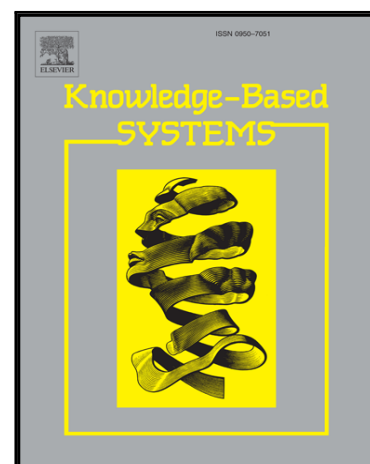


Accepted Manuscript

Unsupervised Geographically Discriminative Feature Learning for
Landmark Tagging

Xiaoming Zhang, Zhonghua Zhao, Haijun Zhang, Senzhang Wang,
Zhoujun Li

PII: S0950-7051(18)30122-9
DOI: [10.1016/j.knosys.2018.03.005](https://doi.org/10.1016/j.knosys.2018.03.005)
Reference: KNOSYS 4254



To appear in: *Knowledge-Based Systems*

Received date: 27 July 2017
Revised date: 28 February 2018
Accepted date: 2 March 2018

Please cite this article as: Xiaoming Zhang, Zhonghua Zhao, Haijun Zhang, Senzhang Wang, Zhoujun Li, Unsupervised Geographically Discriminative Feature Learning for Landmark Tagging, *Knowledge-Based Systems* (2018), doi: [10.1016/j.knosys.2018.03.005](https://doi.org/10.1016/j.knosys.2018.03.005)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Unsupervised Geographically Discriminative Feature Learning for Landmark Tagging

Xiaoming Zhang^a, Zhonghua Zhao^b, Haijun Zhang^{c,*}, Senzhang Wang^d, Zhoujun Li^e

^a*School of Cyber Science and Technology, Beihang University, China*

^b*National Computer Network Emergency Response Technical Team/Coordination Center of China*

^c*School of Information, Beijing Wuzi University, China*

^d*School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, 210016, China*

^e*Beijing Key Laboratory of Network Technology, Beihang University, Beijing, 100191, China*

Abstract

Recently, a large number of geo-tagged landmark images have been uploaded through various social media services. Usually, these geo-tagged images are annotated by users with GPS and tags related to the landmarks where they are taken. Landmark tagging aims to automatically annotate an image with the tags to describe the landmark where the image is taken. It has been observed that the images and tags show strong correlation with the geographical locations. The widely used assumption by many existing tagging methods is that images are independently and identically distributed is not effective to capture the geographical correlation. In this paper, we study the novel problem of utilizing the geographical correlation among images and landmarks for better tagging landmark images. In particular, we propose an unsupervised feature learning approach to learn the geographically discriminative features across geographical locations, by integrating latent space learning and geographically structural analysis (LSGSA) into a joint model. A latent space learning model is proposed to effectively fuse the heterogeneous features of visual content and tags. Meanwhile, the geographical structure analysis and group sparsity are applied to learn the geographically discriminative features. Then, a geo-guided sparse reconstruction method is proposed to tag images by utilizing the discriminative information of features, in which the landmark-specific tags are boosted by a weighting method. Experiments on the real-world datasets demonstrate the superiority of our approach.

Keywords: Image tagging, Feature learning, Landmark image, Landmark features

*Corresponding author

Email address: zhanghaijun@bnu.edu.cn (Haijun Zhang)

Preprint submitted to Journal of Knowledge-Based Systems

March 5, 2018

1. Introduction

In the social network sites, many geo-tagged images are annotated with GPS information and tags about the landmarks where they are taken. It has been observed that these images and their tags show strong correlation with the locations where the images are taken [1]. For example, images about Eiffel Tower cover entirely different patterns, i.e., vision characteristic of images and linguistic preference of tags, compared to those about Forbidden City. Landmark tagging is to annotate an image with the tags to describe the landmark where the image is taken [1], [35], [49]. The high availability of geo-tagged images and the geographical characteristic of images provide both challenges and opportunities to landmark tagging.

Most existing image tagging approaches recommend tags to the query image based on their relation to the visual content of the image, i.e., the search-based approaches and the model-based approaches. The search-based approaches assume that two visually similar images are usually assigned with similar tags [5], [36], [55], [58]. Given a query image, tags are recommended from the tags associated with its visually nearest neighbours. The model-based approaches require models to be learned for a set of predefined concepts [45], [50], [51], which limits their extension to an uncontrolled tag set produced by Web users. A major limitation of the above mentioned methods is that they assume the images are identically distributed without considering their geographical distributions. This assumption does not hold in the geo-based social media since two visually similar images may be taken in different locations. Usually, the geo-tagged social images are inherently correlated with each other via geographical relations. For example, the landmark images in Beijing cover the specific vision patterns and tags about the historical spots of Qing Dynasty, while the images in Paris have both the visual and textual pattern about Gothic scenery. Recently, some methods have been proposed to tag images by utilizing the geographical information, which assumes that the geographically adjacent and visually similar images have similar tags [35], [48]. However, the images about the same landmark usually cover a large diversity of visual content, which affects the measure of visual similarity. Moreover, the assumption of identical distribution may lead to that many common tags that are irrelevant to the landmark are recommended.

In this paper, we study how the geographical distribution of the multi-modal contents of image could be used to tackle the problem of landmark tagging. This is a non-trivial task because of the distinct characteristics of geo-tagged images. First, social images are freely produced and

can be geo-tagged with different locations. There is no prior knowledge about the structured organization of images over the geographical regions, and manually relating the geo-tagged images with the landmarks is intellectually expensive. Second, since the images in different regions present a great diversity of viewpoints, the visually similar images may be taken in different landmarks. The noisy and redundant features of images affect the effectiveness of the similarity-based methods. Finally, the multi-modal contents of images, i.e., visual content and textual content, are correlated with the landmarks simultaneously [2]. However, the visual content and the textual content are represented in two heterogeneous spaces. These inconsistent structures make it difficult to discover the inherent relation between locations and images from the raw features separately.

Currently, discriminative feature learning is an effective way to learn a subset of important features from the high dimensional feature set for a more accurate data representation [52]. It has been widely used in various tasks to capture the latent relation between the raw features and the high-level concepts, such as classification, pattern recognition, and information retrieval [43], [57], [68]. The improved performance is largely attributed to that the effect of noisy features is alleviated and the important features are reinforced. On the other side, each location has its own characteristics of vision and linguistic model. For example, some color features might be more important to distinguish the natural landmarks with specific tags like “desert” or “mountain”, while some edge features might be more important to identify the artificial view landmarks with specific tags like “White House” or “Eiffel”. Therefore, motivated by recent development in the feature learning domain, we propose to learn the discriminative features for the locations to improve landmark tagging.

To address the new challenges posed by landmark images, we propose to take advantage of the correlation between geographical locations and the multi-modal contents for landmark tagging. In this paper, we study the problem of learning the geographically discriminative features for landmark tagging. In essence, we investigate: (1) how to model geographical knowledge in discriminative feature learning; and (2) how to seamlessly combine the textual content and visual content for the problem we study. Our solutions to the challenges result in a new framework for unsupervised learning of geographically discriminative features for landmark tagging. In particular, we employ a geo-guided feature learning model to integrate latent space learning and geographically structural analysis (LSGSA) into a joint framework. Then, the query im-

age is tagged based on the reconstruction coefficients of the dictionary images by leveraging the discriminative information of features and the geographical distribution.

65 The main contributions of this paper are summarized as follows:

- A novel feature learning framework is proposed by integrating multi-modal feature fusion with geographically structural analysis, which takes advantage of the robust encoding ability of group sparse coding, the prior knowledge about the geographical relation among images, and the latent correlation among multi-modal contents;
- 70 • To tag landmark images, a geo-guided group sparsity model is proposed to utilizing the discriminative information of image features and the geographical distribution of images, and a weighting method is proposed to boost the landmark-specific tags;
- Extensive experiments are conducted on the public image datasets to present the effectiveness of our approach.

75 The remainder of this paper is organized as follows. In the next section, we introduce the related works. We formulate the problem in Section 3, and we introduce LSGSA in Section 4. Section 5 describes how to tag landmark images, and Section 6 presents the experimental results and analysis. Finally, the paper is concluded in Section 7.

2. Related Works

80 With the explosive growth of geo-tagged images, landmark tagging is an emerging research topic in multimedia application and computer vision. The related works include image tagging and geographic mining of social images.

2.1. Image tagging

85 There have been many image tagging approaches. The generic image tagging approaches are to assign relevant tags to an image based on the relation between tags and visual content, which is independent of the geographical factor. These approaches can be roughly categorized into search-based methods and model-based methods. First, the search-based methods assume that two visually similar images should be assigned with similar tags. For example, the tags for the query image are usually ranked by the votes from a subset of the training images [5],

[36], [54], [55], [58], or the tags are naturally ranked according to their probabilities conditioned on the query image using topic models [6], [7], [8], [9] and mixture models [10], [11], [12]. These methods are greatly depended on the visual similarity measure, or impose some statistic assumptions on the models. Second, the model-based methods require models to be learned for a set of predefined concepts [45], [50], [65] or learn the intricate dependencies between image content and annotations [69]. For example, classification based on SVM [13], ranking SVM [14], [56], and boosting [15] are used to recommend tags based on the visual features. To improve the classification, Shen et al. [50] proposes a multi-task structured SVM algorithm to leverage both the inter-object correlations and the loosely-tagged images. [45] proposes to tag image using subspace-sparsity collaborated feature selection methods. Based on convolutional neural network, a large scale image annotation model MVAIACNN is constructed in [63]. [64] proposes a image annotation method that makes use of Convolutional Neural Network features extracted from images and word embedding vectors to represent their associated tags. Usually, these approaches learn a object function to mapping from visual feature to tags, which is not effective to capture the different importance of various image features and tags. Overall, these image tagging approaches ignore the geographical correlation of images and tags, which affects the performance on tagging landmark image.

Recently, with the popularity of location-based social web services, there are some works focusing on tagging images by exploring geographical location information. Similar to the search-based approaches, the query image is annotated by the voting of the constrained k nearest neighbors [16], [17], [35], [48], where the visual neighbors are retrieved from the geo region of the query image. Besides the geographical information, the user's profile is also taken into account in [42]. The main work of [18] learns a tag list to post-filter the result of tag result, which categorizes tags as landmarks and visual descriptors. [49] annotates geo-tagged images with descriptive tags by exploring the redundancy over the large volume of annotations available at online repositories with other geo-tagged images. [19] focuses on improving the visual search by fusing geographical context with visual concept detection. It assumes that the geographically adjacent and visually similar images have similar tags as [1], [35], [48]. However, these approaches are not effective to explore the geographical distribution of images and tags, since they assume that images and tags are identically distributed.

120 2.2. Geographic mining of images

There are also many works on exploring the geographical information for other applications such as geographic referencing of images. Geographic referencing of images is to determine the landmark or geographical location of the query image. It mainly includes the data-driven methods and the model-based methods. First, the data-driven methods conduct the geographic
 125 referencing by retrieving the nearest neighbours from a pre-built database. [21] presents a feature matching approach to return the K nearest neighbours with respect to the query landmark image, which represents the query images and the images in database by aggregating a set of low-level features to perform landmark retrieval. [22] retrieves the visually similar candidates by considering their geo-visual neighbors which are both geographically nearby and visually similar
 130 to the query image. The search method [23], [24], [25] and the language model [26] are also used for landmark recognition. [62] proposes an unsupervised image GPS location estimation approach with hierarchical global feature clustering and local feature refinement. Usually, the performance of these data-driven methods is affected by the selection of neighbour images.

Second, the model-based methods attempt to build models to extract the geographical pat-
 135 terns or discriminative features for location recognition. In [28], a region based recognition method is proposed to detect discriminative landmark regions at the patch level, such as a set of stylistic of visual elements to characterize a city, e.g., windows and street signs. [29] presents an approach, namely GIANT, to discover both discriminative and representative mid-level attributes for landmark retrieval. [20] proposes to select a set of images and patches to represent a
 140 landmark, and then the SVM classifiers are learned to classify landmark images. [30] proposes a multi-query expansions method to retrieve landmark images, which learns the discriminative patterns from the query expansion images. However, these methods learn the landmark model mainly based on the visual content, which ignores the plentiful source of textual information and the geographical preference of tags. [27] and [31] propose to classify landmark images by lin-
 145 early combining the visual features and textual tags. [32] proposes a ranking method to fuse the multiple evidences derived from textual features and visual features for image location estimation. [59] presents an image location estimation approach based on multisaliency enhancement. [60] proposes to mine the salient image feature for GPS location estimation, which mines the salient region of the input image by exploring its relation with the k nearest images. [61] esti-
 150 mates the location based on content-based image retrieval, in which spatial constraint is utilized

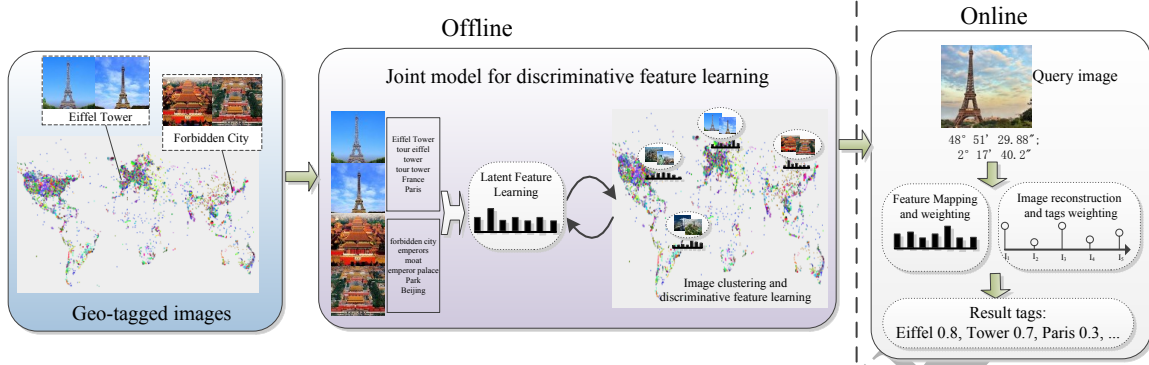


Figure 1: The framework of our approach. Our framework comprises two main components: 1) the discriminative feature learning component, which is joint model to learn the latent space and discriminative information simultaneously; 2) the landmark tagging component, which tag image by exploring both the discriminative information of image feature and geographical distribution via a sparse coding model.

to code the relative position of visual words. The ranking method on hypergraphs learning is proposed for simultaneous image tagging and geo-location prediction in [70]. These approaches handle different features independently and identically, which can not effectively explore the inherent correlation between different types of features and the discriminative information of image features.

3. Problem Statement

In this section, we first introduce the notations used in the paper and then formally define the problem which we study.

Notation: Matrices are denoted by boldface uppercase letters, vectors by boldface lowercase letters, and scalars by lower case letters. For a matrix $\mathbf{A} \in \mathcal{R}^{n \times m}$, \mathbf{A}^T denotes its transpose, \mathbf{A}_i denotes its i th row except specific declaration, and $\|\mathbf{A}\|_{2,1}$ denotes the $l_{2,1}$ -norm regularization, i.e., $\|\mathbf{A}\|_{2,1} = \sum_{i=1}^n \sqrt{\sum_{j=1}^m (\mathbf{A}_{ij})^2}$ [4]. \mathbf{I} denote the training dataset with n geo-tagged images. Each image $I_i = \{\mathbf{x}_i, \mathbf{y}_i, \mathbf{g}_i\}_{I_i \in \mathbf{I}}$ consists of three atoms: $\mathbf{x}_i \in \mathcal{R}^d$ is the visual feature vector; $\mathbf{y}_i \in \{0, 1\}^{v \times 1}$ is the tag indicator vector, where v is the size of tag vocabulary and $y_{ij}=1$ if the i th image is tagged with the j th tag, and $y_{ij}=0$ otherwise; \mathbf{g}_i is a 2-D vector containing the latitude and longitude where the image is taken. Let $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$ denotes the visual feature matrix, $\mathbf{G} = [\mathbf{g}_1, \dots, \mathbf{g}_n]$ denotes the location matrix, and $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_n]$ denotes the tag label

matrix.

Then, the problem of landmark tagging is defined as:

170 *Given a set of geo-tagged social images I in which each image contains text features, tags, and geo-coordinate, we aim to automatically tag the query landmark image I_q with only visual features and location information by exploring the discrimination information of the features and the geographical distribution of images.*

The framework of our approach is shown in Fig. 1. It contains two components, i.e., the
175 discriminative feature learning component and the landmark tagging component. In the first component, we simultaneously learn a latent feature space and the discriminative information of image features via a joint model, which exploits both the data structure and geographical structure. The latent feature space learned from the visual content is more consist with the semantic space and is also an ideal surrogate for discriminative feature learning. Meanwhile, a unsuper-
180 vised method with $l_{2,1}$ -norm regularization is adopted to cluster the images with geographical information and the discriminative information of features is derived from the clustering function accordingly. In the second component, the query image is tagged with the landmark tags by exploring the discriminative information of the features via the $l_{2,1}$ -norm regularization. To improve the recall of the rare tag which is mostly appear in the related landmarks, a weighting
185 method is proposed to boost the tag by exploring its geographical distribution information and the visual similarity between the associated images and the query image.

4. Geo-Guided Discriminative Feature Learning

In this section, we first introduce how to fuse the multi-modal contents, and then we introduce a joint model to learn the geographically discriminative features from the multi-modal contents.
190 Finally, we present the optimization algorithm.

4.1. Fusing Multi-modal Contents

Usually, an social image contains both the visual content and textual content. The semantic gap problem leads to that two visually similar images may be semantically different, which affects the performance of the similarity-based image tagging approaches [48], [54], [58]. It
195 is necessary to learn the discriminative features by effectively fusing the multi-modal contents. Though there are some studies on fusing multi-modal features for image tagging [42], they focus

on learning two feature spaces separately to make the text features consist with the visual features, which is ineffective to capture the geographical correlation among different images. We propose a latent feature space for the visual content by fusing the semantic structure information of the text content. We expect this latent feature is an ideal surrogate for discriminative feature learning.

Specifically, we use a linear transformation matrix $\mathbf{P} \in \mathcal{R}^{z \times d}$ to convert the visual features to z -dimension features. We also expect that the transformed space is consistent with the textual space with semantic structure, i.e., the similarity of two objects in the latent space should be consistent with their similarity in the textual space. We propose a cost function to measure the difference between the structures of the two spaces as follows:

$$\begin{aligned} \min_{\mathbf{P}} \Upsilon(\mathbf{M}) &= Tr[\mathbf{M}\mathbf{L}^f\mathbf{M}^T] + Tr[\mathbf{M}\mathbf{L}^o\mathbf{M}^T] + \alpha\|\mathbf{P}\|_F^2 \\ s.t. \quad \mathbf{M} &= \mathbf{P}\mathbf{X} \end{aligned} \quad (1)$$

where \mathbf{L}^f and \mathbf{L}^o are the normalized graph Laplacian matrices [44], and $Tr[\cdot]$ denotes the trace operation.

To preserve the spatial structure of the original features, we assume that the visually similar images should be located closely in the transformed space, which is measured by the first term of Eq. (1). In addition, to make the latent features consistent with the textual representation, we suggest that two textually similar images are also located closely in the latent space, which is measured by the second term of Eq. (1). Let $\mathbf{F} \in \mathcal{R}^{n \times n}$ and $\mathbf{O} \in \mathcal{R}^{n \times n}$ denote the similarity matrices in the original vision and text space respectively. We apply Gaussian kernel similarity and cosine similarity to calculate the similarity matrices \mathbf{F} and \mathbf{O} , respectively. Then, $\mathbf{L}^f = \mathbf{D}^{-1/2}(\mathbf{D} - \mathbf{F})\mathbf{D}^{-1/2}$, where $\mathbf{D} = diag(D_{11}, D_{22}, \dots, D_{nn})$ is a diagonal matrix and $D_{ii} = \sum_{j=1}^n F_{ij}$. Similarly, the Laplacian matrix \mathbf{L}^o is constructed from the similarity matrix in the textual space \mathbf{O} .

4.2. A Joint Model for Unsupervised Feature Learning

Normally, each image is represented by high-dimensional features. However, some of the features are often correlated or redundant to each other, and sometimes noisy [43]. Meanwhile, each location has its own characteristics of vision and linguistic model. That is, each location has the specific correlation between the visual features and tags. Most existing approaches handle

different features identically [42], [48] or learn the discriminative information from the distribution of visual features only [65], [66], which affects the performance of landmark tagging. On the other side, there is no direct relationship between the multimodal features and locations. It is expensive to manually label a large number of images with the landmarks for discriminative features learning.

To address these problems, we propose an unsupervised method to learn the geographically discriminative features by combining the geographical knowledge and latent space learning jointly, which exploits the structure information of both images and locations. That is, the discriminative information of features is learned from the distribution over the clusters, where the images in the same cluster are located nearby with each other. Meanwhile, the sparse structural analysis is employed to remove the redundant features in the clustering process. Assume that n images are sampled from c geographical regions, with the region cluster indicator matrix denoted as $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n]^T \in \{0, 1\}^{n \times c}$. $S_{ij} = 1$ if I_i is assigned to the j th cluster, and $S_{ij} = 0$ otherwise. The scaled cluster indicator matrix \mathbf{R} is defined as follows [52]:

$$\mathbf{R} = \mathbf{S}(\mathbf{S}^T \mathbf{S})^{-1/2}, \mathbf{R}^T \mathbf{R} = \mathbf{I}_c \quad (2)$$

where \mathbf{I}_c is an identity matrix. To conduct discriminative feature learning, we impose a joint model with image region prediction and discriminative feature learning on the regularization term as follows:

$$\min_{\mathbf{P}, \mathbf{R}, \mathbf{q}} J(\mathbf{R}) + \sum_{i=1}^c (\varepsilon \sum_{j=1}^n l(q_i(x_j), \mathbf{R}_i) + \Omega(q_i)) + \lambda \Upsilon(\mathbf{M}) \quad (3)$$

where $J(\mathbf{R})$ is a clustering criterion which is used guarantee that the closely located images have the similar clustering results, $l(\cdot, \cdot)$ is the loss function for the cluster label prediction, $q_i(\cdot)$ is a predictive function to map a image to the i th cluster, and $\Omega(\cdot)$ is a regularization function to guarantee sparsity. The fundamental design principle of this joint model is that the latent space learning module and the discriminative feature learning module should form a mutually-reinforcing learning loop. The latent features should be well explored to improve the prediction of geographical region and hence help the learning of the discriminative information. Meanwhile, the discriminative feature learning is supposed to guide a better learning of latent space from the multi-modal contents in return.

An effective cluster indicator matrix is capable to reflect the discriminative information of

the features. Two images that are adjacent to each other in the geographical space should have similar indicators of cluster labels. Therefore, we use the local geographical structure to conduct clustering, which can be modelled by a nearest neighbour graph in the geographical space with the affinity graph \mathbf{E} defined as:

$$E_{ij} = \begin{cases} \exp(-\|\mathbf{g}_i - \mathbf{g}_j\|^2) & x_i \in N_k(I_j) \text{ or } x_j \in N_k(I_i) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where $N_k(I_i)$ is the k nearest neighbours of I_i in the geographical space. The local geographical correlation structure is modelled as follows:

$$\min_{\mathbf{R}} J(\mathbf{R}) = \text{Tr}[\mathbf{R}^T \mathbf{L}^e \mathbf{R}] \quad (5)$$

where \mathbf{L}^e is a Laplacian matrix built on the affinity matrix \mathbf{E} . Then, we adopt a linear model to predict the pseudo labels of image I_j as follows:

$$q_i(I_j) = \mathbf{w}_i^T \mathbf{M}_j = \mathbf{w}_i^T \mathbf{P} \mathbf{x}_j \quad (6)$$

where $\mathbf{w}_i \in \mathcal{R}^z$ is the weight vector. $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_c]$ is used to denote the weight matrix which also indicates the importance of each feature to different clusters. We use the least square loss for $l(\cdot, \cdot)$ corresponding to label prediction, and the $l_{2,1}$ -norm regularization is adopted for \mathbf{W} . Finally, the formula for discriminative feature learning is

$$\begin{aligned} \min_{\mathbf{P}, \mathbf{R}, \mathbf{W}} \mathcal{L} = & \text{Tr}(\mathbf{R}^T \mathbf{L}^e \mathbf{R}) + \varepsilon \|\mathbf{R} - \mathbf{M}^T \mathbf{W}\|_F^2 \\ & + \beta \|\mathbf{W}\|_{2,1} + \lambda \Upsilon(\mathbf{M}) \\ \text{s.t. } & \mathbf{R}^T \mathbf{R} = \mathbf{I}_c, \mathbf{R} \geq 0 \end{aligned} \quad (7)$$

where the $\|\mathbf{W}\|_{2,1}$ is adopted to guarantee that \mathbf{W} is sparse in rows [46], which constrains the number of features to be selected since some features are unhelpful. Note that each element R_{ij} indicates the relationship between the i th image and the j th cluster, which is nonnegative by nature. However, the optimal \mathbf{R} has mixed signs, which violates its definition and makes it difficult to get the cluster labels. Therefore, \mathbf{R} is constrained to be nonnegative. When both nonnegative and orthogonal constraints are satisfied, only one element in each row of \mathbf{R} is greater than zero and all of the others are zeros, which makes the results more appropriate for clustering. Once \mathbf{W} is learned, the discriminative information of each feature is reflected by $\|\mathbf{W}_i\|_2$. This model is inspired by the feature selection works (e.g., [45], [46]). However, our work is different

from these works in several aspects. First, they learn the discriminative features based on the data structure only, while we learn the discriminative features using both the data structure and geographical structure. Second, they learn the discriminative features from the raw features of the visual content directly, while we encode the correlation between the multi-modal contents to the learning of discriminative features.

4.3. Optimization

By substituting equation (1) into (7) and relaxing the equation condition in equation (1) with minimization of the loss, we obtain

$$\begin{aligned} \min_{\mathbf{P}, \mathbf{R}, \mathbf{W}, \mathbf{M}} \mathcal{L} = & Tr(\mathbf{R}^T \mathbf{L}^e \mathbf{R}) + \varepsilon \|\mathbf{R} - \mathbf{M}^T \mathbf{W}\|_F^2 + \beta \|\mathbf{W}\|_{2,1} + \\ & \lambda (Tr[\mathbf{M} \mathbf{L}^f \mathbf{M}^T] + Tr[\mathbf{M} \mathbf{L}^o \mathbf{M}^T]) + \|\mathbf{M} - \mathbf{P} \mathbf{X}\|_F^2 + \alpha \|\mathbf{P}\|_F^2 \\ & s.t. \mathbf{R}^T \mathbf{R} = \mathbf{I}_c, \mathbf{R} \geq 0 \end{aligned} \quad (8)$$

The optimization problem (8) involves the $l_{2,1}$ norm which is non-smooth and the objective function is not convex over \mathbf{P} , \mathbf{R} , \mathbf{M} , and \mathbf{W} simultaneously. Therefore, we propose an iterative optimization algorithm. That is, we update a matrix by fixing other matrices at each iterative step.

First, by setting the derivative $\frac{\partial \mathcal{L}}{\partial \mathbf{P}} = 0$, we obtain

$$\mathbf{P} \mathbf{X} \mathbf{X}^T - \mathbf{M} \mathbf{X}^T + \alpha \mathbf{P} = 0 \quad (9)$$

$$\mathbf{P} = \mathbf{M} \mathbf{X}^T (\mathbf{X} \mathbf{X}^T + \alpha \mathbf{I}_d)^{-1} \quad (10)$$

By setting the derivative $\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = 0$, we obtain

$$\varepsilon \mathbf{M} \mathbf{M}^T \mathbf{W} - \varepsilon \mathbf{M} \mathbf{R} + \beta \mathbf{D} \mathbf{W} = 0 \quad (11)$$

$$\mathbf{W} = \mathbf{Q} \mathbf{M} \mathbf{R} \quad (12)$$

where $\mathbf{Q} = (\mathbf{M} \mathbf{M}^T + \frac{\beta}{\varepsilon} \mathbf{D})^{-1}$, and $\mathbf{D} \in \mathcal{R}^{z \times z}$ is a diagonal matrix with $D_{ii} = \frac{1}{2\|\mathbf{W}_i\|_2}$ [41].

Then, by calculating $\frac{\partial \mathcal{L}}{\partial \mathbf{M}}$ and following [38], [39], we introduce the multiplicative updating rules

as:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{M}} = 2(\lambda \mathbf{M}(\mathbf{L}^f + \mathbf{L}^o + \mathbf{I}_n) + \varepsilon \mathbf{W}(\mathbf{W}^T \mathbf{M} - \mathbf{R}^T) - \lambda \mathbf{P} \mathbf{X}) \quad (13)$$

$$M_{ij} = M_{ij} \frac{(\varepsilon \mathbf{W} \mathbf{R}^T + \lambda \mathbf{P} \mathbf{X})_{ij}}{(\varepsilon \mathbf{W} \mathbf{W}^T \mathbf{M} + \lambda \mathbf{M}(\mathbf{L}^f + \mathbf{L}^o + \mathbf{I}_n))_{ij}} \quad (14)$$

The optimization of \mathbf{R} has constraint conditions. We employ the Lagrange function method to update \mathbf{R} and obtain the following optimization problem *w.s.t.* \mathbf{R} .

$$\begin{aligned} \min_{\mathbf{R}} \mathcal{L} = & Tr(\mathbf{R}^T \mathbf{L}^e \mathbf{R}) + \varepsilon \|\mathbf{R} - \mathbf{M}^T \mathbf{W}\|_F^2 \\ \text{s.t. } & \mathbf{R}^T \mathbf{R} = \mathbf{I}_c, \mathbf{R} \geq 0 \end{aligned} \quad (15)$$

By relaxing the orthogonal constraint, the above optimization problem can be rewritten as follows:

$$\min_{\mathbf{R} \geq 0} \mathcal{L} = Tr(\mathbf{R}^T \mathbf{L}^e \mathbf{R}) + \varepsilon \|\mathbf{R} - \mathbf{M}^T \mathbf{W}\|_F^2 + \kappa \|\mathbf{R}^T \mathbf{R} - \mathbf{I}_c\|_F^2 \quad (16)$$

where κ is a parameter to control the orthogonality condition. In practice, κ should be large enough to insure the condition satisfied. Let ψ_{ij} be the Lagrange multiplier for the constraint $R_{ij} \geq 0$, and then the Lagrange function is:

$$\begin{aligned} & Tr(\mathbf{R}^T \mathbf{L}^e \mathbf{R}) + \varepsilon \|\mathbf{R} - \mathbf{M}^T \mathbf{W}\|_F^2 \\ & + \kappa \|\mathbf{R}^T \mathbf{R} - \mathbf{I}_c\|_F^2 + Tr[\Psi \mathbf{R}^T] \end{aligned} \quad (17)$$

By substituting \mathbf{W} with Eq. (12) and setting the derivative of the Lagrange function with respect to \mathbf{R} to 0, we obtain:

$$2(\mathbf{H} \mathbf{R} + \kappa \mathbf{R}(\mathbf{R}^T \mathbf{R} - \mathbf{I}_c)) + \Psi = 0 \quad (18)$$

where $\mathbf{H} = \mathbf{L}^e + \varepsilon(\mathbf{I}_n - \mathbf{M}^T \mathbf{Q} \mathbf{M})^2$. Applying the Karush-Kuhn-Tuckre (KKT) condition [37] $\psi_{ij} R_{ij} = 0$, a updating rule can be obtained:

$$2(\mathbf{H} \mathbf{R} + \kappa \mathbf{R}(\mathbf{R}^T \mathbf{R} - \mathbf{I}_c))_{ij} R_{ij} + \Psi_{ij} R_{ij} = 0 \quad (19)$$

$$R_{ij} = R_{ij} \frac{(\kappa \mathbf{R})_{ij}}{(\mathbf{H} \mathbf{R} + \kappa \mathbf{R} \mathbf{R}^T \mathbf{R})_{ij}} \quad (20)$$

Then, \mathbf{R} is normalized with $(\mathbf{R}^T \mathbf{R})_{ii} = 1$. Finally, the iterative algorithm to solve the proposed formulation is summarized in Algorithm 1. The convergence analysis is presented in the following section. The complexity of the proposed algorithm is briefly discussed as follows. In our model, $c \ll d \ll n$, and $z \leq d$. The complexity of calculating the inverse of a few matrices

for \mathbf{P} and \mathbf{Q} are $O(d^3)$ and $O(z^3)$ respectively. In each iteration step, the complexity for updating \mathbf{H} is $O(z^2n + zn^2)$, the complexity for updating \mathbf{P} is $O(zdn + nd^2)$, the complexity for updating \mathbf{M} is $O(z^2n + zdn + zn^2)$, the complexity for updating \mathbf{R} is $O(z^2n + zdn + zn^2)$, and the complexity for updating \mathbf{W} is $O(z^2n)$. The overall cost is $O(t(z^2n + zdn + zn^2 + nd^2))$, where t is the number of iterations.

Algorithm 1 LSGSA

Require: Matrices \mathbf{X} , \mathbf{G} , and \mathbf{Y} of the geo-tagged images; Parameters ε , α , β , λ , c , z .

Ensure: \mathbf{W} , \mathbf{P} , \mathbf{R} .

- 1: Calculate the Laplacian matrices \mathbf{L}^f , \mathbf{L}^o , \mathbf{L}^e based on visual similarity graph, text similarity graph, and geographical neighbour graph;
 - 2: Initialize \mathbf{R}_0 and the diagonal matrix \mathbf{D}_0 ; Set the iteration step $t=0$;
 - 3: **Repeat**
 - 4: $\mathbf{P}_{t+1} = \mathbf{M}_t \mathbf{X}^T (\mathbf{X} \mathbf{X}^T + \alpha \mathbf{I}_d)^{-1}$;
 - 5: $(M_{ij})_{t+1} = (M_{ij})_t \frac{(\varepsilon \mathbf{W} \mathbf{R}^T + \lambda \mathbf{P} \mathbf{X})_{ij}}{(\varepsilon \mathbf{W} \mathbf{W}^T \mathbf{M} + \lambda \mathbf{M} (\mathbf{L}^f + \mathbf{L}^o + \mathbf{I}_n))_{ij}}$;
 - 6: $\mathbf{Q}_t = (\mathbf{M}_{t+1} \mathbf{M}_{t+1}^T + \frac{\beta}{\varepsilon} \mathbf{D}_t)^{-1}$;
 - 7: $\mathbf{H}_t = \mathbf{L}^e + \varepsilon (\mathbf{I}_n - \mathbf{M}_{t+1}^T \mathbf{Q}_t \mathbf{M}_{t+1})^2$;
 - 8: $(R_{ij})_{t+1} = (R_{ij})_t \frac{(\kappa \mathbf{R}_t)_{ij}}{(\mathbf{H}_t \mathbf{R}_t + \kappa \mathbf{R}_t \mathbf{R}_t^T \mathbf{R}_t)_{ij}}$;
 - 9: $\mathbf{W}_{t+1} = \mathbf{Q}_t \mathbf{M}_{t+1} \mathbf{R}_{t+1}$;
 - 10: $\mathbf{D}_{t+1} = \text{diag}(\frac{1}{2\|(\mathbf{W}_{t+1})_1\|_2}, \dots, \frac{1}{2\|(\mathbf{W}_{t+1})_d\|_2})$;
 - 11: $t=t+1$;
 - 12: **until** Convergence criterion satisfied
-

4.4. Convergence Analysis

The proposed iterative procedure in Algorithm 1 can be verified to converge to the optimal solutions by the following theorem.

Theorem 1. The alternative updating rules in Algorithm 1 monotonically decrease the objective function value of equation (8) in each iteration.

Proof. In the iterative procedure, for \mathbf{P} , \mathbf{M} , \mathbf{R} , and \mathbf{W} we update one while keeping the other

two fixed. Let us denote

$$\begin{aligned} \mathfrak{R}(\mathbf{P}, \mathbf{M}, \mathbf{R}, \mathbf{W}) = & Tr(\mathbf{R}^T \mathbf{L}^e \mathbf{R}) + \varepsilon \|\mathbf{R} - \mathbf{M}^T \mathbf{W}\|_F^2 + \\ & \lambda(Tr[\mathbf{M} \mathbf{L}^f \mathbf{M}^T] + Tr[\mathbf{M} \mathbf{L}^o \mathbf{M}^T]) + \|\mathbf{M} - \mathbf{P} \mathbf{X}\|_F^2 + \alpha \|\mathbf{P}\|_F^2 \\ & + \kappa \|\mathbf{R}^T \mathbf{R} - \mathbf{I}_c\|_F^2 + \beta \|\mathbf{W}\|_{2,1} \end{aligned} \quad (21)$$

320 From the analysis, equation (8) can be relaxed into the following problem

$$\min_{\mathbf{P}, \mathbf{M}, \mathbf{R}, \mathbf{W} \geq 0} \mathfrak{R}(\mathbf{P}, \mathbf{M}, \mathbf{R}, \mathbf{W}) \quad (22)$$

In the t -th step, with \mathbf{W}_t , \mathbf{M}_t , and \mathbf{R}_t fixed, we can obtain from equation (10)

$$\mathbf{P}_{t+1} = \arg \min_{\mathbf{P}} \mathfrak{R}(\mathbf{P}, \mathbf{M}_t, \mathbf{R}_t, \mathbf{W}_t) \quad (23)$$

Thus we obtain

$$\mathfrak{R}(\mathbf{P}_{t+1}, \mathbf{M}_t, \mathbf{R}_t, \mathbf{W}_t) \leq \mathfrak{R}(\mathbf{P}_t, \mathbf{M}_t, \mathbf{R}_t, \mathbf{W}_t) \quad (24)$$

With \mathbf{W}_t , \mathbf{R}_t , and \mathbf{P}_{t+1} fixed, by introducing an auxiliary function of $\mathfrak{R}(\mathbf{P}_{t+1}, \mathbf{M}_t, \mathbf{R}_t, \mathbf{W}_t)$ as [38] and [39], it is easy to prove

$$\mathfrak{R}(\mathbf{P}_{t+1}, \mathbf{M}_{t+1}, \mathbf{R}_t, \mathbf{W}_t) \leq \mathfrak{R}(\mathbf{P}_{t+1}, \mathbf{M}_t, \mathbf{R}_t, \mathbf{W}_t) \quad (25)$$

325 Similarly, with \mathbf{W}_t , \mathbf{M}_{t+1} , and \mathbf{P}_{t+1} fixed, by introducing an auxiliary function of $\mathfrak{R}(\mathbf{P}_{t+1}, \mathbf{M}_{t+1}, \mathbf{R}_t, \mathbf{W}_t)$ as [38] and [39], it is also easy to prove

$$\mathfrak{R}(\mathbf{P}_{t+1}, \mathbf{M}_{t+1}, \mathbf{R}_{t+1}, \mathbf{W}_t) \leq \mathfrak{R}(\mathbf{P}_{t+1}, \mathbf{M}_{t+1}, \mathbf{R}_t, \mathbf{W}_t) \quad (26)$$

With \mathbf{P}_{t+1} , \mathbf{M}_{t+1} , and \mathbf{R}_{t+1} fixed, it can be easily verified that equation (12) is the solution to the following problem

$$\min_{\mathbf{W}} \varepsilon \|\mathbf{R}_{t+1} - \mathbf{M}_{t+1}^T \mathbf{W}\|_F^2 + \beta Tr[\mathbf{W}^T \mathbf{D}_t \mathbf{W}] \quad (27)$$

For the ease of representation, let us denote $h(\mathbf{W}) = \varepsilon \|\mathbf{R}_{t+1} - \mathbf{M}_{t+1}^T \mathbf{W}\|_F^2$. It can be verified that

$$\begin{aligned}
 \mathbf{W}_{t+1} &= \arg \min_{\mathbf{W}} h(\mathbf{W}) + \beta \text{Tr}[\mathbf{W}^T \mathbf{D}_t \mathbf{W}] \\
 &\Rightarrow h(\mathbf{W}_{t+1}) + \beta \text{Tr}[\mathbf{W}_{t+1}^T \mathbf{D}_t \mathbf{W}_{t+1}] \\
 &\leq h(\mathbf{W}_t) + \beta \text{Tr}[\mathbf{W}_t^T \mathbf{D}_t \mathbf{W}_t] \\
 &\Rightarrow h(\mathbf{W}_{t+1}) + \beta \sum_i \frac{\|(\mathbf{W}_{t+1})_i\|_2^2}{2\|(\mathbf{W}_t)_i\|_2^2} \\
 &\leq h(\mathbf{W}_t) + \beta \sum_i \frac{\|(\mathbf{W}_t)_i\|_2^2}{2\|(\mathbf{W}_t)_i\|_2^2} \\
 &\Rightarrow h(\mathbf{W}_{t+1}) + \beta \|\mathbf{W}_{t+1}\|_{2,1} - \\
 &\quad \beta \left(\|\mathbf{W}_{t+1}\|_{2,1} - \sum_i \frac{\|(\mathbf{W}_{t+1})_i\|_2^2}{2\|(\mathbf{W}_t)_i\|_2^2} \right) \\
 &\leq h(\mathbf{W}_t) + \beta \|\mathbf{W}_t\|_{2,1} - \\
 &\quad \beta \left(\|\mathbf{W}_t\|_{2,1} - \sum_i \frac{\|(\mathbf{W}_t)_i\|_2^2}{2\|(\mathbf{W}_t)_i\|_2^2} \right)
 \end{aligned} \tag{28}$$

330 According to the Lemmas in [46], $\|\mathbf{W}_{t+1}\|_{2,1} - \sum_i \frac{\|(\mathbf{W}_{t+1})_i\|_2^2}{2\|(\mathbf{W}_t)_i\|_2^2} \leq \|\mathbf{W}_t\|_{2,1} - \sum_i \frac{\|(\mathbf{W}_t)_i\|_2^2}{2\|(\mathbf{W}_t)_i\|_2^2}$, we have

$$h(\mathbf{W}_{t+1}) + \beta \|\mathbf{W}_{t+1}\|_{2,1} \leq h(\mathbf{W}_t) + \beta \|\mathbf{W}_t\|_{2,1} \tag{29}$$

Therefore, we obtain

$$\mathfrak{R}(\mathbf{P}_{t+1}, \mathbf{M}_{t+1}, \mathbf{R}_{t+1}, \mathbf{W}_{t+1}) \leq \mathfrak{R}(\mathbf{P}_{t+1}, \mathbf{M}_{t+1}, \mathbf{R}_{t+1}, \mathbf{W}_t) \tag{30}$$

Finally, based on equations (26), (27), (28), and (32), we can arrive at

$$\begin{aligned}
 \mathfrak{R}(\mathbf{P}_{t+1}, \mathbf{M}_{t+1}, \mathbf{R}_{t+1}, \mathbf{W}_{t+1}) &\leq \mathfrak{R}(\mathbf{P}_{t+1}, \mathbf{M}_{t+1}, \mathbf{R}_{t+1}, \mathbf{W}_t) \\
 &\leq \mathfrak{R}(\mathbf{P}_{t+1}, \mathbf{M}_{t+1}, \mathbf{R}_t, \mathbf{W}_t) \leq \mathfrak{R}(\mathbf{P}_{t+1}, \mathbf{M}_t, \mathbf{R}_t, \mathbf{W}_t) \\
 &\leq \mathfrak{R}(\mathbf{P}_t, \mathbf{M}_t, \mathbf{R}_t, \mathbf{W}_t)
 \end{aligned} \tag{31}$$

Thus, the objective function monotonically decreases using the updating rules in Algorithm 1.

335 5. Geo-guided Landmark Tagging

After the learning process, all the images are clustered into c clusters, and the geographically discriminative information is reflected by the learned matrix \mathbf{W} . Instead of assuming all

the features are identical in the existing approaches [48], [42], we leverage the discriminative information and the geographical structure analysis for landmark tagging. Let the visual feature matrix of the training images be $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_c]$, where \mathbf{X}_i denotes the i th cluster of images. Given a query landmark image I_q with the visual feature \mathbf{x}_q , the reconstruction with the Group Lasso [53] can be formulated as follows:

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \|\tilde{\mathbf{W}}(\mathbf{P}\mathbf{x}_q - \sum_{g=1}^c \mathbf{P}\mathbf{X}_g \boldsymbol{\theta}_g)\|_F^2 + \eta \sum_{g=1}^c h_g \|\boldsymbol{\theta}_g\|_2 \quad (32)$$

where $\tilde{\mathbf{W}} = \text{diag}(\|\mathbf{W}_1\|_2, \dots, \|\mathbf{W}_z\|_2)$ is a diagonal matrix for weighting the reconstruction residual for each entry in the feature vector, $\boldsymbol{\theta}_g$ denotes the reconstruction coefficient corresponding to \mathbf{X}_g , and h_g is the weight of each cluster which is inversely proportional to the importance of the image cluster.

The first term of equation (32) measures the reconstruction error, and the second term provides the constraint of geo-guided group sparsity. Usually, geographically adjacent images are more likely to cover the same scene. Therefore, the images located nearby the query image are more important in determining the result tags. We use the distance between the locations of the query image and the centroid of each cluster as the weight h_g . The iterative algorithm [46] is used to solve the joint $l_{2,1}$ -norm minimization problem.

On the other side, in most of the generic image tagging approaches, all the tags are considered equally, which leads to that the recommended tags might be dominated by the popular tags. However, in the circumstance of landmark tagging, there are many rare tags that are mainly related to the specific landmarks. These landmark-specific tags may have a low recall in the tagging result. It has been found that dealing with rare tags can improve tagging performance [5], [33]. We introduce a weighting method to highlight the rare tags. For the query landmark image I_q , the weight of the i -th tag t_i is computed as following:

$$e_i = \frac{1}{|C(i)|} \sum_{u \in C(i)} \frac{\sum_{I_k \in \Psi(u, t_i)} \text{sim}(I_q, I_k)}{|\Psi(u, t_i)|} \exp(-\text{dist}(u, I_q)/a) \quad (33)$$

where $C(i)$ denotes the set of image clusters that contain the tag t_i , $\Psi(u, t_i)$ denotes the set of images tagged with t_i in u , $\text{sim}(I_q, I_k)$ denotes the visual similarity between images I_q and I_k , and a is a constant to control the effect of the distance value $\text{dist}(u, I_q)$ between the location of I_q and the centroid of u . The intuition of this formula includes several aspects. First, the tag is more important if its associated images are more similar to the query image. Second, the tag

365 which is mostly appear in the regions nearby the query image is more important to the query
 image. Then we predict the tags of the query image based on the reconstruction coefficients and
 the weight values. Suppose $\hat{\theta}$ is the solution of Eq. (32), the tag indicator vector \mathbf{y}_q of the query
 image can be obtained as: $\mathbf{y}_q = \mathbf{B}\mathbf{Y}\hat{\theta}$, where $\mathbf{B} = \text{diag}(e_1, e_2, \dots, e_v)$. We can select the top- k
 tags to tag the query image by sorting all tags according to the value of $y_{qi}(i = 1, \dots, m)$ in
 370 descending order.

6. Experiments

In this section, we conduct experiments to assess the effectiveness of the proposed model
 LSGSA. Through the experiments, we aim to answer:

- How effective is the discriminative information of image features?
- 375 • How effective is the proposed image tagging approach compared with existing approaches?
- How sensitivity are the parameters to the performance of the approach?

We begin by introducing the experiment preparation and then analyse the experiments.

6.1. Experiment Preparation

We use two open benchmark image datasets in our experiments, i.e., MediaEval2012 [47] and
 380 NUS-WIDE [3]. MediaEval2012 is a community-driven benchmark and is run by the MediaEval
 organizing committee. MediaEval2012 contains geo-tagged Flickr images randomly sampled
 with a method that attempts to maintain coverage of the globe. Since it doesn't include the
 raw images and many of the images have been removed in the network, we download about one
 million images with their tag lists using the links in the meta-data of MediaEval2012 from Flickr.
 385 To evaluate tagging performance, we remove the images with less than 5 tags. Consequently, we
 obtain a dataset with the remaining 409,787 images. The second dataset NUS-WIDE is created
 by the lab for media search in National University of Singapore. It contains about 50 thousands
 of geo-tagged images, and each image is annotated with user tags in a uncontrolled manner.
 Tagging performances are also evaluated based on the 81 concepts manually defined in NUS-
 390 WIDE, where the ground-truth of these concepts for all images are provided for evaluation.

For the visual features, we adopted the 4096-D DeCAF generic visual features [34], which is
 the activations of the 6-th layer of a deep CNN trained in a fully supervised fashion on ImageNet

[40]. The features have been demonstrated to be effective on image benchmark datasets. For the images of our dataset, we normalized the visual features into a zero-mean unit-variance Gaussian distribution. For the textual content, we remove the noisy and misspelt image tags, e.g., the tags which are assigned to less than 10 images or more than 5% of the total images. In the experiments, we use 10-fold cross validation method to evaluate the performance, and the traditional criterion are used, i.e., the average precision (AP), recall (AR), and $F1$ at top- k annotated tags [48]:

$$\begin{aligned} AP &= \frac{1}{N_{test}} \sum_{i=1}^{i=N_{test}} \frac{T_k \cap H}{T_k} \\ AR &= \frac{1}{N_{test}} \sum_{i=1}^{i=N_{test}} \frac{T_k \cap H}{H} \\ F1 &= \frac{2 \times AP \times AR}{AP + AR} \end{aligned} \quad (34)$$

where N_{test} denotes the total number of images used to evaluated the performance, T_k denotes the set of top- k results returned by the approach, and H denotes the ground-truth result. On the other side, since the tags provided by Web users may contains many noisy ones, we also use the manually labelled tags in NUS-WIDE to measure the performance more precisely. We average the APs over the 81 manually labelled concepts in NUS-WIDE to create the Mean Average Precision (MAP), which is an overall performance measure.

We compare LSGSA against two types of state-of-the-art approaches for image tagging. The first type of approaches are the geo-based image tagging methods, i.e., **GeoTag** [48], **MFGSTag** [42], and **GeoSVM** [71]. **GeoTag** uses a probability model to select the tags that frequently appear in the visually similar and geographically adjacent images. **MFGSTag** propose two spaces to make the visual features of an image similar to its text features, and the similarity-based method is used to select geo-specific tags from geographically adjacent images. **GeoSVM** [71] learns the geo-aware tag features from the geo-tagged images. The second type of approaches annotate images without considering the geographical information, i.e., **GSTag** [54] which is based on feature selection with regularization of group sparsity, **NMF-KNN** [36] which is based on multi-view non-negative matrix factorization, and **VNN** [58] which recommends the tags that obtains the most number of votes from the neighbours.

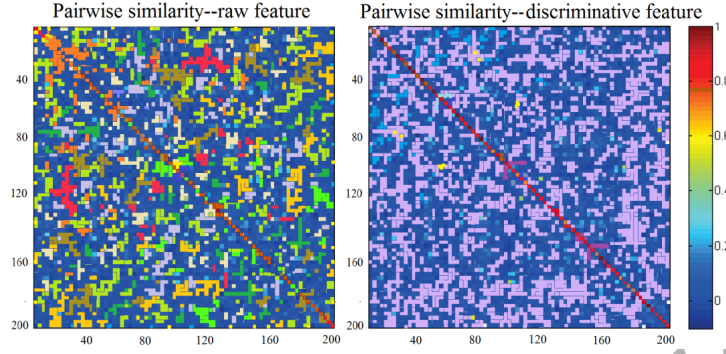


Figure 2: Comparison between the pairwise similarity of two types of features

6.2. Effectiveness of Discriminative Features

First, we study the effectiveness of \mathbf{W} learned by our model in discriminating the geographical clusters and accordingly answer the first question asked in the beginning of this section. We use the experiment result of NUS-WIDE to illustrate the effectiveness of the learned features, in which the parameter c which denotes the number of image clusters is set to 200. Each cluster denotes a geographical region. To evaluate the performance, we compute the intra-cluster similarity and inter-cluster similarity of two types of clustering results, i.e., clustering on the raw features and the learned features respectively. The similarity is measured by Gaussian kernel method. Fig. 2 shows the examples of intra-cluster similarity and inter-cluster similarity of a subset of the 200 clusters. Different colour denotes different similarity values. We can see that the difference between the intra-cluster similarity and inter-cluster similarity of the clusters derived from the learned features is more significant than that of the clusters obtained from the raw features. It verifies that the features learned by LSGSA are more effective in discriminating the regions.

Next, we further conduct the following experiment to evaluate the effectiveness of the discriminative features in landmark tagging. We first conduct image tagging without considering the discriminative information of the features by removing the diagonal matrix $\tilde{\mathbf{W}}$ and the transformation matrix \mathbf{P} , and we name this method as NWSparse. To further illustrate the effectiveness of integrating the text information to learn the discriminative features in equation (3), we evaluate the tagging performance by removing the latent space learned by LSGSA. That is, we remove $\Upsilon(\mathbf{M})$ from equation (3) to learn the geographically discriminative features from the visual content directly. Then, the learned discrimination information of the visual features is used

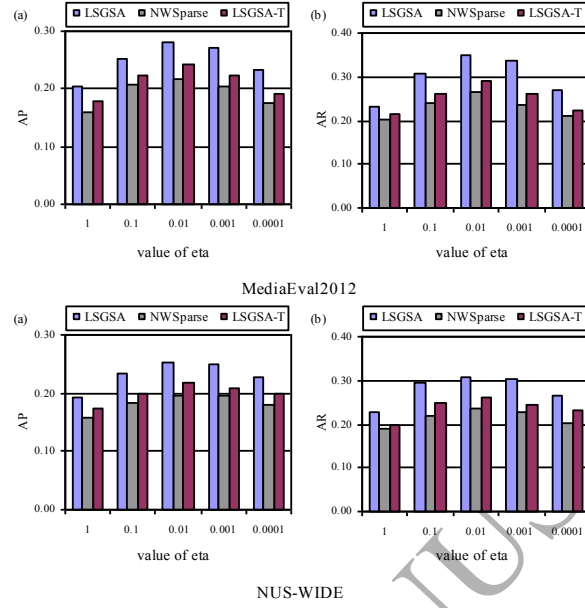


Figure 3: Effectiveness of discriminative feature learning

for landmark tagging, which is named LSGSA-T.

Fig. 3 shows the precision and recall at the top-5 tags with different values of the parameter η in Eg. (32). From this figure, several conclusions can be derived. First, it is clear that LSGSA outperforms NWSparse, which verifies the effectiveness of the discriminative features learned by our model. NWSparse employs a reconstruction to the visual content directly. Due to the heterogeneous feature spaces, it is not ensured that the image which is important to reconstruct the visual content of the query image is also similar to the query image in the semantic space. LSGSA introduces a latent space, in which the representation is more consistent with the semantic representation and more effective to model the relation between the visual content and tags. In addition, with the discriminative information, the geographical correlation between visual feature and tags can be exploited to improve landmark tagging. Second, LSGSA also outperforms LSGSA-T significantly. It indicates that the latent space can improve the performance of feature learning, and only the visual feature is not sufficient to reflect the geographical characteristic of images. Thus, a more effective way is needed to integrate the visual content and text content for discriminative feature learning.

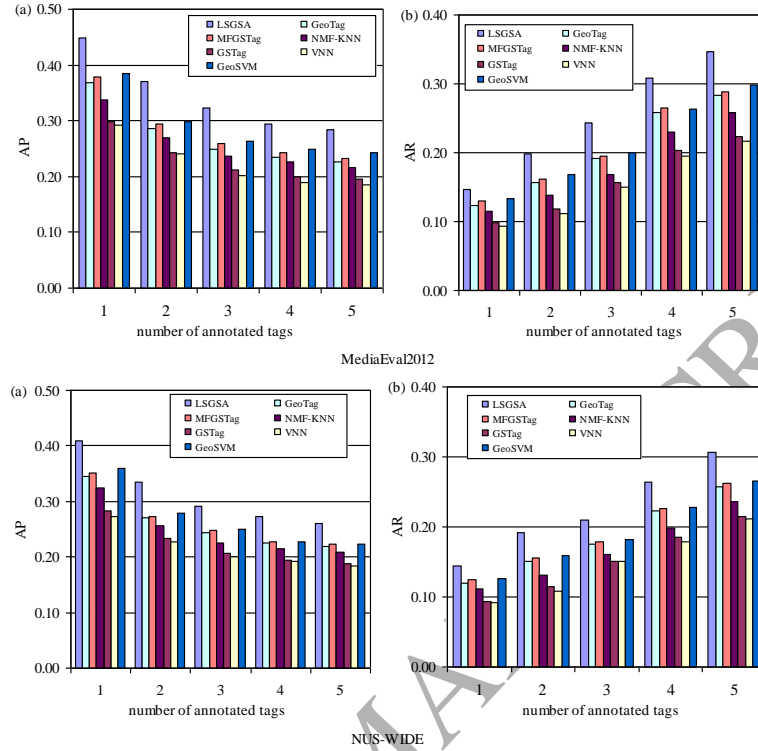


Figure 4: Performance comparison of top- k tags

6.3. Landmark Tagging Performance Evaluation

To answer the second question, we compare LSGSA against the baseline approaches. Since many users in social network upload a very small set of geotagging images, our goal is to annotate any images even the uploader has no more other images. Therefore, we implement MFGSTag by computing the relevance scores of the images to tags based on geo-location only. As for the method GeoSVM, we use it to learn the geo-aware tag features instead of the discriminative features learned by our method.

Fig. 4 shows the average precision and recall on the two datasets directly, and Fig. 5 present the MAP of the 81 concepts in NUS-WIDE. We also give some examples of the result tags in Fig. 6. From these figures, we can see that our approach achieves encouraging performance. The improvement is supposed to stem from the fact that our model exploits the geographically discriminative information of image features and use a weighting method to boost the landmark-specific tags. GeoTag mainly selects the most frequent tags weighted by visual similarity from the geo-

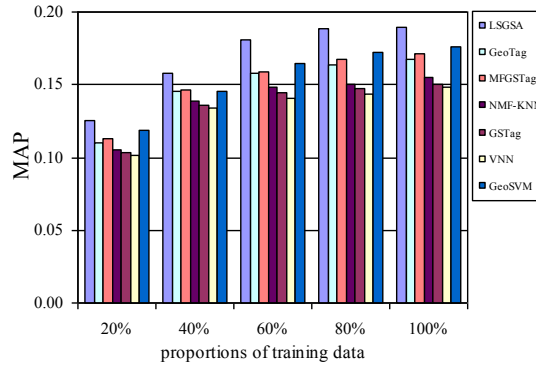


Figure 5: The comparison of MAPs of the 81 concepts using different proportions of training data

graphically neighbor images as the annotated tags, which handles all features and tags equally. MFGSTag learns two feature spaces for visual content and text content separately, without considering the geographically discriminative information of different features. Thus, many images are annotated with the common tags that might be unrelated to the landmark of the query image by these approaches. For example, the image about the Summer Palace in Beijing may be tagged with “Imperial Palace” which is frequently used to tag the images located in Beijing. GSTag proposes a feature selection method based on group sparsity, which exploits the pairwise similarity on visual content. VNN selects the tags that are mostly voted by the visual neighbours. NMF-KNN learns a generative model for the query image on the features of the nearest-neighbors in visual space and text space, respectively. These approaches learn the common relations between visual features and tags without considering the location-specific correlation between images and tags, which results in that many landmark-specific tags are absent in the result tags. Therefore, they mainly assign the query image with the tags that are associated with other visually similar images. For example, the image about Menara KL tower will be tagged with “CCTV Tower” which is a landmark located in Beijing and is visually similar to Menara KL tower. Though GeoSVM obtains better performance than other approaches by learning the geo-aware features, it mainly depended on the tag lists, which is affected by the problem of sparse representation. Our model can reduce the chance of selecting the “CCTV Tower” tag by exploiting the geographical correlation between images and tags. It demonstrates the requirement to effectively exploit geographical knowledge for landmark tagging. In Fig. 6, the results on the 81 manually labelled concepts in NUS-WIDE show that our approach also improve the performance on annotating the






Image	Groud Truth	GeoTag	MFGSTag	GSTag	NMF-KNN	VNN	GeoSVM	LSGSA
	Eiffel Tower, Paris, People, France, Building, Landermark.	Pairs, Eiffel Tower, Arc de Triomphe, Travel, People.	Pairs, Eiffel Tower, France, Travel, Street.	Landscape, Sky, travel, Eiffel Tower, Crowd.	Building, People, Sky, Eiffel Tower, Paris.	Building, Street, clouds, Eiffel Tower, Paris.	Clouds Paris, France, Eiffel Tower, People.	Eiffel Tower, Paris, France, Building, People.
	Tower bridge, Building, London, River Thames, Building.	London, UK, Tower bridge, Tower London, River.	London, Building, Tower bridge, Tower London, UK.	River, Water, Tower Bridge, Building, Cloud.	River, Water, Tower, Boat, Street.	Bridge, Water, Tower, Boat, River.	London, River Thames, Building, Boat, Water.	Tower bridge River Thames, Building, London, Water.
	Golden Gate Bridge, Hill, Seascape, Sunset, San Francisco.	Beach, Bridge, San Francisco, Fisherman, Sausalito, USA.	Golden Gate, Bridge, San Francisco, Fisherman, Water, USA.	Bridge, Sunset, Hill, Water, Bay.	Bridge, Sunset, Golden Gate, Water, Bay, Hill	Bridge, Water, Golden Gate, Car, Bay, Bridge	San Francisco, Golden Gate Bridge, Building, Water, Sunset	Golden Gate Bridge, Building, Seascape, Sunset, San Francisco
	Summer Palace, Park, Longevity Hill, Beijing, Kunning Lake.	Summer Palace, YuanMingYuan, Xiang Shan, Imperial Palace, Beijing.	Summer Palace, Beijing, YuanMingYuan, Great Wall, China.	Hill, Tree, tourists, water, Sky.	Hill, Tree, lake, Park, Boat.	Beijing, Tree, lake, Park, Boat.	Beijing, Longevity Hill, Boat, Water, Park.	Summer Palace, Longevity Hill, Boat, Beijing, Park.
	Menara, Kuala, Malaysia, KL Tower, Lumpur.	Menara, KL Tower, Malaysia, Building, Twin Towers.	Sky, Malaysia, KL Tower, Twin Towers, Building.	Building, Cloud, CCTV Tower, KL Tower, Sky.	Cloud, Tree, KL Tower, Building, CCTV Tower.	Tower, Cloud, Tree, Building, airplane.	Tower, Menara, Kuala, Malaysia, Cloud.	KL Tower, Menara, Kuala, Malaysia, Cloud.

Figure 6: Examples of landmark tagging.

manually labelled tags. Meanwhile, the approaches that exploit the geo-coordinates outperform the other approaches. Therefore, it further proves that geo clue is helpful for improving image tagging.

6.4. Parameter Sensitivity Analysis

The proposed model LSGSA requires some parameters to be set in advance, i.e., the trade-off parameters α , ε , β , and λ in Eq. (8). In the experiments, we find that the parameter α in Eq. (8) has less effect on the performance. Therefore, we focus on discussing the sensitiveness of these parameters in this subsection.

The sensitivity analysis of the parameters ε , β , and λ are studied by tuning these parameters in the corresponding ranges. The experiment results in term of $F1$ of top-5 annotated tags are illustrated in Fig. 6 and 7. By analysing these results, we have the following observations: (1) We can see that the performance change has a small difference among the two datasets. Thus, the optimal setting of the parameter values is affected by the dataset in some degree. (2) The result obtained on a given dataset indicates that both small β and large β degrade the performance as

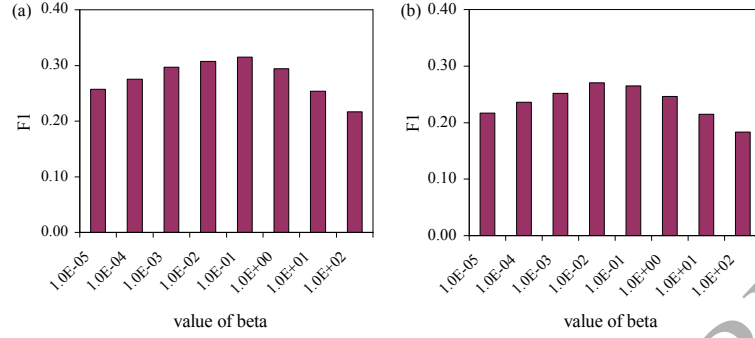


Figure 7: The sensitiveness of parameter β : (a) MediaEval2012, (b) NUS-WIDE.

shown in Fig. 6. This is because β is used to constrain how to select the features. When β is small, the noisy and correlative features can not be removed effectively, while the informative features might be removed when β is large. (3) In the proposed model, ε controls the impact of the discriminative features used for location prediction, and λ controls the impact of the learned latent space. From the results in Fig. 7, it can be observed that the best results are obtained when the two parameters are in the middle interval of the value range. That is, when the values of the two parameters are not too large or too small, the performance changes slightly. This demonstrates that the geographical information and latent space are both useful.

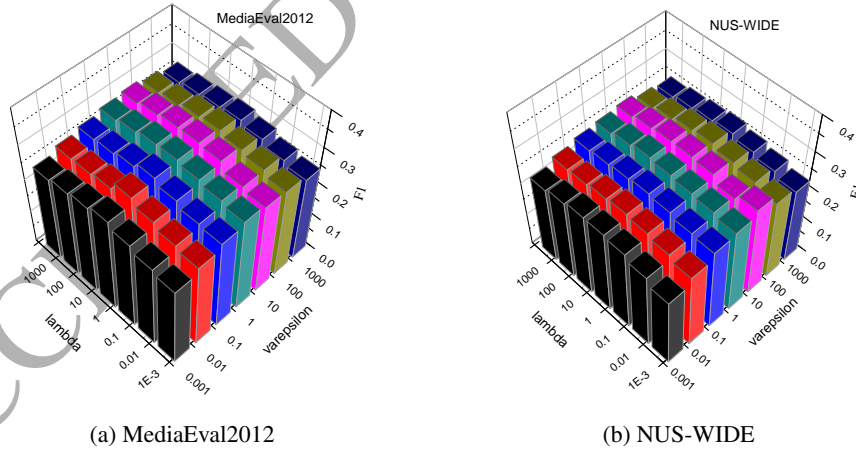


Figure 8: The sensitiveness of ε and λ .

7. Conclusion and Future Work

In this paper, we proposed a geo-guided feature learning model LSGSA for landmark tagging. A latent space is introduced to make a consistence between visual features and text representation, and geo-guided clustering combined with group sparsity analysis is proposed to learn the discriminative information of features. Then, a geo-guided sparse reconstruction method is proposed to tag images by utilizing the discriminative information of features, in which the landmark-specific tags are also boosted by a weighting method. Experimental results on real-world datasets demonstrate the effectiveness of LSGSA for landmark tagging. The novelty of this work is to tackle the analysis and application in geo-tagged multi-modal data with discriminative features automatically learned from geographically structural analysis and multi-modal data fusion. This improves the current research of image tagging which are mainly based on the raw features of visual content directly and ignores the discriminative information of features.

Our model can be used in other environment of social images even without geo-information, e.g., learning the concept-specific features by exploiting the semantic structure of image collection. There are also many potential extensions, e.g., investigating other information, like user's social activities and travel routes, for landmark tagging.

Acknowledgement

This work was supported in part by Beijing Natural Science Foundation of China (No. 4182037), in part by the National Natural Science Foundation of China (No. U1636210, No. U1636211), in part by Natural Science Foundation of Jiangsu Province of China (No. BK20171420), and in part by the Fund of the State Key Laboratory of Software Development Environment (No. SKLSDE-2017ZX-19).

References

References

- [1] J. Luo, D. Joshi, J. Yu, and A. Gallagher. Geotagging in multimedia and computer vision—a survey. *Multimedia Tools and Applications*, 51(1):187-211, 2011.
- [2] X. Zhang, X. Hu, and Z. Li. Learning Geographical Hierarchy Features for Social Image Location Prediction. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2401-2407, 2015.

- [3] T. S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proceedings of ACM Conference on Image and Video Retrieval*, Article No. 48, 2009.
- 540 [4] C. Ding, D. Zhou, X. He, and H. Zha. R1-PCA: Rotational invariant L1-norm principal component analysis for robust subspace factorization. In *Proceedings of International Conference on Machine Learning*, 281-288, 2006.
- [5] M. Guillaumin, T. Mensink, J. Verbeek, C. Schmid, Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation, In *Proceedings of IEEE International Conference on Computer Vision ICCV*, 309-316, 2009.
- 545 [6] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *The Journal of Machine Learning Research*, 3:11071135, 2003.
- [7] F. Monay and D. Gatica-Perez. PLSA-based image autoannotation: constraining the latent space. In *Proceedings of the 12th annual ACM international conference on Multimedia*, 348351, 2004.
- 550 [8] O. Yakhnenko and V. Honavar. Annotating images and image objects using a hierarchical dirichlet process model. In *Proceedings of the 9th International Workshop on Multimedia Data Mining: held in conjunction with the ACM SIGKDD*, 17, 2008.
- [9] Z. Niu, G. Hua, X. Gao, and Q. Tian. Semi-supervised relational topic model for weakly annotated image recognition in social media. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 42334240, 2014.
- 555 [10] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, 119126, 2003.
- [11] A. Tariq and H. Foroosh. Feature-Independent Context Estimation for Automatic Image Annotation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 19581965, 2015.
- 560 [12] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2:II1002, 2004.
- [13] C. Cusano, G. Ciocca, and R. Schettini. Image annotation using svm. In *Proc. Electronic Imaging*, 330-338, 2004.
- 565 [14] D. Grangier and S. Bengio. A discriminative kernel-based approach to rank images from text queries. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(8):1371-1384, 2008.
- [15] T. Hertz, A. Bar-Hillel, and D. Weinshall. Learning distance functions for image retrieval. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2:II-570, 2004.
- 570 [16] E. Moxley, J. Kleban, and B. S. Manjunath, SpiritTagger: Age-aware tag suggestion tool mined from Flickr. In *Proceedings of ACM Conference on Multimedia Information Retrieval*, 24-30, 2008.
- [17] J. Kleban, E. Moxley, J. Xu, and B. Manjunath, Global annotation on georeferenced photographs. In *Proceedings of ACM International Conference on Image and Video Retrieval*, Article 12, 2009.
- [18] E. Moxley, J. Kleban, and B. S. Manjunath, Not all tags are created equal: Learning Flickr tag semantics for global annotation. in *Proceedings of IEEE International Conference on Multimedia and Expo*, 1452-1455, 2009.
- 575

- [19] X. Li, C. G. M. Snoek, M. Worring, and A.W. M. Smeulders, Fusing concept detection and geo context for visual search. In *Proceedings of ACM International Conference on Multimedia Retrieval*, Article 4, 2012.
- [20] L. Zhu, J. Shen, H. Jin, L. Xie, R. Zheng. Landmark Classification With Hierarchical Multi-Modal Exemplar Feature. *IEEE Transactions on Multimedia*, 17(7):1-1, 2015.
- 580 [21] J. Hays, and A. A. Efros. IM2GPS: estimating geographic information from a single image. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1-8, 2008.
- [22] X. Li, M. Larson, and A. Hanjalic. Global-Scale Location Prediction for Social Images Using Geo-Visual Ranking. *IEEE Transactions on Multimedia*, 17(5):1-1, 2015.
- [23] D. M. Chen, G. Baatz, K. Koser, S. S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk. City-scale landmark identification on mobile devices. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 737-744, 2011.
- 585 [24] Y. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua, H. Neven. Tour the world: Building a web-scale landmark recognition engine. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1085-1092, 2009.
- 590 [25] Z. Cheng, J. Ren, J. Shen, and H. Miao. Building a large scale test collection for effective benchmarking of mobile landmark search. *Advances in Multimedia Modeling*, 36-46, 2013.
- [26] P. Serdyukov, V. Murdock, and R. Van Zwol. Placing Flickr Photos on a Map. In *proceedings of International Acm Sigir Conference on Research and Development in Information Retrieval*, 484-491, 2009.
- [27] Y. Li, D. J. Crandall, and D. P. Huttenlocher. Landmark classification in large-scale image collections. In *Proceedings of IEEE 12th International Conference on Computer Vision*, 1957-1964, 2009.
- 595 [28] C. Doersch, S. Singh, H. Mulam, J. Sivic, and A. Efros. What makes paris look like paris? *Acm Transactions on Graphics*, 31(4):13-15, 2012.
- [29] Q. Fang, J. Sang, and C. Xu. Discovering Geo-Informative Attributes for Location Recognition and Exploration. *ACM Trans. Multimedia Comput. Commun. Appl.*, Vol. 11, No. 1s, Article 19, 2014.
- 600 [30] Y. Wang, X. Lin, L. Wu, and W. Zhang. Effective Multi-Query Expansions: Robust Landmark Retrieval. In *Proceedings of ACM Multimedia*, 2015.
- [31] David J. Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon M. Kleinberg. Mapping the worlds photos. In *Proceedings of the 18th International Conference on World Wide Web*, 761-770, 2009.
- [32] J. Cao, Z. Huang, Y. Yang. Spatial-aware Multimodal Location Estimation for Social Images. In *Proceedings of ACM Multimedia*, 119-128, 2015.
- 605 [33] M. Chen, A. Zheng, and M. Redmond. Fast Image Tagging. In *Proceeding of the 30th International Conference on Machine Learning*, III-1274, 2013.
- [34] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *Computer science*, 50(1):815-830, 2013.
- 610 [35] G. J. F. Jones, D. Byrne, M. Hughes, N. E. O'Connor, and A. Salway. Automated annotation of landmark images using community contributed datasets and web resources. In *Proceedings of the 5th SAMT international conference on Semantic and digital media technologies*, 111-126, 2010.
- [36] M. M. Kalayeh, H. Idrees, M. Shah, NMF-KNN: image annotation using weighted multi-view non-negative matrix factorization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 184-

- 191, 2014.
- [37] H. Kuhn, and A. Tucker. Nonlinear programming. In *Berkeley Symposium on Mathematical Statistics and Probabilistics*, 1951.
- [38] L. D. Lee, and S. H. Sebastian. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401:788-791, 1999.
- [39] L. D. Lee, and S. H. Sebastian. Algorithms for nonnegative matrix factorization. *Advances in Neural Information Processing Systems*, 13(6):556-562, 2001.
- [40] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li. Imagenet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 248-255, 2009.
- [41] J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient $l_{2,1}$ -norm minimization. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, 339-348, 2009.
- [42] 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.
- [43] H. Liu, X. Wu, and S. Zhang. Feature selection using hierarchical feature clustering. In *Proceedings of ACM International Conference on Information and Knowledge Management*, 979-984, 2011.
- [44] U. V. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395-416, 2007.
- [45] Z. Ma, F. Nie, Y. Yang, J. Uijlings, and N. Sebe. Web Image Annotation Via Subspace-Sparsity Collaborated Feature Selection. *IEEE Transactions on Multimedia*, 14(4):1021-1030, 2012.
- [46] F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selection via joint $l_{2,1}$ -norms minimization. In *Proceedings of the 24th Conference on Neural Information Processing Systems*, 1813-1821, 2010.
- [47] A. Rae, and K. Pascal. Working notes for the Placing Task at MediaEval 2012. In *Proceedings of MediaEval 2012 Workshop*, 2012.
- [48] H. M. Sergieh, G. Gianini, M. Doller, H. Kosch, E. Egyed-Zsigmond, and J. M. Pinon. Geo-based automatic image annotation. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, Article 46, 2012.
- [49] A. Silva, and B. Martins. Tag recommendation for georeferenced photos. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, 57-64, 2011.
- [50] Y. Shen, and J. Fan. Leveraging Loosely-Tagged Images and Inter-Object Correlations for Tag Recommendation. In *Proceedings of ACM Conference on Multimedia*, 5-14, 2010.
- [51] K. Tang, M. Paluri, F. F. Li, R. Fergus, L. Bourdev. Improving Image Classification with Location Context. In *Proceedings of International Conference on Computer Vision*, 1008-1016, 2015.
- [52] Y. Yang, H.-T. Shen, Z. Ma, Z. Huang, and X.-F. Zhou. $l_{2,1}$ -norm regularized discriminative feature selection for unsupervised learning. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence*, 1589-1594, 2011.
- [53] M. Yuan, and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49-67, 2006.
- [54] S. Zhang, Huang, J.; Huang, Y.; Yu, Y.; Li, H.; and Metaxas, D. N. Automatic image annotation using group sparsity. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3312-3319, 2010.
- [55] X. Zhang, Z. Huang, H.-T. Shen, and Y. Yang. Automatic tagging by exploring tag information capability and

- correlation. *World Wide Web Journal*, 15(3): 233-256, 2012.
- 655 [56] Y. Zhang, B. Gong, and M. Shah. Fast Zero-Shot Image Tagging. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 5985-5994, 2016.
- [57] Cheng, Q.; Zhou, H.; and Cheng, Jie. The Fisher-Markov selector: Fast selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data. *IEEE Trans. Pattern Anal.* 33(6):1217-1233, 2011.
- 660 [58] X. Li, C. G. M. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia*, 11(7):1310-1322, 2009.
- [59] X. Qian, H. Wang, Y. Zhao, X. Hou, R. Hong, M. Wang, and Y. Tang. Image Location Inference by Multi-Saliency Enhancement. *IEEE Transactions on Multimedia*, 19(4):813-821, 2017.
- [67] J. Li, X. Qian, K. Lan, P. Qi, and A. Sharma. Improved image gps location estimation by mining salient features. *Signal Processing Image Communication*, 38, 141-150, 2015.
- 665 [61] X. Qian, Y. Zhao, and J. Han. Image Location Estimation by Salient Region Matching. *IEEE Transactions on Image Processing*, 24(11):4348, 2015.
- [62] J. Li, X. Qian, Y. Y. Tang, L. Yang, and T. Mei. Gps estimation for places of interest from social users' uploaded photos. *IEEE Transactions on Multimedia*, 15(8), 2058-2071, 2013.
- 670 [63] R. Wang, Y. Xie, J. Yang, L. Xue, M. Hu, and Q. Zhang. Large scale automatic image annotation based on convolutional neural network. *Journal of Visual Communication and Image Representation*, 49:213-224, 2017.
- [64] V. N. Murthy, S. Maji, and R. Manmatha. Automatic Image Annotation using Deep Learning Representations. In *Proceedings of ACM on International Conference on Multimedia Retrieval*, 603-606, 2015.
- [67] Z. Li, J. Liu, J. Tang, and H. Lu. Robust structured subspace learning for data representation. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 37(10), 2085-2098, 2015.
- 675 [66] Z. Li and J. Tang. Weakly Supervised Deep Matrix Factorization for Social Image Understanding. *IEEE Transactions on Image Processing*, 26(1):276-288, 2017.
- [67] Z. Li and J. Tang. Weakly Supervised Deep Metric Learning for Community-Contributed Image Retrieval. *IEEE Transactions on Multimedia*, 17(11):1989-1999, 2015.
- 680 [68] A. Habibian, T. Mensink, C. G. M. Snoek. Discovering Semantic Vocabularies for Cross-Media Retrieval. In *Proceedings of ACM International Conference on Multimedia Retrieval*, 131-138, 2015.
- [69] L. Ballan, T. Uricchio, L. Seidenari, and A. D. Bimbo. A Cross-media Model for Automatic Image Annotation. In *Proceedings of ACM International Conference on Multimedia Retrieval*, Article 73, 2014.
- [70] K. Pliakos and C. Kotropoulos. Simultaneous image tagging and geo-location prediction within hypergraph ranking framework. In *Processing of IEEE International Conference on Acoustics, Speech and Signal*, 6894-6898, 2014.
- 685 [71] S. Liao, X. Li, H. T. Shen, Y. Yang, X. Du. Tag Features for Geo-Aware Image Classification. *IEEE Transactions on Multimedia*, 17(7):1058-1067, 2015.