LAST: Location-Appearance-Semantic-Temporal Clustering Based POI Summarization

Xueming Qian, *Member, IEEE*, Yuxia Wu, Mingdi Li, Yayun Ren, Shuhui Jiang, and Zhetao Li

***Abstract*—When planning a trip, users tend to browse Place-of-Interest (POI) information on the Internet and then depart. Many works aimed at summarizing POIs by visual and textual analysis, while many of them ignored the inter-relationship between different views offered by the community-contributed information. In this paper, we propose a City-POI-LOI (CPL) summarization method to automatically mine POIs from the city-level landmark images. And a Location-Appearance-Semantic-Temporal (LAST) clustering method is proposed to mine the popular viewpoints termed Location-Of-Interest (LOI) in each POI by taking location, appearance, semantic, and temporal feature into consideration. We perform text and image summarization for each LOI, and we further summarize the POIs based on season. We conduct a series of experiments based on DIV400 and ATCF Dataset. Experimental results show the effectiveness of the proposed POI summarization approach.**

***Index Terms*—Clustering, feature extraction, multimedia, POI summarization, social media.**

I. Introduction

With the popularity of smart terminal and the rapid development of social media, more and more users are willing to share their lives on social network sites, including their travel, shopping etc.

Thus, there emerges a large amount of social media information online, including text, image, video, audio, etc. And with the popularity of smart phone, images uploaded by users always have much useful information, e.g., the time of taking photo, location, tags, viewing times. For example, Flickr had a total of 87 million registered members and more than 3.5 million new images uploaded daily in March 2013 [34]. This large amount of data not only facilitates the big media management but also gives us a wealth of resources to carry out travel planning [4], [5], [36]–[38], [43], [44], [47], [63].

A lot of existing works are devoted to excavate POIs from the massive social media information. As we know, there are many popular locations where people usually go and take photos. We call these interesting locations as Location-Of-Interest (LOI). When planning a trip, users often browse many representative images provided by websites and then determine where to go. The representative images are convenient for the visitors to fetch more detailed information about the POIs. Automatic POI summarization will be time-saving for users to obtain the first-hand materials of the POI and convenient for them to make plan.

Images shared in social media usually have much supplementary information, e.g., tags, location, image taken time etc. They are very valuable information for LOI mining. For instance, images with similar tags are more likely to be similar [45]. And locations where many pictures are shot and uploaded are more likely to be LOIs in POI [4], [20], [36], [37]. Fusing multimodal information can help us better learn the comprehensive representations of POIs and the preferences of users [60], [61]. In this paper, we treat location, visual, semantic and temporal features as different views in our multi-view clustering framework. In [6]–[8], researchers aimed at finding the optimal weights for the different views, or the best weights between groups within individual view. However, they ignored the closeness of the groups and the diversity of different groups in calculating the clustering quality within one view and the clustering consistency cross different views.

Many researchers mined landmark from the large amount of community-contributed information [1]–[5], [59] and performed personalized POI recommendation based on users' interests [5], [16], [17], [25], [30]. Zheng *et al.* [42] mined a comprehensive list of landmarks based on 20 million GPS-tagged photos and online tour guide web pages. It should be noted that there are two challenges in POI summarization: 1) how to summarize POI accurately by utilizing multi-modality information available from social media. Many existing works consider the visual, textual, and location feature independently in LOI mining [6], [33]. But for the images, the visual distribution is complicated due to luminance variations, and shooting locations angles changing. The pictures shot in the day are generally bright, while the pictures shot at night are usually dark. Besides, the location and textual descriptions also imply the latent popular viewpoints.

Manuscript received December 21, 2017; revised March 5, 2019, December 28, 2019, and February 13, 2020; accepted February 18, 2020. Date of publication March 2, 2020; date of current version December 17, 2020. This work was supported in part by the NSFC under Grants 61732008 and 61772407 and in part by Guangdong Provincial Science and Technology Plan under Grant 2016A010101005. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Pradeep K. Atrey. *(Corresponding author: Xueming Qian.)*

Xueming Qian is with the Ministry of Education Key Laboratory for Intelligent Networks and Network Security, School of Information and Communication Engineering, and SMILES LAB, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: qianxm@mail.xjtu.edu.cn).

Yuxia Wu, Mingdi Li, and Yayun Ren are with the School of Information and Communication Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: wuyuxia@stu.xjtu.edu.cn; limingdi@stu.xjtu.edu.cn; renyy@stu.xjtu.edu.cn).

Shuhui Jiang is with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 USA (e-mail: shjiang@ece.neu.edu).

Zhetao Li is with the College of Information Engineering, Xiangtan University, Hunan 411105, China (e-mail: liztchina@gmail.com).

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

Digital Object Identifier 10.1109/TMM.2020.2977478

1520-9210 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on September 06,2022 at 12:43:08 UTC from IEEE Xplore. Restrictions apply.

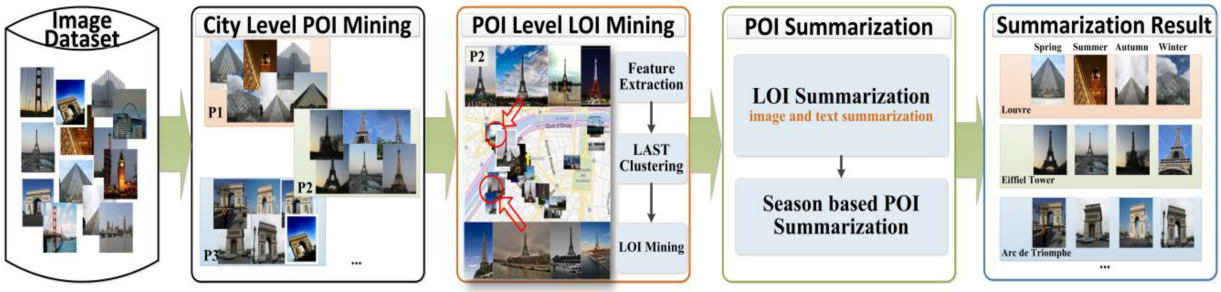


Fig. 1. The framework of the proposed City-POI-LOI summarization.

2) How to summarize the POI by multimodality source information rather the only visual feature based approach. There are many existing works that only provide representative images of the POIs. Nevertheless, representative textual descriptions such as tags are also important for users' understanding of the POIs. In addition, users are usually concerned about the golden time for travel. For example, spring is the best time for the cherry blossom in Japan.

Facing with the above two challenges, we proposed a coarse to fine POI summarization approach named City-POI-LOI (in short CPL). The framework of CPL is shown in Fig. 1. It consists of the following steps. Firstly, we remove the irrelevant images and reserve the relevant images. Secondly, we mine famous city-level POIs in each city. Then, a location-appearance-semantic-time (LAST) clustering approach is proposed to find LOIs for a POI. Finally, a season based POI summarization approach is proposed by selecting representative tags and images.

The main contributions of our work are as follows:

1) We propose a LAST clustering method that fuses the location, appearance, semantic, and temporal information into POI summarization system. LAST clustering method not only finds the best clustering of images, but also improves the performance of LOI mining. Furthermore, we learn different weights for these four views for multi-view clustering. Different from traditional clustering method, we further consider the clustering quality within view and the consistency across views. Specially, for the clustering quality within view, we try to guarantee the closeness of intra-group and the diversity of inter-group to get better results.
2) We fuse visual representativeness, significance, and season relevance to rank images for each POI. Besides, TF-ILF (Tag Frequency-Inverse LOI Frequency) and season relevance are proposed to rank texts of each LOI. In this way, we can choose the most representative images and texts for users to explore interesting, appealing and important LOIs for trip planning. And we also specially and innovatively propose a season based POI summarization approach to show the representative images and texts of LOIs in different season.

II. Related Work

In this paper, we are aimed to perform POI summarization on the crowd Dataset. Meanwhile, we propose a new multi-view clustering method named LAST, so we review two fields of related works.

A. POI Summarization

Recently, a great number of researches have been dedicated to visual summarization [1]–[4], [52], [54], [55]. Many of them use image clustering to classify images, and the information they use including geographic locations, canonical views, and scenic themes. For instance, Zhao *et al.* [31] annotated POIs with Geo-tagged Tweets. Ying *et al.* [32] conduct POI recommendation by personalized geographical ranking. Kennedy and Naaman [1] used the number of users, visual and temporal information to choose the representative clusters. Among them, the temporal information is used to calculate the standard deviation of dates in each cluster. Then the clusters with higher variability in dates are more likely to be representative clusters. They successfully implemented their method in some aspects such as clustering on visual features, ranking clusters and ranking representative images (denoted CRR). However, they only consider visual information in clustering stage and they only show the representative images for users.

Simon *et al.* [2] used multiuser collections on the Internet to construct scene summarization. On the base of images with specified tags, they obtained canonical views (denoted CV) by clustering of images' visual properties, and extracted representative tags for each cluster. Qian *et al.* [3] modeled viewpoint within an image in four aspects including horizontal, vertical, scale and orientation. Then, they used 4-D vectors to construct the viewpoint vectors for each image. They selected identical semantic points (ISP) from SIFT points of the image to capture major and unique parts of a landmark. Finally, they chose diverse viewpoints in four angles to carry out visual summarization. However, the location information of different images is ignored. Jiang *et al.* [4] proposed a location based high frequency shooting location (HFSL) mining method and POI summarization approach. On the base of this method, they also proposed an automatic city-level POI mining method in [5] that not only considered the location information, but also the visual appearance. Chum proposed a randomized data mining method relies on the min-Hash algorithm to speed up the clustering process [41]. Crandall *et al.* used traditional mean-shift to perform geo-clustering and to mine high-density locations that correspond to popular places [40]. Qian *et al.* [52] summarize the POI images by combing the aesthetics and diversity by

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on September 06,2022 at 12:43:08 UTC from IEEE Xplore. Restrictions apply.

exploring the saliency from its 3D reconstruction. The representative images are selected if the photos contain more salient regions which are inferred from salient regions in 3D space. Goel *et al.* [62] proposed a visual event summarization method by extracting mid-level visual elements from images associated with social media events on Twitter.

It is obvious that many existing summarization methods only show users the visual information of the POIs, excepted for [1], which considers 3 aspects of features: the number of users, visual and temporal information. As mentioned before, they only use the visual features of images for clustering based on k-means method and they only show the images to users. It is not enough to provide more rich and comprehensive information for users' to better understand the LOIs from different views. In our work, we aim to gain famous POIs automatically for a given city by a three level City-POI-LOI summarization method, and generate text and image summarization for a POI. We combine the appearance, semantic, and temporal information both for clustering and summarization. In this way, we can choose the most representative images and texts for users to explore interesting, appealing and important LOIs for trip planning. And we also specially and innovatively propose a season-based POI summarization approach to show the representative information in different season.

B. Multi-View Clustering

As the community-contributed information includes many modalities such as location, appearance, semantic and time, here we mainly introduce some existing works about multi-view clustering. Especially, the term "view" in this paper refers to different modalities and different features.

The earlier studies [11] estimated parameters of mixture components, thereby group the data into subsets. Thus, the two images generated by the same mixture component will be assigned to the same cluster. At the same time, the examples generated by different components will be assigned to different clusters. However, they simply look the weights of different views as equal. For instance, Cao *et al.* [9] proposed a diversity-induced multi-view subspace clustering which is extended from the existing subspace clustering approach. Wang *et al.* [27] proposed to combine information from multiple social media websites to enhance the co-clustering performance of two types of objects (social media objects and users) in one social network. Zhang *et al.* [28] proposed weighted multi-view online competitive clustering approach. It simultaneously exploited the variable weighting strategy and the online competitive learning and cast the multi-view clustering problem into an optimization problem. Quack *et al.* [46] proposed an approach for mining touristic POIs from community photo collections in an unsupervised fashion. They used several modalities information, such as visual, textual and location. One of the highlights was that they used a large Dataset consisting of 200,000 photos. Many works that merged concept learning [13], belief propagation [10], or nonnegative matrix factorization [12] with multi-view clustering method are also proposed for clustering, and made very good performance. Images are similar in some aspects such as e.g., semantic, time, and location. If we cluster the images into groups in advance, the multi-view clustering will be faster and more effective. Wang *et al.* [51] aimed to study the semantics of point-of-interest by exploiting the abundant heterogeneous user generated content (UGC). However, it ignored the relationship between different modalities UGC. Nie *et al.* [50] aimed to solve the problem of uncorrelation between the text and user-generated content (UGC, such as images) in location-based social networks. They mainly focused on topic modeling for each image. They also used a graph clustering method to detect the latent topics for each venue. However, they only consider the visual and textual information for clustering.

Mean-shift clustering is a very common method for location based image clustering [4], [21], [22]. A number of approaches only use mean-shift with an appropriate bandwidth to find the locations where so many photos are taken [21], [22]. However, the visual appearances are often ignored in clustering. Li *et al.* [39] proposed a method to speed up reconstruction by a hierarchical approach.

Some works [6]–[8], [29], [56]–[58] learned the underlying clustering structure from multiple views by regularizing each view towards a common consensus. Besides, Cai *et al.* [6] first constructed a graph for each view, then they fused all graphs into a better one. According to the importance, the weight of every graph is assigned. They showed that their methods were more immune to the ineffective views. What's more, some of approaches conjunctively learned the optimal combination of each single view graph and the clustering results [7], [8]. Since the affinity matrix was not required to be positive semi-definite, the graph based multi-view clustering methods obtained wider applicability and made good performance. However, MMSC [6] learns the good features combination in advance and then perform clustering, but it does not differentiate the contribution of different views. MVSpec [8] only finds the optimal weights for different views, and the weight for the same view is invariable. GOMES [7] ignores the clustering consistency across views.

Recently, Luo *et al.* [56] proposed a consistent and specific multi-view subspace clustering method. They learned the shared consistent representation of all views and the specific representation for every view. However, they didn't consider the different weights of different views during clustering. Wang *et al.* [57] proposed a new graph-based method named Graph-Based System (GBS). It constructed the graphs of all views, and then learned the weights of different views to fuse them into a unified graph. However, they ignore the weights for the same view and the clustering consistency across views.

Besides, the datasets of many exist works only contain one modality. For example, [7] use 5 types of features for images: LBP, GIST, CENTRIST, Dog-SIFT and HOG. As mentioned above, social media offers us much useful information not only including the visual appearance. The multi-view clustering methods only consider visual information is not suitable.

Therefore, we propose to take the location, semantic, temporal, time features into consideration. In addition, we both consider the clustering quality within view and the consistency across views to learn the optimal weight between two groups.

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on September 06,2022 at 12:43:08 UTC from IEEE Xplore. Restrictions apply.

Then we use this weight to cluster the community contributed images and mine LOIs.

III. City-POI-LOI Summarization System

As shown in Fig. 1, the City-POI-LOI summarization system mainly consists of three parts: 1) City level POI mining, 2) POI level LOI mining, and 3) POI summarization. The following will introduce the City-POI-LOI summarization system in detail.

A. City Level POI Mining

For a given image set collected from social media, we utilize the coarse to fine POI mining approach proposed by Jiang *et al.* [4]. The detailed procedures are as follows: firstly we use tag and geo-tags constraints to obtain the city-related image set. Then we use mean-shift based geographical clustering approach to mine the candidate POIs. Finally, we use visual merging to determine the final POIs of each city. Thus, images in a POI are geographically near and visually similar.

On the basis of [4], we further use TF-IPF (Tag Frequency-Inverse POI Frequency) to choose suitable tags to describe the POI. We calculate the total numbers of tags that occur in each POI. Let PW_{ij} represent the times of tag w_j that occurs in the *i*-th POI. Then the frequency of tag w_j in the *i*-th POI can be represented as follows:

$$pf_{ij} = \frac{PW_{ij}}{\sum_k PW_{kj}} \tag{1}$$

Furthermore, in order to guarantee that the tag occurs in a specific POI with high frequency, and rarely appears in other POIs, we use the inverse POI frequency that is computed as traditional IDF as follows:

$$idf_j = \log \frac{|P|}{|\{i : w_j \in P_i\}|} \tag{2}$$

where $|P|$ is the total POI number, and idf_j is the inverse POI frequency of w_j. TF-IPF of each tag is computed as follows:

$$pfidf_{ij} = pf_{ij} \times idf_j \tag{3}$$

After that, we obtain the TF-IPF value of each tag in each POI. Then we choose the top ranked tags as the representative tag for POI.

B. POI-Level LOI Mining

After obtaining the POIs in every city, we further mine LOIs in each POI. These LOIs constitute the popular viewpoints of the POIs. Our LOI mining method includes three parts: feature extraction, LAST clustering, and LOI determination.

1) Feature Extraction: In our proposed method, we mainly use four types of features: location, visual, semantic, and temporal features.

a) Location feature: Location consists of latitude and longitude. So the location feature of every image can be represented by a 2-dimensional vector.

b) Visual feature: As for visual appearances, we unsupervised integrate different descriptors such as global features, e.g., HOG, color texture feature [19], and local features including SIFT and GIST. Meanwhile, we also use CNN feature which calculated from VGG16 Caffe model [23]. We select the output of the last full-link layer as the CNN feature, and the dimension is 4096.

c) Semantic feature: As for semantic feature, we use the tags belong to images. In order to eliminate the noise and meaningless tags such as "Nikon" and "Canon", we use TF-IDF in advance [14]. It should be noted that tags like "Nikon" and "Canon" could be very important in other fields. However, they are noise in our system. Assume that the total number of the reserved tags is *m*. The semantic feature of an image *i* can be represented as a *m* dimensional binary word vector $s_i = [s_{ik}]_{k=1}^m$, i.e., one-hot representation. $s_{ik}= 1$ represents the tag *k* appears in the image *i*, and $s_{ik}= 0$ represents it doesn't appear. In spite of the one-hot representation of the semantic feature, we also utilize word2vec to represent each POI [48]. More detailed comparison is discussed in experiment.

d) Temporal feature: Because season in the southern hemisphere is in order like autumn, winter, spring, summer, we first need to split a year into different seasons. But season in the northern hemisphere is in order like spring, summer, autumn, winter, and season in the southern hemisphere is different. Here we use an automatic division method for season division [15].

It should be noted that, different time quantum in a season may have big differences. For instance, many people prefer to view the sunrise of Mountain Huang in the early morning but admire the meteor shower in the evening. So we further apply the same method to cut a day into four time quantum: morning, noon, afternoon, evening [15]. Thus, we get four seasons with 16 time quantum. For simplicity, the temporal vector can be presented as a sixteen dimensional binary vector. We set S_q as the season vector of image *q*, and S_{L_i} as the season representation of LOI L_i.

2) LAST Clustering: We propose a multi-view fusion approach that fuses the location, appearance, semantic and temporal features to mine LOIs. Our goal is to find the optimal weight among groups within different views by maximizing the clustering quality within view and the clustering consistency across views.

Assume that an image set consisting of *n* images, and we denote it as X. The total view number in *X* is *H*. Let $x_i = \{x_i^h, i = 1, \ldots, n\}$ denote the feature set of image *i*, and x_i^h is the feature of the *h*-th view. Then we construct graph for each view, and correspondingly we build H graphs in total. In each graph, images belong to the vertex set and similarity of two different images constructs the edge set. The similarity of image *p* and image *q* is measured by the Euclidean distance of their visual features as follows:

$$W_{pq}^h = \exp\left(-\frac{\left\|x_p^h, x_q^h\right\|^2}{\delta^2}\right), 1 \le p, q \le n, h = 1, \ldots, H. \tag{4}$$

where δ is the parameter to control the spread of neighbors in k-nearest neighbor graph [8].

The similarity matrices are used in LAST clustering. We use classical clustering method, e.g., K-means or spectral clustering to segment the graph into *K* different groups $G_i^h, i = 1, \ldots, K$

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on September 06,2022 at 12:43:08 UTC from IEEE Xplore. Restrictions apply.

within different views $h = 1, \ldots, H$, and G_i^h is the image set that is assigned into the *i*-th group within the *h*-th view. We also record the corresponding clustering center C_i^h. In this paper, we set $K = 40$. More detailed discussion of *K* to the summarization performance is given in our experiment.

Based on the initial clustering results, we further measure the cluster quality within views and clustering consistency across views, and then we adaptive determine the optimal fusion weight for each view in POI summarization.

a) Clustering quality within views: To maximize the clustering quality within an individual view, each image should be assigned to the most suitable group with its similar neighbors, and images with dissimilar features should be assigned into different groups. The goal we want to achieve is to make the clustering results cross different views as consistent as possible.

Here we use two criterions to represent the clustering quality within views: closeness and diversity. Based on the obtained the groups G_i^h and their group center C_i^h, we measure the closeness of group G_i^h within *h*-th view by the average distance from the group center to all the images belong to the group. And the smaller the average distance is, the closer the group is. So we use the average distance to represent the closeness of the group as follows.

$$clo\left(h,i,i\right) = \frac{1}{\left|G_i^h\right|} \sum_{p:\{p|x_p \in G_i^h\}} dist\left(x_p^h, C_i^h\right) \tag{5}$$

where $|G_i^h|$ is the total number of images that belong to G_i^h, C_i^h is the center of group G_i^h. x_p^h is the feature of image *p* within *h*-th view. $dist(\cdot)$ is the Euclidean distance between two elements.

Meanwhile, many centers are obtained after clustering, and every center can represent the average attribute of the group. So the diversity of group G_i^h and G_j^h within the *h*-th view is computed by the distance between their group centers as follows:

$$div\left(h,i,j\right) = dist\left(C_i^h, C_j^h\right) \tag{6}$$

We aim at making the images within the same group close enough and the images in different groups diverse enough. So we use a linear combination of the closeness and diversity [35] to measure the clustering quality within views *WV* as follows:

$$WV\left(h,i,j\right) = \left(1-\beta\right) div\left(h,i,j\right) - \beta clo\left(h,i,i\right) \tag{7}$$

where the first term is the diversity of group G_i^h and G_j^h within the h-th view. It is computed by the distance between their group centers. The second term is to remove bias by subtracting the average. $\beta \in [0,1]$ is the trade-off parameter between closeness and diversity within views.

b) Clustering consistency across views: We not only focus on the clustering quality within views, but also the clustering consistency across views. In order to measure the clustering consistency across views, we need to define the group consistency within the *h*-th view and the *w*-th view as follows:

$$GC\left(h,w,i,j\right) = \frac{\left|G_i^h \cap G_j^w\right|}{\left|G_i^h \cup G_j^w\right|} \tag{8}$$

where $GC(h,w,i,j)$ is the group consistency between the *i*-th group within the *h*-th view and the *j*-th group within the *w*-th view. $|G_i^h \cap G_j^w|$ denotes the number of image in the intersection of groups G_i^h and G_j^w, and $|G_i^h \cup G_j^w|$ represents the corresponding image number of the union of the two groups.

The consistency within the *h*-th view can be measured by the average group consistency between it and all the other views.

$$CV\left(h,i,j\right) = \frac{1}{H-1} \sum_{w=1,w\neq h}^{H} GC\left(h,w,i,j\right) \tag{9}$$

Large $CV(h,i,j)$ means that the groups G_i^h and G_j^h have high consistency within the same view.

c) Objective function: By considering the clustering quality within view and clustering consistency across views, the cost function of LAST clustering can be written as:

$$\begin{aligned} Q(a) &= \sum_{i=1}^{K}\sum_{j=i}^{K}\sum_{h=1}^{H}(a_{ij}^h)^r(\alpha WV(h,i,j) + (1-\alpha)CV(h,i,j)) \\ &= \sum_{i=1}^{K}\sum_{j=i}^{K}\sum_{h=1}^{H}\left({a_{ij}}^h\right)^r(\alpha(\beta \cdot clo(h,i,i) \\ &\quad + (1-\beta)div(h,i,j)) + (1-\alpha)CV(h,i,j) \\ \text{s.t.} &\sum_{h=1}^{H} a_{ij}^h = 1, a_{ij}^h \geq 0, 1 \leq i \leq j \leq K, h = 1, \ldots, H \end{aligned} \tag{10}$$

where $WV(h,i,j)$ is the clustering quality of the group *i* and the group *j* within the *h*-th view, $CV(h,i,j)$ is the cross view clustering consistency of the *i*-th group and the *j*-th group between the *h*-th view and all the other views, $\alpha \in [0,1]$ is the trade-off between the clustering quality within view and clustering consistency across views, $r \in (1,\infty)$ is the parameter to control the sparseness of the solution.

We aim to find the optimal weights $a_{ij}^h, h = 1, \ldots, H$. Maximizing the objective function by Lagrangian multiplier method as follows:

$$a^* = \text{argmax}\left(Q\left(a\right)\right) \tag{11}$$

We set the lagrange function with inequality constraint as

$$L(a,\lambda,\upsilon) = Q(a) + \lambda\left(\sum_{h=1}^{H} {a^h}_{ij} - 1\right) + \sum_{k=1}^{i\cdot j} \upsilon_k {a^h}_{ij} \tag{12}$$

where *Q(a)* is the objective function,λ and $\upsilon_k$are the constraint factor. We solve the problem using KKT condition (Karush Kuhn Tucker), which means the following three conditions must be true when the optimum values are obtained.

$$\frac{\partial L(a,\lambda,\upsilon)}{\partial a} = 0, H(a) = \sum_{h=1}^{H} {a^h}_{ij} - 1 = 0, \upsilon_k \cdot {a^h}_{ij} = 0 \tag{13}$$

We unite the above three equality to an equation set and solve the. Solutions of the equation set are the optimal values we need.

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on September 06,2022 at 12:43:08 UTC from IEEE Xplore. Restrictions apply.

Algorithm1: Location-Appearance-Semantic-Temporal Clustering Algorithm (LAST)

Input:
location, visual, semantic and temporal feature,
the group number *K*,
weight parameters: α, β ,
the number of iterations: M

Output: image clusters (LOI), weight parameters: a^h_{ij}

Initialize parameter a^h_{ij}
Calculate the initial similarity matrix W
Get initial cluster based on K-means clustering method
for i =1 to M **do**
Get the new graph: $W_{pq} = \sum_{h=1}^{H} a_{ij}^h \times W_{pq}^h, p \in G_i^h, q \in G_j^h$
Segment the graph into *K* groups by K-means method
Calculate the clustering quality according to Eq.(10)
Update a^h_{ij} according to Eq. (11~13)
if converges **then**
break;
end if
end for

After optimizing a_{ij}^h, we get the final graph whose vertexes are the images and edges are the comprehensive weight between the *p*-th image and the *q*-th image that can be represented as $W_{pq} = \sum_{h=1}^{H} a_{ij}^h \times W_{pq}^h, p \in G_i^h, q \in G_j^h$. We segment the final graph into *K* different groups and then obtain the corresponding clustering results. We show the algorithm of LAST clustering in Algorithm 1.

3) LOI Determination: After LAST clustering, the images are grouped into different clusters. We further conduct a pruning for the mined clusters. We only reserve the clusters (i.e., LOIs) with sufficient images and users. In this paper, we set image threshold as 10 to determine which clusters will be reserved.

C. POI Summarization

Based on the determined LOIs, each POI can be represented by the representative texts and images from the summarization of LOIs.

1) LOI Summarization: We choose representative text (i.e., tag) and image to represent each LOI.

a) Representative image selection: We use three criterions such as representativeness, significance, and season relevance to select representative images for a LOI. The detailed calculations are as follows:

Representativeness: the images that are similar to most of the other images within LOI have more possibility to be representative. Specifically, visual representativeness can be measured as the average visual similarity from one image to all the others within the LOI. The visual representativeness of image q is calculated as:

$$vp_q = \frac{1}{|L_i| - 1} \sum_{p=1, p \neq q}^{|L_i|} W_{pq} \tag{14}$$

where L_i denotes the *i*-th LOI, $|L_i|$ represents its image number.

Significance: Social media website, such as Flickr, has recorded the browsing/view times of each image which implies its popularity to some extent. We measure the significance of an image as follows:

$$sig_q = \log\left(v\_times_q + 1\right) \tag{15}$$

where v_times_q is the view times of image q.

Season Relevance: Different images are captured in different seasons. After the LAST clustering, we have clustered the similar image into the same group. Here we use the average season factor of all the images in a LOI to represent its season vector [15]. Thus, for each LOI, we get a 1×16 dimension season vector. Different seasons have different scenes, and the images that are shot near the season representation are more likely to represent the LOI. So, we measure the season relevance of two images based on their season vectors. The season relevance of image q is calculated as:

$$sr_q = 1 - \frac{dist\left(S_q, S_{|L_i|}\right)}{\max_{q, L_i}\left(dist\left(S_q, S_{|L_i|}\right)\right)} \tag{16}$$

where S_q is the season vector of image q, and S_{L_i} is the season representation of LOI L_i.

Finally, we use a simple linear combination of vp_q, sig_q and sr_q to calculate the final score of each image[4], [5], [53],

$$score_q = vp_q + sig_q + sr_q \tag{17}$$

We rank all the images in the LOI L_i and we choose the top ranked images as its representative images.

b) Representative text selection: Here we select the representative tags for every LOI to construct the text summarization of the POIs. We use TF-ILF (Tag Frequency-Inverse LOI Frequency) and season relevance to calculate the representative value of different tags.

TF-ILF: We have used the TF-IPF to choose representative tags for each POI. Here we use the similar method to calculate the TF-ILF value of different tags in different LOIs. Details are available in III.A City Level POI Mining.

Season Relevance: Different tags are corresponding to different seasons. The season relevance value of an image implies the season relevance of its affiliated tags to some extent. So we use the image's season to denote its affiliated tags. We propose a tag-season relevance measurement approach for tags selection by linear combination of the TF-ILF value and season relevance. We choose the top ranked tags to summarize the LOI.

2) Season Based POI Summarization: As we have obtained the LOIs summarization both with representative texts and images, we put forward a season based LOI ranking method that helps users understand well about the POIs' season changes. Based on the season representation of each LOI, we utilize the season with maximal numbers of images as the best matched season for the LOI. Then we divide the LOIs in the order of spring, summer, fall and winter. Since the number of the mined LOIs varies along with POIs, we can get one or more LOIs in a season. We perform further ranking for the candidate images and tags, specifically for the images and tags with much higher significance.

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on September 06,2022 at 12:43:08 UTC from IEEE Xplore. Restrictions apply.

IV. Experiment

In order to display the advantage of our proposed LAST clustering method, we compare it with some baselines: Kmeans, SC, MMSC [6], MVSpec [8], and GOMES [7] and GBS [57].

LAST: our proposed multi-view clustering method, i.e., LOI mining method in this paper.

Kmeans: this approach first merges different views into a feature, and then utilizes Kmeans to group images.

Spectral Clustering (SC): this approach first merges different views together as above Kmeans, and utilizes spectral clustering to get image clusters.

Multi-Modal Spectral Clustering (MMSC): a multi-view spectral clustering method [6]. It first learns the combination of different views, and then finds the optimal spectral clustering.

Multi-View spectral clustering (MVSpec): a multi-view spectral clustering method [8]. It aims to find the optimal weight of different views.

GOMES: a group-aware multi-view fusion approach for real world image clustering [7]. The biggest difference from GOMES to our proposed method is the criterions of clustering quality within views and clustering consistency across views. On the base of diversity, we add closeness to measure the clustering quality within views.

GBS: a graph-based multi-view clustering method [57]. It constructed the graphs of all views, and then learned the weights of different views to fuse them into a unified graph. However, they ignore the weights for the same view and the clustering consistency across views.

In order to measure the effectiveness of our proposed method, especially the introduction of location, semantic, and temporal features, we perform some comparison experiments on the ATCF Dataset [4], [5], [36] and DIV400 Dataset [24], [26]. We use normalized mutual information (NMI), adjusted mutual information (AMI), and adjusted rand index (ARI) to evaluate the clustering performance [18]. Given two sets G and C, their NMI is defined as:

$$NMI = \frac{I(G,C)}{\sqrt{H(G) \cdot H(C)}} \tag{18}$$

where $H(\cdot)$ denotes the entropy, $I(G,C)$ is the mutual information between *G* and *C*. It equals 1 when the two sets are identical and 0 when the two sets are independent.

A. Datasets

ATCF Dataset is an image Dataset crawled form Flickr, and mainly includes the images' affiliated information, e.g., tags, geo-tags, image taken time and uploading time, views, etc. [4], [5], [36]. The total number of POIs and LOIs of each city are shown in Table I.

DIV400 Dataset (The Social Image Retrieval Result Diversification Dataset) consists of Creative Commons data related to 396 landmark locations and contains 43,418 Flickr photos together with their Wikipedia and Flickr metadata, and some content descriptor information (visual and text) [24], [26]. Data is annotated for the relevance and the diversity of the photos (both expert and crowd annotations are provided).

TABLE I
Details of ATCF Dataset

City	images	Number of POI/LOI	Number of LOI per POI
#1 Barcelona	4,183	31/109	3.5
#2 Belin	5,273	42/161	3.8
#3 Chicago	4,661	34/147	4.3
#4 London	5,073	18/168	9.3
#5 Los Angeles	3,961	14/108	7.7
#6 New York	4,423	39/132	3.4
#7 Paris	4,944	13/157	12.1
#8 Rome	5,388	17/68	4
#9 San Francisco	5,175	10/139	14

TABLE II
Performance Comparisons of Different Methods

Method	DIV400 Dataset			ATCF Dataset		
	NMI	AMI	ARI	NMI	AMI	ARI
SC	0.3580	0.2649	0.1388	0.3603	0.1200	0.1014
Kmeans	0.3681	0.2737	0.1419	0.3567	0.1429	0.0920
MMSC	0.3398	0.2589	0.1200	0.3437	0.0979	0.0744
MVSpec	0.3590	0.2609	0.1356	0.3594	0.0991	0.0971
GOMES	0.5729	0.4748	0.2419	0.4191	0.1331	0.1006
GBS	0.3738	0.2985	0.1653	0.3821	**0.1972**	0.1509
LAST(SC)	**0.7176**	0.5515	**0.4284**	0.4766	0.1683	**0.1650**
LAST(KM)	0.7060	**0.5587**	0.4196	**0.5105**	0.1503	0.1645

B. Clustering Performance Comparison

The clustering performances on different Dataset are illustrated in Table II, and the best performance is shown in bold. We can observe that LAST gains the better performance whether using K-means clustering method or spectral clustering method, in short LAST (KM) and LAST (SC). This is due to our approach both considers the clustering quality within views and the clustering consistency across views when clustering. In addition, we also use the group-based weight learning to get the optimal clustering result. In order to recommend POI with the most suitable season, the LAST takes season information into consideration. In addition, the GOMES and GBS perform better than other compared methods. That's because they both learn different weights for different views. Besides, the performance on DIV400 Dataset is better than that on ATCF dataset. It is because there are less noisy images in DIV400 Dataset than ATCF Dataset.

C. Parameter Discussions

1) Discussion of Group Number K: Due to space limitation, all the discussions in our paper are conducted on ATCF dataset.

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on September 06,2022 at 12:43:08 UTC from IEEE Xplore. Restrictions apply.

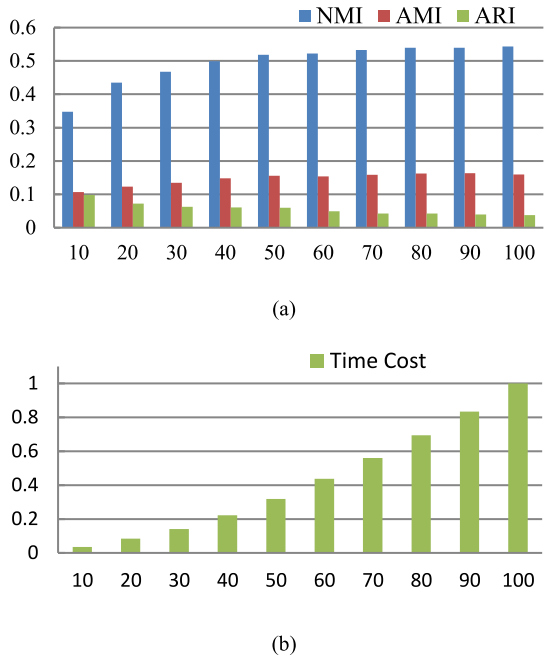


Fig. 2. Discussion of group number. (a) Performance discussion of group number. (b) Time Cost of different group numbers.

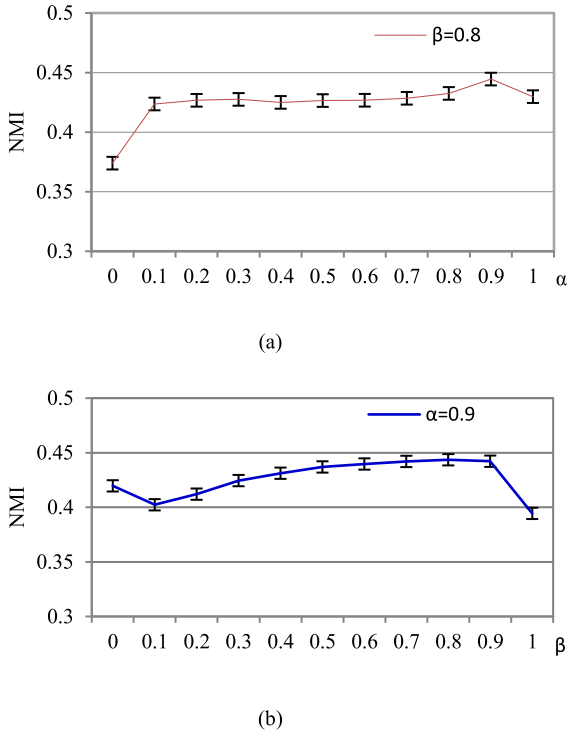


Fig. 3. Influence of different parameters on the performance of NMI. (a) varying α while setting $\beta = 0.8$, (b) varying β while setting $\alpha = 0.9$.

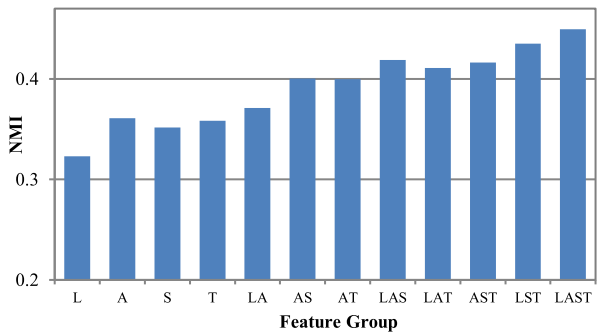


Fig. 4. Performances of LAST clustering on different features combinations.

The group number K is used in LAST clustering when mining LOIs. Here we set K varies from 10 to 100 with the step 10, and display their NMI, AMI, and ARI of clustering results on ATCF Dataset in Fig. 2(a). The normalized time cost of different group numbers is shown in Fig. 2(b).

It can be seen that with the increase of K, the result of clustering is becoming better and better. But after K = 40, the growth becomes slow. That's because when the group number is too small, the images in the same cluster are quite different, which will lead to low mutual information among these images. When the group number is too large, we will obtain more fine-grained clusters. And the images in the same cluster are very similar, leading to high mutual information. However, as shown in Fig. 2(b), the larger the number of clusters, the longer the time cost.

To make a trade-off between performance and cost time, in this paper we set $K = 40$.

2) Discuss of α and β: The object function of LAST clustering is a weighted linear combination of two criteria: clustering quality within views and clustering consistency across views, and the former is also a weighted linear combination of closeness and diversity.

In this part, we examine the effects of the corresponding weighted parameters α and β to achieve the optimal parameter setting ATCF Dataset.

We fix one of the parameters and then change the other one from 0 to 1 with an interval of 0.1. From Fig. 3(a), we can observe that when α increases from 0 to 0.1, the NMI value also increases, and then maintains a stable trend. When α reaches to 0.9, the performance shows the best. When $\alpha = 1$, it is equal to only considering the clustering quality under different perspectives without considering the consistency of clustering in different perspectives. At this time, the performance shows a significant decline. From Fig. 3(b), we can see that when β increases from 0.1 to 0.9, the NMI value increases continuously. The best performance is obtained when $\beta = 0.8$. Therefore, $\alpha = 0.9$ and $\beta = 0.8$ are selected as the best parameters in ATCF dataset.

D. Effects of Multiple Features

In this paper, we systematically fuse the location, semantic, appearance and temporal features in the proposed LAST based POI summarization approach. In order to show the contribution of each feature in POI summarization, we conduct experiments on ATCF Dataset using the same parameters $K = 40$, $\alpha = 0.9$ and $\beta = 0.8$. We show the recommendation performances of twelve type features in Fig. 4, The symbols L, A, S, T respectively represent the location, appearance, semantic and temporal feature. The combinations of different factors such as LA, AS, AT, LAS, LAT, AST, LST and LAST respectively represent

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on September 06,2022 at 12:43:08 UTC from IEEE Xplore. Restrictions apply.

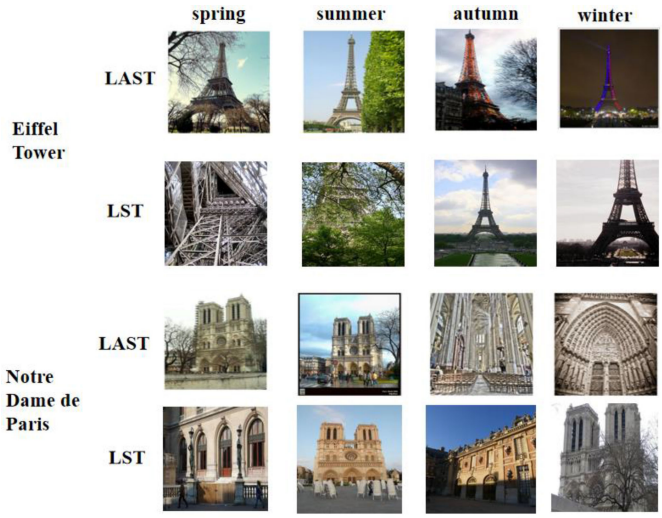


Fig. 5. The comparison of the summarization result of LST and LAST.

feature is the most indispensable one. We can find that when fusing four modalities features, the better performances will be achieved. When fusing all four modalities of features, we get the best performances that NMI = 0.5105. It proves the superiority of LAST.

To further demonstrate the effectiveness of the four features, we observe the summarization results of different combination. Due to the space limitation, we only show the comparison between LST and LAST considering that they have the closed NMI value. We take the Eiffel tower and Notre Dame de Paris as examples to explain the visualization of summarization results. The result can be seen in Fig. 5.

We can observe that if we ignore the appearance of images, we may choose lower quality images as representative images, such as the images of spring for Eiffel tower. That's because there are many images with different views and angles in one POI. If we don't consider the visual feature, the model may be difficult to choose the proper representative images. In addition, we may also obtain images of other POIs close to the target POI. For example, the real POI of the autumn image of Notre Dame de Paris obtained by LST is Versailles. The reason may be that the Notre Dame and Versailles are close and they are belonging to the same category. Therefore, only consider the location, textual and temporal information is not enough to divide them into the right clusters. Thus, the visual feature is also important to get better summarization results.

E. POI Summarization Performance Comparison

We have proposed a new framework of POI summarization and now we compare our method with some existing methods: Random, CV [2], CRR [1], ISP [3], HFSL [4], and SCCG [20]. Random means that we choose the random selected images/texts as representative results. For the sake of fairness, the settings of all the comparison methods are selected by multiple experiments to choose the best parameters. We invite 10 volunteers to score for the final POIs summarization considering the relevance, diversity and comprehensiveness [21].

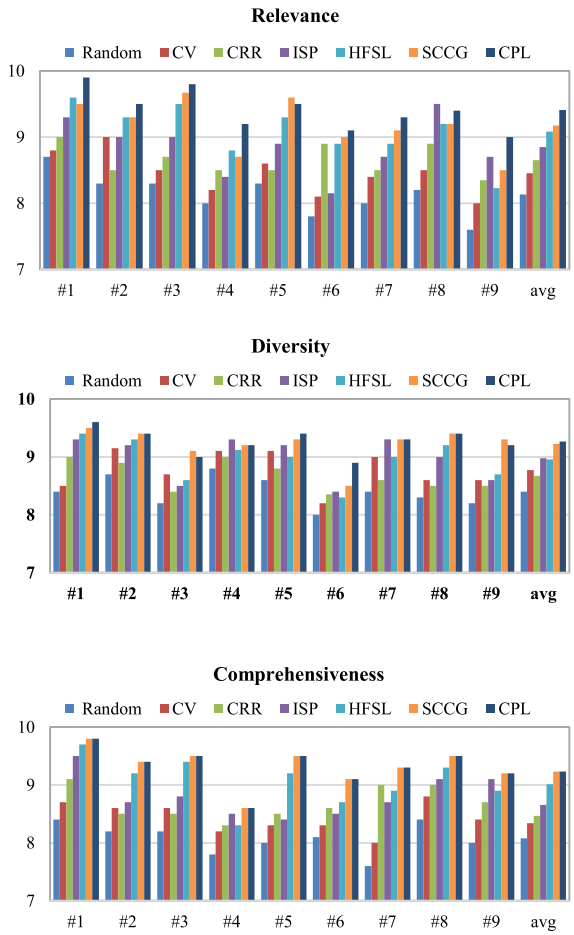


Fig. 6. Average scores for the criteria of "Relevance", "Diversity" and "Comprehensiveness" of different methods on the ATCF dataset.

As for these 10 volunteers, they are all researchers who have worked in the related field for more than 1 years. In addition, they are not belonging to our research group. In order to guarantee the reference and persuasion of result, they are asked to give the score as follows:

Relevance: Are the text and image relevant to the POI? 10-perfect, 0-irrelevant.

Diversity: Are the summarization results diverse to each other? 10-perfect, 0-related.

Comprehensiveness: Can the summarization results describe the POI comprehensively? 10-absolutely, 0-absolutely not.

The relevance is to measure whether the image and text can represent the characters of the POI. The diversity is to measure whether the differences between the result pictures and tags are large enough. The comprehensiveness is to measure whether the results offer a comprehensive view of the POI. The comprehensiveness includes the representativeness and other aspects such as the style, category, seasonality and so on.

We summarize the score in Fig. 6. We can find that the score of relevance, diversity and comprehensiveness present a similar law: although CPL is not the best one in some POI, CPL has the highest average score on all three criteria. And random selected results show the worst performance.

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on September 06,2022 at 12:43:08 UTC from IEEE Xplore. Restrictions apply.

TABLE III
AVERAGE SCORES FOR THE CRITERIA OF P@N, CR@N, AND F1@N OF DIFFERENT METHODS ON THE ATCF DATASET

Method	P@N	CR@N	F1@N
CV	0.4475	0.7923	0.6199
CRR	0.4525	0.7512	0.6018
ISP	0.4721	0.8314	0.6517
HFSL	0.4981	0.8231	0.6606
SCCG	0.4989	0.8771	0.6880
CPL	**0.5326**	**0.9247**	**0.7286**

Finally, we assess the experimental results with three auto evaluation criteria referred from [35], they are as follows:

P@N: The percentage of relevant images in top N images. It represents the relevance standard.

CR@N: The percentage of different topics/aspects retrieved in the top N images. It represents the diversity standard.

F1@N: The harmonic mean of CR@N and P@N. It represents the comprehensiveness standard.

We count P@N, CR@N, F1@N in each POI of 9 cities, and calculate the mean value of 9 cities as the final result. N is alterable for different POIs, because the number of images in different summarization of POIs is different. It dependents on the abundance degree of images that loaded by users in each POI. We summarize the result in Table III, and find that CPL has the highest average score on all three criteria, which is the same as the conclusion made by volunteers. According to the assessment results both of subjective and objective evaluation, our method perform best.

F. Discussion of Semantic Feature

Compared to the semantic feature used in this paper, word2vec [48] are more advanced methods in text encoding. So we carry out contrast test using word2vec, which is more suitable for tag encoding. We use gensim 3.0.1 [49] to obtain a 200 dimension vector for each tag, and use the average value of all tag's vectors that belong to the picture as its semantic feature representation. Table IV shows the experimental results on ATCF.

From the table, we find that most results on three evaluation criterions promote to some extent after changing word2vec. It is because word2vec is more powerful than one-hot in word encoding. Because we first cluster pictures and tags by GPS information, so the number of tags among a certain location is not too large. In our experiment, the number is 28,219.

After city-level POI summarization, we have got 13 POIs in Paris including **Eiffel Tower, Notre Dame de Paris, Musée du Louvre, Palace of Versailles, Paris Disneyland** and so on. Since the number of POIs is large, we only LAST the five POIs mentioned in Fig. 7. And for each season, we select one image to be representative. The final summarizations with texts and images of the five POIs are displayed in Fig. 7.

TABLE IV
PERFORMANCE COMPARISONS OF DIFFERENT SEMANTIC FEATURES ON ATCF DATASET

Method	One-hot			Word2vec		
	NMI	AMI	ARI	NMI	AMI	ARI
SC	0.3603	0.1200	0.1014	0.3947	0.1207	0.1291
Kmeans	0.3567	0.1429	0.0920	0.3602	0.1366	0.0903
MMSC	0.3437	0.0979	0.0744	0.3220	0.0774	0.1028
MVSpec	0.3594	0.0991	0.0971	0.3403	0.0937	0.1124
GOMES	0.4191	0.1331	0.1006	0.4590	0.1609	0.1356
GBS	0.3821	**0.1972**	0.1509	0.4512	**0.2085**	0.1460
LAST(SC)	0.4766	0.1683	**0.1650**	**0.5269**	0.1760	**0.1734**
LAST(KM)	**0.5105**	0.1503	0.1645	0.5109	0.1472	0.1579

As we can see in Fig. 7, taking Eiffel Tower as an example, the four images in one row correspond to the four seasons, which have discrimination not only on the visual appearance, but also on the semantic appearance. In the first image, trees are withered because it is a picture taken in January. As for the second image, trees are very green. The leaves vary yellow in the third image. The hue of the last image is dark because it is shot in winter. All in all, our proposed method for POI summarization can well summarize the city with different POIs, and it can also summarize each POI with images and texts in different seasons.

Besides, the semantic information in our summarization results is representative and helpful. Taking Paris Disneyland as an example, the representative tags in spring include "grass", which can attract users who prefer the lawn. And the "airport" indicates the convenient transportation. The texts of winter include "snow" and "cool", which have big difference compared with spring and can attract users to enjoy the beautiful snow scenes. The differences between different seasons can provide users the comprehensive information to help them make trip plans. In addition, the visual information of the four representative images is diverse that illustrates our proposed CPL system can well cluster the visually similar images.

G. Some Bad Cases of POI Summarization

The LAST clustering method fuse the location, appearance, semantic, and temporal information for LOI clustering. On this basis, we generate a season based visual and textual summarization for each POI. We analysis our clustering and summarization results and find that there are mainly two types of poor clustering results. We show the examples of bad cases in Fig. 8. They are the summarization of **Millenium Bridge** and **Kew Garden** in London. The four images in one row correspond to the four seasons.

We can observe that:

1) It is difficult to distinct the co-occurrence POIs. For example in Fig. 8, the St. Paul's Cathedral and the Millennium Bridge in London appear in the same image. It is difficult to distinguish them into the right clusters. In future work, we will add more constraints to verify such samples.

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on September 06,2022 at 12:43:08 UTC from IEEE Xplore. Restrictions apply.

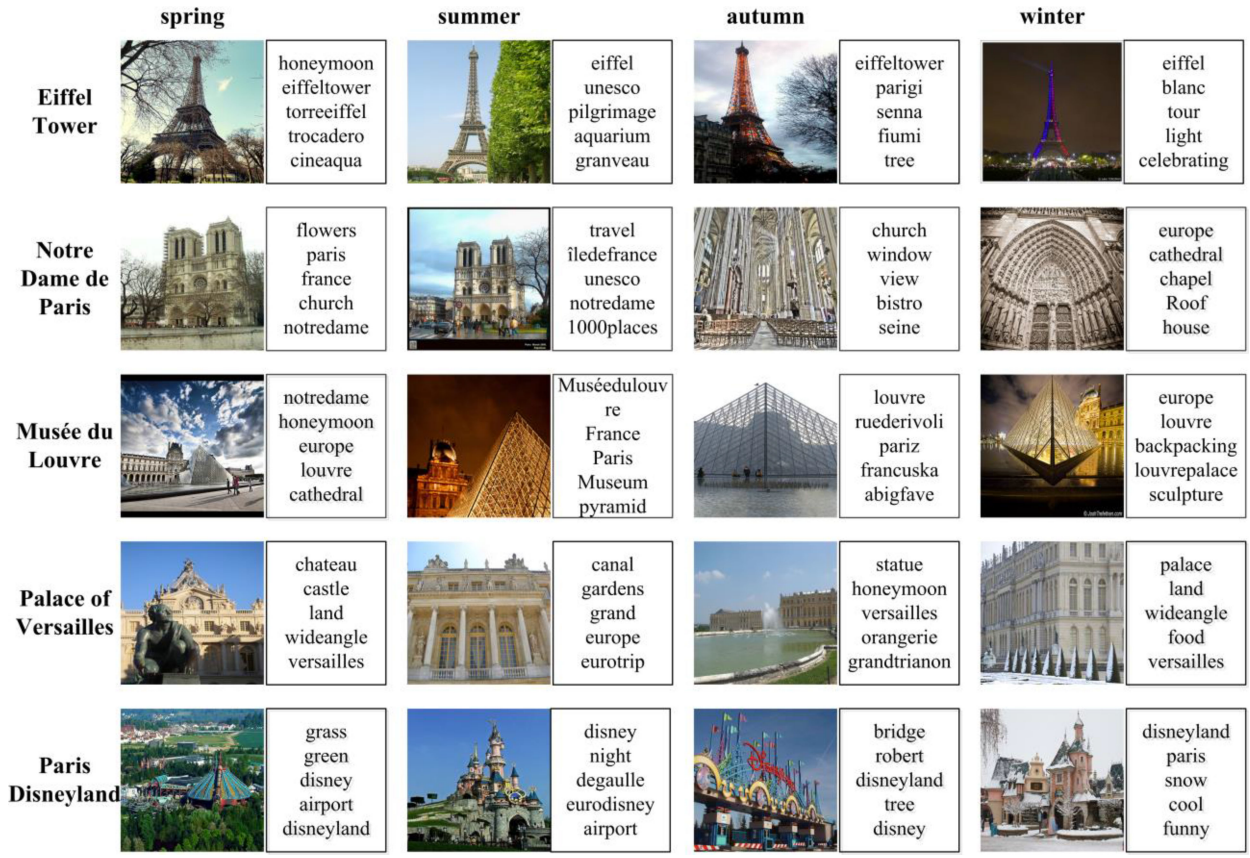


Fig. 7. POI summarization for Eiffel Tower, Notre Dame de Paris, Musée du Louvre, Palace of Versailles, and Paris Disneyland in Paris.

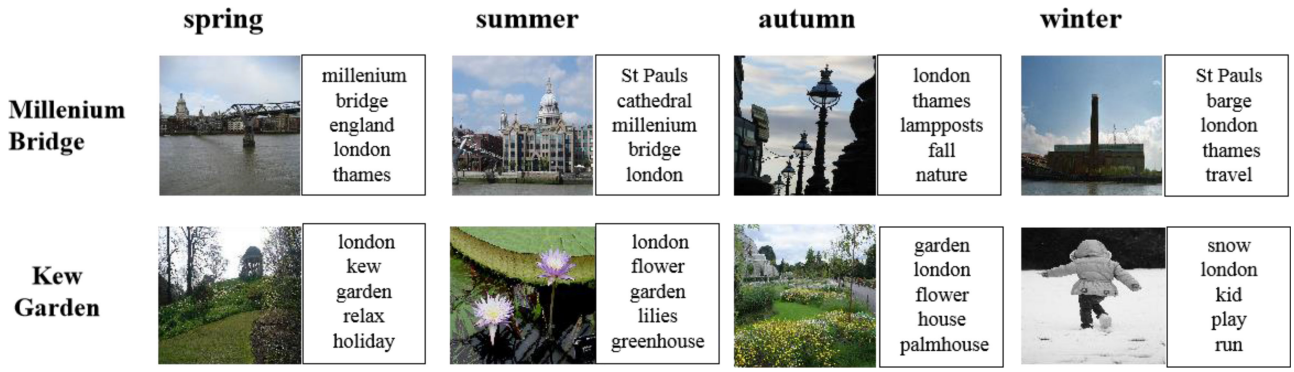


Fig. 8. The examples of some bad cases for Millenium Bridge and Kew Garden in London.

2) There are some irrelevant images about the POIs which are difficult to be filtered, which will appear in the summarization results. For example in Fig. 8, the representative image of Kew Garden for winter includes a running child which is unrelated to the target POI. Although it is the right image belonging to the POI, it is useless to help users understand the POI. Therefore, in future work, we will improve the ranking framework to move the unrelated images.

V. Conclusion

In this paper, we propose a CPL system that consists of three levels to automatically mine famous POIs for a given city, and then mine latent popular LOIs in each POI. Finally, it summarizes the POIs with texts and images. In LAST clustering, the clustering quality within views and clustering consistency across views are considered to evaluate the importance of different views, and an iterative optimization algorithm is proposed to learn the clustering results and fusion weights simultaneously.

Experiments on two realistic image Datasets indicate that LAST improves the clustering performance compared with baseline methods, and the whole CPL framework improves the summarization performance on relevance, diversity and comprehensiveness. In future work, we will try more better fusion and feature extraction methods such as deep learning-based methods.

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on September 06,2022 at 12:43:08 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

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

[2] I. Simon, N. Snavely, and S. Seitz, "Scene summarization for online image collections," in *Proc. IEEE Int. Conf. Comput. Vision*, 2007, pp. 1–8.

[3] X. Qian, Y. Xue, 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. 2015.

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

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

[6] X. Cai, F. Nie, H. Huang, and F. Kamangar, "Heterogeneous image feature integration via multi-modal spectral clustering," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2011, pp. 1977–1984.

[7] Z. Xue *et al.*, "GOMES: A group-aware multi-view fusion approach towards real-world image clustering," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2015, pp. 1–6.

[8] G. Tzortzis and A. Likas, "Kernel-based weighted multi-view clustering," in *Proc. IEEE Int. Conf. Data Mining*, 2012, pp. 675–684.

[9] X. Cao, C. Zhang, H. Fu, S. Liu, and H. Zhang, "Diversity-induced multi-view subspace clustering," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 586–594.

[10] C. Wang, J. Lai, and P. Yu, "Multi-view clustering based on belief propagation," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 4, pp. 1007–1021, Apr. 2016.

[11] S. Bickel and T. Scheffer, "Multi-view clustering," in *Proc. 4th IEEE Int. Conf. Data Mining*, 2004, pp. 19–26.

[12] X. Zhang, L. Zhao, L. Zong, X. Liu, and H. Yu, "Multi-view clustering via multi-manifold regularized nonnegative matrix factorization," in *Proc. IEEE Int. Conf. Data Mining*, 2014, pp. 1103–1108.

[13] Z. Guan, L. Zhang, J. Peng, and J. Fan, "Multi-view concept learning for data representation," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 11, pp. 3016–3028, Nov. 2015.

[14] Y. Yang, Z. Gong, and L. Hou, "Identifying points of interest using heterogeneous features," in *Proc. ACM Trans. Intell. Syst. Technol.*, 2014, pp. 1–29.

[15] U. Fayyad and K. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, 1993, pp. 1022–1029.

[16] Y. Zhao, Y. Zheng, X. Zhou, and T.-S. Chua, "Generating representative views of landmarks via scenic theme detection," in *Proc. Int. Conf. Multimedia Model.*, 2011, pp. 392–402.

[17] Q. Liu *et al.*, "A cocktail approach for travel package recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 2, pp. 278–293, Feb. 2014.

[18] N. Xuan, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Is a correction for chance necessary?" in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 1073–1080.

[19] X. Qian, X. Liu, X. Ma, D. Lu, and C. Xu, "What is happening in the video?–Annotate video by sentence," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 9, pp. 1746–1757, Sep. 2016.

[20] 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.*, 2015, pp. 1–6.

[21] Q. Hao *et al.*, "Generating location overviews with images and tags by mining user-generated travelogues," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 801–804.

[22] Y. Pang *et al.*, "Summarizing tourist destinations by mining user-generated travelogues and photos," *Neurocomput.*, vol. 115, pp. 352–363, 2010.

[23] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Multimedia*, pp. 675–678, 2014

[24] B. Ionescu, A. Popescu, H. Müller, M. Menéndez, and A. Radu, "Benchmarking result diversification in social image retrieval," in *Proc. ICIP,* 2014, pp. 3072–3076.

[25] M. Li, Z. Bao, L. Song, and H. Duh, "Social-aware visualized exploration of tourist behaviours," in *Proc. Int. Conf. Big Data Smart Comput.*, 2016, pp. 289–292.

[26] B. Ionescu *et al.*, "Div400: A social image retrieval result diversification dataset," in *Proc. ACM Multimedia Syst.*, Mar. 2014, pp. 29–34.

[27] F. Wang, S. Lin, and P. Yu, "Collaborative co-clustering across multiple social media," in *Proc. IEEE Int. Conf. Mobile Data Manage.*, 2016, pp. 142–151.

[28] G. Zhang, D. Huang, C. Wang, and W. Zheng, "Weighted multi-view on-line competitive clustering," in *Proc. IEEE Int. Conf. Big Data Comput. Service Appl.*, 2016, pp. 286–292.

[29] X. Zhang, Z. Wang, L. Zong, and H. Yu, "Multi-view clustering via graph regularized symmetric nonnegative matrix factorization," in *Proc. IEEE Int. Conf. Cloud Comput. Big Data Anal.*, 2016, pp. 109–114.

[30] Y. Liu, C. Liu, B. Liu, M. Qu, and H. Xiong, "Unified point-of-interest recommendation with temporal interval assessment," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1015–1024.

[31] K. Zhao, G. Cong, and A. Sun, "Annotating points of interest with geotagged tweets," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, 2016, pp. 417–426.

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

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

[34] A. Jeffries, "The man behind Flickr on making the service 'awesome again'," [Online]. Available: https://www.theverge.com/2013/3/20/4121574/flickr-chief-markus-spiering-talks-photos-and-marissa-mayer. Accessed on: Jan. 12, 2018.

[35] E. Xioufis *et al.*, "Improving diversity in image search via supervised relevance scoring," in *Proc. Int. Conf. Multimedia Retrieval*, 2015, pp. 323–330.

[36] 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, Mar. 2016.

[37] Y. Chen, A. Cheng, and W. Hsu, "Travel recommendation by mining people attributes and travel group types from community-contributed photos," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1283–1295, Oct. 2013.

[38] X. Yang, T. Zhang, and C. Xu, "Cross-domain feature learning in multimedia," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 64–78, 2015.

[39] X. Li, C. Wu, C. Zach, S. Lazebnik, and J. Frahm, "Modeling and recognition of landmark image collections using iconic scene graphs," in *Proc. Eur. Conf. Comput. Vision*, 2008, pp. 427–440.

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

[41] O. Chum and J. Matas, "Large-scale discovery of spatially related images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 371–377, Feb. 2010.

[42] Y. T. Zheng *et al.*, "Tour the world: Building a web-scale landmark recognition engine," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 1085–1092.

[43] P. Zhou, Y. Zhou, D. Wu, and H. Jin, "Differentially private online learning for cloud-based video recommendation with multimedia big data in social networks," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1217–1229, Jun. 2016.

[44] M. Jiang, P. Cui, F. Wang, W. Zhu, and S. Yang, "Scalable recommendation with social contextual information," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 11, pp. 2789–2802, Nov. 2014.

[45] S. Papadopoulos, C. Zigkolis, Y. Kompatsiaris, and A. Vakali, "Cluster-based landmark and event detection for tagged photo collections," *IEEE MultiMedia*, vol. 18, no. 1, pp. 52–63, Jan. 2011.

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

[47] S. Kaur *et al.*, "Recommendation generation using typicality based collaborative filtering," in *Proc. Int. Conf. Cloud Comput. Data Sci. Eng.*, 2017, pp. 210–215.

[48] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. NIPS,* 2013, pp. 3111–3119.

[49] G. Zhao *et al.*, "Personalized reason generation for explainable song recommendation," *ACM TIST*, vol. 10, no. 4, pp. 41:1–41:21, 2019.

[50] W. Nie, W. Peng, X. Wang, Y. Zhao, and Y. Su, "Multimedia venue semantic modeling based on multimodal data," *J. Vis. Commun. Image Representation*, vol. 48, pp. 375–385, 2017.

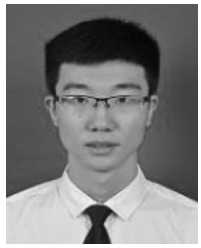
[51] X. Wang *et al.*, "Semantic-based location recommendation with multimodal venue semantics," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 409–419, Mar. 2015.

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on September 06,2022 at 12:43:08 UTC from IEEE Xplore. Restrictions apply.

[52] X. Qian *et al.*, "POI summarization by aesthetics evaluation from crowd source social media," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1178–1189, Mar. 2018.
[53] X. Qian, M. Li, Y. Ren, and S. Jiang, "Social media based event summarization by user-text-image co-clustering," *Knowl.-Based Syst.*, vol. 164, pp. 107–121, 2019.
[54] A. Psyllidis, J. Yang, and A. Bozzon, "Regionalization of social interactions and points-of-interest location prediction with geosocial data," *IEEE Access*, vol. 6, pp. 34334–34353, 2018.
[55] G. McKenzie and K. Janowicz, "The effect of regional variation and resolution on geosocial thematic signatures for points of interest," in *Proc. Annu. Int. Conf. Geograph. Inf. Sci.*, 2017, pp. 237–256.
[56] S. Luo, C. Zhang, W. Zhang, and X. Cao, "Consistent and specific multi-view subspace clustering," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 3730–3737.
[57] H. Wang, Y. Yang, B. Liu, and H. Fujita, "A study of graph-based system for multi-view clustering," *Knowl.-Based Syst.*, vol. 163, pp. 1009–1019, 2019.
[58] J. Xu, J. Han, F. Nie, and X. Li, "Re-weighted discriminatively embedded k-means for multi-view clustering," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 3016–3027, Jun. 2017.
[59] X. Qian, X. Lu, J. Han, B. Du, and X. Li, "On combining social media and spatial technology for POI cognition and image localization," *Proc. IEEE*, vol. 105, no. 10, pp. 1937–1952, Oct. 2017.
[60] Y. Wang *et al.*, "Position focused attention network for image-text matching," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 3792–3798.
[61] G. Zhao, X. Lei, X. Qian, and T. Mei, "Exploring users' internal influence from reviews for social recommendation," *IEEE Trans. Multimedia*, vol. 21, no. 3, pp. 771–781, Mar. 2019.
[62] S. Goel, S. Ahuja, A. Subramanyam, and P. Kumaraguru, "VisualHashtags: Visual summarization of social media events using mid-level visual elements," in *Proc. ACM Multimedia*, 2017, pp. 1434–1442.
[63] K. Li, Y. Wu, Y. Xue, and X. Qian, "Viewpoint recommendation based on object oriented 3D scene reconstruction," *IEEE Trans. Multimedia*, to be published, doi: 10.1109/TMM.2020.2981237.

Xueming Qian (Member, IEEE) received the B.S. and M.S. degrees from the 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 with Xi'an Jiaotong University. 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. He received the Microsoft Fellowship in 2006. He received outstanding doctoral dissertations of Xi'an Jiaotong University and Shaanxi Province, in 2010 and 2011, respectively.

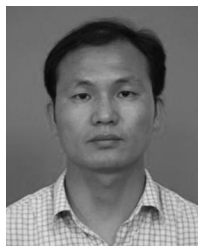
Yuxia Wu received the B.S. degree from Zhengzhou University, Henan, China, in 2014, the M.S. degree from Fourth Military Medical University, Xi'an, China, in 2017. She is currently working toward the Ph.D. degree with Xi'an Jiaotong University, Xi'an, China. She is mainly engaged in the research of multimedia mining and recommender systems.



Mingdi Li received the B.S. and M.S. degrees from Xi'an Jiaotong University, Shannxi, China, in 2016 and 2019, respectively. He is mainly engaged in the research of multimedia mining and recommender systems.

Yayun Ren received the B.S. degree from the Zhengzhou University of Aeronautics, Henan, China, in 2013, and the M.S. degree from Xi'an Jiaotong University, Xi'an, China, in 2016. Her research interests include social multimedia mining and recommendation.

Shuhui Jiang received the B.S. and M.S. degrees from Xi'an Jiaotong University, Xi'an, China, in 2007 and 2011, respectively. She received the Ph.D. degree from the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA, in 2018. Her research interests include machine learning, multimedia, and computer vision. She was the recipient of the Dean's Fellowship of Northeastern University from 2014. She was a research intern with Adobe research laboratory, San Jose, US, in summer 2016. She was as the reviewers for IEEE journals IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS etc. She was the recipients of the Best Paper of TOMM 2019 and Best Paper Candidate of ACM MM 2017.



Zhetao Li was born in Hunan province, China. He received the B.E. degree in electrical information engineering from Xiangtan University, Xiangtan, China, in 2002, the M.E. degree in pattern recognition and intelligent system from Beihang University, Beijing, China, in 2005, and the Ph.D. degree in computer application technology from Hunan University, Changsha, China, in 2010. He is a Professor with the College of Information Engineering, Xiangtan University. From December 2013 to December 2014, he was a Post-Doc in wireless network with Stony Brook University. From December 2014 to December 2015, he was an invited Professor with Ajou University. His research interests include wireless communication and multimedia signal processing. For his successes in teaching and research he was the recipient of the Second Prize of Fok Ying Tung Education Foundation Fourteenth Young Teachers Award in 2014.

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on September 06,2022 at 12:43:08 UTC from IEEE Xplore. Restrictions apply.