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

Jiaming Zhang[†], Shuhui Wang[†], Qingming Huang[*†]
[†]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
[*]University of Chinese Academy of Sciences, Beijing, 100049, China
jiaming.zhang@vipl.ict.ac.cn wangshuhui@ict.ac.cn qmhuang@ucas.ac.cn

ABSTRACT

Benefit from tremendous growth of user-generated content, social annotated tags get higher importance in organization and retrieval of large scale image database 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, existing studies only utilize the relations between visual content and tags to construct global information completion models. In this paper, 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 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. 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 database.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

ICMR'15, June 23–26, 2015, Shanghai, China.

DOI: http://dx.doi.org/10.1145/2671188.2749353.

General Terms

Algorithms, Experimentation, Performance

Keywords

tag matrix completion, geo-location information, social image retrieval

1. INTRODUCTION

The Online Sharing Websites(OSW), such as Flickr[1] and Panoramio[2], have experienced vigorous evolution in Web 2.0 era. Benefit 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 towards the need of efficient large scale online image retrieval.

However, according to the principle of least effort [10, 13], 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 certain level of incompleteness and noise existing in the manually annotated tags of the ever-growing images on OSW. Therefore, it gives rise to a challenging research problem 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 [7, 1, 12], which formulate the problem with a standard multi-class classification [2] or multi-label classification [11, 29], and the missing tags are obtained via image annotation process [31, 2]. 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, 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 [28, 32, 21, 15] and enhancing the number of informative tags [4, 25] by modeling the relation between visual content and tags. Generally, the tag refinement and completion can be achieved by information averaging [14, 4] and latent factor learning [25]. For example, neighborhoods [14] and

[1]www.flickr.com
[2]www.panoramio.com

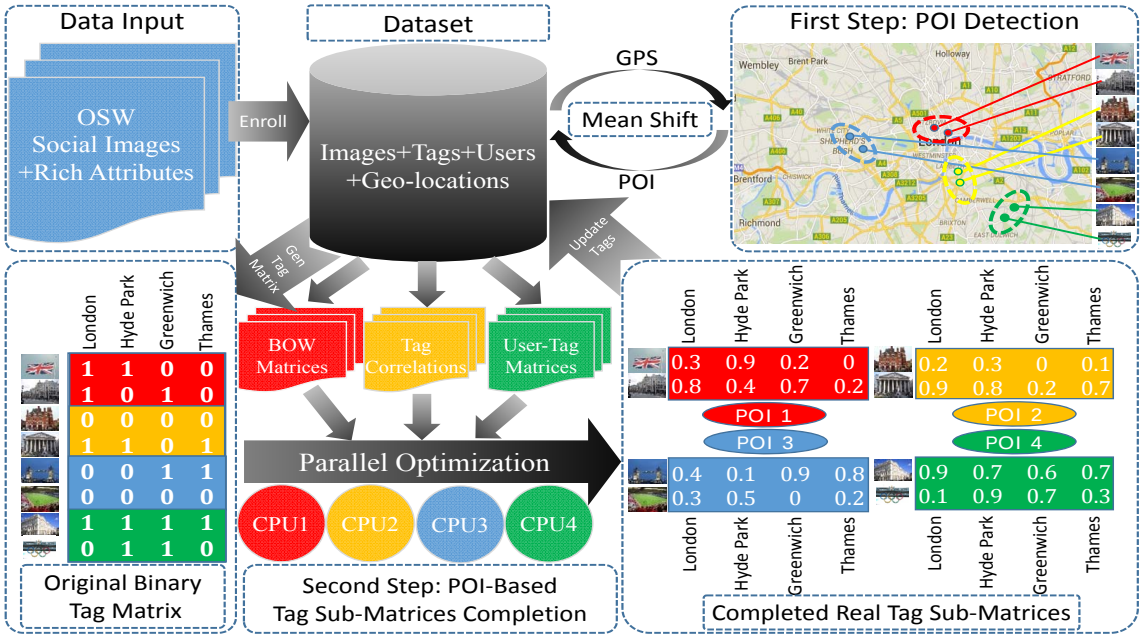


Figure 1: Framework of Location-Based Parallel Tag Completion

graphs are usually exploited, and the tagging information from visually similar social images is borrowed to construct probabilistic description on the uncertainty of the social tags. Wu et al. [25] propose an *image-tag matrix* completion framework based on 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 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 diversified backgrounds and preference styles of online users also lead to complementary expressions in social tags to 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 towards the real world applications.

In this paper, 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 Interests(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 problem 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. 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 paper are summarized as below:

- 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 learning social tags.
- 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.
- 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 performance promotion in automatic image annotation.
- Experimental results demonstrate that our approach achieves higher performance in social tag completion on real world social media images.

The rest of this paper 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

Many algorithms for automatic image annotation have been proposed in the past decades. Both global [7] and local visual features [1, 12] are taken into account for feasible solutions of image annotation. Moreover, several recent works [26, 20] focus on spatial structure of visual content for performance promotion. 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.

Besides general content-based image annotation techniques, many recent works exploit multilabel learning techniques to deal with image annotation as multilabel classification problem. Desai et al. [6] introduce a discriminative model in multilabel learning. Hariharan et al. [11] combine Support Vector Machine with multilabel learning to manage large scale data collection. Zha et al. [29] propose a graph-based multilabel learning approach for image annotation. These multilabel 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 does not match the requirement of multilabel learning approaches.

Meanwhile, several researchers choose other tag-based approaches to solve image annotation problem, such as image retagging, tag recommendation, tag propagation, etc. Li et al. [14] propose a neighbor voting method for social tagging. Guillaumin et al. [9] propose a tag propagation (TagProp) method to transfer tags through a weighted nearest neighbor graph. Chen et al. [4] propose an image retagging approach processing in batch-mode. Liu et al. [16] propose a graph-based algorithm including both tag-specific visual similarity graphs and tag semantic similarity graph to handle 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.

In recent years, matrix completion techniques are also introduced to address poor initial image annotation problem, getting brilliant experiment results. Goldberg et al. [8] propose a matrix completion for transductive classification. Wu et al. [25] 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 completed and observed tag matrix.

Besides numerous research works on image annotation, there are many research works focusing on combining geographical attributes and visual content [19, 5, 18, 17, 30]. Moxley et al. [19] adopt 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. [5] formularize the image location estimation as a classification problem by classifying images into POI categories. Liu et al. [17] propose a unified framework using subspace learning in personalized and geo-specific tag recommendation for images on OSW.

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×m}, T_k^1 = T̂_k^1, t = 1;
while ||L^{t+1} − L^t|| ≤ εL^t do
    Step size δ_t = δ_0/t;
    {The loop below is executed in parallel}
    for k = 1 to p do
        Calculate T̄_k^{t+1} : Equation 13
        Update T_k^{t+1}: Equation 17
        t = t + 1;
    end for
    Calculate W̄^{t+1} : Equation 14
    Update W^{t+1}: Equation 18
end while
return T_k^t, k ∈ {1, ..., p};
```

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 mean shift procedure, which aims at matrix partition preprocessing. The second step is the proposed POI-based tag matrix completion model. Detail 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 abundant annotated tags, how to 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 i-th 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 r-th 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 i-th 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 i-th column vector of $\widehat{T}$, as in [17].

3.2 Finding POIs using Mean Shift

In general, large value of n makes the original tag matrix T very large. Therefore, the computational burden for di-

Table 1: 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	771099	16225	47.52	1000	5.2	63	4016.4	481957
New York	1,210,094	732555	15344	47.74	1000	5.5	71	4031.5	299867

rectly handling such large matrices is costly and prohibitive. As discussed in previous section, geographically adjacent images may have similar visual content or semantic information with higher probability. Therefore, we utilize the geolocations 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 Point of Interests (POI). Therefore, we apply 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 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 POI center and the i-th 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 submatrices 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 $\widehat{T}_k \in \mathbb{R}^{n_k \times m}$ $(k \in \{1, 2, \cdots, p\}, \sum_{k=1}^{p} n_k = n)$. For example, image 1, 2 belong to user I, image 3, 4, 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 to tags for semantic representation of 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 tend 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 r_k images lack of user information in POI k by assigning an anonymous user to each image.

Similar as the i-th row of T_k depicting the actual tag distribution of image i, the i-th 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 term compare similarities differences among different images.

The second type is tag-wise constraint terms. Tag co-occurrence is proved to be effective in image tagging [24]. 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 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 regularization terms. To avoid dense solution of T_k, we require that only a small number of entries of T_k are nonzero, i.e. several unique tags are attached to each image. As studied in many sparse coding literatures, we consider to introduce an ℓ_1-norm regularization term $\|T_k\|_1$ for a sparse solution of T_k. For the shared

Table 2: Performance comparison about MAP for Automatic Image Annotation

London	MAP@5			MAP@10			MAP@15			MAP@20		
	TMC	PTC-U	PTC	TMC	PTC-U	PTC	TMC	PTC-U	PTC	TMC	PTC-U	PTC
r=1	82.37	83.15	85.68	73.42	74.32	78.12	70.45	71.27	73.98	66.87	67.55	71.18
r=2	82.49	83.78	87.50	73.61	75.74	79.70	71.74	71.89	75.22	67.93	68.86	72.47
r=3	82.79	84.03	88.17	73.57	76.57	80.28	69.60	72.43	75.75	67.61	69.46	73.04
r=4	82.87	84.36	88.64	73.65	76.28	80.67	69.62	72.39	76.12	67.75	69.53	73.38
r=5	83.61	85.67	**89.01**	74.60	77.71	**81.75**	71.53	73.96	**77.54**	68.49	70.68	**74.78**
New York	MAP@5			MAP@10			MAP@15			MAP@20		
	TMC	PTC-U	PTC	TMC	PTC-U	PTC	TMC	PTC-U	PTC	TMC	PTC-U	PTC
r=1	77.13	78.06	83.39	68.70	71.79	77.21	65.07	67.62	73.56	62.48	64.30	70.36
r=2	77.41	78.38	85.58	68.90	71.99	77.57	65.09	68.01	72.81	62.81	64.34	69.97
r=3	77.68	78.66	87.33	68.85	71.74	78.52	65.07	67.94	73.90	62.92	64.39	**70.92**
r=4	77.60	78.70	86.84	68.38	71.72	78.84	65.15	67.77	**74.07**	63.00	64.21	70.90
r=5	78.24	79.35	**87.50**	68.90	71.78	**79.17**	65.11	67.73	73.52	62.98	64.27	70.46

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 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-convex. While subgradient descent approach is one of the commonly used iterative methods dealing with non-convex optimization problems. Its remarkable less calculation time per iteration makes it more practical in processing large scale image datasets. So we adopt subgradient descent approach to solve the non-convex 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 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 [3]. In particular, we construct an auxiliary function as follows:

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

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

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

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:

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

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

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

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

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

while δ_t is the step size.

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

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

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

Combined with the immediate solutions, we obtain the solution as follows:

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

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

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 progress 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 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.

4. EXPERIMENTS

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

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 1 shows some statistical information about this dataset.

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

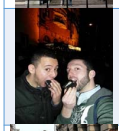
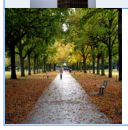
Photos	Ground Truth	TMC	PTC-U	PTC
	london united kingdom westminster big ben parliament palace	london united kingdom biorhythms big ben palace silhouette	united kingdom westminster big ben southbank parliament palace	london united kingdom westminster big ben parliament palace
	london england united kingdom 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

Figure 2: Examples of image annotation results by different methods

Users' tagging behavior in specific POI may become uncertain as the geographical scope goes larger. According to '*landmark-scale*' POI defined in [5], 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 subset from this large dataset by geographical restriction. We choose two famous international metropolises, *London* and *New York* (also used as the name of the subset), 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 totally obtain 1,026,345 and 924,707 images for *London* and *New York*, respectively. As studied in [27], 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 its number of annotated images and select the top 1000 to serve as the vocabulary in experiment. After this operation, the size of *London* shrinkages to 771,099, and *New York* to 732,555 respectively. The second and third row in Table 1 show some statistical information about *London* and *New York*.

For both *London* and *New York*, we extract dense SIFT [22] descriptor as local visual feature. Then we cluster randomly chosen 1,000,000 descriptor samples into 1000 visual words. Each local feature descriptor is quantized to one of these 1,000 visual words for Bag-Of-Words representation. After cross-validation procedure, we determine that $\alpha = 100, \beta = 10, \gamma = 1, \lambda = 1, \eta = 1$. For the TMC method, we adopt the parameter settings reported in their paper.

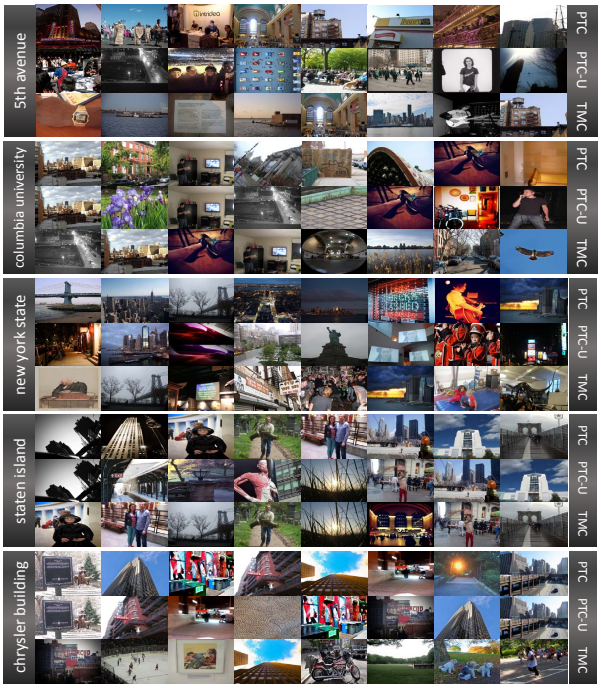


Figure 3: Illustration of some examples in TBIR with single tag queries

The initial step size δ_0 is set as 10^{-6} according to experience.

4.2 Comparison Methods

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

- **Tag Matrix Completion (TMC)** [25], 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.
- **PTC**, our method containing both user-related prior constraint term in loss function and POI-based matrix partition strategy for parallel processing.
- **PTC-U**, our proposed method without user-related prior constraint term, corresponding to the case of $\alpha = 0$ in Eq. 8.

4.3 Automatic Image Annotation

Given a query image q in automatic image annotation task, we simply rank all the tags in descending order of their probability scores attached to image q, corresponding to the q-th 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 i-th row of the observed tag matrix in 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) at top s ($s \in \{5, 10, 15, 20\}$) of completed tags to measure the performance of different algorithms.

As shown in Table 2, the annotation accuracy goes up along with the increase of the number of initial tags for all

Table 3: Performance of Tag-Based Image Retrieval with Single-Tag Queries

London	MAP@5			MAP@10			MAP@15			MAP@20		
	TMC	PTC-U	PTC	TMC	PTC-U	PTC	TMC	PTC-U	PTC	TMC	PTC-U	PTC
r=1	64.03	64.03	64.03	63.89	64.02	64.03	63.54	63.84	64.00	63.42	63.77	63.88
r=2	72.92	72.93	72.94	72.84	72.93	72.94	72.63	72.74	72.87	72.62	72.68	72.78
r=3	80.86	80.86	80.86	80.77	80.82	80.86	80.49	80.65	80.79	80.39	80.61	80.78
r=4	87.75	87.79	87.79	87.60	87.71	87.79	87.36	87.64	87.75	87.12	87.59	87.75
r=5	90.21	90.24	90.26	90.08	90.17	90.25	89.80	90.11	90.23	89.61	90.03	90.20
New York	MAP@5			MAP@10			MAP@15			MAP@20		
	TMC	PTC-U	PTC	TMC	PTC-U	PTC	TMC	PTC-U	PTC	TMC	PTC-U	PTC
r=1	88.58	88.58	88.58	88.50	88.58	88.58	88.46	88.57	88.57	88.42	88.56	88.56
r=2	93.49	93.49	93.49	93.48	93.49	93.49	93.47	93.46	93.46	93.44	93.46	93.46
r=3	95.43	95.43	95.43	95.43	95.43	95.43	95.43	95.43	95.43	95.43	95.43	95.43
r=4	97.14	97.15	97.15	97.13	97.15	97.14	97.13	97.13	97.13	97.13	97.13	97.13
r=5	98.74	98.74	98.74	98.74	98.74	98.74	98.74	98.74	98.74	98.74	98.73	98.73

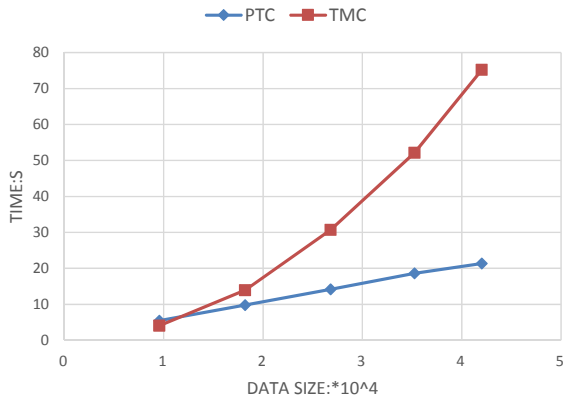


Figure 4: CPU Execution Time of different methods

methods in vertical comparison. It is in line with our expectation because more initial tags for each image means lower systemic risk. In horizontal comparison, we observe that the proposed PTC approach outperforms its weakening variant PTC-U method and 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 performance promotion. The matrix partition strategy makes locality consistency more compact in tag space. And the user-related prior constraint term makes coincidence with common sense. Figure 2 shows several annotation results selected from *London* in condition of $e = 5$.

4.4 Tag-Based Image Retrieval

In tag-based image retrieval task, we consider a simple scenario that the query is a single tag. We rank all of the gallery images according to their relevance scores to the given query tag in descending order as the retrieval results. The relevance score of each image to the given query tag is represented by the corresponding column in the completed tag matrix T. 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 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 [23] [27] we adopt in the experiment is that: an image is relevant if its annotation contains the query.

Table 3 shows the MAP at top s ($s \in \{5, 10, 15, 20\}$) 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. It means that our method has little promotion compared with the TMC method. We attribute this phenomenon to poor original annotation which we used as ground truth.

However, we observe that the MAP value of our PTC method decreases more slowly than the TMC method as the number of recall results goes up. It reveals that our proposed method is more robust than the TMC method to a certain extent. Figure 3 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 make 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 run on the Intel (R) Core (TM) i7-4770K CPU @3.50 GHz and 32 GB RAM PC. Figure 4 reveals the running time per iteration of both PTC and TMC method. 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 matrix partition strategy is the key point of efficiency improvement.

5. CONCLUSIONS AND FUTURE WORK

In this paper, 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 reality scenario. And we would like to improve our method to handle variant tag vocabulary.

6. ACKNOWLEDGMENTS

This work was supported in part by National Basic Research Program of China (973 Program) : 2012CB316400 and 2015CB351802, National Natural Science Foundation of China : 61025011, 61332016, 61390511 and 61303160, 863 program of China : 2014AA015202, Postdoctoral Science Foundation of China : 2014T70126, and Basic Research Program of Shenzhen: JCYJ20140610152828686.

7. REFERENCES

[1] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[2] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(3):394–410, 2007.

[3] C. Cartis, N. I. Gould, and P. L. Toint. On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *SIAM Journal on Optimization*, 21(4):1721–1739, 2011.

[4] L. Chen, D. Xu, I. W. Tsang, and J. Luo. Tag-based web photo retrieval improved by batch mode re-tagging. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3440–3446. IEEE, 2010.

[5] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *WWW'09*, pages 761–770. ACM, 2009.

[6] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. *International Journal of Computer Vision*, 95(1):1–12, 2011.

[7] K.-S. Goh, E. Y. Chang, and B. Li. Using one-class and two-class svms for multiclass image annotation. *Knowledge and Data Engineering, IEEE Transactions on*, 17(10):1333–1346, 2005.

[8] A. Goldberg, B. Recht, J. Xu, R. Nowak, and X. Zhu. Transduction with matrix completion: Three birds with one stone. In *Advances in neural information processing systems*, pages 757–765, 2010.

[9] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 309–316. IEEE, 2009.

[10] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *Proceedings of the 16th international conference on World Wide Web*, pages 211–220. ACM, 2007.

[11] B. Hariharan, L. Zelnik-Manor, M. Varma, and S. Vishwanathan. Large scale max-margin multi-label classification with priors. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 423–430, 2010.

[12] Y.-G. Jiang, C.-W. Ngo, and J. Yang. 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*, pages 494–501. ACM, 2007.

[13] M. E. Kipp and D. G. Campbell. Patterns and inconsistencies in collaborative tagging systems: An examination of tagging practices. *Proceedings of the American Society for Information Science and Technology*, 43(1):1–18, 2006.

[14] X. Li, C. G. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *Multimedia, IEEE Transactions on*, 11(7):1310–1322, 2009.

[15] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. Snoek, and A. Del Bimbo. Socializing the semantic gap: A comparative survey on image tag assignment, refinement and retrieval. *arXiv preprint arXiv:1503.08248*, 2015.

[16] D. Liu, S. Yan, X.-S. Hua, and H.-J. Zhang. Image retagging using collaborative tag propagation. *Multimedia, IEEE Transactions on*, 13(4):702–712, 2011.

[17] J. Liu, Z. Li, J. Tang, Y. Jiang, and H. Lu. Personalized geo-specific tag recommendation for photos on social websites. *IEEE Transactions on Multimedia*, 16(3):588–600, 2014.

[18] S. Liu, Y. Liu, L. M. Ni, J. Fan, and M. Li. Towards mobility-based clustering. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 919–928. ACM, 2010.

[19] E. Moxley, J. Kleban, and B. Manjunath. Spirittagger: a geo-aware tag suggestion tool mined from flickr. In *ACM MIR*, pages 24–30. ACM, 2008.

[20] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173, 2008.

[21] J. Sang, C. Xu, and J. Liu. User-aware image tag refinement via ternary semantic analysis. *Multimedia, IEEE Transactions on*, 14(3):883–895, 2012.

[22] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *ACM Multimedia*, pages 1469–1472. ACM, 2010.

[23] L. Wu, S. C. Hoi, R. Jin, J. Zhu, and N. Yu. Distance metric learning from uncertain side information with application to automated photo tagging. In *ACM Multimedia*, pages 135–144. ACM, 2009.

[24] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li. Flickr distance. In *ACM Multimedia'08*, pages 31–40. ACM, 2008.

[25] L. Wu, R. Jin, and A. K. Jain. Tag completion for image retrieval. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(3):716–727, 2013.

[26] L. Wu, M. Li, Z. Li, W.-Y. Ma, and N. Yu. Visual language modeling for image classification. In *Proceedings of the International Workshop on MIR*, pages 115–124. ACM, 2007.

[27] L. Wu, L. Yang, N. Yu, and X.-S. Hua. Learning to tag. In *WWW'09*, pages 361–370. ACM, 2009.

[28] H. Xu, J. Wang, X.-S. Hua, and S. Li. Tag refinement by regularized lda. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 573–576. ACM, 2009.

[29] Z.-J. Zha, T. Mei, J. Wang, Z. Wang, and X.-S. Hua. Graph-based semi-supervised learning with multiple labels. *Journal of Visual Communication and Image Representation*, 20(2):97–103, 2009.

[30] J. Zheng, S. Liu, and L. M. Ni. User characterization from geographic topic analysis in online social media. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 464–471. IEEE, 2014.

[31] N. Zhou, W. K. Cheung, G. Qiu, and X. Xue. A hybrid probabilistic model for unified collaborative and content-based image tagging. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(7):1281–1294, 2011.

[32] G. Zhu, S. Yan, and Y. Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *Proceedings of the international conference on Multimedia*, pages 461–470. ACM, 2010.