * n_k + l_k * m^2 + m^3 + n_k * l_k^2 + n_k^2 * l_k + n_k^2 * m + n_k * m^2)$$
$$\nabla_W B_k. : O(d * n_k^2 * p + d^2 * n_k * p + d^2 * m * p)$$

In our experiment we fix m and d as constants, and p can also be regarded as a constant. Additionally, we make an approximation for l_k by $l_k \approx n_k/\omega$, where ω is a

constant. Then, the time complexity of our framework is $O(n_k^3)$, where n_k is the number of photos in POI k.

5. CONCLUSIONS AND FUTURE WORK

In this article, we propose an efficient POI-based parallel tag matrix completion method for social image tagging and retrieval. By using geo-location information, we exploit clustering results as auxiliary clustering labels to make the framework easily processed in parallel. Then, by using image-user relationship, we introduce a localized prior constraint term to improve the performance for tag prediction. In order to evaluate our method, we conduct experiments on two applications: automatic image annotation and tag-based image retrieval. Extensive experiments on two subsets of a new large-scale social image dataset illustrate that the proposed method not only achieves better accuracy for automatic image annotation than the state-of-the-art method, but also enhances the computational efficiency. In future work, we combine our method with stream clustering techniques to handle streaming social images according to real application scenario. And we would like to improve our method to handle tag vocabularies from more domains beyond the geographical one.

REFERENCES

Kobus Barnard, Pinar Duygulu, David Forsyth, Nando De Freitas, David M. Blei, and Michael I. Jordan. 2003. Matching words and pictures. *Journal of Machine Learning Research* 3 (2003), 1107–1135.

Gustavo Carneiro, Antoni B. Chan, Pedro J. Moreno, and Nuno Vasconcelos. 2007. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 3 (2007), 394–410.

Coralia Cartis, Nicholas I. M. Gould, and Philippe L. Toint. 2011. On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *SIAM Journal on Optimization* 21, 4 (2011), 1721–1739.

Lin Chen, Dong Xu, Ivor W. Tsang, and Jiebo Luo. 2010. Tag-based web photo retrieval improved by batch mode re-tagging. In *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*. IEEE, 3440–3446.

Lin Chen, Dong Xu, Ivor W. Tsang, and Jiebo Luo. 2012. Tag-based image retrieval improved by augmented features and group-based refinement. *IEEE Transactions on Multimedia* 14, 4 (2012), 1057–1067.

David J. Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. 2009. Mapping the world's photos. In *Proceedings of the 18th International Conference on World Wide Web*. ACM, 761–770.

Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. 2010. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the 4th ACM Conference on Recommender Systems*. ACM, 39–46.

Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S Mirrokni. 2004. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the Twentieth Annual Symposium on Computational Geometry*. ACM, 253–262.

Chaitanya Desai, Deva Ramanan, and Charless C. Fowlkes. 2011. Discriminative models for multi-class object layout. *International Journal of Computer Vision* 95, 1 (2011), 1–12.

Yue Gao, Meng Wang, Huanbo Luan, Jialie Shen, Shuicheng Yan, and Dacheng Tao. 2011. Tag-based social image search with visual-text joint hypergraph learning. In *Proceedings of the 19th ACM International Conference on Multimedia*. ACM, 1517–1520.

K.-S. Goh, Edward Y. Chang, and Beitao Li. 2005. Using one-class and two-class SVMs for multiclass image annotation. *IEEE Transactions on Knowledge and Data Engineering* 17, 10 (2005), 1333–1346.

Andrew Goldberg, Ben Recht, Junming Xu, Robert Nowak, and Xiaojin Zhu. 2010. Transduction with matrix completion: Three birds with one stone. In *Advances in Neural Information Processing Systems*. 757–765.

Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid. 2009. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision*. IEEE, 309–316.

Harry Halpin, Valentin Robu, and Hana Shepherd. 2007. The complex dynamics of collaborative tagging. In *Proceedings of the 16th International Conference on World Wide Web*. ACM, 211–220.

Bharath Hariharan, Lihi Zelnik-Manor, Manik Varma, and Svn Vishwanathan. 2010. Large scale max-margin multi-label classification with priors. In *Proceedings of the 27th International Conference on Machine Learning (ICML'10)*. 423–430.

Choochart Haruechaiyasak and Chaianun Damrongrat. 2010. *Improving Social Tag-based Image Retrieval with CBIR Technique*. Springer.

Yang Hu, Mingjing Li, and Nenghai Yu. 2008. Multiple-instance ranking: Learning to rank images for image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*. IEEE, 1–8.

Rongrong Ji, Ling-Yu Duan, Jie Chen, Tiejun Huang, and Wen Gao. 2014. Mining compact bag-of-patterns for low bit rate mobile visual search. *IEEE Transactions on Image Processing* 23, 7 (2014), 3099–3113.

Rongrong Ji, Yue Gao, Wei Liu, Xing Xie, Qi Tian, and Xuelong Li. 2015. When location meets social multimedia: A survey on vision-based recognition and mining for geo-social multimedia analytics. *ACM Transactions on Intelligent Systems and Technology (TIST)* 6, 1 (2015), 1.

Yu-Gang Jiang, Chong-Wah Ngo, and Jun Yang. 2007. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*. ACM, 494–501.

Margaret EI Kipp and D. Grant Campbell. 2006. Patterns and inconsistencies in collaborative tagging systems: An examination of tagging practices. *Proceedings of the American Society for Information Science and Technology* 43, 1 (2006), 1–18.

Noam Koenigstein, Parikshit Ram, and Yuval Shavitt. 2012. Efficient retrieval of recommendations in a matrix factorization framework. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. ACM, 535–544.

Sihyoung Lee, Wesley De Neve, and Yong Man Ro. 2014. Visually weighted neighbor voting for image tag relevance learning. *Multimedia Tools and Applications* 72, 2 (2014), 1363–1386.

Xirong Li, Cees G. M. Snoek, and Marcel Worring. 2009. Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia* 11, 7 (2009), 1310–1322.

Xirong Li, Tiberio Uricchio, Lamberto Ballan, Marco Bertini, Cees GM Snoek, and Alberto Del Bimbo. 2016. Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval. *ACM Computing Surveys (CSUR)* 49, 1 (2016), 14.

Zijia Lin, Guiguang Ding, Mingqing Hu, Jianmin Wang, and Xiaojun Ye. 2013. Image tag completion via image-specific and tag-specific linear sparse reconstructions. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1618–1625.

Dong Liu, Xian-Sheng Hua, Meng Wang, and HongJiang Zhang. 2009a. Boost search relevance for tag-based social image retrieval. In *IEEE International Conference on Multimedia and Expo*. IEEE, 1636–1639.

Dong Liu, Xian-Sheng Hua, Linjun Yang, Meng Wang, and Hong-Jiang Zhang. 2009b. Tag ranking. In *Proceedings of the 18th International Conference on World Wide Web*. ACM, 351–360.

Dong Liu, Shuicheng Yan, Xian-Sheng Hua, and Hong-Jiang Zhang. 2011. Image retagging using collaborative tag propagation. *IEEE Transactions on Multimedia* 13, 4 (2011), 702–712.

Jing Liu, Zechao Li, Jinhui Tang, Yu Jiang, and Hanqing Lu. 2014. Personalized geo-specific tag recommendation for photos on social websites. *IEEE Transactions on Multimedia* 16, 3 (2014), 588–600.

Siyuan Liu, Yunhuai Liu, Lionel M. Ni, Jianping Fan, and Minglu Li. 2010. Towards mobility-based clustering. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 919–928.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. 2010. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research* 11 (2010), 19–60.

Deepshikha Mishra, Uday Pratap Singh, and Vineet Richhariya. 2014. Tag relevance for social image retrieval in accordance with neighbor voting algorithm. *International Journal of Computer Science and Network Security (IJCSNS)* 14, 7 (2014), 50.

Emily Moxley, Jim Kleban, and B. S. Manjunath. 2008. Spirittagger: A geo-aware tag suggestion tool mined from Flickr. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*. ACM, 24–30.

X. Qian, X. S. Hua, Y. Y. Tang, and T. Mei. 2014. Social image tagging with diverse semantics. *IEEE Transactions on Cybernetics* 44, 12 (2014), 2493–2508.

Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. 2008. LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision* 77, 1–3 (2008), 157–173.

Jitao Sang, Changsheng Xu, and Jing Liu. 2012. User-aware image tag refinement via ternary semantic analysis. *IEEE Transactions on Multimedia* 14, 3 (2012), 883–895.

Anshumali Shrivastava and Ping Li. 2014. Asymmetric LSH (ALSH) for sublinear time maximum inner product search (MIPS). In *Advances in Neural Information Processing Systems*. 2321–2329.

Andrea Vedaldi and Brian Fulkerson. 2010. VLFeat: An open and portable library of computer vision algorithms. In *Proceedings of the 18th ACM International Conference on Multimedia*. ACM, 1469–1472.

Zhi Wang, Wenwu Zhu, Peng Cui, Lifeng Sun, and Shiqiang Yang. 2013. Social media recommendation. In *Social Media Retrieval*. Springer, 23–42.

Shikui Wei, Dong Xu, Xuelong Li, and Yao Zhao. 2013. Joint optimization toward effective and efficient image search. *IEEE Transactions on Cybernetics* 43, 6 (2013), 2216–2227.

Shikui Wei, Yao Zhao, Ce Zhu, Changsheng Xu, and Zhenfeng Zhu. 2011. Frame fusion for video copy detection. *IEEE Transactions on Circuits and Systems for Video Technology* 21, 1 (2011), 15–28.

Shikui Wei, Yao Zhao, Zhenfeng Zhu, and Nan Liu. 2010. Multimodal fusion for video search reranking. *IEEE Transactions on Knowledge and Data Engineering* 22, 8 (2010), 1191–1199.

Di Wu, Jun Wu, Ming-Yu Lu, and Chun-Li Wang. 2014. A two-step similarity ranking scheme for image retrieval. In *Proceedings of the 2014 6th International Symposium on Parallel Architectures, Algorithms and Programming (PAAP'14)*. IEEE, 191–196.

Lei Wu, Xian-Sheng Hua, Nenghai Yu, Wei-Ying Ma, and Shipeng Li. 2008. Flickr distance. In *Proceedings of the 16th ACM International Conference on Multimedia*. ACM, 31–40.

Lei Wu, Rong Jin, and Anil K. Jain. 2013. Tag completion for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 3 (2013), 716–727.

Lei Wu, Linjun Yang, Nenghai Yu, and Xian-Sheng Hua. 2009. Learning to tag. In *Proceedings of the 18th International Conference on World Wide Web*. ACM, 361–370.

Hao Xu, Jingdong Wang, Xian-Sheng Hua, and Shipeng Li. 2009. Tag refinement by regularized LDA. In *Proceedings of the 17th ACM International Conference on Multimedia*. ACM, 573–576.

Yang Yang, Yue Gao, Hanwang Zhang, Jie Shao, and Tat-Seng Chua. 2014. Image tagging with social assistance. In *Proceedings of International Conference on Multimedia Retrieval*. ACM, 81.

J. Yu, D. Tao, M. Wang, and Y. Rui. 2015. Learning to rank using user clicks and visual features for image retrieval. *IEEE Transactions on Cybernetics* 45, 4 (2015), 767–779.

Zheng-Jun Zha, Tao Mei, Jingdong Wang, Zengfu Wang, and Xian-Sheng Hua. 2009. Graph-based semi-supervised learning with multiple labels. *Journal of Visual Communication and Image Representation* 20, 2 (2009), 97–103.

Jiangchuan Zheng, Siyuan Liu, and Lionel M. Ni. 2014. User characterization from geographic topic analysis in online social media. In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'14)*. IEEE, 464–471.

Ning Zhou, William K. Cheung, Guoping Qiu, and Xiangyang Xue. 2011. A hybrid probabilistic model for unified collaborative and content-based image tagging. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 7 (2011), 1281–1294.

Guangyu Zhu, Shuicheng Yan, and Yi Ma. 2010. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *Proceedings of the 18th ACM International Conference on Multimedia*. ACM, 461–470.

Received July 2015; revised February 2016; accepted September 2016