

Landmark Image Retrieval by Jointing Feature Refinement and Multimodal Classifier Learning

Xiaoming Zhang, Senzhang Wang, Zhoujun Li, and Shuai Ma

Abstract—Landmark retrieval is to return a set of images with their landmarks similar to those of the query images. Existing studies on landmark retrieval focus on exploiting the geometries of landmarks for visual similarity matches. However, the visual content of social images is of large diversity in many landmarks, and also some images share common patterns over different landmarks. On the other side, it has been observed that social images usually contain multimodal contents, i.e., visual content and text tags, and each landmark has the unique characteristic of both visual content and text content. Therefore, the approaches based on similarity matching may not be effective in this environment. In this paper, we investigate whether the geographical correlation among the visual content and the text content could be exploited for landmark retrieval. In particular, we propose an effective multimodal landmark classification paradigm to leverage the multimodal contents of social image for landmark retrieval, which integrates feature refinement and landmark classifier with multimodal contents by a joint model. The geo-tagged images are automatically labeled for classifier learning. Visual features are refined based on low rank matrix recovery, and multimodal classification combined with group sparse is learned from the automatically labeled images. Finally, candidate images are ranked by combining classification result and semantic consistence measuring between the visual content and text content. Experiments on real-world datasets demonstrate the superiority of the proposed approach as compared to existing methods.

Index Terms—Image classification, image geo-tagging, landmark retrieval.

I. INTRODUCTION

WITH the rising popularity of camera devices and mobile terminals, the amount of user-contributed social images with rich content like textual tags, description, and visual content is increasing rapidly. Many of these images are geo-tagged and related to landmarks, e.g., flickr.com and Picasa Web

Manuscript received September 20, 2016; revised February 13, 2017 and April 8, 2017; accepted May 27, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant U1636210, Grant U1636211, Grant 61370126, Grant 61672081, and Grant 61602237, and in part by the State Key Laboratory of Software Development Environment under Grant SKLSDE-2015ZX-11. This paper was recommended by Associate Editor Q. Zhao. (Corresponding author: Xiaoming Zhang.)

X. Zhang, Z. Li, and S. Ma are with the State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China.

S. Wang is with the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China.

This paper has supplementary downloadable multimedia material available at <http://ieeexplore.ieee.org> provided by the authors.

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TCYB.2017.2712798

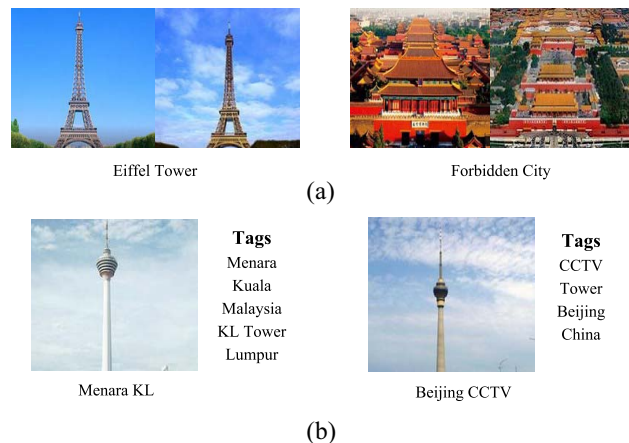


Fig. 1. Illustration of landmark images. (a) Two group of visually similar images taken at Eiffel Tower and Forbidden City. (b) Two visually similar images corresponding to different landmark but tagged with different tags.

Album. It is challenging and promising to leverage the overwhelming amount of context data and geometry information for social image applications [1], such as landmark retrieval which returns a set of images with their landmarks highly similar to that of the query image.

Comparing to the image data in traditional content-based image retrieval (CBIR) systems, landmark images have a few distinguishing characteristics. In particular, though each landmark has its own characteristics, many landmarks have similar visual content as shown in Fig. 1(b). Landmark retrieval is different from conventional image retrieval: while we can retrieve a set of similar image within the low-level feature spaces (e.g., color and texture), it is very difficult to retrieve the images with similar landmark due to the high diversity of the low-level features of each landmark. Therefore, the geometry information of landmarks is usually exploited to conduct landmark retrieval. A number of works [7], [8], [9], [59] have been proposed to conduct landmark retrieval, including data-driven or instance-based methods [7], [21], [39], and model-based methods [9], [18]. Data-driven methods propose to retrieve the most visually similar images in the landmark database. Although simple and effective, the performance of these methods depends on the similarity measure and the quality of the database. The model-based methods learn the discriminative features or intrinsic geographical patterns for landmark retrieval. Compared with data-driven methods, model-based methods has better generalization. However, they also suffer from some problems. First, most of existing methods

retrieve the landmark images based on the visual content, which requires images to contain highly discriminative visual patterns of landmarks. Although it is well recognized that images from one landmark share some similar patterns, such as the two groups of images taken at Eiffel Tower and Forbidden City in Fig. 1(a), there are also many visually similar images taken at different landmarks as shown in Fig. 1(b). Second, many methods need a well-labeled landmark dataset to training the model. Usually, the geo-tagged images are tagged with GPS coordinates which are related to the location of the landmark, since the geographically adjacent images may cover the same landmark or viewpoint. Therefore, it is desirable to automatically learning the inherent and discriminative correlation between landmarks and images with the geo-tagging information, which can largely overcomes the limitations of existing methods.

Meanwhile, the distribution of images across geographical landmarks presents unique characteristics which can contribute to landmark retrieval. First, geo-tagged images contain visual content and textual description. Both of the two types of content are related to the landmark. For example, the visual content of the images taken in landmarks of Eiffel Tower and Forbidden City present distinctive patterns as shown in Fig. 1(a). Though the visual content of the images taken in Menara KL and Beijing CCTV are similar, their image tags of these two landmarks present distinctive patterns as shown in Fig. 1(b). It has been found that language modeling approaches are particularly suitable for location recognition [2], [3]. Therefore, the combination of visual content and text content could potentially improve the performance of landmark retrieval. Second, there are many spatial-aware terms that are most indicative of geographical locations from a given text collection [4]. Many geo-tagged images are annotated with the landmark-specific tags, such as landmark name. These tags can be utilized to recognize landmark, which can then contribute to learn the latent relation between images and landmarks.

To learn the geographical relation from multimodal data for landmark image retrieval, it is quite challenging due to the following reasons. First, the multimodal contents of geo-tagged images are represented in different spaces with inherently different structures. Second, the images are freely produced by users, which results in redundant or noisy visual contents of many images. Third, image tags are also annotated manually. Not all the geo-tagged images are annotated with the unique tags of the associated landmark.

To tackle these challenges, we propose to take full advantage of the inherent relation between landmarks and the multimodal content for landmark retrieval. In particular, we investigate: 1) how to automatically label landmark images with geo-tagging information and the landmark-specific tags and 2) how to seamlessly combine both text content and visual content for the problem we are studying. Our solutions to these questions result in a new approach for landmark retrieval based on multimodal landmark classification (MMLC). In particular, the approach includes three components: 1) image labeling; 2) MMLC learning; and 3) image ranking. The spatial-aware language model and geo-based clustering method are used

to automatically label the images. To learn the latent relation between landmarks and images, a multimodal classifier is learned from the multimodal contents of images for landmark recognition. Then, the retrieval results are ranked based on the classification result and consistence measure on the text content and visual content of the candidate images. The main contributions of this paper are summarized as follows.

- 1) We explore the intrinsic nature of the geographical distribution of social images. A novel landmark retrieval method is proposed to exploit the latent relation between landmarks and multimodal content of social images, based on an effective MMLC paradigm.
- 2) A novel feature refinement module based on low-rank matrix recovery is proposed to refine visual features. An effective image landmark recognition module is proposed to learn the reliable image classifiers by exploring the latent correlation between multimodal contents of social image.
- 3) A unified model is proposed to explore the in-depth reinforcement between the two modules. Then, the landmark retrieval result is ranked based on classification result and consistence measure between visual features and text features.

The rest of this paper is structured as follows. We review the related work in Section II. Then, the problem is formally defined in Section III, and image retrieval with images containing multimodal contents is presented in Section IV. We experimentally validate the performance of our approach in Section V and conclude this paper in Section VI.

II. RELATED WORKS

With the explosive growth of geo-tagged images, landmark retrieval is an emerging research topic in multimedia application and computer vision. The related works includes geographic referencing of images, CBIR [5], [42], [43], and ensemble learning [62].

A. Geographic Referencing of Images

Geographic referencing of images include data-driven methods and model-based methods. Data-driven method determines the landmark or geographical location of the query image by retrieving the nearest neighbors from a prebuilt database. Some works construct the image databases with tree-based structure [21] or a 3-D model [19], [22] to preserve retrieval efficiency. Hays and Efros [7] presented a feature matching approach to return the K nearest neighbors with respect to the query landmark image, which represents the query images and images in database by aggregating a set of low-level features to perform landmark retrieval. Li *et al.* [11] retrieved the visually similar candidates by considering their geo-visual neighbors which are both geographically nearby and visually similar. There are some methods which estimate image location at city-scale or global scale [14], and the search method is also used for landmark recognition or classification [15], [16], [18]–[20]. Serdyukov *et al.* [2] estimated the location of Flickr image by using a language model purely based on the tags assigned by users. There are also location

estimation on videos [41], in which the videos are assigned to different cities by matching the embedded audio against the typical sounds of ambulance sirens. However, textual tags were not used in this method. Usually, these data-driven methods suffer from huge storage cost, and the performance is affected by the quality of database.

Model-based methods attempt to build models to extract the geographical patterns or discriminative features for location recognition [12], [13], [17]. In [8], a region-based recognition method is proposed to detect discriminative landmark regions at patch level, such as a set of stylistic of visual elements to characterize a city such as windows, street signs, etc., which are seen as the features for landmark retrieval. Fang *et al.* [9] presented an approach, namely GIANT, to discover both discriminative and representative mid-level attributes for landmark retrieval. Wang *et al.* [10] proposed a multiquery expansions method to retrieve landmark, which learns the discriminative patterns from the query expansion images. The multimodal hypergraph (MMHG) is proposed to combine different types of visual features for landmark image retrieval. However, these methods learn the landmark model mainly based on the visual content, which neglect the plentiful source of textual information of social images. Li *et al.* [17] and Crandall *et al.* [23] proposed to classify landmark images by linearly combining the visual features and textual tags. Cao *et al.* [24] proposed a ranking method to fuse the multiple evidences derived from textual features and visual features for image location estimation. These approaches consider different types of features independently, which cannot exploit the inherent correlation between different types of features effectively.

B. Classical Technologies Related to Image Retrieval

Traditional image retrieval techniques include text-based image retrieval and CBIR which is most related to this paper. CBIR is used to find images based on the visual content of the images such as color, texture, and shape, and the retrieved images will have visually similar appearance to the query image. The most important problem of CBIR is how to bridge the gap between low-level feature layout and high-level semantic concepts. Feature extraction and image ranking are the bases of CBIR.

Different CBIR techniques have adopted different feature extraction methods. Some CBIR technologies are mainly based on global features (e.g., color, texture, edges, and spatial information), in which color and texture can be ultimately combined together, such as texton co-occurrences matrix [42], micro-structure descriptor [45], color difference histogram [44], etc. Some other CBIR technologies are mainly based on local features. There are many famous keypoints detectors and descriptors, where SIFT is the most popular local feature representation [46]. SURF [47] or ORB [48] can be considered as an efficient alternative to SIFT. Recently, bag-of-visual words model has been used for object-based image retrieval [49]. As opposite to the classical approach that extracts an image descriptor from the original image, some approaches employ image features derived from the

compressed data stream [50], [51]. There are some studies in utilizing both visual features and text features for image retrieval. For example, the kernelized version of canonical correlation analysis (CCA) is proposed to learn a common representation for the multimodal content of image, and then the similarity-based method is used to retrieve images. However, developing an efficient method to learn the geographically discriminative information of features within CBIR framework needs to be further studied.

Unlike the extensive researches in text information retrieval, image ranking has been seldom explored in CBIR. Some approaches apply the classical batch learning [52], [53] and online learning [54] to image ranking algorithm. Some other approaches [55]–[57] propose to apply machine learning techniques (supervised or unsupervised learning) to learn a good ranking function on a single type of features or some combined features. In this paper, we propose to combine multiple types of image features and geographic referencing information to rank images.

C. Ensemble Learning

With the popular of multimodal data, classifier ensemble methods, such as bagging [63], boosting [64], random forest [66], and random subspace [65] have attracted more and more research attention. These methods have achieved good performances in areas of image and video processing [67], [68]. As compared with a single classifier, the classifier ensemble approach aims to integrate the predicted results from multiple classifiers into a unified predicted result. The classifier ensemble approaches mainly can be categorized into three classes [62].

The first class aims to design a new ensemble of classifier. For example, Garcia-Pedrajas [69] constructed a new ensemble of classifiers by means of weighted instance selection and also uses a nonlinear projection technique to construct an ensemble. A graph-based transductive multilabel ensemble classifier is proposed by Yu *et al.* [70]. To address the problems raised from the special datasets, Yu *et al.* [71] proposed a random subspace ensemble framework based on hybrid k -nearest neighbor classification to perform classification on the datasets with noisy attributes in the high-dimensional space, and Yu *et al.* [75] designed a noise immune cluster ensemble framework to tackle the challenges raised by noisy datasets. The second class of approaches try to theoretically exploring the properties of a classifier ensemble [72]. For example, Yu *et al.* [73] investigated the problem of how to select the suitable cluster structures in the ensemble which will be summarized to a more representative cluster structure. Kuncheva [74] studied how to use a kappa-error diagram to analyze the performance of classifier ensemble approaches. Yu *et al.* [62] designed a general hybrid adaptive ensemble learning framework, and applied it to address the limitations of random subspace-based classifier ensemble. It consists of two adaptive processes, i.e., base classifier competition and classifier ensemble interaction, which adjusts the weights of the base classifiers in each ensemble and to explore the optimal random subspace set simultaneously. The approaches in the third

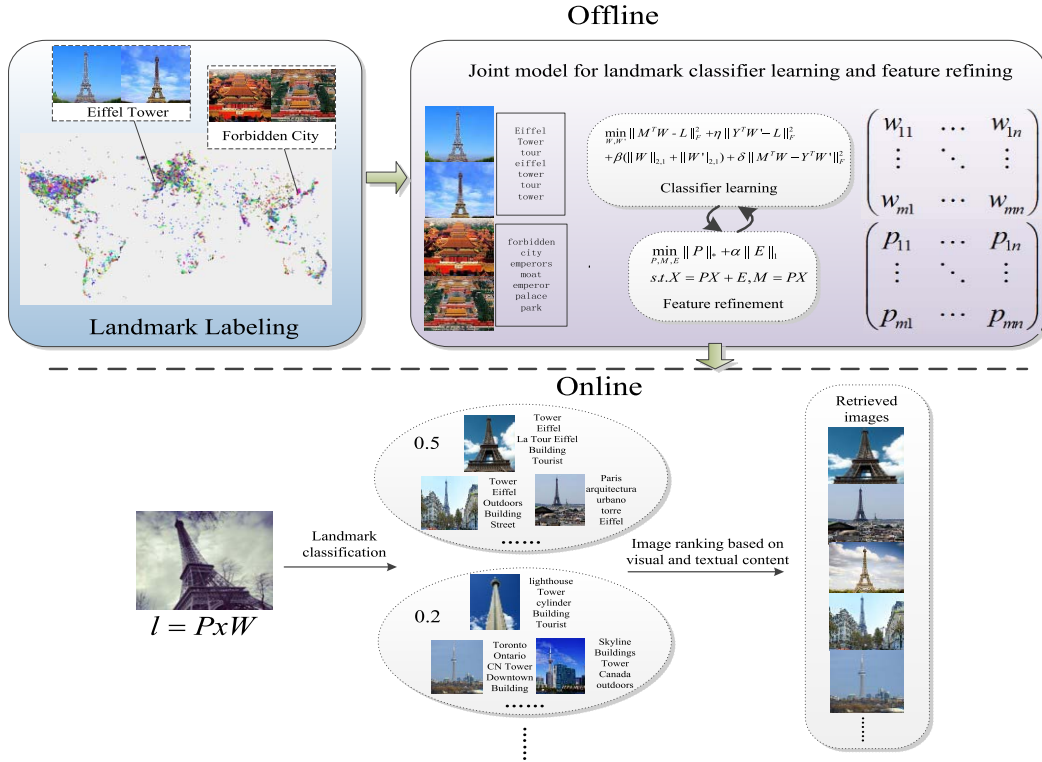


Fig. 2. Framework of our approach, which contains three main components: 1) landmark labeling for the training images; 2) a joint model to learn the landmark classifier and refine the feature of image based on the multimodal contents of images; and 3) with the predicted label information, ranking the image result list by exploiting the textual information of images. The first two components are trained offline, and the third component retrieves landmark images for the query image online.

class focus on how to apply the classifier ensemble approaches to different domains. For example, Song *et al.* [67] proposed a localized multiple kernel learning algorithm for realistic human action recognition in videos. Kuncheva *et al.* [68] investigated the suitability of the random subspace ensemble method for classification of brain images obtained through functional magnetic resonance imaging. Our approach is similar to the second class of methods, which adopts a joint model to reinforce the classifiers of visual features and textual features with each other through the correlation.

III. PROBLEM FORMULATION

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

A. Notation

The following notations are used. Matrices are denoted by boldface uppercase letters, vectors by boldface lowercase letters, and scalars by lowercase letters. For a matrix $\mathbf{A} \in \mathcal{R}^{n \times m}$, \mathbf{A}^T denotes its transpose, \mathbf{A}_i and \mathbf{A}^j denote its i th row and j th column, respectively, $\|\mathbf{A}\|_{2,1}$ denotes the $l_{2,1}$ -norm regularization [28], i.e., $\|\mathbf{A}\|_{2,1} = \sum_{i=1}^n \sqrt{\sum_{j=1}^m (\mathbf{A}_{ij})^2}$, and $\|\mathbf{A}\|_*$ denotes the nuclear norm (i.e., the sum of the singular values of \mathbf{A}).

Let \mathbf{I} denotes the training dataset with n geo-tagged images. Each geo-tagged 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 of the visual content; $\mathbf{y}_i \in \{0, 1\}^{v \times 1}$ 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 real-valued 2-D vector containing the latitude and longitude where the image is taken. Specifically, let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$ denotes the visual feature matrix, $\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_n]$ denotes the location matrix, and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n]$ denotes the tag label matrix.

With the above given notations, we formally define the problem of landmark retrieval as follows.

Given a set of geo-tagged social images \mathbf{I} in which each image contains visual features, tags, and geo-coordinate, we aim to retrieve a set of images that describe the same landmark as that of the query image I_q which contains only visual content, by exploring the geographical relation among the multimodal features.

The framework of the proposed approach is illustrated in Fig. 2, which contains three components: 1) landmark labeling; 2) multimodal classifier learning; and 3) image ranking. The first step is to automatically label each training image with the corresponding landmark based on the geo-tagging information and image tag lists. Based on the labeling information, an MMLC is learned with the low rank matrix recovery and group sparse coding. Then, for a query image, the candidate images are ranked based on the classification result and semantic consistency measure between different types of features. Note that the first two steps are offline and the third step is online.

IV. MMLC FOR LANDMARK RETRIEVAL

In this section, we give a detailed description on the proposed landmark retrieval framework, including its three components: landmark labeling, MMLC learning, and image ranking.

A. Landmark Labeling

To retrieve landmark images, we should recognize the landmark of the query image first. Therefore, we first label the images in the dataset \mathbf{I} with the corresponding landmark for landmark classifier learning. To correctly label the images, it is important to avoid potential biases and incompletely labeling. For instance, the methods [10], [16] which search for images tagged with manually selected keywords are prone to bias because one might inadvertently choose keywords corresponding to objects that are amenable to a particular image task. Also problematic is using geo-tagging information to clustering images with mean shift algorithm directly [26], where each cluster represents a landmark [23]. This is because that not all the images are geo-tagged in the same spatial scale. For example, many images are geo-tagged in the city scale while others are geo-tagged in a much smaller scale. We thus advocate an automatic technique for landmark labeling based on both textual information and geo-tagged coordinates.

First, we use mean shift algorithm to cluster the images for which the precision of the geotags is better than about a city block. In particular, we consider the latitude-longitude coordinates as a point in the plane, and then conduct a mean shift clustering on the points to identify local peaks in the image density distribution, as in [23]. The radius of the distance used in mean shift is about 100 m. Then, we select the landmark-specific tags from each landmark cluster to label other geo-tagged images which are not included in the clustering procedure. Given a landmark cluster s , for each tag t in s , the weight of t is calculated as

$$w(t, s) = N_{t,s} \cdot \log \frac{N_c}{N_t + 1} \quad (1)$$

where $N_{t,s}$ is the number of occurrences of tag t in cluster s , N_t is the number of clusters in which tag t appear, and N_c is the total number of clusters. The weight $w(t, s)$ is calculated in the way similar to “tf-idf” [60]. This is, the landmark-specific tags appear frequently in the landmark cluster and occur rarely in other clusters. We rank all the tags for each cluster based on the weight values, and top-ranked tags are selected as the landmark-specific tags. Then, a vector is constructed based on the selected tags for each cluster, where each element is represented by the weight value. To label the images which are not clustered, the cosine similarity is calculated on the textual features and each image is assigned to the most similar landmark cluster if the corresponding similarity is larger than a threshold value.

B. Multimodal Landmark Classification

With the landmark label information, we can then learn a classifier for landmark recognition. Note that each geo-tagged social image contains multiple types of content, i.e., visual

content and text description. Besides, there exist many near-duplicate or duplicate images on the social sites, and thus a more effective feature representation is needed. Therefore, we first refine the visual features, and then a landmark classifier on the multimodal content is proposed.

1) *Refining Image Features*: Usually, the near-duplicate or duplicate images have identical semantics, and their visual content should be represented by similar features. In other words, the rank of the corresponding feature matrix should be low enough. In many previous works [10], [18], [25], the landmark recognition is conducted on the raw feature matrix directly, which ignores the low-rank property of the feature matrix. One drawback of this method is that latent relation between image and landmark cannot be well explored. Based on this analysis, we propose a feature space transformation module which simultaneously explores and preserves the endowed low-rank nature of visual feature matrix

$$\begin{aligned} \min_{\mathbf{P}, \mathbf{M}, \mathbf{E}} \quad & \|\mathbf{P}\|_* + \alpha \|\mathbf{E}\|_1 \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{P}\mathbf{X} + \mathbf{E}, \quad \mathbf{M} = \mathbf{P}\mathbf{X} \end{aligned} \quad (2)$$

where \mathbf{M} is the transformed feature matrix, \mathbf{P} is the transformation matrix, and \mathbf{E} is a matrix of the error on \mathbf{M} and α is a tradeoff parameter. $\|\cdot\|_*$ denotes the nuclear norm which helps to explore the low-rank property of the transformed feature matrix. With this transformation module, we can refine the feature matrix and cleanse the noisy features for the training images. Next, we will detail an effective classification method for landmark recognition with the refined training data.

2) *Landmark Recognition*: To retrieve landmark images, an essential prerequisite is that the landmark where the query image is taken can be recognized. Therefore, in this section we focus on introducing how to learn an effective image classifier for landmark recognition. Due to the multimodal contents of social image, it may become more difficulty to learn an effective landmark classifier with the multimodal features. Many of the existing landmark retrieval approaches learn a multiclass classifier based on the visual features [9], [14], [16] or linear combination of different types of feature [17], [23], which is ineffective to exploit the geographical correlation between different types of feature.

In order to effectively explore the latent correlation among different types of features for landmark classification, we expect the learning of classifier can capture the correlation information. First, we adopt a linear model to predict the landmark label of image I_j as follows:

$$f_i(I_j) = \mathbf{w}_i^T \mathbf{P}\mathbf{x}_j = \mathbf{w}_i^T \mathbf{M} \quad (3)$$

where $\mathbf{w}_i \in \mathcal{R}^d$ is the weight vector. $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_c]$, where c is the number of landmark labels which are constructed with the mean shift algorithm as discussed above. \mathbf{W} is used to denote the weight matrix which also indicates the importance of each feature to different landmarks. Let the ground truth label matrix be $\mathbf{L} \in \{0, 1\}^{n \times c}$. Then, the classifier based on the visual features is formulated as follows:

$$\min_{\mathbf{W}} \|\mathbf{M}^T \mathbf{W} - \mathbf{L}\|_F^2 + \beta \|\mathbf{W}\|_{2,1} \quad (4)$$

where the least square loss is used for the landmark label prediction, as the project matrix \mathbf{W} is used for feature selection, a sophisticated regularizer is needed to make \mathbf{W} able to reflect the importance of different features. Thus, the $\|\mathbf{W}\|_{2,1}$ is adopted to guarantee that \mathbf{W} is sparse in rows [28], [40], which constrains the number of features to be selected since some features are unhelpful.

Besides, the textual features are also used to complement landmark recognition. A similar classifier based on textual feature matrix \mathbf{Y} is defined as follows:

$$\min_{\mathbf{W}'} \|\mathbf{Y}^T \mathbf{W}' - \mathbf{L}\|_F^2 + \beta \|\mathbf{W}'\|_{2,1} \quad (5)$$

where \mathbf{W}' is the weight matrix for the textual features.

Meanwhile, the classification results based on both types of these features should be equal to each other. We propose to use $\|\mathbf{M}^T \mathbf{W} - \mathbf{Y}^T \mathbf{W}'\|_F^2$ to penalize the diversity of the labels predicted based on different types of feature. Next, we build up the connection between the predicted labels on different features. To this end, we propose the following object function for landmark recognition with discriminative feature learning based on both visual and textual features:

$$\min_{\mathbf{W}, \mathbf{W}'} \|\mathbf{M}^T \mathbf{W} - \mathbf{L}\|_F^2 + \eta \|\mathbf{Y}^T \mathbf{W}' - \mathbf{L}\|_F^2 + \beta (\|\mathbf{W}\|_{2,1} + \|\mathbf{W}'\|_{2,1}) + \delta \|\mathbf{M}^T \mathbf{W} - \mathbf{Y}^T \mathbf{W}'\|_F^2 \quad (6)$$

where η is a tradeoff parameter. In this way, the module in (6) provides us a powerful and flexible tool for training an effective classifier to recognize the landmark. Once \mathbf{W} is learned, the discriminative information of each visual feature can also be reflected by $\|\mathbf{W}_i\|_2$. In the next part, we will elaborate a joint model which effectively combines the visual feature refinement module and the landmark classifier learning module by exploring their correlation.

3) *Joint Model*: As discussed above, we have proposed two independent modules to undertake the tasks of refining visual feature matrix as well as learning the landmark classifier, respectively. In order to perform effective image classification, a straightforward two-step approach can be used, i.e., first refine the visual features of image and then feed it together with the textual features into the learning module for landmark classification. One limitation of this approach is that the intrinsic correlation between the two modules is not well explored to reinforce the performance of each other.

Therefore, we propose a novel joint landmark classification model, termed MMLC, which simultaneously conducts the feature refinement and landmark classifier learning by exploring their intrinsic correlations. The fundamental design principle of MMLC lies in that the feature refinement module and landmark classifier learning module should form a mutually reinforcing learning loop. The refined features should be well explored to better embedded and help the learning of the landmark classifier, while the classifier learning process with both types of feature is supposed to guide a better feature refinement in return. Based on the analysis, we formulate the landmark classifier under the social image

circumstance as follows:

$$\begin{aligned} \min_{\mathbf{P}, \mathbf{M}, \mathbf{E}, \mathbf{W}, \mathbf{W}'} \quad & \|\mathbf{P}\|_* + \alpha \|\mathbf{E}\|_1 \\ & + \frac{\gamma}{2} \left(\|\mathbf{M}^T \mathbf{W} - \mathbf{L}\|_F^2 + \eta \|\mathbf{Y}^T \mathbf{W}' - \mathbf{L}\|_F^2 \right. \\ & \quad \left. + \beta (\|\mathbf{W}\|_{2,1} + \|\mathbf{W}'\|_{2,1}) \right. \\ & \quad \left. + \delta \|\mathbf{M}^T \mathbf{W} - \mathbf{Y}^T \mathbf{W}'\|_F^2 \right) \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{P}\mathbf{X} + \mathbf{E}, \quad \mathbf{M} = \mathbf{P}\mathbf{X} \end{aligned} \quad (7)$$

where γ is a tradeoff parameter which controls the balance between the two modules. The objectives of the two modules are simultaneously obtained under the joint model with their correlation substantially explored. In the next part, we will present an effective solution for the optimization of MMLC.

4) *Optimization*: It is difficulty to solve (7) directly since it is nonconvex with respect to all the variables at the same time, and the nonsmooth property of regularization makes it nontrivial to optimize the problem as a whole. To address these challenges, we devise an iterative optimization algorithm to optimize the model.

We first need to introduce a variational formulation for the $l_{2,1}$ norm. If we define $\phi(x) = \sqrt{x^2 + \varepsilon}$, the $l_{2,1}$ norm $\|\mathbf{W}\|_{2,1}$ and $\|\mathbf{W}'\|_{2,1}$ can be replaced with $\sum_i^d \phi(\|\mathbf{W}_i\|_2)$ and $\sum_i^v \phi(\|\mathbf{W}'_i\|_2)$, respectively, where d and v denote the number of rows of \mathbf{W} and \mathbf{W}' , respectively, according to the analysis for $l_{2,1}$ in [29]. ε is a smoothing term which is usually set to be a small value. It can be proved that $\phi(x)$ satisfies all the conditions as follows:

$$\begin{aligned} x &\longrightarrow \phi(x) \text{ is convex on } \mathcal{R} \\ x &\longrightarrow \phi(\sqrt{x}) \text{ is concave on } \mathcal{R}_+ \\ \phi(x) &= \phi(-x), \quad \forall x \in \mathcal{R} \\ \phi(x) &\text{ is } C^1 \text{ on } \mathcal{R} \\ \phi''(x) &> 0, \quad \lim_{x \rightarrow \infty} \phi(x)/x^2 = 0 \end{aligned} \quad (8)$$

where C^1 indicates that $\phi(x)$ is a first-order differentiable function. Then $\phi(\cdot)$ can be optimized in a half-quadratic way [30] according to the following lemma [29].

Lemma 1: Let $\phi(\cdot)$ be a function satisfying all the conditions in (8), for a fixed $\|\mathbf{x}\|_2$, there exists a dual potential function $\varphi(\cdot)$, such that

$$\phi(\|\mathbf{x}\|_2) = \inf_{e \in \mathcal{R}} \left\{ e \|\mathbf{x}\|_2^2 + \varphi(e) \right\} \quad (9)$$

where e is determined by the minimizer function $\varphi(\cdot)$ with respect to $\phi(\cdot)$, and $\inf\{\cdot\}$ denotes the inferior function.

According to Lemma 1, the object function (7) can be reformulated as follows:

$$\begin{aligned} \min_{\mathbf{P}, \mathbf{M}, \mathbf{E}, \mathbf{W}, \mathbf{W}'} \quad & \|\mathbf{P}\|_* + \alpha \|\mathbf{E}\|_1 \\ & + \frac{\gamma}{2} \left(\|\mathbf{M}^T \mathbf{W} - \mathbf{L}\|_F^2 + \eta \|\mathbf{Y}^T \mathbf{W}' - \mathbf{L}\|_F^2 \right. \\ & \quad \left. + \beta (\text{tr}(\mathbf{W}^T \mathbf{D}\mathbf{W}) + \text{tr}(\mathbf{W}'^T \mathbf{D}'\mathbf{W}')) \right. \\ & \quad \left. + \delta \|\mathbf{M}^T \mathbf{W} - \mathbf{Y}^T \mathbf{W}'\|_F^2 \right) \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{P}\mathbf{X} + \mathbf{E}, \quad \mathbf{M} = \mathbf{P}\mathbf{X} \end{aligned} \quad (10)$$

where $\mathbf{D} = \text{Diag}(\mathbf{d})$ and $\mathbf{D}' = \text{Diag}(\mathbf{d}')$. \mathbf{d} and \mathbf{d}' are auxiliary vectors of the two $l_{2,1}$ norms, respectively. The elements of \mathbf{d} and \mathbf{d}' are computed, respectively, as follows:

$$\begin{cases} d_i = \frac{1}{2\sqrt{\|\mathbf{W}_i\|_2^2 + \varepsilon}} \\ d'_i = \frac{1}{2\sqrt{\|\mathbf{W}'_i\|_2^2 + \varepsilon}} \end{cases} \quad (11)$$

where ε is a smoothing term, which is usually set to be a small constant value. We further rewrite (10) as follows:

$$\begin{aligned} \min_{\mathbf{F}, \mathbf{P}, \mathbf{M}, \mathbf{E}, \mathbf{W}, \mathbf{W}'} \quad & \|\mathbf{F}\|_* + \alpha \|\mathbf{E}\|_1 \\ & + \frac{\gamma}{2} \left(\|\mathbf{M}^T \mathbf{W} - \mathbf{L}\|_F^2 + \eta \|\mathbf{Y}^T \mathbf{W}' - \mathbf{L}\|_F^2 \right. \\ & \quad + \beta (\text{tr}(\mathbf{W}^T \mathbf{D} \mathbf{W}) + \text{tr}(\mathbf{W}'^T \mathbf{D}' \mathbf{W}')) \\ & \quad \left. + \delta \|\mathbf{M}^T \mathbf{W} - \mathbf{Y}^T \mathbf{W}'\|_F^2 \right) \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{P} \mathbf{X} + \mathbf{E}, \quad \mathbf{M} = \mathbf{P} \mathbf{X}, \quad \mathbf{F} = \mathbf{P}. \end{aligned} \quad (12)$$

Note that \mathbf{D} and \mathbf{D}' are actually depended on \mathbf{W} and \mathbf{W}' , respectively. To handle this problem, we design an iterative algorithm, which updates \mathbf{D} and \mathbf{D}' in each iteration with \mathbf{W} and \mathbf{W}' of the previous iteration. Then, the problem in (12) can be solved via exact or inexact augmented Lagrange multiplier method [31]

$$\begin{aligned} \min_{\mathbf{F}, \mathbf{P}, \mathbf{M}, \mathbf{E}, \mathbf{W}, \mathbf{W}', \mathbf{G}_1, \mathbf{G}_2, \mathbf{G}_3} \quad & \|\mathbf{F}\|_* + \alpha \|\mathbf{E}\|_1 \\ & + \frac{\gamma}{2} \left(\|\mathbf{M}^T \mathbf{W} - \mathbf{L}\|_F^2 + \eta \|\mathbf{Y}^T \mathbf{W}' - \mathbf{L}\|_F^2 \right. \\ & \quad + \beta (\text{tr}(\mathbf{W}^T \mathbf{D} \mathbf{W}) + \text{tr}(\mathbf{W}'^T \mathbf{D}' \mathbf{W}')) \\ & \quad \left. + \delta \|\mathbf{M}^T \mathbf{W} - \mathbf{Y}^T \mathbf{W}'\|_F^2 \right) \\ & + \text{Tr}(\mathbf{G}_1^T (\mathbf{X} - \mathbf{P} \mathbf{X} - \mathbf{E}) + \mathbf{G}_2^T (\mathbf{M} - \mathbf{P} \mathbf{X}) + \mathbf{G}_3^T (\mathbf{F} - \mathbf{P})) \\ & + \frac{\theta}{2} \left(\|\mathbf{X} - \mathbf{P} \mathbf{X} - \mathbf{E}\|_F^2 + \|\mathbf{M} - \mathbf{P} \mathbf{X}\|_F^2 + \|\mathbf{F} - \mathbf{P}\|_F^2 \right) \end{aligned} \quad (13)$$

where \mathbf{G}_1 , \mathbf{G}_2 , and \mathbf{G}_3 are the Lagrange multipliers and θ is a tradeoff parameter. That is, we update a matrix by fixing other matrices at each step.

- 1) *Update \mathbf{W} and \mathbf{W}' by Fixing Others:* When we fix all others except \mathbf{W} and \mathbf{W}' , we have the following subproblem:

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{W}'} \quad & \|\mathbf{M}^T \mathbf{W} - \mathbf{L}\|_F^2 + \eta \|\mathbf{Y}^T \mathbf{W}' - \mathbf{L}\|_F^2 \\ & + \beta (\text{tr}(\mathbf{W}^T \mathbf{D} \mathbf{W}) + \text{tr}(\mathbf{W}'^T \mathbf{D}' \mathbf{W}')) \\ & + \delta \|\mathbf{M}^T \mathbf{W} - \mathbf{Y}^T \mathbf{W}'\|_F^2. \end{aligned} \quad (14)$$

By setting the derivative of the above object function with respect to \mathbf{W} and \mathbf{W}' to zero, respectively, we have

$$\mathbf{W} = ((1 + \delta)\mathbf{M}\mathbf{M}^T + \beta\mathbf{D})^{-1}(\mathbf{M}\mathbf{L} + \delta\mathbf{M}\mathbf{Y}^T\mathbf{W}') \quad (15)$$

$$\mathbf{W}' = ((\eta + \delta)\mathbf{Y}\mathbf{Y}^T + \beta\mathbf{D}')^{-1}(\eta\mathbf{Y}\mathbf{L} + \delta\mathbf{Y}\mathbf{M}^T\mathbf{W}). \quad (16)$$

- 2) *Update \mathbf{M} by Fixing Others:* When we fix all others except \mathbf{M} , the problem becomes

$$\begin{aligned} \min_{\mathbf{M}} \quad & \frac{\gamma}{2} \|\mathbf{M}^T \mathbf{W} - \mathbf{L}\|_F^2 + \frac{\gamma\delta}{2} \|\mathbf{M}^T \mathbf{W} - \mathbf{Y}^T \mathbf{W}'\|_F^2 \\ & + \text{Tr}(\mathbf{G}_2^T (\mathbf{M} - \mathbf{P} \mathbf{X})) + \frac{\theta}{2} \|\mathbf{M} - \mathbf{P} \mathbf{X}\|_F^2. \end{aligned} \quad (17)$$

Then, we can update \mathbf{M} with the following closed-form solution:

$$\mathbf{M} = (\gamma(1 + \delta)\mathbf{W}\mathbf{W}^T + \theta\mathbf{I}_d)^{-1}(\gamma\mathbf{W}\mathbf{L}^T + \gamma\delta\mathbf{W}\mathbf{W}'^T\mathbf{Y} + \theta\mathbf{P}\mathbf{X} - \mathbf{G}_2) \quad (18)$$

where \mathbf{I}_d is a $d \times d$ identity matrix.

- 3) *Update \mathbf{P} by Fixing Others:* When updating \mathbf{P} , we need to solve the following subproblem:

$$\begin{aligned} \min_{\mathbf{P}} \quad & \text{Tr}(\mathbf{G}_1^T (\mathbf{X} - \mathbf{P} \mathbf{X} - \mathbf{E}) + \mathbf{G}_2^T (\mathbf{M} - \mathbf{P} \mathbf{X}) + \mathbf{G}_3^T (\mathbf{F} - \mathbf{P})) \\ & + \frac{\theta}{2} \left(\|\mathbf{X} - \mathbf{P} \mathbf{X} - \mathbf{E}\|_F^2 + \|\mathbf{M} - \mathbf{P} \mathbf{X}\|_F^2 + \|\mathbf{F} - \mathbf{P}\|_F^2 \right). \end{aligned} \quad (19)$$

Then, we can update \mathbf{P} with the following closed-form solution:

$$\mathbf{P} = (\Gamma\mathbf{X}^T + \theta\mathbf{F} + \mathbf{G}_3)(\theta\mathbf{X}\mathbf{X}^T + \theta\mathbf{I}_d)^{-1} \quad (20)$$

where $\Gamma = \theta\mathbf{X} + \theta\mathbf{M} - \theta\mathbf{E} + \mathbf{G}_1 + \mathbf{G}_2$.

- 4) *Update \mathbf{F} by Fixing Others:* When we update \mathbf{F} with other variables fixed, the problem reduces to

$$\min_{\mathbf{F}} \|\mathbf{F}\|_* + \frac{\theta}{2} \left\| \mathbf{F} - \left(\mathbf{P} - \frac{\mathbf{G}_3}{\theta} \right) \right\|_F^2 \quad (21)$$

which can be solved by the singular value thresholding algorithm [32] as follows:

$$\mathbf{F} = \mathbf{U} \Theta_{\frac{1}{2\theta}} \Sigma \mathbf{V}^T \quad (22)$$

where $\mathbf{U}\Sigma\mathbf{V}^T$ is the singularly valuable decomposition of $\mathbf{P} - (\mathbf{G}_3/\theta)$ and $\Theta_{\tau}(\cdot)$ is the singular value thresholding (SVT) operator defined by

$$\Theta_{\tau}(\Sigma) = \text{diag}(\text{sgn}(\Sigma_{ii})(|\Sigma_{ii}| - \tau)). \quad (23)$$

- 5) *Update \mathbf{E} by Fixing Others:* When updating \mathbf{E} , we have to solve the following subproblem:

$$\min_{\mathbf{E}} \alpha \|\mathbf{E}\|_1 + \frac{\theta}{2} \left\| \mathbf{E} - \left(\mathbf{X} - \mathbf{P} \mathbf{X} + \frac{\mathbf{G}_1}{\theta} \right) \right\|_F^2. \quad (24)$$

The solution to the above problem can be obtained by the soft-thresholding (shrinkage) operator [33].

Finally, we summarized the algorithm for solving the problem in Algorithm 1. Similar to the works in [31] and [34], it can be proven that the algorithm converges to the problem in (7) by iteratively solving the problem in (10). The complexity of the proposed algorithm is briefly discussed as follows. The complexity of calculating the inverse of a few matrices for \mathbf{W} , \mathbf{W}' , \mathbf{M} , and \mathbf{P} are $O(d^3)(d \ll n)$ or $O(v^3)(v \ll n)$. In each iteration, SVT is applied to update the low rank matrices whose complexity is $O(rd^2)$, where r is the ranks for \mathbf{P} . The soft-thresholding operator to update the sparse error matrix has a complexity of $O(dn)$. The complexity of matrix multiplication is $O(cdn)(c \ll n)$ for \mathbf{W} , \mathbf{W}' , and $O(d^2n)$ for \mathbf{M} and \mathbf{P} . Therefore, the overall computational complexity is $O(d^2n + cdn + d^3 + v^3 + rd^2)$.

Algorithm 1 MMLC**Input:** Matrices \mathbf{X} and \mathbf{Y} of the geo-tagged images;Parameters $\alpha, \gamma, \beta, \delta, \eta$.**Output:** \mathbf{W}, \mathbf{P} .

- 1: Initialize $\mathbf{M}, \mathbf{P}, \mathbf{F}, \mathbf{E}, \mathbf{W}, \mathbf{W}', \mathbf{G}_1, \mathbf{G}_2, \mathbf{G}_3$;
- 2: Initialize $\theta = 10^{-6}, \theta_{max} = 10^{10}, \rho = 1.1$;
- 3: **Repeat**
- 4: Update \mathbf{D} and \mathbf{D}' according to Eq. (11);
- 5: Update \mathbf{W} according to Eq. (15);
- 6: Update \mathbf{W}' according to Eq. (16);
- 7: Update \mathbf{M} according to Eq. (18);
- 8: Update \mathbf{P} according to Eq. (20);
- 9: Update \mathbf{F} according to Eq. (22);
- 10: Update \mathbf{E} by solving the sub-problem Eq. (24) with the soft-thresholding operator;
- 11: Update the multipliers $\mathbf{G}_1, \mathbf{G}_2$, and \mathbf{G}_3 :

$$\begin{cases} \mathbf{G}_1 = \mathbf{G}_1 + \theta(\mathbf{X} - \mathbf{P}\mathbf{X} - \mathbf{E}) \\ \mathbf{G}_2 = \mathbf{G}_2 + \theta(\mathbf{M} - \mathbf{P}\mathbf{X}) \\ \mathbf{G}_3 = \mathbf{G}_3 + \theta(\mathbf{F} - \mathbf{P}) \end{cases}$$
- 12: Update $\theta = \min(\rho\theta, \theta_{max})$
- 13: **until** Convergence
- 14: **return** \mathbf{W}, \mathbf{P}

C. Image Ranking

Once the landmark classifier with the discriminative information of image feature has been learned, we next calculate the ranking score for the candidate images. A straightforward method is to return the most similar images within the landmark cluster to which the query image is classified. However, there are usually some images that do not describe the same object even they are located near to each other. For example, some images are about the photographer, while some others are about the landmark. Moreover, some images may be misclassified due to the large diversity of vision patterns.

To address these problems, we combine both classification information and similarity measure to further refine the candidate images. Given a query image I_q with visual feature vector \mathbf{x}_q , its landmark classification result is a vector \mathbf{l}_q over the landmark clusters

$$\mathbf{l}_q = \mathbf{P}\mathbf{x}_q\mathbf{W}. \quad (25)$$

The value of each element in the vector \mathbf{l}_q can be considered as how likely the query image relates to the landmark corresponding to the element, which can be considered as a factor for image ranking. Then, we select l landmark clusters which have the largest classification values as the candidate landmarks.

For each candidate cluster, we then rank the images within it based on similarity measure. Due to the “semantic gap” problem [35], the visually similar images may be semantically dissimilar. We combine both textual similarity and visual similarity for ranking. It is assumed that, if two images are similar on both visual content and text content, they are more semantically consistent, which indicates that they are likely to reflect the same semantic object [35], [36]. Therefore, we

prefer the visual neighbors which are also textually similar with the query image. Since the query image does not contain textual content, we expand it with the textual content derived from its visually nearest neighbors. Then, the images which are more textually similar with the visually nearest neighbors of the query image should have priority in ranking. The semantic consistency between the candidate image I_i and the query image I_q is measured as follows:

$$Cs(I_i) = \frac{1}{K} \sum_{I_j \in \text{Nei}(I_q)} \text{sim}_{\text{text}}(I_i, I_j) \cdot \text{sim}_{\text{vis}}(I_q, I_i) \quad (26)$$

where $\text{Nei}(I_q)$ denotes the set of the K nearest neighbors of I_q in the visual space, $\text{sim}_{\text{text}}(\cdot)$ denotes the textual similarity, and $\text{sim}_{\text{vis}}(\cdot)$ denotes the visual similarity. To compute the visual similarity, the visual features are weighted by $\|\mathbf{W}_i\|_2$ which represents the importance of the corresponding features on landmark recognition. Finally, the classification result and semantic consistency information are combined to rank the images in the candidate landmark clusters as follows:

$$\text{score}(I_i, I_q) = \exp(-p\mathbf{l}_{qk}) \cdot Cs(I_i) \quad (27)$$

where \mathbf{l}_{qk} denotes the k th element of \mathbf{l}_q , k denotes the indicator of the landmark cluster that contains I_i , and p ($p > 0$) is a parameter to control the impact of the classification value to the ranking. The output of landmark retrieval is the top-ranked images according to (27).

V. EXPERIMENTS

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

A. Datasets and Features

We use the datasets MediaEval2012 [58] and NUS-WIDE [76] to evaluate our model. Since MediaEval2012 does not include the raw images and some images were removed after the dataset was collected, we download about one million of raw images from Flickr according to the URL randomly. For each image, we download the visual content, the tag list, and the geo-coordinate. We use the latitude–longitude coordinates as a point in the plane, and then cluster the images to train MMLC as discussed in Section IV-A. Since our goal is to analyze the impact of large diversity of images to the performance of landmark retrieval, we remove the clusters whose size are smaller than 100. As a result, there are 378 clusters and about 580 K images remained. Some of the popular clusters have a large number of images and some clusters have a small number of images. The image numbers of most clusters are between the two cases. Fig. 3 shows some examples of images with their landmark label information. 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. For this dataset, we select the 50 greatest clusters obtained by the labeling method proposed in Section IV-A to evaluate the approach.

For the visual features, we adopted the 4096-D DeCAF generic visual features [37], which is the activations of the

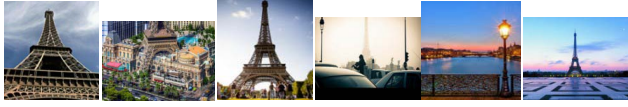





Landmark	Specific tags	Radom images
Eiffel tower	Eiffel Eiffel Tower tour eiffel tower tour tower	
Londoneye	Eye London eye Wheel London UK	
Forbidden city	forbidden city emperors moat emperor palace park	
EmpireStateBuilding	EmpireStateBuilding Cityscape manhattan Grattacielo Architecture	
Colosseum	Colosseum Rome open air theatre architecture building	
Cloudgate	Cloudgate Sculptures Bean Reflections Millenium Park	

Fig. 3. Examples of landmark images from our collected dataset MediaEval2012.

sixth layer of a deep CNN trained in a fully supervised fashion on ImageNet [38]. This feature representation has 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 ten images or more than 5% of the total images. As the raw tag list is very sparse, the textual feature vector is high-dimensional and many of the element entries are zero, which may affects the effectiveness of classification on text features as shown in (5). We include a new representation for text content by combining the geographical information. First, Word2Vec [27] is used to represent each tag with a 500-D vector, which is trained on Wikipedia articles and Web news of about 100 million words. Then, all the image tag lists are clustered into 2000 groups based on the geographical information, and a 2000-D geographical dictionary is obtained. Finally, each tag is represented by a 2000-D sparse codes learned from the geographical dictionary, and the tag list of each image is represented by a 2000-D feature vector by max pooling all the tags in the list.

B. Experimental Settings

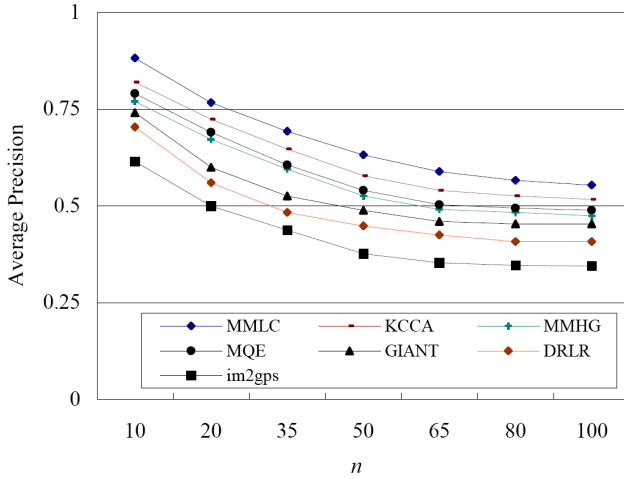
We randomly sample 80% images to form the training set, and 20% for test. Precision-recall is chosen as the metrics to evaluate both our model and state-of-the-art methods over top-100 retrieval ranked list based on each query landmark. For each landmark, we issue 20 queries and an average precision score is computed for all queries, and they are averaged to

obtain a mean average precision (mAP) for the each landmark category. Given one query image and first R top-ranked retrieved data, the average precision is defined as follows:

$$\frac{1}{M} \sum_{r=1}^R p(r) \cdot \text{rel}(r) \quad (28)$$

where M is the number of relevant data in the retrieved result, $p(r)$ is the precision at r , and $\text{rel}(r)$ presents the relevance of a given rank (one if relevant and zero otherwise). mAP is obtained by averaging average precision (AP) of all the queries. For our learning model MMLC, all the trade-off parameters are turned in the range $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ by a “grid-search” strategy. To evaluate the performance, we compare our approach with the following competitors.

- 1) *im2gps* [7]: It applies a feature matching method to return the K nearest neighbors based on visual features with respect to the query landmark image, where the query image and images in database are represented by aggregating a set of visual features to conduct landmark retrieval.
- 2) *GIANT* [9]: It proposes to discover a set of geo-informative attributes that are discriminative and useful for location recognition.
- 3) *DRLR* [8]: It mines a set of stylistic of visual elements to characterize a landmark, where the whole process is at the patch-level of visual content.
- 4) *MQE* [10]: It expands the query image with a multiquery image set from which a compact pattern

Fig. 4. Average precision at top- n images on MediaEval2012.

set is generated. Then the ranking score is calculate based on the matching information of pattern set in the candidate images.

- 5) *MMHG* [59]: Images are modeled as independent vertices and hyperedges contain several vertices. Multiple hypergraphs are constructed independently based on different visual modalities. The retrieved images are ranked based on the similarity scores.
- 6) *KCCA* [61]: It uses kernel CCA to discover the correlation between tags and visual content. The visual feature of a query image is projected onto the semantic space, and all the database images are sorted based on their correlation in that space.

C. Performance Evaluation

We evaluate the average precision and mAP for our approach and competitors. In our experiment, the AP is calculated by averaging the precision values of all the queries of each landmark over the top- n images of the retrieval list, with n varying from 10 to 100. The average precisions of different approaches on the two datasets are shown in Figs. 4 and 5, respectively. It shows that our approach outperforms others consistently on the average precision. When the number of returned images increases, the precision values of all the approaches decrease. This is probably because that more returned images can also lead to more noisy images returned.

Next, we compare the performance by the metric of mAP as shown in Fig. 6. The mAP of some example landmarks in MediaEval2012 is shown in Table I, and the following conclusions can be made. First, our approach outperforms all competitors on the metric of mAP, since we explore the latent correlation between landmarks and the multimodal content of images, and learn the important features for landmark recognition. Second, the large intraclass variance affects the performance of im2gps which is mainly based on visual similarity measuring. Both DRLR and GIANT detect discriminative visual regions from images. Though GIANT exploits the relations among the regions at the photo-level, only the visual content is insufficient to model each landmark since the images from different landmarks may share the common patterns of

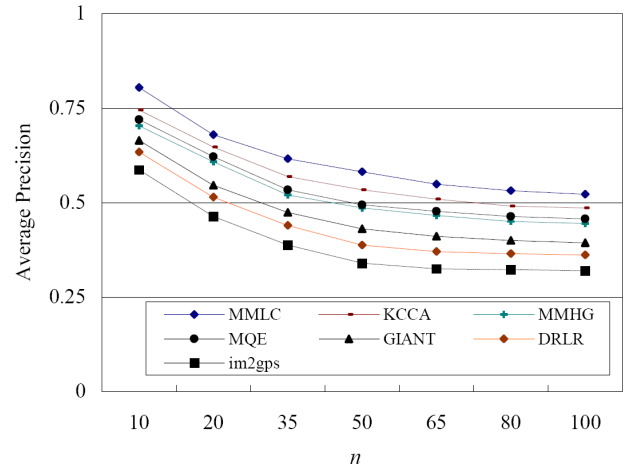
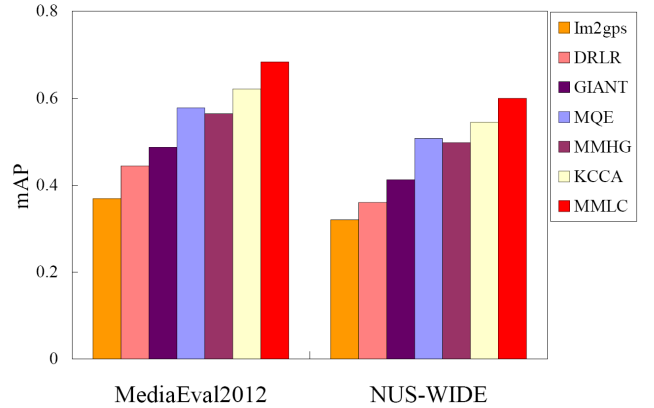
Fig. 5. Average precision at top- n images on NUS-WIDE.

Fig. 6. mAP comparison of all approaches on two datasets.

visual content. MMHG uses hyperedges which contain several images to capture different visual views of landmarks, and landmark images are retrieved using a graph-based search on the hypergraph. Their performance is better than im2gps, but the improvement is limited. MQE expands the query by the images selected from other users. However, the expanded images are expected to be similar with the query image in a coarse grained topic, where the mined pattern may be common to many landmarks. Moreover, the constructed user-image matrix does not contain the co-occurrence information of users, which may affects the performance of non-negative matrix factorization and hence affect the query expansion by introducing noisy images. KCCA combines both visual and text features for image retrieval. However, it assumes that images are identically distributed, which ignores the geographical correlation between images. This result also demonstrates the benefit of integrating textual information into landmark retrieval for the query image with only visual content.

For the visual-content-based retrieval approaches, it is important that there are many similar images in the landmark where the query image is taken. Therefore, these approaches suffer from the queries with relatively few visually similar images [7]. Usually, many users may upload a large number of similar images for the same landmark. Although we wish our approach to leverage patterns in the image dataset, we

TABLE I
COMPARISON OF mAP OF THE OVERALL QUERIES AND SOME EXAMPLE LANDMARKS IN MEDIAEVAL2012

Method	effieltower	londoneye	forbiddencity	empirestatebuilding	colosseum	cloudgate	Overall
Im2gps	38.56	30.56	39.83	33.19	36.16	32.12	36.89
DRLR	47.12	39.98	47.92	37.92	49.91	38.67	44.45
GIANT	50.39	42.75	51.37	41.87	51.32	42.17	48.67
MQE	62.63	53.56	63.63	50.36	60.56	51.34	57.76
MMHG	61.73	51.51	61.78	52.12	58.16	50.57	56.38
KCCA	66.13	56.54	66.98	54.25	63.94	56.78	62.13
MMLC	71.47	58.78	73.45	59.55	70.22	63.21	68.32

TABLE II
COMPARISON OF mAP OF THE OVERALL QUERIES AND SOME EXAMPLE LANDMARKS
IN MEDIAEVAL2012, WITH THE OTHER IMAGES OF QUERY USERS REMOVED

Method	effieltower	londoneye	forbiddencity	empirestatebuilding	colosseum	cloudgate	Overall	Change
Im2gps	31.59	24.51	33.21	26.13	31.26	27.31	30.43	-6.46
DRLR	39.19	33.56	38.46	30.59	43.55	31.61	38.22	-6.23
GIANT	41.36	35.59	43.87	33.85	44.34	35.56	41.76	-6.91
MQE	53.17	47.47	57.15	45.37	52.13	45.39	50.49	-7.27
MMHG	52.23	46.53	58.79	46.12	52.98	46.59	51.23	-5.15
KCCA	59.23	49.12	62.91	48.27	60.35	52.19	57.55	-4.58
MMLC	67.23	53.89	70.13	56.47	67.34	58.12	64.98	-3.34

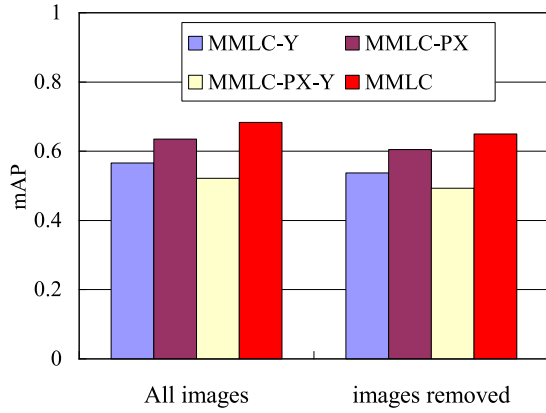


Fig. 7. Results on mAP scores by removing different component of MMLC.

also consider it important that the approach can make a reliable retrieval of the landmark images, without requiring that there are other many similar images taken by the same user in the collection. In order to delve deeper into the impact of visually similar images associated with the landmark of the query image, we next conduct an experiment to analyze that, for each query, the images from the same uploader of the query image are removed from the dataset MediaEval2012. The result of this experiment is depicted in Table II. It reveals that the performance degradation of the visual-content-based approaches is greater than that of our approach, which indicates that these approaches have a relatively stronger dependence on the presence of similar images in the dataset. However, our approach can exploits the textual content as the complementary information to model the patterns of landmarks more effectively. Thus, it is less affected by the absence of similar images uploaded by the query user.

Besides, our approach utilizes multimodal features in the learning process. We further investigate the help from individual components (e.g., visual, textual, and transformed features). We conduct experiments to analyze the performance of MMLC without the textual feature (**Y**, named MMLC-Y),

without the transformation feature (**PX**, named MMLC-PX), and without both textual features and transformed features (named MMLC-PX-Y), respectively. Fig. 7 shows the performance of these implementations based on all the training images and removing the images from the same uploader of the query image, respectively. It indicates that both of the feature transformation and text feature contribute to landmark retrieval.

D. Parameter Analysis

There are some tradeoff parameters to be set in the model MMLC, i.e., α , γ , η , β , and δ . We tune these parameters besides η in the range $\{0.001, 0.01, 0.1, 1, 10, 100\}$ by cross validation. We further analyze four important parameters in our retrieval approach on the dataset MediaEval2012.

- 1) η : The parameter to balance the importance between visual feature and textual feature of images as shown by (6).
- 2) l : The number of top ranked landmark clusters returned by the landmark classifier MMLC as shown by (25).
- 3) K : The number of nearest neighbors in the visual space used to measure the semantic consistency between a candidate image and the query image as shown in (26).
- 4) p : The value to control the impact of the classification result to image ranking as shown in (27).

We tune the parameters via the following strategy: we vary one parameter at a time while fixing the others. mAP is adopted as the evaluation metric. Then, several observations are made from the experiment results.

First, we analyze the importance of different features to landmark retrieval. Fig. 8 shows the mAP values against a variety of η values. From this figure, we can conclude that the textual features are more important than the visual features to landmark retrieval. This is because that the textual description is usually more efficient in reflecting the semantics of images than the visual content.

Fig. 9 shows the mAP values against a variety of values l . The result indicates that the performance of the approach is

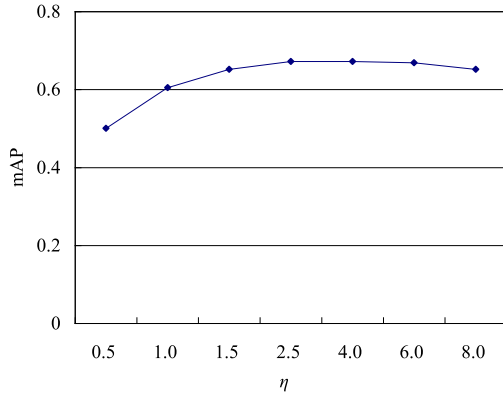


Fig. 8. Results on mAP scores by varying the importance weight of textual feature.

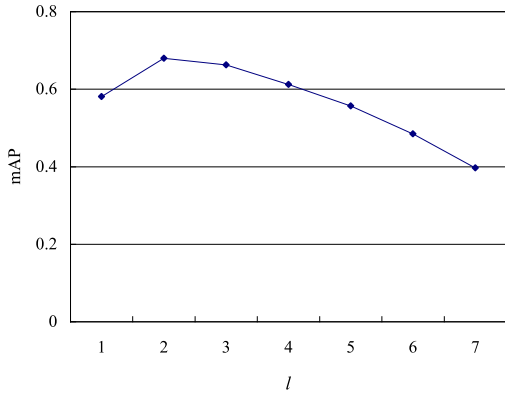


Fig. 9. Results on mAP scores by varying the l .

not proportionate to the number of selected landmark clusters, and the optimal choice of l is 2. When l is too small, the risk of the false classification increases. When l is too large, the chance of selecting more false landmark clusters increases, which results in more noisy images included in the retrieval result. Therefore, there is a tradeoff between the number of selected landmark clusters and the retrieval performance.

Then, Fig. 10 depicts the mAP values against varied values of K . It indicates that, when the value of K is relatively small, the increase of K can also lead to the increase of mAP, since a larger K will enrich the textual description and visual patterns of the landmark. Usually, many landmarks may have a large diversity in both text representation and visual views. Meanwhile, the bias influence of the visually similar images that are semantically dissimilar to the query image will be reduced. However, when the value of K keeps increasing, the performance will degrade gradually. A too large set of visually similar images result in that many noisy tags as well as unrelated visual pattern will be included, which means that the semantic consistence is gradually similar to be measured on a randomly sampled collection. Thus, a large value of K degrades the effectiveness of semantic consistence measure with a large probability.

Finally, the mAP values against varied values of p is reported in Fig. 11. Interestingly, one can see that the mAP value increases when the value of p increases within a relative small beginning point, since it will reduce the effect of false

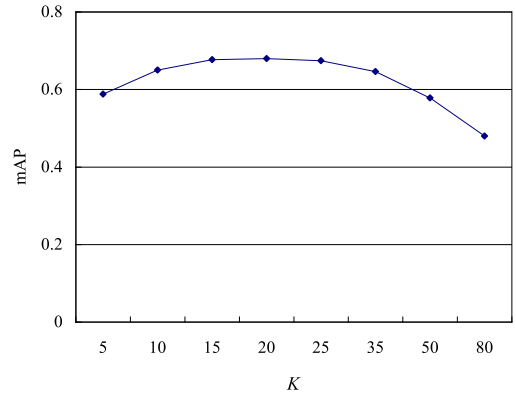


Fig. 10. Results on mAP scores by varying the K .

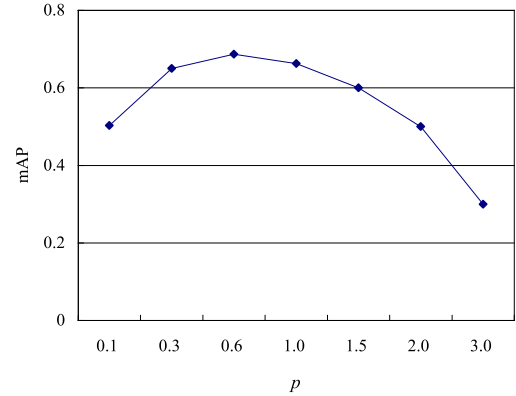


Fig. 11. Results on mAP scores by varying the p .

classification of landmark. However, when p keeps increasing, the performance will be downgraded. That is because that a large value of p means a small impact of landmark classification to image retrieval, which results in that the retrieval is similar to similarity-based search.

VI. CONCLUSION

In this paper, we motivate the problem of landmark retrieval with the geo-tagged images containing multimodal content. For this purpose, we propose an effective MMLC paradigm to leverage the multimodal content of images for landmark retrieval. We mainly address the following challenges, i.e., redundant and noisy visual content of images, and heterogeneous features corresponding to the visual content and text content of image. In particular, visual features are refined based on low rank matrix recovery, and multimodal classification combined with group sparse is learned from the automatically labeled images to recognize landmarks. Extensive experiments on real-world social image datasets demonstrate the superiority of the proposed approach as compared to existing methods. The novelty of this paper is to tackle the landmark retrieval with multimodal classifier automatically learned from geographically structural analysis and group sparse model. This improves the current research of image retrieval which mainly focuses on learning the retrieval model from raw features directly and ignores the latent correlation among multimodal contents.

There are some potential future extensions of this paper. It would be interesting to investigate other social information, like uploader's interests, social activities, and travel routes, for landmark retrieval. Also, conducting local culture and linguistic analysis across social media sites to better understand motivations and characteristics of image geo-tagging and landmark retrieval is a promising direction.

REFERENCES

- [1] H. M. Sergieh *et al.*, "Geo-based automatic image annotation," in *Proc. Annu. ACM Int. Conf. Multimedia Retrieval*, Hong Kong, 2012, Art. no. 46.
- [2] P. Serdyukov, V. Murdock, and R. Van Zwol, "Placing Flickr photos on a map," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Boston, MA, USA, 2009, pp. 484–491.
- [3] O. Van Laere, S. Schockaert, and B. Dhoedt, "Finding locations of Flickr resources using language models and similarity search," in *Proc. Annu. ACM Int. Conf. Multimedia Retrieval*, Trento, Italy, 2011, pp. 1–8.
- [4] O. Van Laere, J. Quinn, S. Steven, and B. Dhoedt, "Spatially aware term selection for geotagging," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 221–234, Jan. 2014.
- [5] Z. Xia *et al.*, "A privacy-preserving and copy-deterrence content-based image retrieval scheme in cloud computing," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 11, pp. 2594–2608, Nov. 2016.
- [6] J. Luo, D. Joshi, J. Yu, and A. Gallagher, "Geotagging in multimedia and computer vision—A survey," *Multimedia Tools Appl.*, vol. 51, no. 1, pp. 187–211, 2011.
- [7] J. Hays and A. Efros, "IM2GPS: Estimating geographic information from a single image," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA, 2008, pp. 1–8.
- [8] C. Doersch, S. Singh, H. Mulam, J. Sivic, and A. Efros, "What makes Paris look like Paris?" *ACM Trans. Graph.*, vol. 31, no. 4, pp. 13–15, 2012.
- [9] 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, 2014, Art. no. 19.
- [10] Y. Wang, X. Lin, L. Wu, and W. Zhang, "Effective multi-query expansions: Robust landmark retrieval," in *Proc. ACM Multimedia*, Brisbane, QLD, Australia, 2015, pp. 79–88.
- [11] X. Li, M. Larson, and A. Hanjalic, "Global-scale location prediction for social images using geo-visual ranking," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 674–686, May 2015.
- [12] C.-Y. Chen and K. Grauman, "Clues from the beaten path: Location estimation with bursty sequences of tourist photos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 42, Colorado Springs, CO, USA, 2011, pp. 1569–1576.
- [13] E. Kalogerakis, O. Vesselova, J. Hays, A. A. Efros, and A. Hertzmann, "Image sequence geolocation with human travel priors," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Kyoto, Japan, 2009, pp. 253–260.
- [14] G. Patterson and J. Hays, "SUN attribute database: Discovering, annotating, and recognizing scene attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, 2012, pp. 2751–2758.
- [15] D. M. Chen *et al.*, "City-scale landmark identification on mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, 2011, pp. 737–744.
- [16] Y.-T. Zheng *et al.*, "Tour the world: Building a Web-scale landmark recognition engine," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, 2009, pp. 1085–1092.
- [17] Y. Li, D. J. Crandall, and D. P. Huttenlocher, "Landmark classification in large-scale image collections," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Kyoto, Japan, 2009, pp. 1957–1964.
- [18] R. Ji *et al.*, "Learning compact visual descriptor for low bit rate mobile landmark search," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, Barcelona, Spain, 2011, pp. 2456–2463.
- [19] X. Xiao, C. Xu, J. Wang, and M. Xu, "Enhanced 3-D modeling for landmark image classification," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1246–1258, Aug. 2012.
- [20] Z. Cheng, J. Ren, J. Shen, and H. Miao, "Building a large scale test collection for effective benchmarking of mobile landmark search," in *Advances in Multimedia Modeling*. Heidelberg, Germany: Springer, 2013, pp. 36–46.
- [21] G. Schindler, M. Brown, and R. Szeliski, "City-scale location recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Minneapolis, MN, USA, 2007, pp. 18–23.
- [22] H. Liu, T. Mei, J. Luo, H. Li, and S. Li, "Finding perfect rendezvous on the go: Accurate mobile visual localization and its applications to routing," in *Proc. ACM Multimedia*, Nara, Japan, 2012, pp. 9–18.
- [23] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. M. Kleinberg, "Mapping the world's photos," in *Proc. 18th Int. Conf. World Wide Web*, Madrid, Spain, 2009, pp. 761–770.
- [24] J. Cao, Z. Huang, and Y. Yang, "Spatial-aware multimodal location estimation for social images," in *Proc. ACM Multimedia*, Brisbane, QLD, Australia, 2015, pp. 119–128.
- [25] R. Ji *et al.*, "When location meets social multimedia: A survey on vision-based recognition and mining for geo-social multimedia analytics," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 1, pp. 1–18, 2015.
- [26] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.
- [27] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. 1st Int. Conf. Learn. Represent.*, 2013, pp. 4089–4114.
- [28] F. Nie, H. Huang, X. Cai, and C. H. Q. Ding, "Efficient and robust feature selection via joint l_{21} -norms minimization," in *Proc. 24th Annu. Conf. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.
- [29] R. He, T. Tan, L. Wang, and W.-S. Zheng, " $L_{2,1}$ regularized correntropy for robust feature selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, 2012, pp. 2504–2511.
- [30] M. Nikolova and M. K. Ng, "Analysis of half-quadratic minimization methods for signal and image recovery," *SIAM J. Sci. Comput.*, vol. 27, no. 3, pp. 937–966, 2005.
- [31] Z. Lin, M. Chen, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," Dept. Elect. Comput. Eng., UIUC, Champaign, IL, USA, Tech. Rep. UILUENG-09-2215, 2009.
- [32] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.
- [33] E. T. Hale, W. Yin, and Y. Zhang, "Fixed-point continuation for l_1 -minimization: Methodology and convergence," *SIAM J. Optim.*, vol. 19, no. 3, pp. 1107–1130, 2008.
- [34] F. Nie, H. Huang, and C. Ding, "Low-rank matrix recovery via efficient Schatten p -norm minimization," in *Proc. 26th AAAI Conf. Artif. Intell.*, Toronto, ON, Canada, 2012, pp. 655–661.
- [35] Y. Lu, L. Zhang, Q. Tian, and W.-Y. Ma, "What are the high-level concepts with small semantic gaps?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA, 2008, pp. 1–8.
- [36] X. Zhang, Z. Huang, H. T. Shen, Y. Yang, and Z. Li, "Automatic tagging by exploring tag information capability and correlation," *World Wide Web J.*, vol. 15, no. 3, pp. 233–256, 2012.
- [37] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 647–655.
- [38] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, 2009, pp. 248–255.
- [39] G. Friedland, J. Choi, and A. Janin, "Video2GPS: A demo of multimodal location estimation on flickr videos," in *Proc. 19th Int. Conf. Multimedia*, Scottsdale, AZ, USA, 2011, pp. 833–834.
- [40] Y. Yang, H. T. Shen, F. Nie, R. Ji, and X. Zhou, "Nonnegative spectral clustering with discriminative regularization," in *Proc. 25th AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, 2011, pp. 555–560.
- [41] G. Friedland, O. Vinyals, and T. Darrell, "Multimodal location estimation," in *Proc. ACM Int. Conf. Multimedia*, Florence, Italy, 2010, pp. 1245–1252.
- [42] G.-H. Liu and J.-Y. Yang, "Image retrieval based on the textron co-occurrence matrix," *Pattern Recognit.*, vol. 41, no. 12, pp. 3521–3527, 2008.
- [43] Z. Zhou, Y. Wang, Q. M. J. Wu, C.-N. Yang, and X. Sun, "Effective and efficient global context verification for image copy detection," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 1, pp. 48–63, Jan. 2017.
- [44] G.-H. Liu and J.-Y. Yang, "Content-based image retrieval using color difference histogram," *Pattern Recognit.*, vol. 46, no. 1, pp. 188–198, 2013.
- [45] G.-H. Liu, Z.-Y. Li, L. Zhang, and Y. Xu, "Image retrieval based on micro-structure descriptor," *Pattern Recognit.*, vol. 44, no. 9, pp. 2123–2133, 2011.
- [46] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

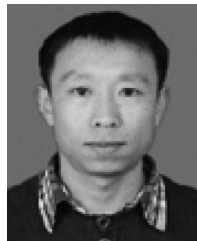
- [47] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.*, Graz, Austria, 2006, pp. 404–417.
- [48] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE Int. Conf. Comput. Vis.*, Barcelona, Spain, 2011, pp. 2564–2571.
- [49] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA, 2008, pp. 1–8.
- [50] P. Poursistani, H. Nezamabadi-Pour, R. A. Moghadam, and M. Saeed, "Image indexing and retrieval in JPEG compressed domain based on vector quantization," *Math. Comput. Model.*, vol. 57, nos. 5–6, pp. 1005–1017, 2013.
- [51] M. E. ElAlami, "A novel image retrieval model based on the most relevant features," *Knowl. Based Syst.*, vol. 24, no. 1, pp. 23–32, 2011.
- [52] F. F. Faria *et al.*, "Learning to rank for content-based image retrieval," in *Proc. MIR*, Philadelphia, PA, USA, 2010, pp. 285–294.
- [53] D. C. G. Pedronette and R. D. S. Torres, "Image re-ranking and rank aggregation based on similarity of ranked lists," *Pattern Recognit.*, vol. 46, no. 8, pp. 2350–2360, 2013.
- [54] J. Wan *et al.*, "Online learning to rank for content-based image retrieval," in *Proc. IJCAI*, Buenos Aires, Argentina, 2015, pp. 2284–2290.
- [55] J. He, M. Li, H.-J. Zhang, H. Tong, and C. Zhang, "Manifold-ranking based image retrieval," in *Proc. ACM MM*, New York, NY, USA, 2004, pp. 9–16.
- [56] S. C. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma, "Learning distance metrics with contextual constraints for image retrieval," in *Proc. CVPR*, vol. 2, 2006, pp. 2072–2078.
- [57] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large scale online learning of image similarity through ranking," *J. Mach. Learn. Res.*, vol. 11, pp. 1109–1135, Jan. 2010.
- [58] A. Rae and P. Kelm, "Working notes for the placing task at MediaEval 2012," in *Proc. MediaEval Workshop*, 2012.
- [59] L. Zhu, J. Shen, H. Jin, R. Zheng, and L. Xie, "Content-based visual landmark search via multimodal hypergraph learning," *IEEE Trans. Cybern.*, vol. 45, no. 12, pp. 2756–2769, Dec. 2015.
- [60] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513–523, 1988.
- [61] S. J. Hwang and K. Grauman, "Learning the relative importance of objects from tagged images for retrieval and cross-modal search," *Int. J. Comput. Vis.*, vol. 100, no. 2, pp. 134–153, 2012.
- [62] Z. Yu, L. Li, J. Liu, and G. Han, "Hybrid adaptive classifier ensemble," *IEEE Trans. Cybern.*, vol. 45, no. 2, pp. 177–190, Feb. 2015.
- [63] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.
- [64] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.
- [65] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [66] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.
- [67] Y. Song *et al.*, "Localized multiple kernel learning for realistic human action recognition in videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 9, pp. 1193–1202, Sep. 2011.
- [68] L. I. Kuncheva, J. J. Rodriguez, C. O. Plumpton, D. E. J. Linden, and S. J. Johnston, "Random subspace ensembles for fMRI classification," *IEEE Trans. Med. Imag.*, vol. 29, no. 2, pp. 531–542, Feb. 2010.
- [69] N. Garcia-Pedrajas, "Constructing ensembles of classifiers by means of weighted instance selection," *IEEE Trans. Neural Netw.*, vol. 20, no. 2, pp. 258–277, Feb. 2009.
- [70] G. Yu, C. Domeniconi, H. Rangwala, G. Zhang, and Z. Yu, "Transductive multi-label ensemble classification for protein function prediction," in *Proc. 18th ACM SIGKDD KDD*, Beijing, China, 2012, pp. 1077–1085.
- [71] Z. Yu *et al.*, "Hybrid k-nearest neighbor classifier," *IEEE Trans. Cybern.*, vol. 46, no. 6, pp. 1263–1275, Jun. 2016.
- [72] Z. Yu *et al.*, "A new kind of nonparametric test for statistical comparison of multiple classifiers over multiple datasets," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2016.2611020.
- [73] Z. Yu *et al.*, "Distribution-based cluster structure selection," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2016.2569529.
- [74] L. I. Kuncheva, "A bound on kappa-error diagrams for analysis of classifier ensembles," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 3, pp. 494–501, Mar. 2013.

- [75] Z. Yu, L. Li, J. Liu, J. Zhang, and G. Han, "Adaptive noise immune cluster ensemble using affinity propagation," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 12, pp. 3176–3189, Dec. 2015.
- [76] T. S. Chua *et al.*, "Nus-wide: A real-world Web image database from national university of Singapore," in *Proc. ACM Conf. Image Video Retrieval*, 2009, Art. no. 48.



Xiaoming Zhang received the B.Sc. degree and the M.Sc. degree in computer science and technology from the National University of Defence Technology, Changsha, China, in 2003 and 2007, respectively, and the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2012.

He is currently with the School of Computer, Beihang University, where he has been a Lecturer, since 2012. He has published over 30 papers, such as the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTION ON CYBERNETICS, *World Wide Web Journal*, *Neurocomputing*, the *Journal of Intelligent Systems*, *Signal Processing*, International Joint Conference on Artificial Intelligence 2015, International Conference on Multimedia Retrieval 2015, Siam International Conference on Data Mining 2014, International Conference on Web-Age Information Management 2014, AAAI Conference on Artificial Intelligence 2013, and Coling 2012. His current research interests include social media analysis, image tagging, and text mining.



Senzhang Wang was born in Yantai, China, in 1986. He received the M.Sc. degree from Southeast University, Nanjing, China, in 2009, and the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2015.

He is currently an Associate Professor with the School of Nanjing University of Aeronautics and Astronautics, Nanjing. He has published over ten papers on the famous international journals and conferences, such as Knowledge and Information Systems, ACM SIGKDD Conference on Knowledge Discovery and Data Mining, AAAI Conference on Artificial Intelligence, Siam International Conference on Data Mining, and International Conference on Database Systems for Advanced Applications. His current research interests include data mining and social network analysis.



Zhoujun Li received the M.Sc. and Ph.D. degrees in computer science from the National University of Defence Technology, Changsha, China, in 1984 and 1999, respectively.

He is currently with the School of Computer, Beihang University, Beijing, China, where he has been a Professor since 2001. He has published over 150 papers on international journals, such as the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *Information Science*, and *Information Processing and Management*, and international conferences, such as ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2014, Siam International Conference on Data Mining (SDM) 2014, ACM International Conference on Information and Knowledge Management (CIKM) 2014, AAAI Conference on Artificial Intelligence 2013, and International ACM SIGIR Conference on Research and Development in Information Retrieval 2013. His current research interests include the data mining, information retrieval, and database.

Dr. Li was a PC Member of several international conferences, such as SDM 2015, CIKM 2013, International Conference on Web-Age Information Management 2012, and Pacific Rim International Conferences on Artificial Intelligence 2012.



Shuai Ma received the Ph.D. degrees from the University of Edinburgh, Edinburgh, U.K., in 2010, and Peking University, Beijing, China, in 2004.

He is a Professor with the School of Computer Science and Engineering, Beihang University, Beijing, China. He was a Post-Doctoral Research Fellow with the Database Group, University of Edinburgh, a Summer Intern with Bell Laboratories, Murray Hill, NJ, USA, and a Visiting Researcher of MRSA, Beijing, China. His current research interests include database theory and systems, social data and graph analysis, and data intensive computing.

Dr. Ma was a recipient of the Best Paper Award for International Conference on Very Large Data Bases 2010 and the Best Challenge Paper Award for International Conference on Web Information Systems Engineering 2013.