Abstract—Social tagging becomes increasingly important to organize and search large-scale community-contributed photos on social websites. To facilitate generating high-quality social tags, tag recommendation by automatically assigning relevant tags to photos draws particular research interest. In this paper, we focus on the personalized tag recommendation task and try to identify user-preferred, geo-location-specific as well as semantically relevant tags for a photo by leveraging rich contexts of the freely available community-contributed photos. For users and geo-locations, we assume they have different preferred tags assigned to a photo, and propose a subspace learning method to individually uncover the both types of preferences. The goal of our work is to learn a unified subspace shared by the visual and textual domains to make visual features and textual information of photos comparable. Considering the visual feature is a lower level representation on semantics than the textual information, we adopt a progressive learning strategy by additionally introducing an intermediate subspace for the visual domain, and expect it to have consistent local structure with the textual space. Accordingly, the unified subspace is mapped from the intermediate subspace and the textual space respectively. We formulate the above learning problems into a united form, and present an iterative optimization with its convergence proof. Given an untagged photo with its geo-location to a user, the user-preferred and the geo-location-specific tags are found by the nearest neighbor search in the corresponding unified spaces. Then we combine the obtained tags and the visual appearance of the photo to discover the semantically and visually related photos, among which the most frequent tags are used as the recommended tags. Experiments on a large-scale data set collected from Flickr verify the effectiveness of the proposed solution.

Index Terms—Personalized tag recommendation, Subspace Learning, Tagging history, User preference, Geo-location preference

I. INTRODUCTION

Due to the rapid popularization of GPS-enable camera devices and mobile phones, recent years have witnessed an explosive growth of personal photos with rich context like tags, geo-locations\(^1\) and visual attributes (colors and textures) \([1]\). Furthermore, many photo-sharing websites, such as Flickr, Corbis and Picasa, facilitate millions of users to upload and share their personal multimedia data by their smart phones or other internet access devices. As a consequence, the volume of community-contributed photos increases drastically whether on personal devices or on the social websites. It is challenging and promising to exploit the overwhelming amount of context data for multimedia applications, such as retrieval, annotation and recommendation \([2]\).

Among these applications, assigning proper tags to photos is the crucial task. Obviously, fully manual tag assignment is very time-consuming and impractical due to the massive photos and the limited screen size of the mobile devices. To make it easier, tag recommendation methods \([3]\), \([4]\), \([5]\), \([6]\), \([7]\) are proposed to suggest some relevant tags to a given photo and allow users to select their preferred tags, which cannot only ease the burden for users to upload and share their photos on social website, but facilitate users to tag and organize their personal images on mobile devices. However, most work attempt to learn the association between tags and photos, while the user preferences are ignored in the recommendation. Users have personal preferences for photos, which can be observed by the following two aspects. First, users favor different types of photos. For instance, some prefer the ‘architectures’ photos, while others are in favor of the ‘natural landscape’ ones. Second, different users have different favorite tags. In other words, if very similar photos would be tagged by two users, different tags may be produced, as shown in Fig. 1 (a).

Fig. 1: An illustration of preference of users and geo-locations. (a): Users have their specific preferences towards tags for similar photos; (b): Two visually similar photos corresponding to different geo-locations are tagged with different tags.

\(^1\)In this work, geo-location refers to the city in which each photo was taken. It can be derived according to its longitude and latitude by Flickr API.

Consequently, personalized tag recommendation can provide more appropriate recommendation results by taking user’s profile into account.
On the other hand, users are used to spend considerable effort to organize their photo albums geographically by describing photos with tags related to locations where they were taken [8], [9], [10]. Hence, the geographical information of photos should be explored in tag recommendation [11], [12], [13], [14], [15], [16]. Besides, some location specific tags (e.g., Eiffel Tower and Forbidden City) and location related tags (e.g., Paris and Beijing) are helpful to disambiguate some visually similar images. As shown in Fig. 1 (b), the two visually similar photos are possibly assigned to the same tags without considering the geographic information even if they are taken by the same user. Therefore, investigating the geolocation preference towards tags from this huge amount of context multimedia data can provide us useful information to recommend the most relevant tags to a given photo.

To address the above issues, we seek to develop a framework of personalized tag recommendation by jointly exploring the tagging resources and the geo-location information in social web context. We propose a subspace learning method to individually mine user preference from her tagging history and analyze geolocation preference towards tags based on the location related tagging resources. During the individual subspace learning process, given the tagging photos specific to a user (or a geo-location), we propose to uncover a common structure to link the visual and textual domains, i.e., a unified subspace shared by the both domains, in which visual features and textual representations of photos are comparable. Considering the visual feature is a much lower level representation on semantics than the textual information, we first map the visual features into an intermediate space, which is required to be structure consistent with the textual space. The introduction of the intermediate space for visual domain is to alleviate the semantic gap between the visual and textual domains. Then we attempt to learn a unified space mapping from the intermediate space and the textual space respectively, in which both the visual- and semantic-geometric structures should also be maintained. We integrate the above learning problems about the intermediate space and the unified space together into a united formulation, and propose an effective optimization algorithm followed with its convergence proof. With the learning outputs, we can map an untagged photo given a user (a geo-location) into the unified space corresponding to the user (the location), and use the nearest neighbor search to obtain some user-preferred tags and geo-location-preference tags individually. We further combine the obtained tags and the visual appearance of the photo to discover the semantically and visually related images, and explore the idea of annotation-by-search to rank tags for the untagged photo. Finally, the top ranked tags are recommended to the user. Experiments on a large-scale data set collected from Flickr demonstrate the outperformance of the proposed solution.

The main contributions of this paper are summarized as follows.

- To the best of our knowledge, it is the first research work to jointly explore the user preference and the geographic preference towards tags for tag recommendation.
- We propose to learn a unified space shared by the visual and textual domains, in which the relevant tags to an untagged photos can be easily obtained by the nearest neighbor search.
- An intermediate space between the visual space and the learned unified space is introduced to be analogous to the textual space with semantic structure, and this can alleviate the semantic gap between the visual and textual domains to some extents.
- The learning problems for the intermediate space and the unified space are formulated into a united form, and an iterative optimizing solution and its convergence proof are also provided.

The rest of this paper is structured as follows. In the next section, we review the related work about tag recommendation. The outline of our solution is introduced in Section III. The proposed methodology for learning user preference and geo-location preference towards tags is elaborated in Section IV, and Section V introduces how to suggest tags to new photos based on the learned results. Our experimental settings and result analysis are introduced in Section VI. Finally, conclusions with future work are drawn in Section VII.

II. RELATED WORK

Generic Tag Recommendation. Generic tag recommendation methods [5], [17], [3], [18], [19], [6], [20] are to predict the same list of tags for the same photo, i.e., it is independent of the user factor. Chen et al. [5] proposed an automatic tag recommendation approach that directly predicts the possible tags with models learned from training data. Shen et. al [3] proposed a multi-task structured SVM algorithm to leverage both the inter-object correlations and the loosely-tagged images. Images are annotated purely based on image visual content [21]. For an image, it first finds its top- $k$ neighboring images from the community image set and then selects the most frequent tags in the neighbor set as the annotated results. In [6], two approaches, based on Poisson Mixture Models and Gaussian process respectively, are proposed to make effective and efficient tag recommendations. In [7], tag concepts derived based on tag co-occurrence pairs are indexed as textual documents. The candidate tags associated with the matching concepts, which are retrieved with the query of user-given tags of an image, are recommended. There are some work focusing on tagging images by exploiting geo-tags [11], [12], [13], [14], [15]. A typical approach as introduced by Moxley et al. [11] and Kleban et al. [12] is to annotate a given image by constrained $k$ nearest neighbor ($k$-NN) voting, where the visual neighbors are retrieved from the geo region of the given image. The fundamental idea in [13] is to learn tag semantics, i.e., categorize tags as places, landmarks, and visual descriptors, in order to post-filter the tag results of tag suggestion. Silva et al. [14] annotated georeferenced photos with descriptive tags by exploring the redundancy over the large volume of annotations available at online repositories with other georeferenced photos. Geo context is fused with visual concept detection in a concept-dependent manner to improve visual search in [15]. However, the above methods ignore the user preference and suggest same tags to visually similar photos of different users. Different from them,
we propose a learning algorithm to effectively uncover user preference form her tagging history.

**Personalized Tag Recommendation.** Personalized tag recommendation has attracted significant attention recently. In [22], tag recommendation is obtained using both a Naive Bayes classifier on user tagging history and TF-IDF based global information. In [8], tag co-occurrence for photos is calculated using tags appearing both in the tagging history of a user and in Flickr website, and used to generate recommended tags. Web browsing behavior of a user is exploited to suggest the tags not only to be added to but also to be deleted from the original tags of a photo in Flickr [23]. In [24], image tag recommendation is formulated as a maximum a posteriori problem using a visual folksonomy. With the assumption that favorite images and their associated tags indicate the visual and topical interests of a user, personalized favorite images and their context are used to perform personalized image tag recommendation[25]. A simple personalized image annotation method is designed in [26], which simply annotates an untagged images with the most frequent tags in the user tagging history. Tensor decomposition models have been exploited for tag recommendation [27], [28], [29], [30], [31], [32]. Rendle et al. [29] propose a special case of the tucker decomposition model, pairwise interaction model, to predict the tag sets. MusicBox [30] tags music based on social tags by capturing the three-way correlations between users-tags-music items using three-order tensors. The low order tensor decomposition is proposed in [31], which include 0-th, 1-st, 2-nd order polynomials to reconstruct the data. In [32], a tensor factorization and tag clustering model is proposed to recommend items in social tagging systems, which can handle the problems of sparsity, cold start and learning problem for tag relevance. A personalized method using cross-entropy is proposed to annotate images [33], which personalizes a generic annotation model by learning from a user’s tagging history. However, the above methods only focus on photos, users and tags but ignore the geographical information of photos. Some other personalized tag recommendation methods generate candidate tags by exploiting geo-tags [16]. In [16], new photos are tagged using users’ own vocabularies by accumulating votes from the candidate images, which are selected in term of three factors: visual features, geographical coordinates and image taken time. Different from the above work, we propose a subspace learning approach to individually uncover user preference by exploiting user’s tagging history and geo-location preference by exploiting the geographic information of photos, and then jointly explore the learned subspaces assisted with the search scheme to recommend user preferred tags to a photo.

**III. OVERVIEW OF OUR SOLUTION**

This work focuses on how to ease personalized photo tagging process by exploiting the community-contributed multimedia data with rich contextual information. The proposed framework is illustrated in Fig. 2, which contains two primary parts: the offline and online processes.

The offline process is made up of three subdivisions: data collection, user preference learning and geo-location preference learning. For data collection, we collect a vast amount of photos with their tags, taggers, geo-locations and some relevant text information from Flickr. With the collected resources, we organize the photos according to different taggers (i.e., users) and geo-locations individually. Given a collection corresponding to a user (or a geo-location), we propose a new subspace learning approach to uncover the user’s preference (or the geo-location’s preference) towards tags. Our goal is to find a unified space for the both visual and textual domains, in which the visual features and the tagging information are comparable, i.e., the correlations between the both heterogeneous representations can be directly constructed. The details about the proposed approach and its optimization will be presented in Section IV.

In the online module (discussed in Section V), given a new photo with a specific user and a specific geo-location, we first find its top-ranked neighboring tags in the user-specific unified space and the geo-specific unified space individually, and combine the both sets of neighboring tags to generate the initial tags, by which semantically relevant photos are chosen from the community-contributed photo set. And then visually similar photos are found by implementing content-based photo retrieval from these semantically relevant photos. Finally, the most frequent tags in the semantically and visually related photos are recommended to the user.

**IV. DISCOVERING INTERMEDIATE SPACE AND UNIFIED SPACE**

In this section, we first present our motivation for user preference and geo-location preference learning in Section IV-A. The proposed algorithm is formulated in Section IV-B, followed by its optimization and convergence analysis in Section IV-C. Finally, we introduce how to represent tags in the learned unified space in Section IV-D.

**A. Motivation**

We address the personalized tag recommendation task with the help of overwhelming community-contributed information, such as user tag and geo-location. The heart is how to learn the latent correlation between visual features and tags for each user (or geo-location). However, there exists the well-known semantic gap making it challenging. Fortunately, the community-contributed photos are associated with rich text information, which can reduce this gap. The content of the text and the photos are highly correlated [34]. However, the text space and visual space have inherently different structures. To address this problem, it is crucial and necessary to discover a common structure to link them. On the other hand, the visual representation and the text representation of a photo should be consistent, i.e., corresponding representations should be same in the common structure. This naturally motivates an approach to discover a latent unified space, in which the corresponding text features and visual features of the same photo are identical. This can be achieved by two transformations from text features and visual feature respectively to a unified space. However, compared with the tag information, visual feature is a much lower level representation on semantics. To reduce the semantic gap, we adopt a progressive way (as shown in Fig.
Problem Formulation

For the ease of exposition, we first introduce the notations used throughout this work. Let \( u \) be a user for whom we aim to suggest tags and \( g \) be a geo-location. Our goal is to discover latent spaces of user \( u \) and geo-location \( g \) individually. In the following, we derive the unified space of user \( u \) in detail, while the unified space of geo-location \( g \) can be easily obtained in the same manner. Let \( \mathbf{X}_{u,te} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \) be a set of photos user \( u \) has already tagged and \( \mathbf{X}_{u,te} \) is a set of untagged photos to which we want to recommend tags. Here \( \mathbf{x}_i \in \mathbb{R}^d \) is the visual feature of the \( i \)-th photo. \( d \) is the dimension of visual feature space. We use \( w \) to denote a word in the text space and \( \mathcal{V} = \{w_1, w_2, \cdots, w_m\} \) for a large vocabulary. \( \mathbf{Y}_u = [y_1, y_2, \cdots, y_n] \in \mathbb{R}^{m \times n} \) is the text representation matrix of photos tagged by user \( u \), where \( (\mathbf{Y}_u)_{ij} \) is used to represent the relevance between tag \( w_i \) and photo \( \mathbf{x}_j \) assigned by user \( u \). We use the binary scheme to define \( \mathbf{Y}_u \) in our experiments, i.e., \( (\mathbf{Y}_u)_{ij} = 1 \) when photo \( \mathbf{x}_j \) is tagged with \( w_i \) and \( (\mathbf{Y}_u)_{ij} = 0 \) otherwise. Motivated by the above idea, the unified space of user \( u \) is discovered based on three linear transformation matrices \( \mathbf{W}_u \in \mathbb{R}^{q \times d} \), \( \mathbf{V}_u \in \mathbb{R}^{p \times q} \) and \( \mathbf{T}_u \in \mathbb{R}^{p \times n} \) as illustrated in Fig. 3. Here \( q \) and \( p \) are the dimensions of the intermediate space and unified space respectively. For brevity, we use \( \mathbf{X}, \mathbf{Y}, \mathbf{W}, \mathbf{V} \) and \( \mathbf{T} \) to denote \( \mathbf{X}_{u,te}, \mathbf{Y}_u, \mathbf{W}_u, \mathbf{V}_u \) and \( \mathbf{T}_u \) respectively from here on.

With transformation matrices \( \mathbf{W} \) and \( \mathbf{V} \), we can easily obtain the latent representations \( \mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_n] \in \mathbb{R}^{q \times n} \) and \( \mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n] \in \mathbb{R}^{p \times n} \) of the original visual features in the intermediate space and the unified space, respectively.

\[
\mathbf{G} = \mathbf{WX} \tag{1}
\]

\[
\mathbf{U} = \mathbf{VG} = \mathbf{VWX} \tag{2}
\]

Similarly, the latent representations \( \mathbf{Z} = [z_1, z_2, \cdots, z_n] \in \mathbb{R}^{p \times n} \) of the original text features are obtained by

\[
\mathbf{Z} = \mathbf{TY}. \tag{3}
\]

In the following, we first introduce our optimization problem without setting any canonical form for the loss functions. The specific forms will be explained with the proposed optimization algorithm in Section IV-C. To derive the effective transformation matrices, the unified space is expected to connect the visual structure and text structure. For this purpose,
we require that
\[ \mathbf{U} = \mathbf{Z} \Rightarrow \mathbf{VWX} = \mathbf{TY}. \] (4)

The above equation is over-determined and usually there is no exact solutions [35]. Therefore, we employ a cost function \( f \) as a measure of disagreement between them
\[ \min_{\mathbf{U}, \mathbf{Z}} f(\mathbf{U}, \mathbf{Z}). \] (5)

We also expect that the intermediate space is analogous to the text space with semantic structure. To this end, we design another cost function \( h \) to measure the semantic difference between the structures of the intermediate space \( \mathbf{G} \) and the text space \( \mathbf{Y} \) as follows.
\[ \min_{\mathbf{W}} h(\mathbf{G}, \mathbf{Y}) \] (6)

Considering the above two aspects simultaneously, we have the following optimization problem to learn the transformation matrices.
\[ \min_{\mathbf{W}, \mathbf{V}, \mathbf{T}} f(\mathbf{U}, \mathbf{Z}) + \alpha h(\mathbf{G}, \mathbf{Y}) + \lambda \Omega(\mathbf{W}, \mathbf{V}, \mathbf{T}) \] (7)

where the third term \( \Omega(\mathbf{W}, \mathbf{V}, \mathbf{T}) \) is the regularization term about these three transformations, which can avoid overfitting problem. \( \alpha \) and \( \lambda \) are the trade-off parameters, which balance the three terms in the objective function.

Besides, the above objective function overlooks the preservation of the geometric structures in the original visual and text spaces. Theoretically, geometric information can benefit the effective discovery of the latent space. Thereupon, it is necessary to add corresponding regularization terms in the objective function in Eq. 7. There are many strategies to preserve the local structures without loss of generality. Let \( \mathbf{S} \in \mathbb{R}^{n \times n} \) and \( \mathbf{C} \in \mathbb{R}^{n \times n} \) denote the similarity matrices in the original visual and text spaces, respectively. We take \( \mathbf{S} \) to sketch the details.

To preserve the geometric structure, we hope that visually similar photos should be located closely in the unified space. A reasonable criterion is to minimize the following objective function:
\[ \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} S_{ij} \left\| \frac{u_i}{D_{ii}^{1/2}} - \frac{u_j}{D_{jj}^{1/2}} \right\|^2 = \text{Tr}[(\mathbf{UL}^S \mathbf{U}^T], \] (8)

where \( D_{ii} = \sum_{j=1}^{n} S_{ij} \) is engaged for normalization purpose. \( \mathbf{D} = \text{diag}(D_{11}, D_{22}, \ldots, D_{nn}) \) is a diagonal matrix and \( \mathbf{L}^S = (\mathbf{D} - \mathbf{S})^{1/2} \) is the normalized Laplacian matrix. \( \text{Tr}[\cdot] \) denotes the trace operation. In the same manner, the corresponding regularization for local semantic geometry preservation is \( \text{Tr}[\mathbf{ZL}^C \mathbf{Z}^T] \). We adopt Gaussian kernel similarity and cosine similarity to calculate \( \mathbf{S} \) and \( \mathbf{C} \) respectively.

Combining the objective function in Eq.7 and the above two regularization terms for preserving local structures, we obtain the following optimization problem.
\[ \min_{\mathbf{W}, \mathbf{V}, \mathbf{T}} f(\mathbf{U}, \mathbf{Z}) + \alpha h(\mathbf{G}, \mathbf{Y}) + \beta \text{Tr}[\mathbf{UL}^S \mathbf{U}^T] + \gamma \text{Tr}[\mathbf{ZL}^C \mathbf{Z}^T] + \lambda \Omega(\mathbf{W}, \mathbf{V}, \mathbf{T}) \] (9)

Here \( \beta \) and \( \gamma \) are nonnegative trade-off parameters. Our goal is to discover the underlying unified space by predicting the three transformation matrices. It can be achieved by optimizing the above objective function, which will be discussed in the next subsection.

### C. Optimization

We first introduce the optimization algorithm for the above objective function and then analyze its convergence. First of all, we make use of the least square loss function \( f(a, b) = (a - b)^2 \) to measure the disagreement between \( \mathbf{U} \) and \( \mathbf{Z} \). For the function \( h \), in this work we only expect that the intermediate space has the same local geometric structure as the text space for simplicity. That is, the difference between \( \mathbf{G} \) and \( \mathbf{Y} \) is constrained by the Laplacian regularization:
\[ h(\mathbf{G}, \mathbf{Y}) = \text{Tr}(\mathbf{GL}^C \mathbf{G}^T), \] which is computed in the same manner as in Eq. 8. For \( \Omega(\mathbf{W}, \mathbf{V}, \mathbf{T}) \), we use the conventional squared norm to regularize \( \mathbf{V} \) and \( \mathbf{T} \) and the \( \ell_{2,1} \) norm to regularize \( \mathbf{W} \). That is,
\[ \Omega(\mathbf{W}, \mathbf{V}, \mathbf{T}) = \|\mathbf{W}\|_{2,1} + \|\mathbf{V}\|_2^2 + \|\mathbf{T}\|_2^2, \] (10)

in which \( \|\cdot\|_F \) denotes the Frobenius norm and the \( \ell_{2,1} \) norm is defined as \( \|\mathbf{W}\|_{2,1} = \sum_{i=1}^{q} \sqrt{\sum_{j=1}^{d} W_{ij}^2} \). The \( \ell_{2,1} \) norm regularization is utilized to use features across all data points with joint sparsity, i.e. each feature either has small scores or large scores over all data points [39], [40]. It is necessary for our task because some visual features are noisy and unhelpful.

Substituting the above expressions, the objective function of Eq. 9 can be rewritten as
\[ \min_{\mathbf{W}, \mathbf{V}, \mathbf{T}} \mathcal{L} = \|\mathbf{U} - \mathbf{Z}\|_2^2 + \alpha \text{Tr}(\mathbf{GL}^C \mathbf{G}^T) + \beta \text{Tr}[(\mathbf{UL}^S \mathbf{U}^T] + \gamma \text{Tr}[(\mathbf{ZL}^C \mathbf{Z}^T] + \lambda (\|\mathbf{W}\|_{2,1} + \|\mathbf{V}\|_2^2 + \|\mathbf{T}\|_2^2) \] (11)

The \( \ell_{2,1} \)-norm regularization term is non-smooth and the objective function is not convex over \( \mathbf{W}, \mathbf{V} \) and \( \mathbf{T} \) simultaneously.

To optimize the proposed objective function, we employ the gradient method to update these transformation matrices iteratively. The derivatives of \( \mathcal{L} \) with respect to \( \mathbf{W}, \mathbf{V}, \mathbf{T} \) can be calculated as follows:
\[ \frac{\partial \mathcal{L}}{\partial \mathbf{W}} = 2(\mathbf{V}^T(\mathbf{U} - \mathbf{Z})\mathbf{X} + \alpha \mathbf{GL}^C \mathbf{X}^T + \beta \mathbf{UL}^S \mathbf{X}^T + \lambda \mathbf{AW}) \] (12)
\[ \frac{\partial \mathcal{L}}{\partial \mathbf{V}} = 2(\mathbf{(U} - \mathbf{Z})\mathbf{G}^T + \beta \mathbf{UL}^S \mathbf{G}^T + \lambda \mathbf{V}) \] (13)
\[ \frac{\partial \mathcal{L}}{\partial \mathbf{T}} = 2(-\mathbf{W}^T\mathbf{Y}^T + \gamma \mathbf{ZL}^C \mathbf{Y}^T + \lambda \mathbf{T}) \] (14)

Here \( \mathbf{A} \in \mathbb{R}^{d \times q} \) is a diagonal matrix with \( A_{ii} = \frac{1}{2\|w_i\|_2^2} \), where \( w_i \) is the \( i \)-th row of \( \mathbf{W} \). Based on the above analysis, we summarize the detailed optimization algorithm in Algorithm 1. Next, the convergence of the proposed iterative procedure will be proved.

\(^2\)In practice, \( \|w_i\|_2 \) could be close to zero but not zero. Theoretically, it could be zeros. For this case, we can regularize \( A_{ii} = \frac{1}{2\|w_i\|_2^2 + \epsilon} \), where \( \epsilon \) is very small constant.
Algorithm 1 Transformations Discovery Algorithm

Input:
Visual Representation $X$ and Text Representation $Y$; Parameters $\alpha, \beta, \gamma, \lambda, p$ and $q$; The step length $\eta$.
1: Calculate $L^S$ and $L^C$;
2: Randomly initialize $W^t, V^t$ and $T^t$; Set $A^t$ as an identity matrix, and the iteration step $t = 1$;
3: repeat
4: Update $W$: $W^{t+1} = W^t - \eta \frac{\partial \mathcal{L}(W^t, V^t, T^t)}{\partial W}$;
5: Update $V$: $V^{t+1} = V^t - \eta \frac{\partial \mathcal{L}(W^t, V^t, T^t)}{\partial V}$;
6: Update $T$: $T^{t+1} = T^t - \eta \frac{\partial \mathcal{L}(W^t, V^t, T^t)}{\partial T}$;
7: Update the diagonal matrix $A$ as
$$A^{t+1} = \begin{bmatrix} \frac{1}{2 \|w^t_1\|_2^2} & \cdots & \frac{1}{2 \|w^t_n\|_2^2} \end{bmatrix};$$
8: $t = t + 1$;
9: until Convergence criterion satisfied

Output:
Transformation matrices $W = W^{t-1}$, $V = V^{t-1}$ and $T = T^{t-1}$.

**Theorem 1:** The alternate updating rules in Algorithm 1 monotonically decrease the objective function value of (11) in each iteration.

**Proof:** In the iterative procedure, for $W$, $V$ and $T$, we update one while keeping the other two fixed. In the $t$-th step, with $V$ and $T$ fixed, for the ease of representation let us denote
$$F(W) = \|V^T G - T^T V\|_F^2 + \alpha \text{Tr}[GL^C G^T] + \beta \text{Tr}[V^T GL^S (V^T G)^T]$$
(15)

It can easily verified that we have
$$W^{t+1} = \min_W F(W) + \lambda \text{Tr}[W^T A^t W]$$
$$\Rightarrow F(W^{t+1}) + \lambda \text{Tr}[(W^{t+1})^T A^t W^{t+1}] \leq F(W^t) + \lambda \text{Tr}[(W^t)^T A^t W^t]$$
(16)

That is to say,
$$F(W^{t+1}) + \lambda \sum_i \|w_i^{t+1}\|_2^2 \leq F(W^t) + \lambda \sum_i \|w_i^t\|_2^2$$
$$\Rightarrow F(W^{t+1}) + \lambda \|W^{t+1}\|_{2,1} - \lambda(\|W^{t+1}\|_{2,1} - \sum_i \|w_i^{t+1}\|_2) \leq F(W^t) + \lambda \|W^t\|_{2,1} - \lambda(\|W^t\|_{2,1} - \sum_i \|w_i^t\|_2)$$
$$\leq F(W^t) + \lambda \|W^t\|_{2,1} - \lambda(\|W^t\|_{2,1} - \sum_i \|w_i^t\|_2).$$
(17)

According to the Lemmas in [39], $\|W^{t+1}\|_{2,1} - \sum_i \|w_i^{t+1}\|_2 \leq \|W^t\|_{2,1} - \sum_i \|w_i^t\|_2$. Thus, we obtain
$$F(W^{t+1}) + \lambda \|W^{t+1}\|_{2,1} \leq F(W^t) + \lambda \|W^t\|_{2,1}.$$ 
(18)

That is, we arrive at
$$\mathcal{L}(W^{t+1}, V^t, T^t) \leq \mathcal{L}(W^t, V^t, T^t).$$
(19)

Fig. 4: The convergence curve of the proposed optimization algorithm in Algorithm 1.

With $W^{t+1}$ and $T^t$ fixed, the gradient descent method guarantees that
$$\mathcal{L}(W^{t+1}, V^{t+1}, T^t) \leq \mathcal{L}(W^{t+1}, V^t, T^t).$$
(20)

In this manner, we have
$$\mathcal{L}(W^{t+1}, V^{t+1}, T^{t+1}) \leq \mathcal{L}(W^{t+1}, V^{t+1}, T^t).$$
(21)

Based on the above three inequalities, we obtain
$$\mathcal{L}(W^{t+1}, V^{t+1}, T^{t+1}) \leq \mathcal{L}(W^{t+1}, V^{t+1}, T^t) \leq \mathcal{L}(W^{t+1}, V^t, T^t).$$
(22)

Thus, $\mathcal{L}(W, V, T)$ monotonically decreases using the updating rules in Algorithm 1 and Theorem 1 is proved.

According to Theorem 1, we can see that the iterative optimization in Algorithm 1 can converge to local optimal values of $W$, $V$ and $T$. Actually, in the proposed learning algorithm, each matrix can be updated column by column. Therefore, the algorithm can be parallelized. In the experiments, we observed that the proposed optimization converges to the minimum after about 20 iterations. Figure 4 shows the changing values of the objective function in the convergence process. We perform our experiments on the MATLAB in a PC with 2.13GHz CPU and 16 GB memory.

**D. Tag Representation in Unified Space**

Thus far, we have discovered the unified space and are able to embed untagged photos with visual features into this space. However, our target is to estimate the correlations between photos and tags. Hence, we also need to obtain the representation of each tag in the unified space, so as to adopt a similarity measure to calculate the correlations directly. For a tag, we propose to represent it with the feature reconstruction of the photos tagged with the tag. In the reconstruction, the textual representations of the photos are considered, because they have more semantic relevance to tags than visual features. Thus, the reconstructed representation $t_i$ of a tag $w_i$ is defined as
$$t_i = \frac{1}{\sum_{j=1}^{n} Y_{ij}} \sum_{j=1}^{n} Y_{ij} z_j$$
(23)
Here, $Y_{ij}$ is the indicator whether the $j$-th photo is tagged with the tag $w_i$, and $z_j$ is the embedded textual representation of the $j$-th photo in the unified space.

V. TAG RECOMMENDATION

In this section, we will introduce how to recommend tags for a new photo with the user and geo-location information. Using the above proposed learning algorithm, we can learn the corresponding underlying space for each user (or geo-location) and obtain the tag representations in the space. For each new uploaded photo $x_{te}$ taken in geo-location $g$ by user $u$, we first map it into the unified space of its corresponding user and geo-location individually, and obtain a set of initial tags relevant to the photo. Then we perform semantic photo retrieval and select the semantically relevant photos from a large scale community-contributed dataset. Finally visual image retrieval is implemented to identify the semantically and visually relevant photos, and top ranked tags are suggested based on the voting strategy.

First, we elaborate the selection of initial tags. We adopt the same strategy in the user-specific unified space and the geo-specific unified space, and detail the process performed in the user-specific one. With the learned transformation matrices $W$ and $V$, we first embed the visual feature $x_{te}$ of an untagged photo into the unified space $U$ and obtain the latent feature vector $u_{te}$.

$$u_{te} = VWx_{te}$$  \hspace{1cm} (24)

Then we use the embedded visual representation to calculate its relevance to each tag, which has been embedded into the unified space as discussed in IV-D. Specially, we adopt the cosine similarity as the relevance measure. The relevance vector of the photo to each tag is denoted as $s_u$. In the same way, we can obtain the relevance scores of the photo to tags in terms of the given geo-location, denoted as $s_g$. The relevant scores of this photo to tags are obtained by

$$s = \rho \frac{s_u}{\sum_i s_u(i)} + (1 - \rho) \frac{s_g}{\sum_i s_g(i)}$$  \hspace{1cm} (25)

where $\rho$ is a non-negative parameter to weight the importance of these two terms. We set it to 0.6 in our experiments for simplicity. Top ranked tags larger than a specified threshold are utilized to perform tag-based photo retrieval in our large scale community-contributed photo set collected from Flickr.

By this procedure, the top $r_s$ semantically similar photos are selected as the database for the next content-based photo retrieval. To perform content-based photo retrieval, we extract visual features of all photos in advance and employ $\ell_2$ distance metric to measure the visual similarities among images. Then the top $r_v$ visually similar images are selected among the top $r_s$ semantically similar photos ($r_v = 100$, $r_s = 1000$ in our implementation). More efficient retrieval approach can also be applicable, such as LSH-based indexing approach [41].

VI. EXPERIMENTS

A. Dataset

We collect a vast amount of photos with rich context information from Flickr using Flickr API. For each photo, we downloaded this photo together with its tag, geo-location (i.e., longitude and latitude), title, description and comments. For each tag, we also kept its tagger. We obtained 3,309,698 photos with 231,662 users and 20,806 geo-locations.\textsuperscript{3} We summarize the detailed statistical information in Fig. 5.

![Fig. 5: The statistic of users and geo-locations. (a): The number of photos per user; (b): The number of photos per geo-location.](image)

**TABLE I: Statistics of our dataset**

<table>
<thead>
<tr>
<th>#[Photo]</th>
<th>#[User]</th>
<th>#[Geo-location]</th>
<th>#[Tag]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,727,312</td>
<td>559</td>
<td>1,351</td>
<td>15,554</td>
</tr>
</tbody>
</table>

\textsuperscript{3}We first derived the longitude and latitude of each photo and then identified the corresponding place name from Flickr using Flickr API. In this work, the city-level place names are extracted. That is, we obtained 20,806 cities finally.
C. Parameter Setting

There are some parameters to be set in advance. For our learning algorithm, we set \( \lambda = 0.005 \) and \( \eta = 0.05 \) for simplicity. The trade-off parameters \( \alpha, \beta \) and \( \gamma \) in our proposed method are tuned in the range \{0.01, 0.01, 0.1, 1, 10\} by cross validation, and we set \( \alpha = 1, \beta = 0.1 \) and \( \gamma = 0.1 \). For the dimensions of the intermediate space and the unified space, we set \( q = 500 \) and \( p = 300 \) to leverage the performance and the cost. We will discuss the parameter sensitivity in the following.

D. Compared Methods

To prove the effectiveness of our work, we compared it with state-of-the-art models, including generic tag recommendation and personalized tag recommendation. The details of the compared schemes are listed as follows.

- **CommunityPreference (CP)**: It is to suggest the most frequent tags within the community set.

- **Visual** [21]: It chooses relevant tags purely based on visual content of photos. Top \( k \) visual similar photos from the community image set are first obtained and then the most frequent tags in the similar set are used as the recommendation result.

- **GeoVisual** [12]: It first identifies a correct geo-tag using the visual content of a new photo and user feedback. Then, it collects images with the same geo-tag and selects tags from the collected images using kNN density estimation based on visual features.

- **PersonalPreference (PP)** [26]: Given a new image uploaded by a user \( u \), it simply recommends the most frequent tags used by the user \( u \) in the past, i.e., suggests the most frequent tags in \( X_{u, tr} \).

- **GTV** [16]: It recommends tags to users by accumulating votes from the candidate images including visually similar images, images captured in the same geographical coordinates or in the same period of time, which are selected from their history images.

- **Tensor** [31]: It adopts a low-order tensor decomposition approach to model the user-image-tag correlation data, and predicts relevant tags of an image for a specific user.

- **CEPIA** [33]: It predicts personalized tags by learning from user tagging history based on a cross-entropy based learning algorithm.
Now we discuss their influence on the performance.

E. Parameter Sensitivity Analysis

There are some parameters involved in the proposed learning model, i.e., three trade-off parameters $\alpha$, $\beta$ and $\gamma$, and the dimensions of the intermediate space and the unified space. Now we discuss their influence on the performance.

First, we discuss the sensitivity analysis of the trade-off parameters $\alpha$, $\beta$ and $\gamma$. We tune these three parameters in the range of $\{0, 0.0001, 0.001, 0.01, 0.1, 1.0, 10, 100\}$. The corresponding results in terms of $F_1@5$ are illustrated in Fig. 7. By examining these results, we have the following observations:

1. The parameter $\alpha$ has the most remarkable impact on the performance. When $\alpha$ is very small (less than 0.01) or very large (more than 10), the performance in term of $F_1@5$ decreases dramatically. It demonstrates that the introduced intermediate space is not only necessary but also supposed to reflect the semantic information, which is coincident with our intuition.

2. It is necessary to preserve the local semantic structure and the local visual structure for the unified subspace learning. When $\beta = 0$ or $\gamma = 0$, i.e., we do not consider the local structure preservation, the results of tag recommendation become worse.

3. The semantic structure preservation is important than the visual structure preservation in the unified subspace learning procedure since the performance has less change with varying the value of $\beta$ than that of $\gamma$.

The dimensions of the intermediate space and the unified space are two important parameters, which affect the user preference and geo-location preference learning. We conduct experiments to test their effects on the performance of tag recommendation and present results in terms of $F_1@5$ in Fig. 8 (a) and (b), respectively. We can see that the performance varies with different values of the dimension $p$ of the unified space and the dimension $q$ of the intermediate space. The performance is improved by increasing $p$ and $q$ to some extent and arrive relatively stable $F_1$ values when $q \geq 500$ and $p \geq 300$, respectively. Considering high dimension corresponds to an expensive computing cost, we set $q = 500$ and $p = 300$ to leverage the performance and the cost.

F. Experimental Analysis

We conduct extensive experiments to compare UGUI with the above algorithms in Section VI-D. For the sake of clarity, we only present results of CP, Visual, PP, CEPIA, Tensor and UGIU in terms of $P@k$, $R@k$ and $F1@k$ in Fig. 9 (a), (b) and (c), respectively. To compare the variations of our proposed methods, we present results in terms of $F1@k$ ($k = 1, 2, \cdots, 10$) in Table II. The performance comparison on averaged APs is shown in Fig. 10. From these results, we can draw the following observations.

First, compared to generic models (i.e., CP, Visual and GeoVisual), personalized models prevalently achieve better results, which is consistent with the observations made by [26], [33]. The simple PP model gains a relative improvement over CP, which reveals that the user tagging history is helpful to personalized tag recommendation. It can be also demonstrated by the comparison of GTV vs. GeoVisual, UGIU vs. IU-G, and UGIU vs. GIU. Besides, visual content cannot be ignored, which is observed from the better performance of Visual by the comparison of GTV vs. GeoVisual, UGIU vs. IU-G, and UGIU vs. GIU.

Second, to demonstrate the necessity of exploiting geographic information and learning geo-location penchant towards tags, we conduct the comparisons of GeoVisual vs. Visual, UIU vs. IU-U, and IU-U vs. UGIU. First, GeoVisual gains the better performance than the Visual model by additionally considering geo-context. Similarly, IU-U has the
better result in terms of \( F_1 \) than UIU. This reveals that the geographic information can provide helpful cues for tag recommendation. Second, from the comparison between IU-U and UGIU, the better results are obtained by UGIU. Thus, we can claim that the geo-location preference is useful to filter tags for recommendation.

Third, by introducing an intermediate space, UGIU considerably surpasses UGU. This is consistent with our earlier presentations that a better latent space can be learned by the proposed progressive manner. The visual features is a much lower level representation on semantics and the gap between the visual features and the semantic concepts is too large. It is not suitable to directly map the visual features into the underlying unified space. By introducing an intermediate space which is analogous to the tag space, the semantic gap is bridged progressively and a better latent space is learned.

Finally, our proposed UGIU achieves the best performance among all the compared method, specially the personalized models. Compared to the three personalize algorithms GTV, Tensor and CEPIA, UGIU obtains 14\%, 12\% and 7\% gains in terms of \( F_1@5 \) respectively. It is noted that GTV and the proposed method UGIU both consider the user history and geographical information. However, UGIU is significantly better than GTV, which demonstrates the superiority of the proposed learning algorithm.

From the above experimental results, we can see that the improvement of the proposed method over the compared methods becomes narrow when \( k \geq 4 \). Thus, to demonstrate the significance of the improvement, we perform the statistical analysis of the variance of GeoVisual, GTV, CEPIA, Tensor and UGIU in terms of \( F_1@k \) (\( k \geq 4 \)), and present the results in Table III. It is clearly observed that the proposed method UGIU statistically outperforms the compared methods.

To present visualized comparisons, we also illustrate some recommended results by UGIU, IU-U and IU-G in shown Fig. 11. From these examples, we can see that UGIU enables to suggest accurate tags that coincide with the user assigned tags (i.e., the ground-truth). Compared with IU-U, UGIU can recommend more tags that are associated with specific geo-locations and related with photo content, which reveals the validity of exploiting geo-location preference. On the other hand, compared with IU-G, UGIU can generate tags that closely meet tag-preference of specific users. In a word, these examples demonstrate that it is necessary and rational to mine user preference and geo-location preference.

**VII. CONCLUSIONS**

In this work, we propose to mine the personalized tags for new updated photos using users’ tagging histories and geographic information. We propose a new subspace learning algorithm to individually discover the user preference and the geo-location preference towards tags. In the proposed method, the visual features and text features of photos are mapped into a unified space by three transformation matrices: two for visual features and one for text features. To bridge the semantic gap, we propose to first map visual features into an intermediate space having the consistent semantic structure with the text space. For an untagged photo, we first map it into the unified space and then map it into the intermediate space.

Fig. 8: Impact of parameters: (a) the dimension \( p \) of the unified space and (b) the dimension \( q \) of the intermediate space.

Fig. 9: Performance comparisons in terms of (a) \( P@k \), (b) \( R@k \) and (c) \( F_1@k \) with \( k = 1, 2, \cdots, 10 \).
TABLE II: Performance comparison in terms of $F^k@k$ ($k = 1, 2, \ldots, 10$).

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>CP</td>
<td>0.0245</td>
<td>0.0271</td>
<td>0.0280</td>
<td>0.0366</td>
<td>0.0366</td>
<td>0.0369</td>
<td>0.0456</td>
<td>0.0460</td>
<td>0.0480</td>
<td>0.0498</td>
</tr>
<tr>
<td>Visual</td>
<td>0.0362</td>
<td>0.0740</td>
<td>0.1116</td>
<td>0.1489</td>
<td>0.1848</td>
<td>0.211</td>
<td>0.2304</td>
<td>0.2424</td>
<td>0.2518</td>
<td>0.2553</td>
</tr>
<tr>
<td>GeoVisual</td>
<td>0.1029</td>
<td>0.1695</td>
<td>0.2328</td>
<td>0.2921</td>
<td>0.3513</td>
<td>0.4027</td>
<td>0.4359</td>
<td>0.4652</td>
<td>0.4859</td>
<td>0.5005</td>
</tr>
<tr>
<td>PP</td>
<td>0.0864</td>
<td>0.1664</td>
<td>0.2310</td>
<td>0.2990</td>
<td>0.3442</td>
<td>0.3672</td>
<td>0.4069</td>
<td>0.4404</td>
<td>0.4725</td>
<td>0.4964</td>
</tr>
<tr>
<td>GTV</td>
<td>0.1067</td>
<td>0.1813</td>
<td>0.2310</td>
<td>0.3142</td>
<td>0.3374</td>
<td>0.3883</td>
<td>0.4303</td>
<td>0.4543</td>
<td>0.4894</td>
<td>0.5250</td>
</tr>
<tr>
<td>CEPIA</td>
<td>0.1210</td>
<td>0.1842</td>
<td>0.2505</td>
<td>0.3145</td>
<td>0.3605</td>
<td>0.3785</td>
<td>0.4260</td>
<td>0.4543</td>
<td>0.4856</td>
<td>0.5092</td>
</tr>
<tr>
<td>Tensor</td>
<td>0.094</td>
<td>0.1157</td>
<td>0.2253</td>
<td>0.2842</td>
<td>0.3439</td>
<td>0.3610</td>
<td>0.4338</td>
<td>0.4518</td>
<td>0.4730</td>
<td>0.5124</td>
</tr>
<tr>
<td>UIU</td>
<td>0.0658</td>
<td>0.1164</td>
<td>0.1967</td>
<td>0.2829</td>
<td>0.3468</td>
<td>0.3741</td>
<td>0.4120</td>
<td>0.4498</td>
<td>0.4670</td>
<td>0.5082</td>
</tr>
<tr>
<td>GIU</td>
<td>0.0128</td>
<td>0.0280</td>
<td>0.0405</td>
<td>0.0506</td>
<td>0.0599</td>
<td>0.0674</td>
<td>0.0768</td>
<td>0.0861</td>
<td>0.0952</td>
<td>0.104</td>
</tr>
<tr>
<td>UGIU</td>
<td>0.0994</td>
<td>0.1843</td>
<td>0.2510</td>
<td>0.3273</td>
<td>0.3694</td>
<td>0.4041</td>
<td>0.4176</td>
<td>0.4829</td>
<td>0.5099</td>
<td>0.5299</td>
</tr>
<tr>
<td>IU-G</td>
<td>0.0905</td>
<td>0.1662</td>
<td>0.2392</td>
<td>0.2998</td>
<td>0.3356</td>
<td>0.3714</td>
<td>0.4126</td>
<td>0.4441</td>
<td>0.4742</td>
<td>0.5010</td>
</tr>
<tr>
<td>IU-U</td>
<td>0.0972</td>
<td>0.1851</td>
<td>0.2590</td>
<td>0.3173</td>
<td>0.3613</td>
<td>0.3876</td>
<td>0.4122</td>
<td>0.4794</td>
<td>0.4893</td>
<td>0.5099</td>
</tr>
<tr>
<td>UGIU</td>
<td>0.1366</td>
<td>0.1990</td>
<td>0.2630</td>
<td>0.3373</td>
<td>0.3851</td>
<td>0.4115</td>
<td>0.4327</td>
<td>0.4885</td>
<td>0.5124</td>
<td>0.5491</td>
</tr>
</tbody>
</table>

Fig. 10: Performance comparison about Averaged AP.

spaces in terms of the user and geographic information to find relevant tags, and then perform semantic and visual photo retrieval to find relevant photos. Finally, the most frequent tags in the relevant photos are suggested to users. Extensive experiments have been conducted to validate the effectiveness of our personalized tag recommendation method.

With the learned user preference and geo-specific preference to tags or semantic concepts, we can develop some verticalized social applications, such as personalized product recommendation, geo-location based traveling suggestion, personalized geo-specific news report, and so on. Besides, how to investigate the joint or partially joint connections among user, geo-location, social tags, and photos to enhance the latent subspace learning performance, is also a potential research topic.

VIII. ACKNOWLEDGMENTS

This work was partially supported by the 973 Program (Project No. 2010CB327905), the National Natural Science Foundation of China (Grant No. 61103059 and 61272329), the Program for New Century Excellent Talents in University under Grant NCET-12-0632, and Open Projects Program of National Laboratory of Pattern Recognition.

REFERENCES


Table III: The variance analysis in terms of $F^1@k$ ($k \geq 4$).

<table>
<thead>
<tr>
<th>GROUP</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeoVisual [12]</td>
<td>0.2921</td>
<td>±0.0064</td>
<td>0.3513</td>
<td>±0.0073</td>
<td>0.4027</td>
<td>±0.0055</td>
<td>0.4359</td>
</tr>
<tr>
<td>GTV [16]</td>
<td>0.3141</td>
<td>±0.0024</td>
<td>0.3734</td>
<td>±0.0029</td>
<td>0.3883</td>
<td>±0.0041</td>
<td>0.4503</td>
</tr>
<tr>
<td>CEPLA [33]</td>
<td>0.3145</td>
<td>±0.0093</td>
<td>0.3605</td>
<td>±0.0112</td>
<td>0.3785</td>
<td>±0.0084</td>
<td>0.4260</td>
</tr>
<tr>
<td>Sensor [31]</td>
<td>0.2842</td>
<td>±0.0047</td>
<td>0.3439</td>
<td>±0.0044</td>
<td>0.3610</td>
<td>±0.0052</td>
<td>0.4338</td>
</tr>
<tr>
<td>Tensor</td>
<td>0.3574</td>
<td>±0.0059</td>
<td>0.3851</td>
<td>±0.0117</td>
<td>0.4115</td>
<td>±0.0048</td>
<td>0.4327</td>
</tr>
<tr>
<td>UUGI</td>
<td>0.3734</td>
<td>±0.0059</td>
<td>0.3851</td>
<td>±0.0117</td>
<td>0.4115</td>
<td>±0.0048</td>
<td>0.4327</td>
</tr>
</tbody>
</table>


Zechao Li (corresponding author) received the Ph.D degree from National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences in 2013. He received the B.E. degree from University of Science and Technology of China (USTC) in 2008. He is currently an assistant professor in School of Computer Science, Nanjing University of Science and Technology, China. His research interests include machine learning, subspace learning, multimedia understanding, etc. He received the 2013 President Scholarship of Chinese Academy of Science.

Jinhui Tang received the BE and PhD degrees both from the University of Science and Technology of China in July 2003 and July 2008, respectively. He is a professor in the School of Computer Science and Technology, Nanjing University of Science and Technology. From July 2008 to December 2010, he was a research fellow with the School of Computing, National University of Singapore. During that period, he visited the School of Information and Computer Science, UC Irvine, as a visiting research scientist, from January 2010 to April 2010. His current research interests include large scale multimedia search, social media mining, and computer vision. He has authored more than 60 journal and conference papers in these areas. He received the 2008 President Scholarship of Chinese Academy of Science, and a corecipient of the best paper award at ACM Multimedia 2007, PCM 2011 and ICIMCS 2011.

Yu Jiang received his B.E. degree from Beihang University in 2009. He is currently a Ph.D. Candidate in the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include subspace learning, multiview learning, and recommender system, etc.

Hanqing Lu received his B.E. degree in 1982 and his M.E. degree in 1985 from Harbin Institute of Technology, and Ph.D. degree from Huazhong University of Sciences and Technology, Wuhan, China in 1992. Currently he is a Professor at National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include computer vision, object tracking, recognition and image retrieval, etc. He has published more than 300 papers in those areas. He is a senior member of the IEEE.