On Combining Social Media and Spatial Technology for POI Cognition and Image Localization

This paper presents a comprehensive overview for the technologies combining social media and spatial technology for place-of-interest cognition and image geographical localization.

By Xueming Qian, *Member IEEE*, Xiaoqiang Lu, *Senior Member IEEE*, Junwei Han, *Senior Member IEEE*, Bo Du, *Senior Member IEEE*, and Xuelong Li *Fellow IEEE*

ABSTRACT | With fast development of information engineering and social network, people's locations can be conveniently sensed by spatial technology, such as global positioning systems (GPS), base stations, Wi-Fi access points and even from the appearances of the photos they have taken. The social networks and the online shopping platforms have been gathering billions of users, who share a large amount of images taken in places they live in and visit. We can leverage the social networks to express our opinions about the services and places of interest (POIs). The interactions among users, and user and POIs or services generate big social media data, which have rich information for user, location, and service cognition. Many real-time network applications rely heavily on the accurate social users' locations. How to sense the locations from multisource social media data is very important and challenging. Thus, in this paper, we give a systematic review of the works that combine social media and spatial technology for POI cognition and image localization.

KEYWORDS | Geolocalization; global positioning system (GPS); image retrieval; multimedia; place of interests (POI); placing task; recommender system; social media; spatial technology

I. INTRODUCTION

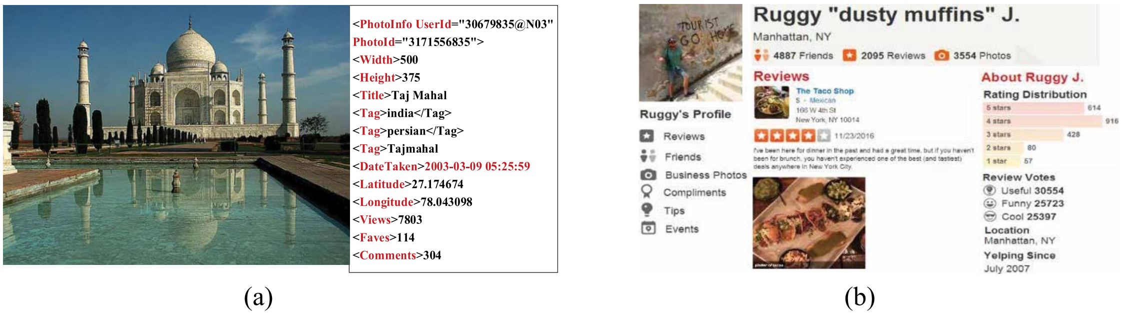
Recently, we have witnessed a boom in smartphones, with which cloud computing and social networks have become main flags of our era. Social networks, such as Instagram, Wechat, Facebook, Flickr, etc., have the magic power to attract billions of users to communicate and share information with each other. Social networks are essentially different from the traditional internet. They change the information transmission and utilization styles. One of the most significant features of social networks is the real-time interaction among users, and that of users and services/items/places [17], [31]. The traditional internet has a passive information acquisition. Usually, a user is in the front of a personal computer (PC) and search information. The location of PC is associated with its IP address. While in social networks, most of users' contextual information can be well sensed from various spatial sensing techniques [3], [9], [172] and the user-generated content. We can locate the user by sensing the locations of a smartphone or from the user's shared photos [74]–[77], GPS trajectories, and the cell tower locations.

One obvious progress in social network applications in recent years has been the introduction of a spatial technique

Manuscript received December 4, 2016; revised May 27, 2017; accepted July 9, 2017. Date of publication August 22, 2017; date of current version September 18, 2017. *(Corresponding authors: Xiaoqiang Lu; Xuelong Li.)*

X. Qian is with the Ministry of Education Key Laboratory for Intelligent Networks and Network Security and SMILES LAB, Xi'an Jiaotong University, Xi'an 710049, China.
X. Lu and **X. Li** are with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, China (e-mail: luxiaoqiang@opt.ac.cn; xuelong_li@opt.ac.cn).
J. Han is with the School of Automation and Information Engineering, Northwestern Polytechnical University, Xi'an 710049, China.
B. Du is with the School of Computer Science, Wuhan University, Wuhan 430072, China.

Digital Object Identifier: 10.1109/JPROC.2017.2731600



```
<PhotoInfo UserId="30679835@N03"
PhotoId="3171556835">
<Width>500
<Height>375
<Title>Taj Mahal
<Tag>india</Tag>
<Tag>persian</Tag>
<Tag>Tajmahal
<DateTaken>2003-03-09 05:25:59
<Latitude>27.174674
<Longitude>78.043098
<Views>7803
<Faves>114
<Comments>304
```

(a)

Ruggy "dusty muffins" J.

Manhattan, NY

4887 Friends 2095 Reviews 3554 Photos

Ruggy's Profile

- Reviews
- Friends
- Business Photos
- Compliments
- Tips
- Events

Reviews

The Taco Shop
$ · Mexican
166 W 4th St
New York, NY 10014

★★★★☆ 11/23/2016

I've been here for dinner in the past and had a great time, but if you haven't been for brunch, you haven't experienced one of the best (and tastiest) deals anywhere in New York City.

About Ruggy J.

Rating Distribution

5 stars	614
4 stars	916
3 stars	428
2 stars	80
1 star	57

Review Votes

Useful 30554
Funny 25723
Cool 25397

Location
Manhattan, NY

Yelping Since
July 2007

(b)

Fig. 1. ***(a) Example of Flickr image information. (b) Example of the Yelp user information. The user's textual descriptions, photo, current location, time, etc., are shown.***

[6], [14], [19], [163]–[164]. From millions or even billions of photos that have been shared in social media communities, we find that more than 30% images have had GPS information [134] before 2012. Today almost all of the shared images taken by smartphones have GPS information [20], [35], [74]–[77], [102], [128], [134]. User's real-time GPS locations are detected and associated with maps in many applications in real-time travel guide [27]. For example, an Uber user can know an accurate location of a booked car based on the urban traffic condition. Thus, he can schedule his travel conveniently.

Today, many applications are built on smartphones. They offer more detailed contextual information, such as GPS, time, image, its textual descriptions, and so on. Based on the supercomputing power, the cloud can return users' instant requirement/query to further guide their actions [5], [8], [38], [99]. With the help of the positioning system and the geographical feature of the location, the cloud end can schedule the detailed route and time for each user.

Social media are rich media. They have multimodality information, such as image, tags, time, views, GPS, etc. As shown in Fig. 1(a), for a Flickr user with a user ID 30678935@N03, the longitude and the latitude denote the GPS coordinate that the photo was taken at. By associating sequential photos' taken dates and locations, we can infer the user's footprints [20], [98], [102] during travel. The dominant locations of users' living, working, and dinning can be inferred from their GPS trajectories [3]–[25], [34], [166], [167]. Social media also provide us the chance to learn more about the places of interest (POIs), location-based services (LBSs), and users' preferences [3]–[25], [169]. By exploring the trajectories of the involved users in a POI, the normal travel routes can be mined from the crowdsource social media [98], [117]. Moreover, the POI characteristics can be mined and their classic/hot viewpoints can be inferred [57], [64]–[95].

Users' active locations can also be well recognized from social media. Thus, from the user interaction with the POIs (or the LBS), we can learn more about the user and POI simultaneously [107], [109], [113], which is the foundation of personalized recommendation [2]–[5], [13], [15]–[18], [28], [35], [37]–[43], [168], [169]. Fig. 1(b) shows a user and his information on Yelp. The probability of a user visiting a location is inversely proportional to the distance from its nearest center [11], [12]. We find that users have close interactions with their neighboring places, which can be learned about more from the temporal–spatial–contextual information extracted from social media.

Social-media-based POI recommendation typically gathers users' check-in records, venue information including categories, and users' social relationships to recommend a list of POIs where users would most likely visit later [3]–[6], [24]–[26]. POI recommendation can bring benefits to advertising agencies since it can construct effective launching advertisements to the potential consumers. Near restaurants and downtown shopping malls can also be explored. GIS data are the important source for POI recommendation system. Actually, 3S spatial technology has provided a great potential for various recommendation systems [3]–[6], including POI [23]–[26], [28], [29], and restaurant recommendation. Remote sensing image analysis has been demonstrated to be efficient in extracting information from Earth's surface. It has been widely acknowledged that remote sensing has become the most important way to observe the important places in real time, on a rapid and wide scale. Among these methods, deep-network-based scene extraction has provided the greatest potential [145]–[146], [151]. The geographical condition can also be an important factor in influencing users' traveling behaviors. When we have enough information about the spatial patterns on an urban street, it would be beneficial to combine it with users' history records to obtain accurate POI recommendation.

How to fully utilize the social media data and the sensed user position to carry out deep cognition of the place is a fundamental problem of personalized recommendation [1]–[2], [13], [15]–[18], [21]–[26], [28]–[38]. However, there are still several challenges listed below in social media and spatial techniques for POI cognition and image localization.

First, many images are without or with wrong GPS information in social media. Although there are many spatial technologies to sense users' locations, sometimes the signals are poor and the GPS devices of smartphones are not ready, or users are not allowed to utilize the GPS information in online services. The precision of localization is not satisfied in some

LBSN applications that require accurate GPS trajectories. Thus, how to estimate the GPS or improve localization accuracy from user's photos taken at a POI by fusing multimodal social media information is a challenging problem [42]–[63], [144].

Second, for automatic POI discovery, although the total volume of data for a POI in social media is huge, there is a lot of noisy and irrelevant content generated by users. We still have the information loss problem. For example, the textual description of a place can be obtained from various aspects by different users. Moreover, users may have different sharing behaviors [21], [130]–[135], [140], images captured at the same POI may be taken from different viewpoints, and images taken during different seasons may look quite different. Due to those ambiguities existing in POI description, it is a challenge to automatically differentiate between the true and false information from the big social media. Furthermore, how to discover the POI adaptively from the orderless social media is quite challenging as well.

Third, comprehensive POI cognition is confusing with the huge size of social media data. There is a lot of noise in social media, and that information often contaminates the data for POI cognition. It is a challenging problem to fully use social media data to represent the POI adaptively. There is a lot of irrelevant and vastly relevant content for the POI. However, as that content comes from various aspects, how to organize it and differentiate it orderly is very important in POI cognition.

The rest of this paper is organized as follows. In Section II, we provide some notations for social media. The related work on POI cognition from social media is presented in Section III. Image localization from social media is reviewed in Section IV. Finally, conclusion and future directions are given in Section V.

II. NOTATIONS FOR SOCIAL MEDIA

Let $\{I_i, u_i, v_i, \text{view}_i, l_i, \text{text}_i, \text{time}_i, \text{camera}_i\}_{i=1}^N$ denote the multisource information of a collected set of social media, including image, user information, location, timestamps, views, etc., from various social networks. Here, N is the total number of images. The corresponding explanations for the multisource information are as follows.

I_i denotes an image that was captured by photographical devices, such as cameras, smartphones, remote sensing satellites, and aerial imagery devices [72]. In general, the images captured by normal cameras are within the range of 1 km, while images captured by remote sensing satellites may have various resolutions ranging from several hundred kilometers to several meters [72], [145]–[148]. The normal optical camera captures images only with three components: R, G, and B channels. However, the images acquired by the remote sensing satellites usually have many more channels/sub-bands [145]–[148].

u_i denotes the social media user who shares image I_i. The corresponding user information in social media can be associated with the unique user ID, such as 30679835@ No3 shown in Fig. 1(a) and user's name "Ruggy" as shown in Fig. 1(b). The user's information is important in many online applications, social recommendations, and image retrieval. High-level spatial–temporal–social context can be explored from big social user's history information, such as social circle [137], sentiments [2], [28], [141], sharing behavior [21], [130]–[135], [140], activity ranging, and preferences.

v_i represents the visual feature of image I_i. The feature can be in various formats, global or local, and low level or high level. The global feature can have the following types: 1) the color feature extracted from an HSV space; the traditional color features are color moment, color histogram, and color distribution; 2) the texture feature extracted by local binary patterns (LBPs) and their extensions [27], and the wavelet transform domain, such as gist; and 3) an edge histogram, such as HOG, etc.; the local feature is successfully utilized in image matching, recognition, and retrieval. The SIFT feature is often utilized [36], which also has some simplified forms such as SURF [39]. A bag of visual words often adopts k-means clustering for the local feature to carry out fast processing and indexing [61], [121], [124]–[129], [155], [156]. The middle-level and high-level feature extracted from the deep convolutional neural network can serve as high-level visual content annotation, classification, and recognition [171], [172]. Except for the raw feature representation of the image, enhanced feature representations, such as Fisher vector and VLAD, are often utilized [1], [42], [47], [52].

view_i denotes how many times image I_i has been browsed/viewed by social users from the moment it was shared [35]. The view of a social image is also an important factor to indicate the popularity/interest/aesthetics of the image. In general, well-photographed images can attract users' attention and encourage them to communicate with other users about the photo content, such as shooting styles, color, illumination, and so on. The view information can be utilized in social image ranking and summarization [109], and social image retrieval [135], [170].

$l_i = (\text{lat}_i, \text{lon}_i)$ denotes the GPS location (coordinates) that the image was taken at. The corresponding latitude and longitude are sensed by the satellites. Some of the location information can be determined by the embedded GPS devices and electronic compasses in user's smartphones or cameras, or determined by cell towers. The GPS is a coarse location representation form for the photo [61]. It denotes the location of image devices rather than the real location of image content. Because of the lack of the pose and angle information, the accurate location of the visual content cannot be determined. Imagine that from the top of a mountain you take a photo of a tower located on the top of another mountain which is far away from the mountain you are on. The visual content of the photo is the tower (also with its real GPS location) but the location of the photo is the mountain you are on. So, the image content and the place where the image was taken should both be utilized to estimate the real GPS of the image.

text_i denotes textual descriptions of an image, including a title, tags, textual comments from other users/friends, etc. For example, as shown in Fig. 1(a), the title of the image is

Taj Mahal . The textual descriptions have high correlation about the visual content. It is also helpful to image content understanding, POI recommendation, image localization, and retrieval [135], [141]–[144], [170]–[171].

time_i is a timestamp of when the image was taken. This information is somewhat valuable, providing information about the season and time of day by combining the GPS information [122], [135].

camera_i is the corresponding camera information about the image. Although some photographs contain additional information embedded as metadata in EXIF format obtained via a GPS device or cell tower triangulation [71], this information is still not available in the majority of social images. Many digital cameras embed focal length and other camera related information in the EXIF tags of image files [80].

III. POI COGNITION

From the big user contributed social media, POIs can be discovered automatically. Assume that P POIs can be explored from a given collection of social media with N images. We denote the corresponding POI discovery as follows:

$$\{I_i, u_i, v_i, \text{view}_i, l_i, \text{text}_i, \text{time}_i, \text{camera}_i\}_{i=1}^{N} \rightarrow \{POI_p\}_{p=1}^{P}. \quad (1)$$

The POI can be explored more based on the available multiple contextual information. However, the variance of users' sharing behaviors raises some problems in POI discovery and cognition [134], [135]. In this section, we review related works on POI cognition including POI summarization and 3-D reconstruction by exploring the available multisource social media.

A. POI Discovery Approach

How to discover POIs [7], [98], [110] and learn their attributes from social media has drawn much attention in the past decade. The following approaches can be utilized in POI discovery [64], [65], [70], [71], [98], [117], [129], [143], including GPS clustering, social factor tuning, visual clustering, and POI merging approaches.

1) GPS Clustering: Most existing POI discovering approaches apply density-based location clustering, such as k-means clustering and mean-shift clustering, and geotagging in community-contributed photos [98], [117], [136]. The GPS clustering-based POI discovering approaches directly utilize the GPS coordinates of social images $\{l_i\}_{i=1}^{N} = \{(\text{lat}_i, \text{lon}_i)\}_{i=1}^{N}$ to determine the POIs $\{\text{POI}_p\}_{p=1}^{P}$ in the real world. These approaches are purely big data driven, mainly consisting of the following steps.

First, by assigning the initial starting point l and bandwidth h, the mean shift vector $M_h(l)$ is expressed as follows:

$$\begin{cases} M_h(l) = \frac{1}{k}\sum_{l_i \in S_h}(l_i - l) \\ [0.5pc]\, S_h \equiv \left\{l_i : (l_i - l)^T(l_i - l) \leq h^2\right\} \end{cases} \quad (2)$$

where S_h is the circle whose radius is h, l is a centroid, and k is the number of images with GPS location information that fall in the region S_h. Second, we update the starting point by adding the determined mean-shift vector as follows: $l \leftarrow \hat{u}l + M_h(l)$. Then, the starting point is iteratively updated until all of the points are traversed. Finally, the centroids of the clusters are utilized to represent the GPS locations of the POIs.

2) Social Factor Tuning: Each image from social media is associated with its shared user. Different users have different sharing behavior. In the mean-shift clustering approach, the contribution of each location is viewed to be identical. The user who shares more images at a location normally contributes more in the density-based GPS clustering as shown in (2). This will make the mean-shift vector biased to the user who shares many photos at a POI, especially, for the user who utilizes batch sharing. To balance the user factor in POI discovery, the location frequencies from the same user are taken into account [109] and utilized to modify the mean-shift clustering approach as follows:

$$M_h(l) = \frac{1}{k}\sum_{l_i \in S_h} W_{u_{l_i}}(l_i - l) \quad (3)$$

where $W_{u_{l_i}}$ is the weight of the user at the GPS location l_i. In the traditional mean-shift clustering, $W_{u_{l_i}} = 1$ for all the images, while in the social factor tuning-based mean-shift clustering approach, the weight of the user who shares many photos at a single location is reduced. The larger the number of images that the user shared at the location, the smaller is the weight of each single image. Thus, the weight of an image is an inverse of the total number of images that are shared by the user, and the weight is rewritten as follows:

$$W_{u_{l_i}} = \frac{1}{\sqrt{N_u}} \quad (4)$$

where N_u denotes the total number of photos shared by user u at GPS l_i. This approach can prevent the centroid drift to the users who share more photos in a place by batch sharing [134], [135]. Except for the user factor, the view information can be fused to mine POI. The photos viewed by many users contribute more in POI discovery than that with few users [109], [135].

3) Visual Clustering: The visual-clustering-approach-based POI discovering is based on the visual features of social images as follows:

$$\{I_i, v_i\}_{i=1}^{N} \rightarrow \{\text{POI}_p\}_{p=1}^{P}. \quad (5)$$

Many image clustering approaches can be utilized [121], [124]–[127], for example, the SIFT-feature-matching-based approach, the graph-growth-based approach, the near duplicated group finding approach, and the visual retrieval approaches.

In the graph-growth-based approach, each image is viewed as a node, and a link between two nodes denotes the edge of the graph. The weight of the edge is determined by the similarity of the two images (i.e., nodes) based on their visual features. In the graph-growth-based approaches, the images are first grown from two seed nodes with the smallest distance in the image set. Then, nodes are adaptively added to the graph by verifying that the newly added node has sufficient similarity to existing nodes in the previous iterations [119], [121].

Similar to the graph-growth-based approach, a community-detection- or graph-cut-based approach can also be utilized to determine the visual clusters for a given image collection [121]. Different from the graph-growth-based approach, the graph-cut- or community-detection-based approach first models the whole connected graph for all the images [1].

The visual clustering approaches are often utilized in the image retrieval with diverse viewpoints or semantics reranking [135], [138]. The top ranked images can be selected from different clusters, respectively, rather than from the same cluster.

4) POI Merging: Different POIs have different ranges. In the mean-shift clustering, the corresponding bandwidths are different. Different POIs have different appearances. Some clusters can be merged as unique ones when the clusters are too close and share similar visual content [98], [117]. Any two clusters can be merged when the following conditions are satisfied:

$$\begin{aligned} \mathrm{POI}_j &= \mathrm{POI}_j \cup \mathrm{POI}_k; \\ s.t.\ \mathrm{dist}\left(v_j^r, v_k^r\right) &\geq \mathrm{thr}\ \textit{and}\ \mathrm{dist}\left(\mathrm{POI}_j, \mathrm{POI}_k\right) \leq d_0 \end{aligned} \tag{6}$$

where v_j^r and v_k^r are the visual features of the representative images from POI_j and POI_k, and $\mathrm{dist}(\cdot)$ denotes their geographical distance. thr and d_0 are the two parameters to constrain the POI merging.

Other factors for social media, such as tags [110], [117], sentiments explored from the textual descriptions [111], and socially-aware factors, can be utilized in POI clustering [13], [103], [112].

B. POI Reconstruction

Social networks gather a wide range of images for POIs from various users taken at various time and various viewpoints. Thus, it is possible to reconstruct the POI from crowd contributed photos. There are some challenges in reconstructing high-quality 3-D models for the photos collected from social media. First, pictures from social media are order-less. Second, camera settings are different from each other. Last, the efficiency of the reconstruction algorithm is extremely computationally intensive for a large-scale social image set.

1) Sparse 3-D Model: Based on the clustered photos, structure from motion (SfM) is an efficient tool to generate sparse 3-D models for a POI. SfM takes the relative geometries from the viewing graph as an input and outputs 3-D reconstruction results, consisting of camera poses and sparse 3-D points. SfM aims to recover camera parameters, estimate poses, and generate a set of sparse 3-D POI geometry from social images [80]–[88] by local feature matching. SfM algorithms initialize each image by pose estimation, and select two images to carry out matching. The reconstructed points are further checked by ensuring that they are well conditioned before adding to the model of the POI. The focal length from the EXIF is important information which is utilized in the POI reconstruction. Sometimes, many social images do not have this information. So, we need to estimate it before POI reconstruction.

In general, the SfM-based 3-D reconstruction consists of the following three steps: 1) SIFT feature extraction; 2) SIFT feature matching between pairs of images; and 3) camera parameters recovery by running SfM iteratively. The fundamental matrix for the pair of images is estimated by RANSAC. A candidate fundamental matrix can be estimated by SfM with nonlinear optimization (or bundle adjustment). The noise matches are viewed as outliers and removed before the fundamental matrix recovering.

The bundle adjustment is computationally intensive for SfM. So, some researchers try to utilize the global SfM [83], rather than the incremental SfM [80].

From the POI clustering on the collected social media, a set of POIs can be discovered as described in Section III-A. For each POI, its 3-D point cloud models can be generated by SfM-based matching as follows:

$$\left\{\mathrm{POI}_p\right\}_{p=1}^{P} \rightarrow \left\{M_j\right\}_{i=1}^{J} \tag{7}$$

where $\left\{M_j\right\}_{i=1}^{J}$ denotes the sparse 3-D point clouds, and J is the total number of reconstructed 3-D models.

After completing the process of 3-D reconstruction for the images from a visual album of a POI, we obtain a group of 3-D models, including camera information C and geometric point's information G [80]. We represent the group of 3-D model information as follows:

$$S = \{C; G\}.$$

Geometric information G of a reconstructed POI contains a three-vector describing points of 3-D position, a three-vector describing the RGB color of the point, and a list of views that the point is in. Camera information C contains three-vector camera position, focal length F, three-vector translation T, 3×3 matrix format of rotation R, and parameters of radial distortion [80].

2) Dense 3-D Model From Multiview Stereo: Compared with the sparse 3-D models, the density 3-D models have consistent texture, smooth surface, and vivid color for the POI, which can be utilized in POI-related VR and AR applications and 3-D image retrieval.

From SfM, a set of sparse 3-D models are reconstructed. The 3-D models are shown in Figs. 2(b) and 3(a) and (b),

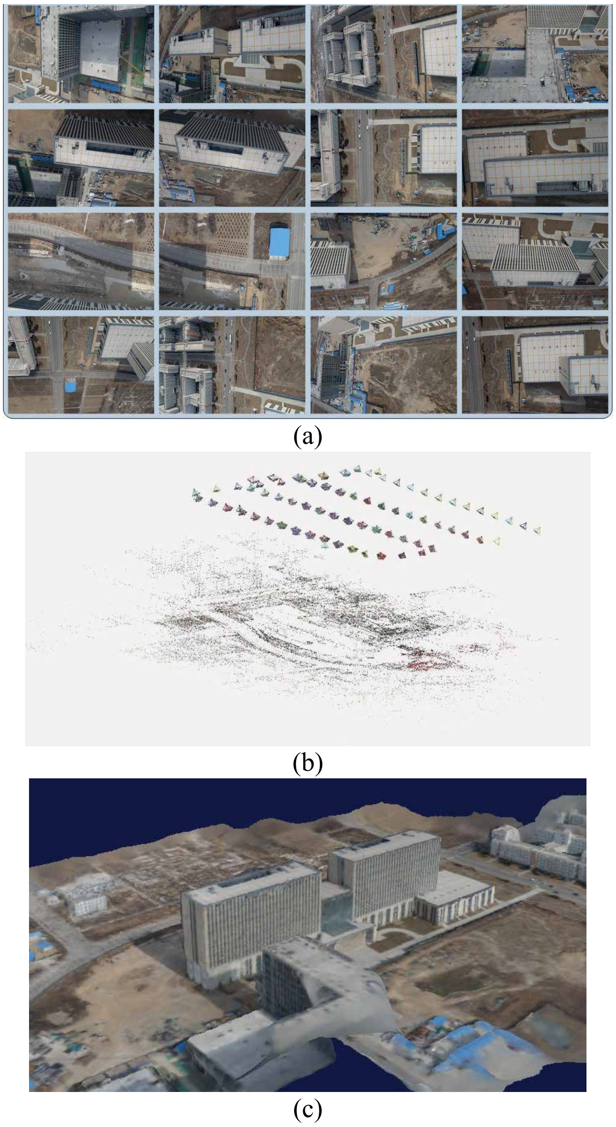


Fig. 2. ***POI reconstruction. (a) Image collections for a POI (to be continued). (b) Sparse 3D point cloud. (c) Dense 3D model..***

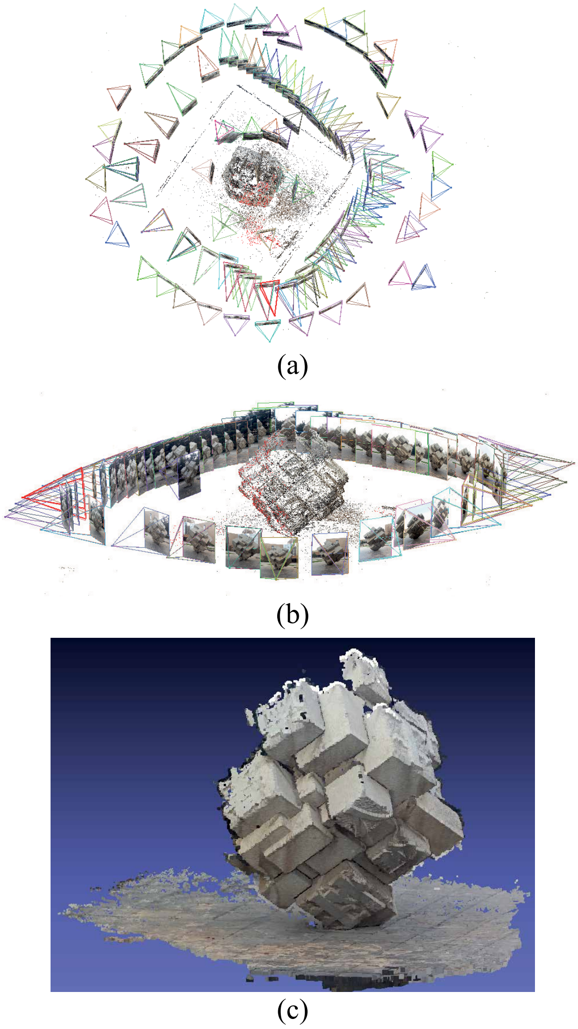


Fig. 3. ***POI reconstruction. (a) Sparse 3D point cloud and the registered images in the overhead view. (b) Sparse 3D point cloud and the registered images in the top-front view. (c) Dense 3D model.***

which mainly consist of the matched points. The sparse 3-D models are composed of a set of isolated points, while the POI itself has consistent texture and surface. Therefore, researchers try to densify the sparse 3-D models to yield the dense model [48], [79]–[81], [84], [86]–[88].

Shen [87] proposed a depth-map merging-based multiple view stereo method to define the 3-D sparse model. A patch-based stereo matching process was used to generate the depth map in each image. A depth-map refinement process is utilized to enforce consistency over neighboring views.

Agarwal *et al.* [79] reconstructed a POI by SfM to infer camera viewpoints and sparse 3-D scene structure from 2-D pictures, and multiview stereo (MVS). The MVS produces dense 3-D geometry based on a set of calibrated photos.

Frahm *et al.* [81] sequentially carried out image clustering, stereo, stereo fusion, and SfM to generate the density model for a POI. The visual clustering is an effective way to remove irrelevant images from the large collected social images. Due to the fact that social images are redundant, selecting some iconic/representative images for 3-D reconstruction is more efficient. These coarse 3-D models reconstructed from iconic images were subsequently extended using additional, noniconic views to improve the coverage of the POI.

Furukawa *et al.* [84] expanded the key points repeatedly to remove false matches from a large sparse collection of matched feature points. They further enforced local photometric consistency and global visibility constraints to improve the quality of the 3-D density model.

Li *et al.* [86] improved the patch-based multiview stereo (PMVS) to reconstruct 3-D models for a POI from the following two aspects. 1) As the reconstructed point cloud is not well consistent with its local geometry in normal reconstructed 3-D model, they utilize a patch adjusting approach through scene geometric information enhancement for the patch normal estimation. 2) As 3-D model reconstruction is computationally intensive, they propose a multiresolution expanding approach to balance the reconstruction accuracy and the computational cost.

Wang *et al.* [88] proposed an improved version of PMVS based on quasi-dense matching. Patch expansion by building a quasi-dense set of initial patches is efficient to reduce the computational cost in dense model reconstruction.

Useful 3-D geometric information that is available in 3-D patches (i.e., oriented points) has not been completely explored. Song *et al.* [48] proposed tensor-based multiview stereo (TMVS) for quasi-dense 3-D reconstruction from a large unstructured collections of social images. This work is based on the PMVS. They fused photograph consistency, visibility, and geometric consistency in MVS to generate high-quality density models.

3) High-Level Semantic Cognition for POI: Each POI has its own characteristics, which can be also inferred from the appearance of the scene related photos, user's check-in patterns [2]–[5], [8]–[10], [115], and the textual descriptions by social users. Thus, from the social media, high-level semantics for the POIs can be mined by visual content analysis and image annotation, for example, learning the POI functionalities [5], [8], [64] from the textual, visual, user behavior, and geographical sources [64], [65], [115].

C. POI Summarization

From the large user contributed photos taken at a POI from various viewpoints, we know the POI well by selecting representative images with diverse viewpoints [121], [130]. The POI summarization can be derived by the following process: 1) selecting representative images from clusters obtained by mean-shift clustering or visual clustering; 2) selecting representative images from the 3-D reconstruction models; and 3) multimodality-based POI summarization.

1) Representative Image Selection from Clusters: From the geographical distribution point of view, the presentative images can be selected from the locations of interest (LOIs), which can form the smaller scale of the POIs. Images selected from the clusters can serve as the top ranked results. For example, based on the GPS or visual clusters, a representative image can be selected by determining the top ranked image for POI summarization. In both of the above cluster-based representative image selections, it is important to verify the top ranked results with diverse viewpoints.

Qian *et al.* proposed a viewpoint-modeling-based POI summarization approach [121]. They first grouped the images into visual album (VA) by SIFT feature matching and identical semantic points (ISPs) detection [124]–[127], [130]. The relative viewpoint can be modeled by exploring the spatial layout of ISPs. Finally, a 4-D vector was utilized to represent the viewpoint of each image from the POI, which captured the horizontal, vertical, scale, and rotation information of the image. Based on the viewpoint of each image, a representative image with diverse viewpoint can be selected for POI summarization.

2) Representative Image Selection from 3-D Models: The dense 3-D models for the POI are an effective way for the POI visualization. Based on the SfM and 3-D reconstruction, the image taken from each POI can be registered/localized in the real 3-D world coordinates as shown in Figs. 2(b) and (c) and 3(a) and (b). The registered images' locations distribute orderly in the scene, which is helpful in discovering the hot spots for each POI. A representative image can be selected from the distribution of the images in 3-D models of the POI. However, there is not much work in this area reported yet.

3) Representative Image Selection by Multisource Fusion: Except for selecting representative images from the clustered locations, there is multisource information available on social media to select representative images for visual summarization of the POI.

Kennedy *et al.* [162] fused contextual and visual features to generate representative images for POI summarization.

Simon *et al.* [69] proposed a greedy clustering technique to partition the image set into groups. They detected the canonical views based on the co-occurrences of visual properties and the representative tags extracted from the multiuser contributed photo collections for POI summarization.

Zhao *et al.* [120] proposed a Dirichlet process Gaussian mixture model to discover latent scenic themes, such as, sunny, night, snowy, foggy view, etc. The top ranked themes are selected to summarize the POI.

Jiang *et al.* [118] proposed a location-based high-frequency shooting locations of POIs. Visual verification is utilized to diversify the top ranked summarization result.

Ren *et al.* [109] proposed an effective POI summarization approach by an improved geoclustering with visual and view verification. It helps to have a representative and comprehensive perception of POI.

IV. IMAGE LOCALIZATION FROM SOCIAL MEDIA

How to determine accurate location for social images is challenging. This is caused by low sensing accuracy in GPS systems and noise in the generated content from social users [45], [54]. Researchers have proposed some solutions by utilizing the visual information and the multimodality information to improve the performance of location estimation [42]–[77].

A. Inference of Location From Photos

In social media, some of images may not contain geotags or only have coarse geotags [42]–[54]. Visual feature matching or learning-based approaches [45], [47], [49], [53], graph-based approach [48], [59], etc., can be utilized to improve image localization performance.

From large-scale geotagged photos, some researchers try to find repeated patterns [43], [53], [56], [57] to estimate accurate image localization and carry out the geographical relevant tag suggestion [71], while in the visual-feature-matching-based geolocalization approaches, the computation is quite expensive. To reduce the computational cost, representative image selection [61], [62], scalable representation [44], [61], [74], and fast indexing techniques [61], [62], [66], [73] are possible solutions. Different features have different contributions in location inference, thus, effective feature selection [52], [53], weighting, ranking, and reranking approaches are often utilized [43], [46], [58], [61]. Some researchers resort to the multimodality and multisource-based geolocalization approaches to suppress noise in social media [76]. The textual descriptions, timestamps of sequential photos, video sources, etc., can be fused in the geolocalization system. Except for the feature-matching-based location inference approach, learning-based approaches are also utilized [45], [49], [52], [54], [77], such as deep-feature-based image location estimation approaches [42], [46], [50], [51], [55].

The existing visual-appearance-based geolocalization approaches can be classified into the following three categories: 1) visual-matching-based approaches [61]–[64], [74], [125]; 2) salient visual feature enhancement approaches [44], [47], [57], [60], [128]; and 3) learning-based approaches [45]–[49], [54], [77]. Next, we discuss these approaches in more details.

1) Visual-Feature-Matching-Based Geolocalization Approach: Images taken at the same location may share similar appearance, i.e., having similar textures, colors, and local patterns. Thus, the straightforward way is to utilize the GPS locations of the closest visual neighbors to represent the location of the input query image q as follows:

$$l_q = (\text{lat}_q, \text{lon}_q) = (\text{lat}_o, \text{lon}_o),\ o = \arg\min_i (\text{dist}(v_i, v_q)) \tag{9}$$

where $\text{dist}(v_i, v_q)$ is often represented by the Euclidean distance of the two visual feature vectors.

IM2GPS [63] estimated image geographic location from a large collection of geotagged social images by utilizing the low-level visual feature similarity. Many improved methods were proposed [61], [62], [64], [125] to overcome the inefficiency in the nearest-neighbor-based approach location estimation.

Instead of utilizing the nearest neighbors, a voting of the locations from the top-k-nearest neighbors (k-NN) is utilized to determine the location of the query image. The majority GPS location that appears in the top-k images is regarded as the location of the input image. When $k = 1$, this approach is identical to the nearest-neighbor-based approaches [61], [63], [74]. Setting appropriate k is important to achieve better performance.

Li *et al.* [61] proposed a hierarchical clustering approach to estimate image GPS location. They utilized global feature clustering to generate coarse GPS locations. Then, the refined GPS location was obtained by local feature matching from the selected representative images. An inverted file structure of bag of visual words (BoW) for representative images was then built to speed up the online positioning process. Finally, k-NN-based approach was applied to estimate the GPS for the input image.

2) Salient-Feature-Enhancement-Based Approach: The visual-feature-matching-based approaches view the contribution of each single feature identical in image geolocalization. Actually, different features have different contributions in image localization estimation. Thus, feature enhancement approaches can be utilized to improve localization accuracy and reduce computational costs, including feature weighting, salient visual structure bundling, salient region enhancement, and salient feature indexing.

a) *Feature-weighting-based approach:* The visual-feature-based approach is to find visually similar neighbors by weighted feature matching, which can be described as follows:

$$\text{dist}(v_i, v_q) = \sum_{j=1}^{D} w_j \left\| v_i(j) - v_q(j) \right\|_2 \tag{10}$$

where w_j is the weighting vector which is utilized to emphasize each feature bin, and D is the dimensionality of the visual feature. For example, the simplest way is to determine the contribution of each feature (or feature bin) extracted from the visual word models. In this case, the weights can be determined by their TF-IDF information, such as utilizing the discriminative power of the BoW or POI-dependent feature selection [60], [128]. So, by assigning the higher weights to the POI discriminative features (visual words), geolocalization performance can be improved. Salient/discriminative feature for a POI is determined by ranking the TF-IDF values of the BoWs or other local descriptors [44]. The extracted salient feature is called a salient visual word (SVW). Based on the extracted salient feature, feature matching or the spatial consistent checking [57], [124] can be adopted to find visually similar neighbors. Then, k-NN can be utilized to determine the final estimated location for the input image.

Spatial consistent checking and verification [155], [156] can be utilized to enhance the feature in image localization. Spatial verification is proved to be an effective way to find duplicated images. Torii *et al.* [47] used the compact vector of locally aggregated descriptors (VLADs) encoding for local descriptors, to carry out efficient compression, storage, and indexing.

b) *Salient visual structure bundling approach:* The direct feature modeling and feature-weighting-based image localization approaches only consider the contribution of features independently. However, in a POI, some features are appeared concurrently. The images taken at a POI have domain structures or unique local patterns [43], [56], [59], [126]. The image localization performance can also be improved by enhancing the salient regions of each

POI. Two salient structure representation approaches can be explored to estimate the location of the image. They are based on the salient visual words: salient visual word pairs, namely visual phrases [124], [126], and salient regions/patches [43], [62], [152].

A salient visual phrase/pair (SVP) consists of two salient visual words, SVW_1 and SVW_2, as follows:

$$\mathrm{SVP} = (\mathrm{SVW}_1, \mathrm{SVW}_2). \tag{11}$$

One way to select the two SVWs is based on their spatial distance [124] as follows:

$$\mathrm{dist}(\mathrm{SVW}_1, \mathrm{SVW}_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{12}$$

where x_i and y_i denote the abscissa and ordinate values of SVW_i. Other methods, such as the visual phrase selection method, can also be utilized, which is based on the stabilities (invariant in scale, rotation, and illumination) of the SVWs [62], [152].

The SVP is a local structure that consists of two SVWs. Different SVPs have different contributions in visual-based image location estimation [152]. In Fig. 4, the SVPs in green ellipses are more stable than those in blue, as those two SVWs in all the three images are the same. The SVPs in red ellipses shown in Fig. 4(a) and (b) are the same, and they are different from those in Fig. 4(c). So, by counting the frequency of SVPs in each POI, we can derive their discrimination ability in image localization. Ranking the SVPs, stable/robust structure can be explored. To represent the SVP, the spatial coding can be utilized [66], [152], [155], [156]. It is an effective feature to find near duplicate images. Thus, the stable ranking order of the three circles (i.e., SVPs in Fig. 4) is green, red, and blue.

Other approaches also can be utilized to determine the visual phrases [155], [157]. Based on SVW and SVP, salient structures can be explored. Sattler *et al.* [43] proposed to utilize the geometric bursts to localize images represented by a set of visual words that co-occur repeatedly in images with similar distributions.

The visual phrase constructs some local regions/patches which are helpful for identifying the local structure of images. The salient local structures often simultaneously appear in a POI. However, SVW and SVP only can capture a small structure/region/patch in an image. Sometimes a large region can be explored. The region can be derived by exploring the visual attention model [158]–[160], the maximally stable region (MSER) [52], [154], regions after mean-shift clustering [62], [152], Harris affine, an edge-based region detector (EBR), an intensity-extrema-based region detector (IBR), and an entropy-based region detector. More description for the salient region detection can be found in [41].

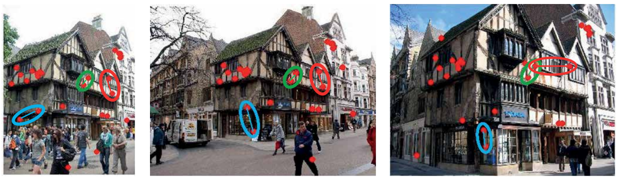

(a) (b) (c)

Fig. 4. ***Salient visual phrase representation for images taken at a POI. Only several salient visual words and salient visual phrases are shown.***

c) *Saliency map:* The saliency map can be also utilized to describe the salient region. Let $S(x, y)$ denote the saliency map of an image. It can be determined by DHSNet [158], RC [159], GBMR [160], and other works [171]. Then, we put those saliency models into our system and recommend pictures. The saliency map information can be fused in finding visually similar neighbors. The strength of the saliency of each pixel or region can be viewed as the weight to constrain the similarity measure.

The similarity can be also measured by some part-model-based approach, for example, the BoW histogram of the region, color, texture patterns, or the spatially consistent information of BoWs [62], [152].

d) *Salient region:* Let $X = \{(x_i, y_i)\}_{i=1}^{n}$ denote the coordinates of SVWs after selection. Mean shift can be utilized to group the salient features into clusters. The sizes of the salient regions after mean-shift clustering are far larger than those of SVWs [62], as shown in Fig. 5. The circles contain a set of SIFT feature points that are closer to each other. The corresponding salient regions after clustering can be utilized to represent the salient structure of the image. In this case, the visual-feature-based image geolocation is converted to measure the similarity of salient structures of images.

For a query image q and a data set image r, we assume that there are m candidate matching regions (patches, structures, or clusters). Let $Str_r^j, (j = 1, 2, \ldots, m)$ denote the corresponding m structure similarity scores.

In [152], the region matching was based on bounded visual words learned from SIFT. In [62], the visual words' geometric information was coded and utilized in similarity measurements.

It is rational to utilize the average score of the candidate matching structure to denote the similarity of those two images. Moreover, the best matched pairs can be represented by the similarity of query image q and refined image r as follows:

$$\mathrm{Score}(r) = (Str_r^j), \ j = 1, 2, \ldots, m. \tag{13}$$

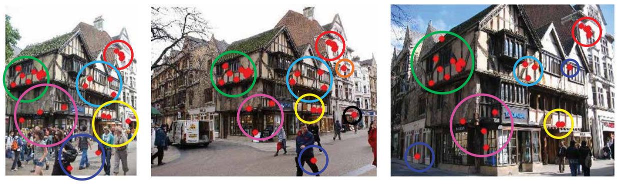


Fig. 5. ***Salient region representation for a POI.***

The maximum similar structure-based similarity measurement approach is robust to the variations of rotation, illumination, scaling, etc. [62], [152]. It is even robust to an occlusion problem in image retrieval.

Li *et al.* [60] exploited geodistinctive visual elements to build a contextual query-specific representation of each candidate location by integrating both the distinctiveness of visual elements and the matching score.

Saurer *et al.* [53] proposed to recognize the mountain by finding the visible skylines in images. The visible skyline was represented by a set of contour words, including their offset angles.

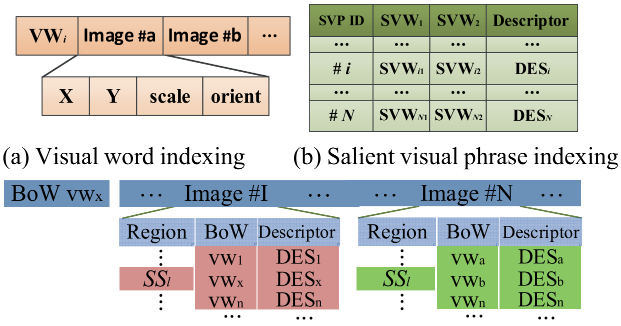
Torii *et al.* [56] utilized a scalable representation for the repeated structures for place recognition in urban environments. They detected and assigned repeated structures in images by exploring the groups of visual words with similar appearance. Furthermore, geometric verification was enforced to suppress ambiguous repeated image patterns.

Shanker *et al.* [59] exploited the inherent structures of urban environments such as man-made buildings that are orderly along streets with geometrically regular intersections for image location estimation. This approach fused the detected roads, intersections, and buildings in an image geolocalization system.

Some researchers also focused their attention on estimating the place where the image was taken at any individual street locations with a few images captured [54]. A cell-based approach was utilized to divide Earth into grids and assign each image to the grids.

e) *Salient region indexing*: The feature-matching-based similarity measurement approach normally is computationally intensive [61]–[65], especially when the size is large. Building a fast indexing structure can speed up the online location estimation process.

The bag of visual models in visual search is also inspired by the fast indexing structure in a text searching engine. An inverted file index for all the images is shown in Fig. 6(a). For each image, the visual word's spatial and description information, such as the coordinates, scale, and orientation of the SIFT feature descriptor, can be recorded. Fig. 6(b) shows the corresponding visual indexing for the visual phrases for an input query image. It only requires to record the two SVWs and their corresponding descriptors. As for each SVW, its indexing structure is identical to that of Fig. 6(a). The salient region indexing structure as shown in Fig. 6(c) is more complicated than those of the visual word and the visual phrase. It builds index for each visual words appearing in the images. Correspondingly, the salient regions in each image are recorded respectively.



(a) Visual word indexing (b) Salient visual phrase indexing

(c) Salient region indexing

Fig. 6. ***Illustration of the inverted file structure for the offline image set. (a) Visual word indexing. (b) Salient visual phrase indexing. (c) Salient region indexing.***

Li *et al.* [61] carried out hierarchical clustering with local refinement to improve image localization performance and reduce computational cost. They selected representative images from each refined cluster to reduce the size of data set and build a fast file structure for representative images to speed up the online image location estimation process.

To improve feature discrimination in image localization, geometric consistent checking, spatial arrangement, and spatial verification for the salient region/structure were often adopted [66], [153]. For example, the spatial coding for the salient visual word [44], [124], the salient visual phrase [126], the visual word group [62], [152], geometric bursts [43], geodistinctive visual elements [60], and so on are effective ways to get better image localization performance.

3) *Learning-Based Geolocalization Approach:* Except for the feature matching and retrieval-based geolocalization approaches [41], [44], [61]–[67], [128], some researchers regarded the image location estimation problem as a classification or location recognition problem [42], [45], [46], [48]–[52], [55].

Given a set of training sets for the locations $\{I_i, u_i, v_i, \mathrm{view}_i, l_i, \mathrm{text}_i, \mathrm{time}_i, \mathrm{camera}_i\}_{i=1}^{N}$, the learning-based approaches utilized the weakly labeled visual features to train the location classifiers. For the input query image q with its visual feature v_q, the learning-based approach directly estimates its location index lq by the trained kernel as follows:

$$lq = \mathrm{sign}\big(K(v_q)\big), \quad l_q \in \{l_n\}_{n=1}^{c} \tag{14}$$

where $K(\cdot)$ is the kernel function, and c is the location number.

Various discriminative learning tools can be utilized in the learning-based approach, including the widely utilized support vector machine (SVM) classifiers [45], [49], [54], [77], graph models [48], and a deep learning approach based on convolutional neural networks (CNNs) [46], [50], [51]. Other learning-based approaches, such as logistic regression (LR), random forest (RF), decision tree (DT), gradient boosting decision tree (GBDT), and SVM cross validation (SVMCV) [161], can also be utilized.

For the SVM classifier, the kernel function is as follows:

$$K(v_q) = \sum_{j=1}^{N} l_j \varepsilon_j \mathrm{SVM}(v_j, v_q) + b \tag{15}$$

where SVM(·) is the trained SVM classifier, and ε_j and b are parameters learned by SVM.

For the CNN-based approach, both the feature from the fully connected layer before softmax and the mapping results of softmax layer can be utilized for location estimation. Moreover, Arandjelovic *et al.* [42] designed a trainable generalized VLAD layer, to aggregate conv5 convolutional features for location estimation. Lin *et al.* [46] used a pair-based network structure to learn deep representations for distinguishing matched and unmatched cross-view image pairs.

B. Multisource-Fusion-Based Image Geolocalization Approach

Social media provide multisource information, which are correlated with each other. Thus, image location can be estimated by multisource fusion. The text information [72], including tags, titles, and textual comments of an image [75], [77], [78], timestamps [76], user information [68], [73], and even other sources of visual content captured by remote sensing satellite, aerial image [46], [51] and video sequences [67], [68], [77] can be fused in a unified system to obtain accurate image localization.

Crandall *et al.* [72] fused visual, textual, and spatial–temporal features of a photo for location estimation. Zheng *et al.* [73] leveraged internet image search engine and the vast amount of multimedia data on the web, to estimate image location. Workman *et al.* [51] aimed to localize ground-level query images by matching against a database of aerial images. They modeled the image localization problem as a cross-view image geolocalization problem. They utilized CNN to learn the similarity between aerial imagery and the corresponding matched ground-level image. Kalogerakis *et al.* [76] modeled the sequence of time-stamped photographs of batch images taken at a certain interval. The process was formulated by the hidden Markov model. The location can be inferred by using the forward–backward algorithm.

C. Three-Dimensional-Modeling-Based Geolocalization Approach

According to the SfM-based 3-D reconstruction from the collected images taken at the POI, a set of 3-D models can be obtained. Correspondingly, the pose of each image/camera can be located in the 3-D word coordinate [89]–[97]. As shown in Fig. 2(b), the image taken locations are registered in the sky. We know that the images were captured by unmanned aerial vehicles, while we can see that positions of the photos are all on the ground for the images shown in Fig. 2(b).

The SfM algorithms estimated the relative camera locations and poses in the 3-D coordinate. The real location estimation process is to align the model with a georeferenced map, such as a remote sensing satellite image. This kind of approach is able to determine the absolute geocentric coordinates of an image (or camera) from an overhead map.

V. CONCLUSION AND FUTURE DIRECTIONS

Social media are full of rich users and services. With more powerful smartphones many latent applications can be carried out at users' mobile end. The giant computing power of cloud computing platform makes it possible for us to provide a quick response to the users' interactions anytime and anyplace. The large volume of social media is convenient for us to learn in detail more about the users' instant requirements. In this paper, we reviewed the existing works on POI localization, cognition, and summarization by exploring the social contextual information. Although temporal, spatial, social, and sentimental aspects have been explored, the bottleneck still exists and there are several directions which can be explored further.

A. Multisource and Multimodality Fusing-Based Localization

In smartphones, there are many applications that can sense users' contextual information and communicate with their local clouds. By gathering the data from various local clouds, we can know more about the users and services by exploring rich multimedia information. How to use these temporal–spatial contextual information and location activity attributes to come up with best personalized services is still a challenge. Most of the recent works only view the GPS information of each photo independently. Their inner correlation is not well explored and fused to achieve better location estimation performance. Moreover, there are various spatial technologies to localize the user. Different spatial technologies have different characteristics and precision. Thus, how to fuse them into a unified framework to improve image localization performances is a future direction.

B. Cross-Media Association-Based POI Cognition

Most of the recent works have focused on only the source from one social medium to carry out POI cognition. However, there are many social media platforms that gather a wide range of sources of the POIs. How to make a full use of multisource social media to obtain complementary cognition for the POI is a future work. However, different social media have different styles. For example, the texts in Weibo are quite short, while the texts in Flickr or other image sharing platforms are much longer. How to fuse them into a unified framework is a challenging problem. We need to balance the mismatching problem of different sources. For example, in some social media, the majority information type is text, while in others, it is image/audio/video. We need to find an effective tool to bridge them and make them contributive in user and POI cognition.

As in social media, different users have different contributions, different influence, and different sharing behaviors, which affect the importance of the source they provided. User cognition and POI cognition can be fused with one another to suppress noise and get comprehensive cognition for the social media. Moreover, in POI summarization, the style can range from various aspects, such as audio, visual, textual, and so on.

C. User Privacy Protection

Applications in users' smartphones can sense their spatial temporal contextual information and infer well the users' working place and location. We can infer users' trajectory with sufficient accuracy through the sequential checkin and cell tower in wireless communications. Thus, with the development of social media and the association within different social media communities, users' privacy may not be safe. How to make sure that the users' data are legally utilized is still a challenging problem. How to detect the existence of illegal information by big social media association and how to remove illegal activity by a data-driven approach are two very urgent issues. How to build a safe defense mechanisms for user data and how to serve the people but not bring them any inconvenience are also challenging topics for the near future.

D. Efficient and Fast 3-D Reconstruction

As for the scene/place, we can get various description information and different languages to illustrate them. In the existing 3-D reconstruction, the order-less social images are utilized to reconstruct the 3-D models for the POI. The images taken from different viewpoints, different seasons, and different periods in a day will construct different 3-D models.

As there are many photos shared by users for the POI, how to select relevant and compensative source to build up the 3-D models efficiently is urgently needed. Not only the redundant images, but also the redundant BoW/SIFT points that are useless for improving the 3-D models can be removed before SfM.

To remove redundant images. 1) As we all know, a fuzzier image can generate many useless features which can affect the quality of the model by noisy 3-D points. 2) Due to a mass of images and lower dimension of global feature, we use a global feature before 3-D reconstruction to remove noisy images. It can generate a visual-similarity images subset quickly.

To remove redundant BoW/SIFT. 1) Many photographers take photos by some criteria to make their photos beautiful and show their subject. There are some well-known heuristic principles which professional photographers follow, including the rule of thirds, diagonal dominance, and sense of balance. So we can extract salient regions in images and just utilize the SIFT points in the regions. 2) Sort SIFT by its importance such as its scale and location and select first *k* features. If they are not matched in the given two images, then the two images are viewed as unmatched. The regular matching is only performed for the matched image pairs.

Directly classifying the photos into different clusters, and then building 3-D models for each cluster, can not only reduce computational costs, but can also improve the reconstruction quality. Moreover, in social media, there are multimodal descriptions for the photos such as tag, title, comments, views, and so on. So making use of them to select photo clusters can carry out fast 3-D reconstruction. ■

REFERENCES

[1] Y. Gu, X. Qian, Q. Li, M. Wang, R. Hong, and Q. Tian, "Image annotation by latent community detection and multikernel learning," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3450–3463, Nov. 2015.

[2] D. Yang, D. Zhang, Z. Yu, and Z. Wang, "A sentiment-enhanced personalized location recommendation system," in *Proc. HT*, 2013, pp. 119–128.

[3] J. Sang, T. Mei, and C. Xu, "Activity sensor: Check-in usage mining for local recommendation," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 3, p. 41, 2015.

[4] C. Biancalana, F. Gasparetti, A. Micarelli, and G. Sansonetti, "An approach to social recommendation for context-aware mobile services," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 1, pp. 871–878, 2013.

[5] V. Zheng, Y. Zheng, X. Xie, and Q. Yang, "Collaborative location and activity recommendations with GPS history data," in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 1029–1038.

[6] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3671, Jun. 2016.

[7] J. Liu, Z. Huang, L. Chen, H. Shen, and Z. Yan, "Discovering areas of interest with geo-tagged images and check-ins," in *Proc. ACM MM*, 2012, pp. 589–598.

[8] N. Yuan, Y. Zheng, X. Xie, Y. Wang, K. Zheng, and H. Xiong, "Discovering urban functional zones using latent activity trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 3, pp. 712–725, Mar. 2015.

[9] H. Hsieh, R. Yan, and C. Li, "Dissecting urban noises from heterogeneous geo-social media and sensor data," in *Proc. ACM MM*, 2015, pp. 1103–1106.

[10] M. Ye, P. Yin, W. Lee, and D. Lee, "Exploiting geographical influence for collaborative point-of-interest recommendation," in *Proc. SIGIR*, 2011, pp. 325–334.

[11] E. Cho, S. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proc. KDD*, 2011, pp. 1082–1090.

[12] C. Cheng, H. Yang, I. King, and M. Lyu, "Fused matrix factorization with geographical and social influence in location-based social networks," in *Proc. AAAI*, 2012, pp. 17–23.

[13] J. Zhang and C. Chow, "GeoSoCa: Exploiting geographical, social and categorical correlations for point-of-interest recommendations," in *Proc. SIGIR*, 2015, pp. 443–452.

[14] G. Cheng, J. Han, L. Guo, Z. Liu, S. Bu, and J. Ren, "Effective and efficient mid-level visual elements-oriented land-use classification using VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4238–4249, Aug. 2015.

[15] H. Yin, Y. Sun, B. Cui, Z. Hu, and L. Chen, "LCARS: A location-content-aware recommender system," in *Proc. KDD*, 2013, pp. 221–229.

[16] B. Liu, Y. Fu, Z. Yao, and H. Xiong, "Learning geographical preferences for point-of-interest recommendation," in *Proc. KDD*, 2013, pp. 1043–1051.

[17] M. Ye, P. Yin, and W. Lee, "Location recommendation for location-based social networks," in *Proc. GIS*, 2010, pp. 458–461.

[18] M. Quezada, V. Araya, and B. Poblete, "Location-aware model for news events in social media," in *Proc. SIGIR*, 2015, pp. 935–938.

[19] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.

[20] Y. Zheng, L. Zhang, X. Xie, and W. Ma, "Mining interesting locations and travel sequences from GPS trajectories," in *Proc. 18th Int. Conf. World Wide Web*, 2009, pp. 791–800.

[21] J. Ying, W. Kuo, V. Tseng, and E. Lu, "Mining user check-in behavior with a random walk

for urban point-of-interest recommendations," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, p. 40, 2014.

[22] H. Yin, B. Cui, L. Chen, Z. Hu, and C. Zhang, "Modeling location-based user rating profiles for personalized recommendation," *ACM Trans. Knowl. Discovery Data*, vol. 9, no. 3, p. 19, 2015.

[23] J. Zhang, C. Chow, and Y. Zheng, "ORec: An opinion-based point-of-interest recommendation framework," in *Proc. CIKM*, 2015, pp. 1641–1650.

[24] Y. Zhao, L. Nie, X. Wang, and T. Chua, "Personalized recommendations of locally interesting venues to tourists via cross-region community matching," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, p. 50, 2014.

[25] X. Li, G. Cong, X. Li, T. Pham, and S. Krishnaswamy, "Rank-GeoFM: A ranking based geographical factorization method for point of interest recommendation," in *Proc. SIGIR*, 2015, pp. 433–442.

[26] J. Bao, Y. Zheng, D. Wilkie, and M. Mokbel, "Recommendations in location-based social networks: A survey," *GeoInformatica*, vol. 19, no. 3, pp. 525–565, 2015.

[27] X. Qian, X. Hua, P. Chen, and L. Ke, "PLBP: An effective local binary patterns texture descriptor with pyramid representation," *Pattern Recognit.*, vol. 44, nos. 10–11, pp. 2502–2515, 2011.

[28] P. Lou, G. Zhao, X. Qian, H. Wang, and X. Hou, "Schedule a rich sentimental travel via sentimental POI mining and recommendation," in *Proc. BigMM*, 2016, pp. 33–40.

[29] E. Kravi, E. Agichtein, I. Guy, Y. Kanza, A. Mejer, and D. Pelleg, "Searcher in a strange land: Understanding Web search from familiar and unfamiliar locations," in *Proc. SIGIR*, 2015, pp. 855–858.

[30] X. Wang *et al.*, "Semantic-based location recommendation with multimodal venue semantics," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 409–419, Mar. 2015.

[31] G. Zhao, X. Qian, and C. Kang, "Service rating prediction by exploring social mobile users' geographical locations," *IEEE Trans. Big Data*, vol. 3, no. 1, pp. 67–78, Mar. 2017.

[32] B. Hu and M. Ester, "Spatial topic modeling in online social media for location recommendation," in *Proc. RecSys*, 2013, pp. 25–32.

[33] J. Zhang and C. Chow, "Spatiotemporal sequential influence modeling for location recommendations: A gravity-based approach," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 1, p. 11, 2015.

[34] Y. Kim, J. Han, and C. Yuan, "TOPTRAC: Topical trajectory pattern mining," in *Proc. KDD*, 2015, pp. 587–596.

[35] V. Zheng, Y. Zheng, X. Xie, and Q. Yang, "Towards mobile intelligence: Learning from GPS history data for collaborative recommendation," *Artif. Intell.*, vols. 184–185, pp. 17–37, Jun. 2012.

[36] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[37] H. Hsieh, T. Yen, and C. Li, "What makes New York so noisy?: Reasoning noise pollution by mining multimodal geo-social big data," in *Proc. ACM Multimedia*, 2015, pp. 181–184.

[38] R. Ji, Y. Gao, W. Liu, X. Xie, Q. Tian, and X. Li, "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, p. 1, 2015.

[39] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. ECCV*, 2006, pp. 404–417.

[40] Y. Zhong, N. Yuan, W. Zhong, F. Zhang, and X. Xie, "You are where you go: Inferring demographic attributes from location check-ins," in *Proc. WSDM*, 2015, pp. 295–304.

[41] K. Mikolajczyk *et al.*, "A comparison of affine region detectors," *Int. J. Comput. Vis.*, vol. 65, nos. 1–2, pp. 43–72, 2005.

[42] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5297–5307.

[43] T. Sattler, M. Havlena, K. Schindler, and M. Pollefeys, "Large-scale location recognition and the geometric burstiness problem," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1582–1590.

[44] R. Arandjelović and A. Zisserman, "DisLocation: Scalable descriptor distinctiveness for location recognition," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2014, pp. 188–204.

[45] R. Gopalan, "Hierarchical sparse coding with geometric prior for visual geo-location," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2432–2439.

[46] T. Y. Lin, Y. Cui, S. Belongie, and J. Hays, "Learning deep representations for ground-to-aerial geolocalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5007–5015.

[47] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1808–1817.

[48] S. Cao and N. Snavely, "Graph-based discriminative learning for location recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 700–707.

[49] B. Zhou, L. Liu, and A. Oliva, "Recognizing city identity via attribute analysis of geo-tagged images," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 519–534.

[50] T. Weyand, I. Kostrikov, and J. Philbin, "PlaNet-photo geolocation with convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 37–55.

[51] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Jun. 2015, pp. 3961–3969.

[52] H. Kim, E. Dunn, and J. Frahm, "Predicting good features for image geo-localization using per-bundle VLAD," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1170–1178.

[53] O. Saurer, G. Baatz, K. Köser, and M. Pollefeys, "Image based geo-localization in the alps," *Int. J. Comput. Vis.*, vol. 116, no. 3, pp. 213–225, 2016.

[54] P. Gronat, G. Obozinski, J. Sivic, and T. Pajdla, "Learning and calibrating per-location classifiers for visual place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 907–914.

[55] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 1. 2014, pp. 487–495.

[56] A. Torii, J. Sivic, M. Okutomi, and T. Pajdla, "Visual place recognition with repetitive structures," *IEEE Trans. Pattern Anal., Mach. Intell.*, vol. 37, no. 11, pp. 2346–2359, Nov. 2015.

[57] P. Panphattarasap and A. Calway, "Visual place recognition using landmark distribution descriptors," *Comput. Res. Repository*, Aug. 2016.

[58] 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.

[59] V. Shankar, J. Zhang, J. Chen, C. Dinh, M. Clements, and A. Zakhor, "Approximate subgraph isomorphism for image localization," *Image Process., Algorithms Syst.*, vol. 2016, no. 15, pp. 1–9, 2016.

[60] X. Li, M. A. Larson, and A. Hanjalic, "Geo-distinctive visual element matching for location estimation of images," *Comput. Res. Repository*, May 2016.

[61] J. Li, X. Qian, Y. Tang, L. Yang, and T. Mei, "GPS estimation for places of interest from social users' uploaded photos," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 2058–2071, Dec. 2013.

[62] X. Qian, Y. Zhao, and J. Han, "Image location estimation by salient region matching," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4348–4358, Nov. 2015.

[63] J. Hays and A. Efros, "IM2GPS: Estimating geographic information from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[64] Y. Li, D. Crandall, and D. Huttenlocher, "Landmark classification in large-scale image collections," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 1957–1964.

[65] T. Quack, B. Leibe, and L. Van Gool, "World-scale mining of objects and events from community photo collections," in *Proc. Int. Conf. Content-Based Image Video Retr.*, 2008, pp. 47–56.

[66] W. Zhou, Y. Lu, H. Li, Y. Song, and Q. Tian, "Spatial coding for large scale partial-duplicate Web image search," in *Proc. Int. Conf. Multimedia*, 2010, pp. 511–520.

[67] M. Trevisiol, J. Delhumeau, and H. Jégou, "How INRIA/IRISA identifies geographic location of videos," in *Proc. Working Notes Mediaeval Workshop*, 2012.

[68] O. Laere, S. Schockaert, and B. Dhoedt, "Ghent University at the 2011 placing task," in *Proc. Working Notes Mediaeval Workshop*, 2011, pp. 385–392.

[69] I. Simon, N. Snavely, and S. Seitz, "Scene summarization for online image collections," in *Proc. ICCV*, 2007, pp. 1–8.

[70] A. Popescu and P. Moëllic, "MonuAnno: Automatic annotation of georeferenced landmarks images," in *Proc. ACM Int. Conf. Image Video Retr.*, 2009, p. 11.

[71] J. Kleban, E. Moxley, J. Xu, and B. Manjunath, "Global annotation on georeferenced photographs," in *Proc. ACM Int. Conf. Image Video Retr.*, 2009, p. 12.

[72] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *Proc. 18th Int. Conf. World Wide Web*, 2009, pp. 761–770.

[73] Y. Zheng *et al.*, "Tour the world: Building a Web-scale landmark recognition engine," in

Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 1085–1092.

[74] J. Li, X. Qian, Y. Tang, L. Yang, and C. Liu, "GPS estimation from users' photos," in *Advances in Multimedia Modeling*. 2013, pp. 118–129.

[75] L. Tzy, J. Almeida, D. Petronette, O. Penatti, and R. Torres, "A multimodal approach for video geocoding at mediaeval 2012," in *Proc. MediaEval Workshop*, 2012.

[76] E. Kalogerakis, O. Vesselova, J. Hays, A. Efros, and A. Hertzmann, "Image sequence geolocation with human travel priors," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Oct. 2009, pp. 253–260.

[77] P. Kelm, S. Schmiedeke, and T. Sikora, "Video2GPS: Geotagging using collaborative systems, textual and visual features," in *Proc. MediaEval Workshop*, 2010.

[78] P. Kelm, S. Schmiedeke, and T. Sikora, "How spatial segmentation improves the multimodal geo-tagging," in *Proc. MediaEval Workshop*, 2012.

[79] S. Agarwal, Y. Furukawa, N. Snavely, B. Curless, S. M. Seitz, and R. Szelisk, "Reconstructing Rome," *Computer*, vol. 43, no. 6, pp. 40–47, 2010.

[80] N. Snavely, S. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3D," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 835–846, Jul. 2006.

[81] J. Frahm *et al.*, "Building Rome on a cloudless day," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 368–381.

[82] R. Raguram, C. Wu, J. Frahm, and S. Lazebnik, "Modeling and recognition of landmark image collections using iconic scene graphs," *Int. J. Comput. Vis.*, vol. 95, no. 3, pp. 213–239, 2011.

[83] C. Sweeney, T. Sattler, T. Höllerer, M. Turk, and M. Pollefeys, "Optimizing the Viewing Graph for Structure-from-Motion," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 801–809.

[84] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Trans. Pattern Anal., Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, Aug. 2009.

[85] C. Song and N. Snavely, "Minimal scene descriptions from structure from motion models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 461–468.

[86] M. Li, S. Fu, and Y. Zhan, "An improved PMVS through scene geometric information," *Acta Automatica Sinica*, p. 37, 2011.

[87] S. Shen, "Accurate multiple view 3D reconstruction using patch-based stereo for large-scale scenes," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1901–1914, May 2013.

[88] L. Wang, R. Chen, and D. Kong, "An improved patch based multi-view stereo (PMVS) algorithm," in *Proc. Int. Conf. Comput. Sci. Service Syst.*, 2014.

[89] Y. Li, N. Snavely, and D. Huttenlocher, "Location recognition using prioritized feature matching," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 791–804.

[90] A. Irschara, C. Zach, J. Frahm, and H. Bischof, "From structure-from-motion point clouds to fast location recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2599–2606.

[91] T. Sattler, M. Havlena, F. Radenovic, K. Schindler, and M. Pollefeys, "Hyperpoints and fine vocabularies for large-scale location recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 2102–2110.

[92] J. Knopp, J. Sivic, and T. Pajdla, "Avoiding confusing features in place recognition," in *Proc. Eur. Conf. Comput. Vis. Comput. Vis.-ECCV*, 2010, pp. 748–761.

[93] T. Sattler, B. Leibe, and L. Kobbelt, "Improving image-based localization by active correspondence search," in *Proc. ECCV*, 2012, pp. 752–765.

[94] T. Sattler, B. Leibe, and L. Kobbelt, "Fast image-based localization using direct 2D-to-3D matching," in *Proc. ICCV*, vol. 24, no. 4, pp. 667–674, 2011.

[95] Y. Li, N. Snavely, H. Dan, and P. Fua, "Worldwide pose estimation using 3D point clouds," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 396–404.

[96] H. Lim, S. Sinha, M. Cohen, and M. Uyttendaele, "Real-time image-based 6-DOF localization in large-scale environments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1043–1050.

[97] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt, "Image retrieval for image-based localization revisited," in *Proc. BMVC*, 2012, p. 4.

[98] S. Jiang, X. Qian, J. Shen, Y. Fu, and T. Mei, "Author topic model-based collaborative filtering for personalized POI recommendations," *IEEE Trans. Multimedia*, vol. 17, no. 6, pp. 907–918, Jun. 2015.

[99] H. Kori, S. Hattori, and T. Tezuka, "Automatic generation of multimedia tour guide from local blogs," in *Advances in Multimedia Modeling.* Berlin, Germany: Springer, 2007, pp. 690–699.

[100] Q. Yuan, G. Cong, and A. Sun, "Graph-based point-of-interest recommendation with geographical and temporal influences," in *Proc. 23rd ACM Int. Conf. Inf. Knowl. Manage.*, 2014, pp. 659–668.

[101] M. Clements, P. Serdyukov, A. Vries, and M. Reinders, "Using flickr geotags to predict user travel behaviour," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, Geneva, Switzerland, Jul. 2010, pp. 851–852.

[102] H. Huang and G. Gartner, "Using trajectories for collaborative filtering-based POI recommendation," *Int. J. Data Mining Model. Manage.*, vol. 6, no. 4, pp. 333–346, 2014.

[103] C. Zhang and K. Wang, "POI recommendation through cross-region collaborative filtering," *Knowl. Inf. Syst.*, vol. 46, no. 2, pp. 369–387, 2016.

[104] J. Bao, Y. Zheng, and M. Mokbel, "Location-based and preference-aware recommendation using sparse geo-social networking data," in *Proc. GIS*, 2012, pp. 199–208.

[105] P. Chitra and H. Girijamma, "Hybrid approach for location based customized POI travel recommendation system," *Int. J. Adv. Trends Comput. Sci., Eng.*, vol. 5, no. 5, 2016.

[106] G. Xu, B. Fu, and Y. Gu, "Point-of-interest recommendations via a supervised random walk algorithm," *IEEE Intell. Syst.*, vol. 31, no. 1, pp. 15–23, Jan./Feb., 2016.

[107] S. Jiang, X. Qian, J. Shen, and T. Mei, "Travel recommendation via author topic model based collaborative filtering," *MultiMedia Modeling.* Springer, 2015, pp. 392–402.

[108] G. Suganeshwari and S. Ibrahim, "A survey on collaborative filtering based recommendation system," in *Smart Innovation, Systems and Technologies*. 2016.

[109] Y. Ren, X. Qian, and S. Jiang, "Visual summarization for place-of-interest by social-contextual constrained geo-clustering," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, Oct. 2015, pp. 1–6.

[110] E. Spyrou, A. Psallas, V. Charalampidis, and P. Mylonas, "Discovering areas of interest using a semantic geo-clustering approach," *Artificial Intelligence Applications and Innovations*, to be published.

[111] C. Huang and D. Wang, "Unsupervised interesting places discovery in location-based social sensing," in *Proc. IEEE Int. Conf. Distrib. Comput. Sensor Syst.*, May 2016, pp. 67–74.

[112] C. Huang and D. Wang, "On interesting place finding in social sensing: An emerging smart city application paradigm," in *Proc. IEEE Int. Conf. Smart City/ Socialcom/Sustaincom*, Oct. 2015, pp. 13–20.

[113] A. Majid, L. Chen, H. T. Mirza, I. Hussain, and G. Chen, "A system for mining interesting tourist locations and travel sequences from public geo-tagged photos," *Data Knowl. Eng.*, vol. 95, pp. 66–86, Jan. 2015.

[114] H. Ying, L. Chen, Y. Xiong, and J. Wu, "PGRank: Personalized geographical ranking for point-of-interest recommendation," in *Proc. Int. Conf. Companion Int. World Wide Web Conf. Steering Committee*, 2016, pp. 137–138.

[115] K. Zhao, G. Cong, and A. Sun, "Annotating points of interest with geo-tagged tweets," in *Proc. ACM CIKM*, pp. 417–426, 2016.

[116] Z. Yu, H. Xu, Z. Yang, and B. Guo, "Personalized travel package with multi-point-of-interest recommendation based on crowdsourced user footprints," *IEEE Trans. Human–Mach. Syst.*, vol. 46, no. 1, pp. 151–158, Jan. 2016.

[117] S. Jiang, X. Qian, T. Mei, and Y. Fu, "Personalized travel sequence recommendation on multi-source big social media," *IEEE Trans. Big Data*, vol. 2, no. 1, pp. 43–56, Jan. 2016.

[118] S. Jiang, X. Qian, Y. Xue, F. Li, and X. Hou, "Generating representative images for landmark by discovering high frequency shooting locations from community-contributed photos," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, Aug. 2013, pp. 1–6.

[119] Y. Xue and X. Qian, "Visual summarization of landmarks via viewpoint modeling," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2012, pp. 2873–2876.

[120] Y. Zhao, Y.-T. Zheng, X. Zhou, and T.-S. Chua, "Generating representative views of landmarks via scenic theme detection," in *Proc. Adv. Multimedia Modeling Int. Multimedia Modeling Conf. (MMM)* Taipei, Taiwan, Jan. 2011, pp. 392–402.

[121] X. Qian, Y. Xue, X. Yang, Y. Y. Tang, X. Hou, and T. Mei, "Landmark summarization with diverse viewpoints," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 11, pp. 1857–1869, Nov. 2014.

[122] S. Jiang, X. Qian, K. Lan, L. Zhang, and T. Mei, "Mobile multimedia travelogue generation by exploring geo-locations and image tags," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2013, pp. 881–884.

[123] H. Chen, B. Guo, Z. Yu, and Q. Han, "Toward real-time and cooperative mobile

visual sensing and sharing," in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Apr. 2016, pp. 1–9.

[124] X. Yang and X. Qian, "Spatial verification for scalable mobile image retrieval," presented at the ACM Int. Conf. Inf. Knowl. Manag., Shanghai, China, Nov. 2014.

[125] Y. Xue, X. Qian, and B. Zhang, "Mobile image retrieval using multiphotos as query," in *Proc. Int. Conf. Multimedia Expo Workshops*, Jul. 2013, pp. 1–4.

[126] X. Yang, X. Qian, and Y. Xue, "Scalable mobile image retrieval by exploring contextual saliency," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1709–1721, Jun. 2015.

[127] X. Yang, X. Qian, and T. Mei, "Learning salient visual word for scalable mobile image retrieval," *Pattern Recognit.*, vol. 48, no. 10, pp. 3093–3101, 2015.

[128] J. Li, X. Qian, K. Lan, P. Qi, and A. Sharma, "Improved image GPS location estimation by mining salient features," *Signal Process., Image Commun.*, vol. 38, pp. 141–150, Oct. 2015.

[129] J. Li, X. Qian, Q. Li, Y. Zhao, L. Wang, and Y. Tang, "Mining near duplicate image groups,"Multimedia Tools Appl. vol. 74, no. 2, pp. 655–669(2015).

[130] X. Lu, C. Wang, J. Yang, Y. Pang, and L. Zhang, "Photo2Trip: Generating travel routes from geo-tagged photos for trip planning," in *Proc. Int. Conf. Multimedia*, 2010, pp. 143–152.

[131] Y. Pang, Q. Hao, Y. Yuan, T. Hu, R. Cai, and L. Zhang, "Summarizing tourist destinations by mining user-generated travelogues and photos," *Comput. Vis. Image Understand.*, vol. 115, no. 3, pp. 352–363, Mar. 2011.

[132] Q. Hao *et al.*, "Equip tourists with knowledge mined from travelogues," in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 401–410.

[133] Q. Hao, R. Cai, X. Wang, J. Yang, Y. Pang, and L. Zhang, "Generating location overviews with images and tags by mining user-generated travelogues," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 801–804.

[134] X. Qian, X. Liu, C. Zheng, Y. Du, and X. Hou, "Tagging photos using users' vocabularies," *Neurocomputing*, vol. 111, pp. 144–153, Jul. 2013.

[135] D. Lu, X. Liu, and X. Qian, "Tag based image search by social re-ranking," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1628–1639, Aug. 2016.

[136] A. Cheng, Y. Chen, Y. Huang, W. Hsu, and H. Liao, "Personalized travel recommendation by mining people attributes from community-contributed photos," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 83–92.

[137] H. Feng and X. Qian, "Mining user-contributed photos for personalized product recommendation," *Neurocomputing*, vol. 129, pp. 409–420, Apr. 2014.

[138] X. Qian, X. Tan, Y. Zhang, R. Hong, and M. Wang, "Enhancing sketch-based image retrieval by re-ranking and relevance feedback," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 195–208, Jan. 2016.

[139] X. Qian, H. Feng, G. Zhao, and T. Mei, "Personalized recommendation combining user interest and social circle," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 7, pp. 1487–1502, Jul. 2014.

[140] G. Zhao, X. Qian, and X. Xie, "User-service rating prediction by exploring social users' rating behaviors," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 496–506, Mar. 2016.

[141] X. Lei, X. Qian, and G. Zhao, "Rating prediction based on social sentiment from textual reviews," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1910–1921, Sep. 2016.

[142] G. Zhao, X. Qian, X. Lei, and T. Mei, "Service quality evaluation by exploring social users' contextual information," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3382–3394, Dec. 2016.

[143] X. Qian, X. Hua, Y. Tang, and T. Mei, "Social image tagging with diverse semantics," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2493–2508, Dec. 2014.

[144] S. Zhao, I. King, and R. Lyu, "A survey of point-of-interest recommendation in location-based social networks," *CoRR*, Jul. 2016.

[145] X. Lu, Y. Yuan, and X. Zheng, "Joint dictionary learning for multispectral change detection," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 884–897, Apr. 2016.

[146] Y. Yuan, H. Lv, and X. Lu, "Semi-supervised change detection method for multi-temporal hyperspectral images," *Neurocomputing*, vol. 148, pp. 363–375, Jan. 2015.

[147] X. Lu, H. Wu, and Y. Yuan, "Double constrained NMF for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2746–2758, May 2014.

[148] X. Lu, H. Wu, Y. Yuan, P. Yan, and X. Li, "Manifold regularized sparse NMF for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2815–2826, May 2013.

[149] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.

[150] F. Zhang, B. Du, and L. Zhang, "Scene classification via a gradient boosting random convolutional network framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1793–1802, Mar. 2016.

[151] F. Zhang, B. Du, L. Zhang, and M. Xu, "Weakly supervised learning based on coupled convolutional neural networks for aircraft detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 9, pp. 5553–5563, Sep. 2016.

[152] X. Qian *et al.*, "Image location inference by multisaliency enhancement," *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 813–821, Apr. 2016.

[153] O. Penatti, F. Silva, and E. Valle, "Visual word spatial arrangement for image retrieval and classification," *Pattern Recognit.*, vol. 47, no. 2, pp. 705–720, 2014.

[154] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate Web image search," in *Proc. CVPR*, 2009, pp. 25–32.

[155] S. Zhang, Q. Tian, G. Hua, Q. Huang, and W. Gao, "Generating descriptive visual words and visual phrases for large-scale image applications," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2664–2677, Sep. 2011.

[156] S. Zhang, M. Yang, X. Wang, Y. Lin, and Q. Tian, "Semantic-aware co-indexing for image retrieval," in *Proc. ICCV*, 2013, pp. 1673–1680.

[157] V. Kumar, A. Namboodiri, and C. Jawahar, "Visual phrases for exemplar face detection," in *Proc. ICCV*, 2015, pp. 1994–2002.

[158] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. CVPR*, 2016, pp. 678–686.

[159] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," in *Proc. CVPR*, 2011, pp. 409–416.

[160] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. CVPR*, 2013, pp. 3166–3173.

[161] A. Astorino and A. Fuduli, "The proximal trajectory algorithm in SVM cross validation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 5, pp. 966–977, May 2016.

[162] L. Kennedy and M. Naaman, "Generating diverse and representative image search results for landmarks," in *Proc. Conf. 19th World Wide Web*, 2008, pp. 297–306.

[163] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[164] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.

[165] H. Deng, L. Zhang, X. Mao, and H. Qu, "Interactive urban context-aware visualization via multiple disocclusion operators," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 7, pp. 1862–1874, Jul. 2016.

[166] F. Grabler, M. Agrawala, R. W. Sumner, and M. Pauly, "Automatic generation of tourist maps," *ACM Trans. Graph.*, vol. 27, no. 3, p. 100, 2008.

[167] L. Zhang, H. Deng, D. Chen, and Z. Wang, "A spatial cognition-based urban building clustering approach and its applications," *Int. J. Geograph. Inf. Sci.*, vol. 27, no. 4, pp. 721–740, 2013.

[168] G. Ference, M. Ye, and W. Lee, "Location recommendation for out-of-town users in location-based social networks," in *Proc. ACM CIKM*, 2013, pp. 721–726.

[169] D. Lian, C. Zhao, X. Xie, G. Sun, E. Chen, and Y. Ru, "GeoMF: Joint geographical modeling and matrix factorization for point-of-interest recommendation," in *Proc. KDD*, 2014, pp. 831–840.

[170] X. Qian, D. Lu, Y. Wang, L. Zhu, Y. Tang, and M. Wang, "Image re-ranking based on topic diversity," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3734–3747, Aug. 2017.

[171] D. Zhang, J. Han, C. Li, J. Wang, and X. Li, "Detection of co-salient objects by looking deep and wide," *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 215–232, 2016.

[172] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2017.2702596.

ABOUT THE AUTHORS

Xueming Qian (Member, IEEE) received the B.S. and M.S. degrees from Xi'an University of Technology, Xi'an, China, in 1999 and 2004, respectively, and the Ph.D. degree from the School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, China, in 2008.

He was a Visiting Scholar with Microsoft Research Asia from 2010 to 2011. He was an Assistant Professor with Xi'an Jiaotong University, where he was an Associate Professor from 2011 to 2014, and is currently a Full Professor. He is also the Director of the Smiles Laboratory at Xi'an Jiaotong University.

Dr. Qian received the Microsoft Fellowship in 2006. He received outstanding doctoral dissertations of Xi'an Jiaotong University and Shaanxi Province, in 2010 and 2011, respectively. His research interests include social media big data mining and search. His research is supported by the National Natural Science Foundation of China, Microsoft Research, and Ministry of Science and Technology.

Xiaoqiang Lu (Senior Member, IEEE) is a full Professor with the Chinese Academy of Sciences, Beijing, China, and an Associate Director of the Research Center. His research interests include machine learning, hyperspectral remote sensing image processing, image/video analysis and computational intelligence. He has published more than 80 research papers on some famous journals (PIEEE, TIP, TCSVT, TNNLS, TCYB, TGRS, JSTAR) and conferences (CVPR, ICCV, ACM MM).

Dr. Lu serves as the associated editor of the IEEE Transactions on Geoscience and Remote Sensing (IEEE TGRS), the editorial board of *Neurocomputing* (Elsevier), *Cognitive Computation* (Springer), and *International Journal of Image and Graphics* (World of Scientific), and the lead guest editor of *Photogrammetry and Remote Sensing* (ISPRS). He was associated with more than 100 IEEE conferences as program committee members and served as a referee of more than 30 journals including IEEE TNNLS, IEEE TCYB, IEEE TCSVT, TMI, TBME, TKDE, TIE, IEEE TIP, IEEE TGRS, PR, CVIU, SP, etc.

Junwei Han (Senior Member, IEEE) is currently a Full Professor with Northwestern Polytechnical University, Xi'an, China. He was a Research Fellow at the Nanyang Technological University, The Chinese University of Hong Kong, and University of Dundee. He was a visiting researcher at the University of Surrey and Microsoft Research Asia. His research interests include computer vision, multimedia processing, and brain imaging analysis. He has published more than 60 papers in top journals and conferences such as IEEE TPAMI, IJCV, TIP, CVPR, ICCV, IJCAI, and so on.

Dr. Han is an Associate Editor of the IEEE Transactions on Human-Machine Systems, *Neurocomputing, Multidimensional Systems and Signal Processing,* and *Machine Vision and Applications.*

Bo Du (Senior Member, IEEE) received the B.S. degree and the Ph.D. degree in photogrammetry and remote sensing from State Key Lab of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2005 and 2010, respectively.

He is currently a Professor with the School of Computer, Wuhan University. He has more than 40 research papers published in the IEEE Transactions on Geoscience and Remote Sensing (TGRS), IEEE Transactions on Image Processing (TIP), IEEE Journal of Selected Topics in Earth Observations and Applied Remote Sensing (JSTARS), IEEE Geoscience and Remote Sensing Letters (GRSL), etc. Five of them are ESI hot papers or highly cited papers. His major research interests include pattern recognition, hyperspectral image processing, and signal processing.

Dr. Du received the best reviewer awards from IEEE GRSS for his service to the IEEE Journal of Selected Topics in Earth Observations and Applied Remote Sensing (JSTARS) in 2011 and ACM rising star awards for his academic progress in 2015. He was the Session Chair for both the International Geoscience And Remote Sensing Symposium (IGARSS) 2016 and the 4th IEEE GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS). He also serves as a reviewer of 20 Science Citation Index (SCI) magazines including IEEE TGRS, TIP, JSTARS, and GRSL.

Xuelong Li (Fellow, IEEE) is a Full Professor with the Center for Optical Imagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China.