POI Summarization by Aesthetics Evaluation From Crowd Source Social Media

Xueming Qian, *Member, IEEE*, Cheng Li, Ke Lan, Xingsong Hou, Zhetao Li, and Junwei Han

***Abstract*— Place-of-Interest (POI) summarization by aesthetics evaluation can recommend a set of POI images to the user and it is significant in image retrieval. In this paper, we propose a system that summarizes a collection of POI images regarding both aesthetics and diversity of the distribution of cameras. First, we generate visual albums by a coarse-to-fine POI clustering approach and then generate 3D models for each album by the collected images from social media. Second, based on the 3D to 2D projection relationship, we select candidate photos in terms of the proposed crowd source saliency model. Third, in order to improve the performance of aesthetic measurement model, we propose a crowd-sourced saliency detection approach by exploring the distribution of salient regions in the 3D model. Then, we measure the composition aesthetics of each image and we explore crowd source salient feature to yield saliency map, based on which, we propose an adaptive image adoption approach. Finally, we combine the diversity and the aesthetics to recommend aesthetic pictures. Experimental results show that the proposed POI summarization approach can return images with diverse camera distributions and aesthetics.**

***Index Terms*— POI summarization, 3D reconstruction, saliency map, aesthetic measurement, crowd source salient feature.**

I. Introduction

IN RECENT years, due to the popularity of mobile imaging devices and social networks, huge numbers of photos of POI are shared on internet by people during their travels [60]. POI summarization can show a set of POI images but it is difficult to obtain a satisfactory result when a user intends to search a landmark by means of text-based image retrieval [24], [63], [74]. There are many irrelevant and unpleasing pictures in the top ranked results. Thus, POI summarization with aesthetics and diversity gives users a better understanding of the given landmark. Estimating aesthetics of images has been attracting much attention [1], [2], [5]–[7], [72], [73], which can provide user-interested photos of a POI.

Fig. 1. Example of Flickr image information.

The motivation of the proposed method can be described in three aspects: 1) considering that the number of photos of POI is huge, it is difficult but interesting to do POI summarization. 2) we tend to show satisfactory POI images to users by considering aesthetic effects and diversity. 3) 3D reconstruction is a better tool for POI summarization

Some kinds of information such as the geo-tags, views, comments, and faves can be exploited to image summarization [54]–[58]. Usually, the representative and beautiful photos can be viewed, commented and forwarded by different users frequently [63], [74]. For instance, one of the representative images of a POI from Flickr is shown in Fig.1. There are several challenges to be solved in POI summarization: 1) how to select relevant images from a large scale user contributed image set; 2) how to evaluate the aesthetics of images; 3) how to model viewpoints of the images.

So far, there is still no consensus on the standard of aesthetics in photography due to the complexity and the subjectivity in this matter. Aesthetics in photography generally derives from many aspects including the composition, colors, illumination, and the theme of the pictures. To carry out photo aesthetic estimation, several multimedia models [5]–[7], [72], [73] have been developed to analyze the content in photographic images and make the quantification of aesthetics possible. Bhattacharya *et al.* [5] present an interactive application to artificially improve the visual aesthetics using spatial re-composition. Cheng *et al.* [6] propose omni-range context modeling based approach to learn the patch/object spatial correlation distribution for the concurrent

Manuscript received December 12, 2016; revised June 8, 2017 and July 27, 2017; accepted September 3, 2017. Date of publication November 2, 2017; date of current version December 5, 2017. This work was supported in part by the NSFC under Grant 61732008, Grant 61772407, Grant 61373113, and Grant u1531141, in part by the National Key Research and Development Program of China under Grant 2017YFF0107700, in part by Guangdong Provincial Science and Technology Plan under Project 2016A010101005 and Project 2017A010101006, and in part by Microsoft Research Asia. The work of X. Qian was supported in part by the National Natural Science Foundation of China, in part by the Microsoft Research, and in part by Ministry of Science and Technology. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Dacheng Tao. *(Corresponding author: Xueming Qian.)*

X. Qian is with the Key Laboratory for Intelligent Networks and Network Security, Ministry of Education, and the Smiles Laboratory, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: qianxm@mail.xjtu.edu.cn).

C. Li and X. Hou are with the School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: chengli3435@yeah.net; houxs@mail.xjtu.edu.cn).

K. Lan is with the School of Software, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: kelan.mail@gmail.com).

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

J. Han is with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: junweihan2010@gmail.com).

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

Digital Object Identifier 10.1109/TIP.2017.2769454

patch/object pair of arbitrary distance. Nishiyama *et al.* [7] use color harmony to distinguish high aesthetic quality and low aesthetic quality of a picture. As the assessment results highly depend on the audience, photo assessment is actually a subjective task. However, with the help of crowd-source social media information, we can reasonably assume that photos favored by more users in the Internet have higher qualities.

Aesthetics is a complex matter. Some standards often utilized in photography can be employed to judge the aesthetics of an image. There are some famous heuristic principles by the professional photographer, including the rule of thirds [12], [72], diagonal dominance and sense of balance [23], which are closely related to the composition of the picture. In those approaches, the rules are obtained based on the key objects or visual saliency of images. However, the objective analysis and recognition of the composition of photos is hard. Hence, the accuracy of measuring the aesthetic quality of photo relies on the precision of salient region detection.

Besides the *rules of thumb* [34] that ordinary photographers follow, however, professional photographers have generalized some knowledge from their experiences. Therefore, we can reach a consensus that guidelines in the community can ensure the quality of photo. Several multimedia models have been developed to evaluate the aesthetic quality of image [1], [2], [10], [11], [72], [73]. With the source from social media, some models [1], [2] employ computational approach to extract visual features and build classifiers between high-quality and low-quality images. In addition, other models [10], [11], [31], [72], [73], utilizing the saliency map, are based on the implementation of some heuristic principles for photographer.

Visual saliency, being bound up with our perception system, has been studied by many disciplines such as cognitive psychology [16], [17], neurobiology [18], [19], and computer vision [20]–[22], [69]–[71]. The saliency map [13]–[15], [65], [66] is a topographically arranged map that can indicate visual saliency of each location in an image. The saliency estimation has been successfully explored in pattern recognition and content based image retrieval [20]–[22], [41], [58], [59], [64], [69]–[71].

In this paper, we propose a new POI image summarization approach by aesthetics evolution from crowd-source social media. Our approach provides more precise aesthetic measurement of an image by taking advantage of the crowd-source to build the 3D model for the POI. Firstly, we generate visual albums and reconstruct 3D model of a POI using structure from motion (SFM) [37]–[40], [60], [72], [73]. Secondly, on the base of 3D to 2D projection relation, we select candidates of aesthetic photos. Thirdly, in order to improve the performance of aesthetic measurement model, we replace the general saliency map with a crowd-sourced saliency map on the basis of the distribution of salient regions in the 3D model. Then, we measure the composition aesthetics of candidate images by applying diagonal dominance. Specially, to improve aesthetic degree precisely, a feature called crowd source salient feature is proposed to select candidate images for POI summarization. Meanwhile, we select images from different visual albums and cluster images into different perspectives to explore the distribution of cameras in each album to ensure the diversity of top ranked images in POI summarization. Finally, we combine the geographical information and the aesthetics to recommend pictures within each perspective. The main contributions of this paper can be summarized as follows:

1) We propose an unsupervised POI summarization method which combines aesthetics and diversity by exploring the saliency from its 3D reconstruction. In our approach, salient region is detected by the crowd source information to measure aesthetics of images and the diversity of summarization is satisfied by mining the position distribution of cameras in the POI.

2) We measure the relevance and aesthetics of each image to the POI based on our aesthetics model. It is effective to filter irrelevant images. We mine salient regions in 3D space based on the density of point cloud. We mine the geographical locations that can explain somewhere in the POI that photographers like very much by crowd source social media information.

3) We propose a crowd source salient feature based photo aesthetic representation approach. We select candidate photos containing more salient regions which are inferred from salient regions in 3D space. The images with high aesthetic values and their content appeared frequently in many people's cameras are selected for POI summarization.

4) We propose a crowd source saliency map based automatic image adoption approaches. This approach is based on the dynamic balance of the salient regions to achieve better visualization results for the final images.

The main differences of this paper to that of our previous approaches [72] are summarized as follows: 1) we describe details for the salient region detection; 2) we discuss the advantages of the proposed crowd source social media driven saliency detection approach to the state-of-the-art content based approaches; 3) we systematically evaluate the different crowd-source saliency detection approaches to the POI summarization performances; 4) a crowd source saliency map based image adopting approach is proposed to get better visualization results for the final ranked image list; 5) we provide comprehensive experimental results.

The remainder of this paper is organized as follows. Section II gives a brief overview of the related work. The overview of the proposed POI summarization by aesthetics evaluation is introduced in Section III. Section IV describes the details of our approach. In Section V, we give the experimental results. Finally, in Section VI, we make a conclusion on our system.

II. Related Works

In this section, we briefly overview the related work on some basic aesthetic guidelines, photo aesthetic evaluation, and POI summarization approaches.

A. Basic Aesthetic Guidelines

In a general sense, image aesthetic measurement task is to evaluate the composition aesthetics of a given image bymeasuring several well-grounded composition guidelines. There exist various basic guidelines for analyzing aesthetics

of images, such as rule of thirds [12], [53], [72], diagonal dominance [12], and visual balance [53], [73] and so on.

1) Rule of Thirds: This rule is one of the most familiar photo composition guideline. It first divides an image into nine equal parts by two equidistant horizontal lines and two such vertical lines. The four intersections are generated by these lines. Photographers are encouraged to place the important subjects around these points, for example, at the center of the image. By this composition rule, strong vertical and horizontal components in the image should be aligned with those lines.

2) Diagonal Dominance: In addition to the lines that mark the thirds, the diagonals of the image are also aesthetically significant. A salient diagonal element creates a dynamic emphasizing effect [12]. Indeed, one of the most common and effective usages for the diagonal dominance is as a leading line – a line that captures the eyes of the viewers to fixate on the subjects along it.

3) Visual Balance: Visual balance is a crucial component to the harmony of an image composition [53]. In a visually balanced photo, the salient objects are distributed evenly around the image center. Similar to a balanced weighing scale, when balanced, the center of the "visual mass" is near the center of the image. The visual mass takes into account both the area and the degree of saliency of each visually salient region in an image.

B. Photo Aesthetic Evaluation

Photo aesthetic evaluation is the analysis of image composition and aesthetics. In a general sense, this task is devoted to solving the problem of analyzing the relations between compositions of a photo and its aesthetics. To recognize the composition of a photo, early researchers in this domain tend to build an abstract map of the photo based on the visual content. Lok *et al.* [32] build a weight map to analyze the layout of pictures. Moreover, other researchers often utilize visual attention model to achieve this goal. In order to employ the prior knowledge of human in aesthetics, a number of researchers have studied techniques for quantification of the aesthetics of photographs. Liu *et al.* [34] create an evaluation criterion, which fuses the rules of thirds, visual balance dominance and diagonal dominance, to produce a maximally-aesthetic version of the input image. There are also other methods, which attempt to optimize the aesthetics of photograph through machine learning [2], [6], [35], [36]. Ni *et al.* [2] and Cheng *et al.* [6] create context modeling to estimate the joint spatial distributions of visual words based on Gaussian mixture models. Liu *et al.* [35] collect images from social media and use a regression model to mine composition knowledge of specific position. Meanwhile, Datta and Wang [36] define several indicators for photograph and use SVM (support vector machine) to learn potential rules in the field.

However, in this paper we make full use of the crowd-source from social media to evaluate the aesthetics of images taken at a POI for summarization.

C. POI Summarization

Many visual summarization approaches [55]–[60] require carrying out image categorization to classify images into geographic locations, canonical views, or scenic themes. Some approaches apply geographical clustering to geo-referenced images to construct image clusters so that images within each cluster are more likely to share a common landmark due to closeness in geographical distribution.

Jiang *et al.* [44] and [55] summarize landmarks by exploring high-frequency shooting locations based on the geo-tag information of photos posted to social networks. The system selects images from intra and inter high frequency shooting locations. Simon *et al.* [56] propose an unsupervised method for finding canonical views to form the POI summary. This approach examines the distribution of images to select a set of canonical views via visual feature clustering. Then it decomposes images into groups using a greedy algorithm to expose canonical views, and uses the likelihood criteria to summarize views. Qian *et al.* [58] model the viewpoint of an image taken at a POI in horizontal, vertical, scale and orientation aspects. They use a 4-D vector to construct the viewpoint for each image. They select identical semantic points (ISPs) from the raw SIFT points of the images to capture the unique parts of a POI [58], [64], [65]. Wang *et al.* [67] study prior researches of feature extraction in aesthetic evaluation of photos. They classify the approaches into four groups: low level, rule based, information theory, and visual attention. Afterwards, they propose a comprehensive feature set, which includes 16 novel features and 70 well proved features. Then they classify the aesthetic of an image, by an SVM based classifier, into two categories: high aesthetics or low aesthetics. Chang and Wang [68] propose method to select images from image collection for visual summarization. They produce different sets of summarized images, and each set corresponds to a particular image style. They carry out unsupervised clustering on images within and across landmark categories to discover the common photographic styles from image collection.

From above, we find that the existing image summarization methods use image content clustering and explore multi-modality information of web images [60], [78], such as text, community, and temporal and geographical information. In the content clustering-based approaches, representative images are selected in terms of image feature diversity. The difference in our approach is that we consider both the aesthetics of each images and diversity of top ranked recommended images for POI summarization by exploring the crowd source social media information.

III. System Overview

In this section, we will give an introduction to POI summarization by aesthetics evaluation from social media. The flowchart is shown in Fig.2. Our method mainly consists of the following five steps: 1) POI visual album mining, 2) 3D representation for POI by SFM, 3) salient region detection, 4) diverse perspectives generation, 5) aesthetics evaluation for POI summarization. Firstly, we generate visual albums by coarse to fine clustering. Secondly, we get the POI model of each visual album by 3D reconstruction. Then we consider two aspects: aesthetics and diversity. We get aesthetics by salient region detection. As for diversity, we select aesthetic

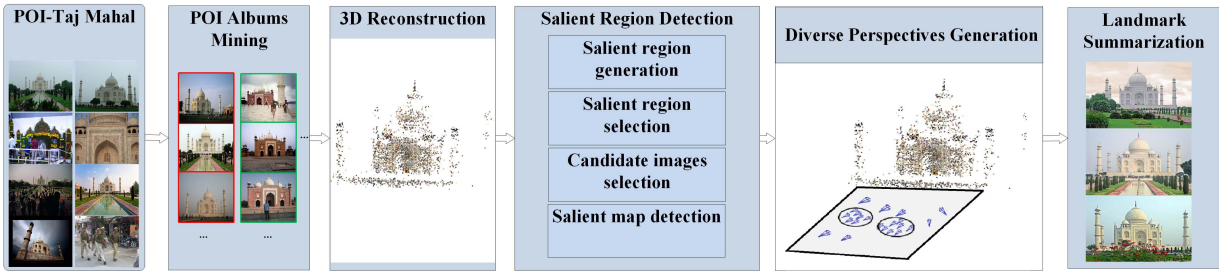


Fig. 2. Flow chart of the proposed POI summarization with aesthetics evaluation from crowd-source social media.

images from different visual albums of one POI and diverse perspectives in each visual album. In each of perspective, there are several POI images with similar perspective. Finally, we carry out POI summarization by aesthetics based image ranking. The following section will explain each part in detail.

IV. POI Summarization With Aesthetics Evaluation

Given a set of collected photos for a POI from social media websites, we only keep the images with geo-tags. We aim at building 3D models of a POI and we carry out aesthetics evaluation for POI summarization.

A. Visual Albums Mining

As for a POI, there exit several positions where photographers can take aesthetic images. As for images of a POI, they are likely to have several styles, for example, images taken at night and daytime are with different styles. Thus we utilize visual album generation approach to classify those miscellaneous images into different albums [58]. There are two steps in album generation for each POI: coarse POI clustering using GPS information (i.e. the longitude and latitude that the image was taken) and image content based fine POI mining.

1) Coarse POI Clustering Using GPS Information: We utilize mean-shift algorithm to cluster images of a POI using GPS information [41], [44], [45], [51], [55]. After mean-shift clustering, a set of coarse POI clusters is generated. The mean-shift based image clustering approach is based on the geo-graphical distribution of image taken place rather than the content similarity. Actually, images with identical GPS, may not share similar content. For example, we stand at a fixed location, and take photos from the different directions and views.

2) Image Content Based Fine POI Clustering: Considering that images in a cluster are the same geographical location but with different appearances, we propose a fine POI clustering approach to refine the coarse clustering results. It consists of the following three steps:

a) Feature extraction: Each image is represented by a set of local features, i.e. SIFT feature.

b) Similarity measurement: The similarity of two photos is measured by SIFT point matching [58] or some feature descriptors [75]–[77]. Each image is compared with the rest of the images in its cluster.

c) Graph growth based albums generation: we group the images in each POI to obtain visual albums based on graph growth [58]. This algorithm first finds two most similar images, and puts them into a visual album, then finds other similar images according to the visual similarity. The visual album stops growing until non-similar image is found. Then we turn to get another visual album. Finally, we sort the albums in descending order by the number of images in it. Actually, some albums are with very little number of images. These albums can be removed from final POI summarization.

B. 3D Reconstruction

As in the websites, there are many duplicated images. The duplicated images do not provide any complimentary information for building 3D structures for the scene but increase the computational costs. So, we first carry out duplicated image removing to speed up the reconstruction process. We generate a 3D model for each POI album based on the SFM algorithm [37]–[40], which consists of the following six steps:

a) we extract SIFT feature points for the images in a visual album.

b) we find the best matched pair in a visual album by SIFT features matching.

c) we estimate fundamental matrix for the pair using RANSAC [58]. We get the camera information C that we describe in detail. During each iteration of RANSAC, we calculate a coarse fundamental matrix using the 8-point algorithm [64], and we get a point cloud for the pair.

d) we select one image from the pair and find the best matched image in the rest images in the visual album. These two images further construct a pair. Based on the step c) we get their corresponding point cloud.

e) we get a set of point clouds from pairs of images by carrying out step c) and d) iteratively.

f) we merge the point clouds of the pairs using bundle adjustment [49] to get an integrated point cloud for each visual album.

After completing the process of 3D reconstruction for a visual album, we get a group of 3D models, including their camera information C and geometric point's information G [39]. We represent the group of 3D models information as follows:

$$S = \{C; G\} \tag{1}$$

Geometric information G of the reconstructed POI contains the information such as a 3-vector describing points of the 3D position, a 3-vector describing the RGB color of the

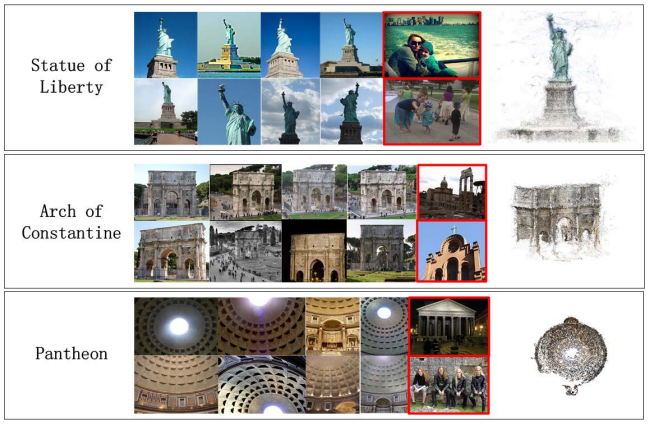


Fig. 3. 3D reconstruction for three different POIs: In each POI, the eight pictures from the first column to the fourth column are the dominant kind of pictures in the cluster and the two noise images in the last column are marked with red border.

point, and a list of views the point is in. Camera information C contains 3-vector camera position, focal length F, 3-vector translation T, 3×3 matrix format of rotation R and the parameters of radial distortion k_1 and k_2[39].

The results of 3D reconstruction in three different POIs are shown in Fig.3. For each POI, we select 8 representative images and 2 noisy images, which are irrelevant to the POI. Correspondingly, the reconstructed 3D cloud points of the POIs are shown in the last column. From Fig.3, we find that density of data points depends on the regions appearing in images in reconstruction. We also find that the influence of noisy images is limited in 3D reconstruction, because pairwise matching can remove noise images effectively.

C. Salient Region Detection

After 3D reconstruction, we get a set of point clouds of the POI. We use each point cloud to detect salient region based on the 'heat' level of regions (which have positive relationships with the photos taken at) in 3D space, which are derived from the crowd-source social media information. If more people like to take more photos at the region, then its "heat" level will be higher. Our motivation is that if a region of a POI appearing in most of images, then it is more popular to visitors.

According to the SFM [37]–[40], there are several point clouds for a POI as shown in Fig.4 (a) and (b). So we need to detect salient regions in each point cloud. The detailed approach consists of the following four steps.

1) Salient Region Generation and Representation: In order to eliminate some noise points in 3D models [72], [73], an improved mean-shift algorithm is employed on the point cloud as follows:

$$\begin{cases} M_h(X) = \dfrac{1}{k} \sum_{X_i \in S_h(X)} (X_i - X) \\ S_h = \{Y : (Y - X)^T (Y - X) \leq h^3\} \end{cases} \tag{2}$$

where $S_h(X)$ is the sphere whose radius is h, k is the number of 3D points falling within the region $S_h(X)$, X_i is a 3D point in $S_h(X)$, and X denotes a cluster center in 3D space.

After salient region clustering, we get a set of clusters in 3D space. Each cluster corresponds to a region containing a

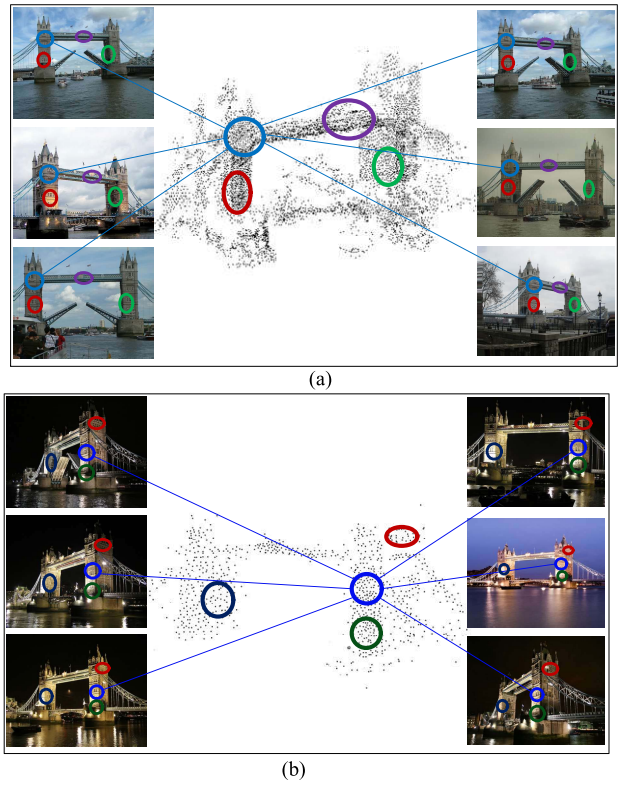


Fig. 4. Projection of salient regions from 3D point clouds to 2D images. There are several point clouds of Tower Bridge and each of them should be detected as salient regions. (a) Point cloud and the salient regions of a visual album. (b) Point cloud and the salient regions of another visual album.

set of matched points. The regions with high saliency indicate their attractiveness to people. Thus, it is reasonable to represent the saliency of a region in terms of its frequency that it appears on photographers' cameras. We denote the saliency of the region r as D_r and we represent it by the mean value of the point frequency as follows:

$$D_r = \frac{1}{n} \sum_{i=0}^{n} f_i \tag{3}$$

where n is the number of points in a cluster, f_i is the frequency of the 3D point X_i. In this paper, we define the frequency f_i as

$$f_i = N_i / N_a \tag{4}$$

where N_i is the number of images with their corresponding SIFT points that can be mapped to the X_i in the 3D model, and N_a is number of images in a visual album.

2) Salient Region Selection: To remove noise points in the point cloud and unimportant regions in 3D models, we sort regions by D_r, and we remove the small regions which contain less points. In this paper, we only keep the top ranked 80% of regions as salient regions to generate saliency map and the rest 20% regions are removed as noise. The kept regions are utilized in image aesthetic measurement for POI summarization. Assume that the total number of the left salient regions is Z for a visual album with Q images. This will be utilized in selecting candidate images and evaluating their aesthetics.

3) Crowd Source Saliency Map Detection: In order to get more accurate description of aesthetics, we utilize crowd source salient feature to select candidate images for POI summarization. We project all the Z salient regions in 3D into each 2D image to get the corresponding crowd-source saliency map. Taking into account the distortion of the camera [37], [49], the projection process of a specific camera can be described as:

$$P_C = P_W R + T \tag{5}$$

where P_W is world coordinate of a point in point clouds, P_C is camera coordinate of a point in 3D space, R is the 3×3 rotation matrix and T is the 3×1 translation vector in 3D space.

After the projection, we obtain the conversion from world coordinates to camera coordinates. Before we convert camera coordinates to image coordinates, we further carry out perspective division for dimension reduction as follows:

$$p = red[-P_C / P_C.z] \tag{6}$$

where $red\,[*]$ is the extraction of (x, y) from (x, y, z), $P_C.z$ is the third (z) coordinate of P_C, and p is the dimension reduction result.

Let P_{Im} denote the position in an image which the world coordinate P_W can be projected into. We get P_{Im} as follows:

$$P_{Im} = (P_{Im}.x, P_{Im}.y) = F(1 + k_1 N_{Im}^2 + k_2 N_{Im}^4) p \tag{7}$$

where F is the camera focal length. $N_{Im} = \sqrt{\|p\|^2 + 1}$ and $\|p\|$ is defined as the norm of matrix p, $(P_I.x, P_{Im}.y)$ is the coordinate in the image. In the projection from 3D to 2D, the parameters of radial distortion is considered, which are represented by k_1 and k_2 respectively for correcting the radial distortion. Each camera corresponds to a set of (k_1, k_2). The value of k_1 and k_2 can be estimated from 3D reconstruction.

Fig.4 (a) and (b) show some representative images from the two visual albums and their corresponding 3D point clouds generated by SFM. We label the salient regions in 3D and we mark their corresponding projected regions in images by ellipses in identical colors. We can see from Fig.4 that a salient region (in 3D point cloud) can project into many images. In general, an image has more than one salient region if no occlusion existed. From Fig.4, we find that the main structure of the tower bridge has denser points than that in the background. For example, the matched points in the water are rather sparse, which can be viewed as noise for POI summarization. Compared to the traditional saliency detection approach, our saliency map detection approach considers the crowd source saliency information which can be well associated with the aesthetic evaluation.

4) Candidate Images Selection: The selected salient regions collaboratively describe photo local aesthetics. We propose a crowd source saliency enhanced feature based POI summarization approach. We select some representative pictures as candidate images from the total Q images in the visual album to carry out POI summarization based on the selected salient regions. The detailed steps are as follows:

a) We project the Z salient regions in 3D into Q images. If more than 80% points in the salient regions can be projected to an image, then we claim that the salient region can be projected to the image, otherwise cannot be.

b) Then we get how many images a salient region can be projected into. We utilize the corresponding image number to denote the importance of the salient region.

c) We project all the selected salient regions in 3D into images in each visual album of a POI.

d) If a salient region can be mapped into an image, we define that the image contains the salient region. We select the images containing more salient regions in 3D point cloud as candidate images for POI summarization.

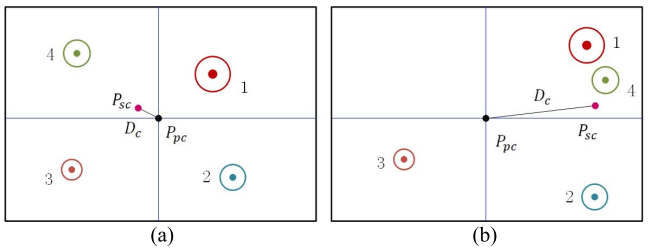


Fig. 5. The diagrams of dynamic balance of two images (a) Image with good visual balance (b) Image with bad visual balance. The radius of each point represents its weight.

D. Aesthetics Measurement

For the selected candidate images, we propose an aesthetic measurement approach by exploring the distribution of salient regions. Motivated by the photographical experiences [42], [43], [62], we build a computational model called dynamic balance to disclose distribution of salient region. The more balance regions, the more aesthetic the image.

Let P_{sc} denote a photo's saliency center (as shown in Fig.5), which is the weighted centroid of salient regions.

$$P_{sc} = (\frac{1}{N}\sum_i^N A_i x_i, \frac{1}{N}\sum_i^N A_i y_i) \tag{8}$$

where N is the number of points that are mapped into an image, and A_i is the weight of 2D point $p_i = (x_i, y_i)$ that is mapped from a 3D point to the coordinate (x_i, y_i) in an image. In this paper, we set A_i as the frequency of the 3D point in the point cloud which is determined by Eq. (3). In our experiment, we use D_r of the salient region to represent the weights of all the 3D points it contains.

We get the offset value D_c from the saliency center to the physical center as follows:

$$D_c = \left| P_{sc} - P_{pc} \right| \tag{9}$$

where P_{pc} is the center of a picture. The dynamic balance of the photo is evaluated by:

$$S_{DB} = 1 - 2arctan\frac{D_c}{\pi} \tag{10}$$

where S_{DB} ranges from 0 to 1. Larger S_{DB} means the better visual balance of the salient regions in the picture. Otherwise, salient regions cannot keep good visual balance and the balance is broken. For simplicity, we use an example to show this. In both Fig.5 (a) and (b), the size of the point p_i

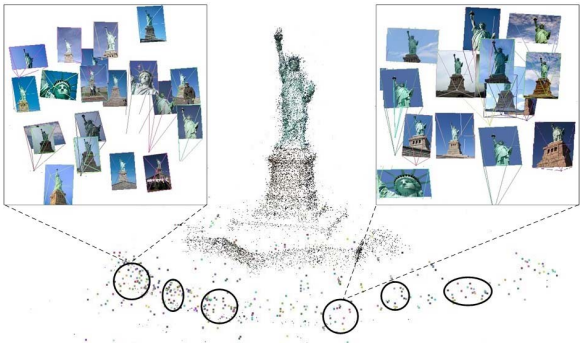


Fig. 6. Perspective clustering for a visual album of a POI.

indicates its weight A_i. The larger the size is, the big the weight is. Fig.5 (a) is with high balance than that of Fig.5 (b).

E. Aesthetic Based Image Ranking for POI Summarization

Our POI summarization approach takes into account both the aesthetics of each image and the diversity among the top ranked images. Firstly, we consider the diversity of POI summarization results by recommending images from different visual albums. We recommend images from each visual album by exploring the camera position distribution estimated from 3D reconstruction process [37], [49]. Secondly, we combine aesthetics and diversity for POI summarization. Our method guarantees aesthetics of each image and makes sure the diversity of top ranked results.

In general, the images in some POI can generate several visual albums in the coarse-to-fine clustering, and images in the same visual album are with high similarity. So, images selected from different visual albums are with high diversity than that from the same visual album.

Images in the same album have much content overlap but they have different perspectives such as the difference of distance (far or near) to a POI and the difference of angle to view a POI.

Mean-shift algorithm is applied here to cluster cameras (the coordinates in 3D space) in a visual album into different perspectives. After the mean-shift clustering, cameras nearby can be grouped into one perspective. As shown in Fig.6, different perspectives are labeled by different ellipses. In Fig.6, a point corresponds to a camera. The enlarged version of two perspectives is given. A camera in 3D reconstruction system corresponds to picture. We can see that pictures in one cluster have a similar perspective.

Based on the ranked 3D point clouds and their corresponding visual albums, we pick out images with high aesthetics scores from each perspective in each visual album for POI summarization.

F. Crowd Source Saliency Map Based Image Adoption

Based on the obtained saliency map, we propose an adaptive image content adoption approach. In this approach, we first put the saliency center P_{sc} determined by Eq.(8) to the picture centroid P_{pc}. Then, we crop the image to the optimal sizes that makes the final image with better aesthetics. For the two

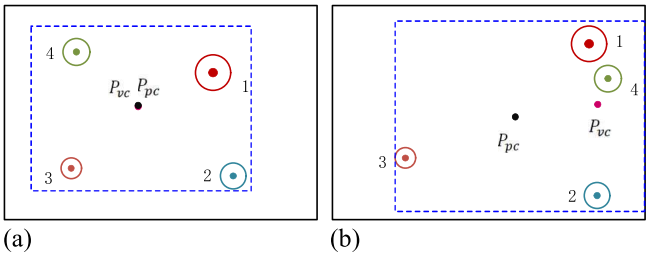


Fig. 7. Crowd source saliency map based image content adoption. The blue virtual boxes represent the images after adoption.

images in shown in Fig.5, after the content adoption based on the crowd source saliency map, the cropped images are shown by the virtual blue boxes as shown in Fig.7. From by comparing Fig.5 and Fig.7, we find that the saliency center P_{sc} and the picture centroid P_{pc} of the first image is identical, while the dynamic balance of the second image is improved somewhat.

V. Experiment

In order to evaluate the effectiveness of the proposed method, some comprehensive comparisons are made with existing approaches, including Canonical Views (denoted as CV) [56], Clustering, Ranking and Ranking (denoted as CRR) [57], Identical Semantic Points (denoted as ISP) [58], High Frequency Shooting Location (denoted as HFSL) [44] and Social-Contextual Constrained Geo-clustering (denoted as SCCG) [45]. The detailed description for each of these approaches is as follows:

CV: it is an unsupervised method for finding canonical views. The approach examines the distribution of images to select a set of canonical views via visual feature clustering. Its basic idea is that an image selected as a representative image is similar to many other images in the input set.

CRR: it is a content-based method to choose diverse and representative image. The method focuses on statistics for number of users, visual coherence, cluster connectivity and variability in dates. However, statistical method needs large enough accurate data.

ISP: it is an effective model to model an image's viewpoint in horizontal, vertical, scale and orientation aspects. It selects identical semantic points (ISPs) from the raw SIFT points using the 4D vector.

HFSL: it is an author topic model-based collaborative filtering method to facilitate comprehensive points of interest (POIs) recommendations for social users. User preference topics are extracted from the geo-tag constrained textual description of photos via the author topic model instead of only from the geo-tags.

SCCG: it employs visual and views verification to select images from LOIs to summarize the POI. They mine LOIs for each POI by the improved geo-location clustering method, and they employ visual and views verification to select images from LOIs to summarize the POI.

Experiments are conducted on the collected 7 million Flickr images uploaded by 7,387 users and the heterogeneous meta-data associated with the images with Flickr API. We choose eight POIs to evaluate our method. They are: #1) Angkor,

#2) Big Ben, #3) Cologne Cathedral, #4) Colosseum, #5) Eiffel Tower, #6) Golden Gate Bridge, #7) Taj Mahal, and #8) Tower Bridge.

A. Evaluation Criteria

In this part, we utilize a user-driven approach that twenty volunteers are invited to evaluate the POI summarization performances. We use aesthetic score and diverse score for performance evaluation. Aesthetics score measures the aesthetics of the results and the larger the value, the more aesthetic the results. Diverse score measures the diversity of the results and the score donates the number of different images in Top-5 results.

We evaluate the POI summarization results by assigning the aesthetics scores aes_i and diversity scores div for different summarization approaches. aes_i s aesthetics scores of the i-th image. $aes_i \in 0, 1, 2, 3\}$, where the four discrete values are on behalf of inelegant, ordinary, good, and perfect. Let div denote the diversity of the POI summarization result. In this paper, we classify it into four categories: 3-excellent, 2-good, 1-normal, 0-irrelevant as that utilized in [63]. When the top ranked image is irrelevant to the POI, then we set both the aesthetics score aes_i and diversity score div zero.

We utilize the average precision (AP) [24], [26] for performance evaluation. Once we get the value of aes_i and div, the AP of the top-n images is determined as follows:

$$AP@n = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{i}\frac{aes_i}{i}\right) \tag{11}$$

Whether POI summarization is good or not depends on whether it provides aesthetic images and makes users fully understand the POI.

B. Objective Performance Comparison

The average aesthetic scores and diverse scores assigned by the 20 volunteers to eight POIs are shown in Fig.8 respectively. The average (denoted by ave) scores are also given in the last columns. From Fig.8, we find that our method is with better aesthetics for all the eight POIs. The diversity of our approach also outperforms other approaches. As for CV, the method just uses visual feature clustering to select a set of canonical views and doesn't analyze image aesthetics and it shows bad effect in aspect of aesthetics and diversity. As for CRR, it considers the factor of number of users, but this factor needs large amounts of accurate data. Other methods consider the representativeness of results, but they do not consider salient region using crowd source information.

From Fig.8 (b), we find that SCCG, HFSL and ours are with highest diversity scores. This is due to the fact that SCCG fuses multimodality information from social media, such as views (the times that photos have been browsed by different users), geographical distribution, and visual clustering. The HSFL both considers the high shooting frequency and the visual content of images. While in our approach, we take both the location of image and the saliency information mined from the 3D models. This can ensure that the top ranked images are selected from diverse perspectives.

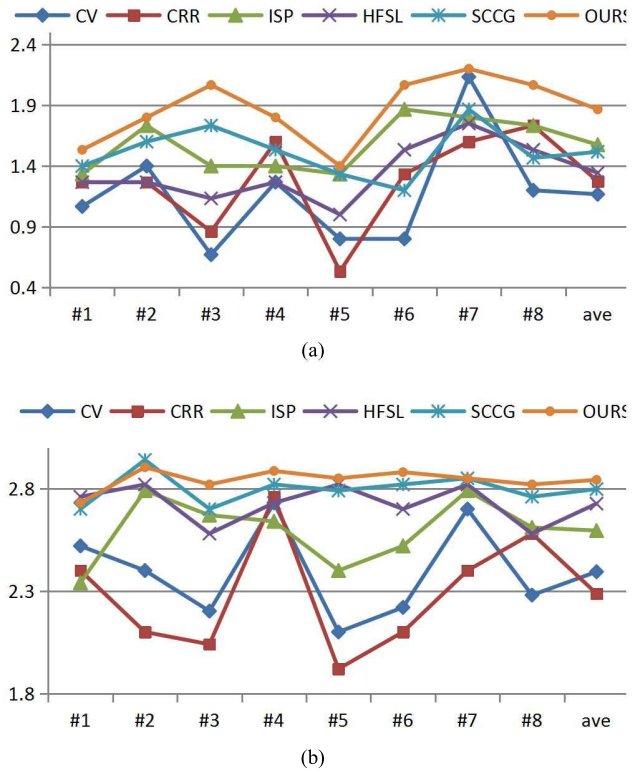


Fig. 8. Average aesthetic scores and diverse scores CV, CRR, ISP, HFSL, SCCG and OURS on 8 POIs. The y-axis is the score and the x-axis is the POI index. (a) the average aesthetic scores. (b) the average diverse scores.

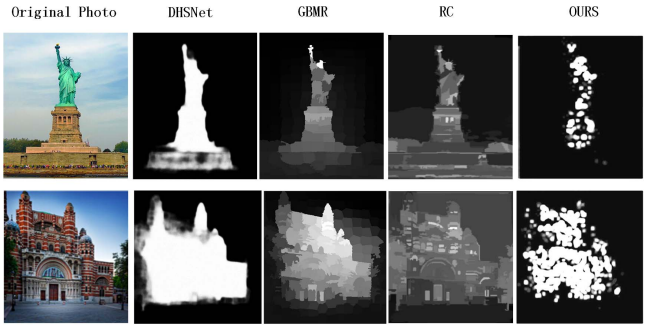


Fig. 9. The comparison of other saliency maps and our saliency map.

C. Discussions

1) Importance of Crowd Source Saliency for POI Summarization: In this part, we compared our saliency map with other saliency map models, including DHSNet [31], RC [65] and GBMR [66]. And we put those saliency map models into our system and recommend images for POI summarization.

We show a few visual comparisons in Fig.9. In our saliency map, we use 3D salient regions generated by crowd source information in 3D models and 3D-2D mapping relationship. Thus, our saliency map is in regional distribution. We observe that our approach not only can distinguish salient objects accurately, but also handle backgrounds well. In our method, we find public attention focuses on the building itself and this is well embodied in our crowd-source saliency map. In Fig.9, our approach distinguishes landmarks in the photos and excludes backgrounds. GBMR and RC cannot distinguish landmarks and background well. Because DHSNet uses object

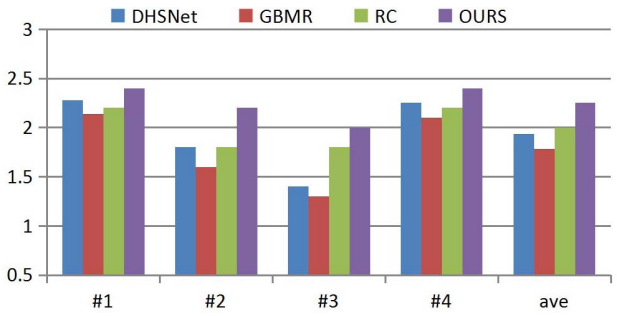


Fig. 10. The average aesthetic scores of four POIs.

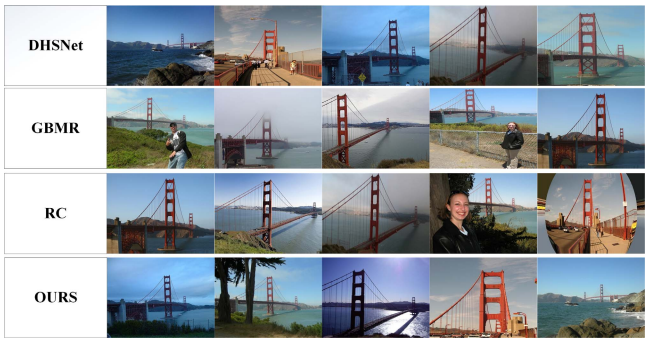


Fig. 11. Example of results of different saliency map methods.

detection, sometimes it cannot make POI salient. For example, a person stands in front of the POI. Our approach eliminates false saliency map in an image, especially, the salient regions that are occluded by pedestrians. For instance, in the second example, the trees around the building are successfully excluded from the salient region of photos, which could cause errors by employing general saliency map. In our salient region detection approach, the fusion of crowd-source social media information is more effective in extracting the POI oriented saliency map than other saliency map models.

We calculate aesthetics of photos using DHSNet, GBMR, RC though dynamic balance and compare their results with OURS. We select four POIs Tower Bridge, Colosseum, Eiffel Tower and Golden Gate Bridge, and the results are shown in Fig.10 and Fig.11. We can see from Fig.10 that our approach shows better results that other saliency maps to select aesthetic images for POI summarization. The main reason is that our approach mines photographers' intention for aesthetics, and selects salient regions to generate saliency map as input of dynamic balance. And there exits another reason that our saliency map detection approach can eliminate disturb objects such as people, trees, cars and so on. Our approach makes full use of feature matching that is effective to remove irrelevant objects from saliency map by SFM based 3D reconstruction. But DHSNet, GBMR, and RC cannot eliminate those parts of photos. In Fig.11, the results of DHSNet, GBMR and RC have some results that contain people that are irrelevant to the POI, but those parts which contain people are also with high saliency scores. The results by our approach can eliminate the influences of the rare objects in the images taken at the POI.

2) Effectiveness of Crowd Source Saliency: In addition, we conduct another experiment to demonstrate the effectiveness of crowd source saliency representation approach in POI

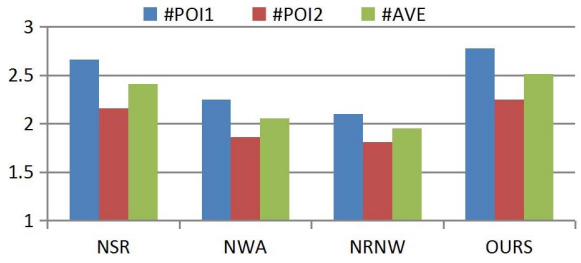


Fig. 12. Effective of crowd source saliency.

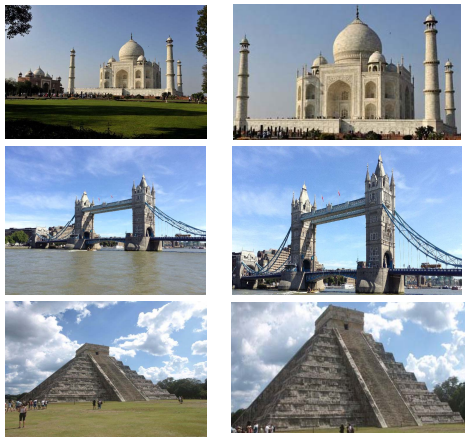


Fig. 13. Crowd source saliency map based image content adoption. (a) original images, (b) images after adoption.

summarization. We take three conditions: the first one (denoted as NSR) is that we do not consider salient regions, and just use the points in 3D models to recommend pictures. The second one (denoted as NWA) is that we don't consider the weight A_i in formula (8). The last one (denoted as NRNW) is that we do not consider salient regions and the weight A_i.

In Fig.12, we show their comparisons on POI1, POI2 and the average aesthetic values on the 8 POIs. We find that, our approach is better than NSR, NRNW and NWA. As for NSR, it brings some noise points into our system because of feature matching of SFM. Using the salient regions can eliminate some noisy points in 3D models in some ways and select the 'heat' level of regions in 3D space. NWA cannot distinguish importance between regions. If a region repeatedly appears in pictures, its importance is significantly higher than others that appear in pictures occasionally. As for NRNW, it not only neglects importance between regions, but also takes some noise points because of feature matching of SFM. Thus, it is with the lowest performances. However, when the region and weight of each point is taken into account, better performance can be achieved.

D. Crowd Source Saliency Based Adaptive Image Adoption

To show the effectiveness of the proposed adaptive image adoption approach based on the obtained crowd source saliency map, three examplar images are shown in Fig.14.

From the comparions results we find that the image adoption contain the salient part of the image and with high aesthetics than the original images. This kind of image visualization

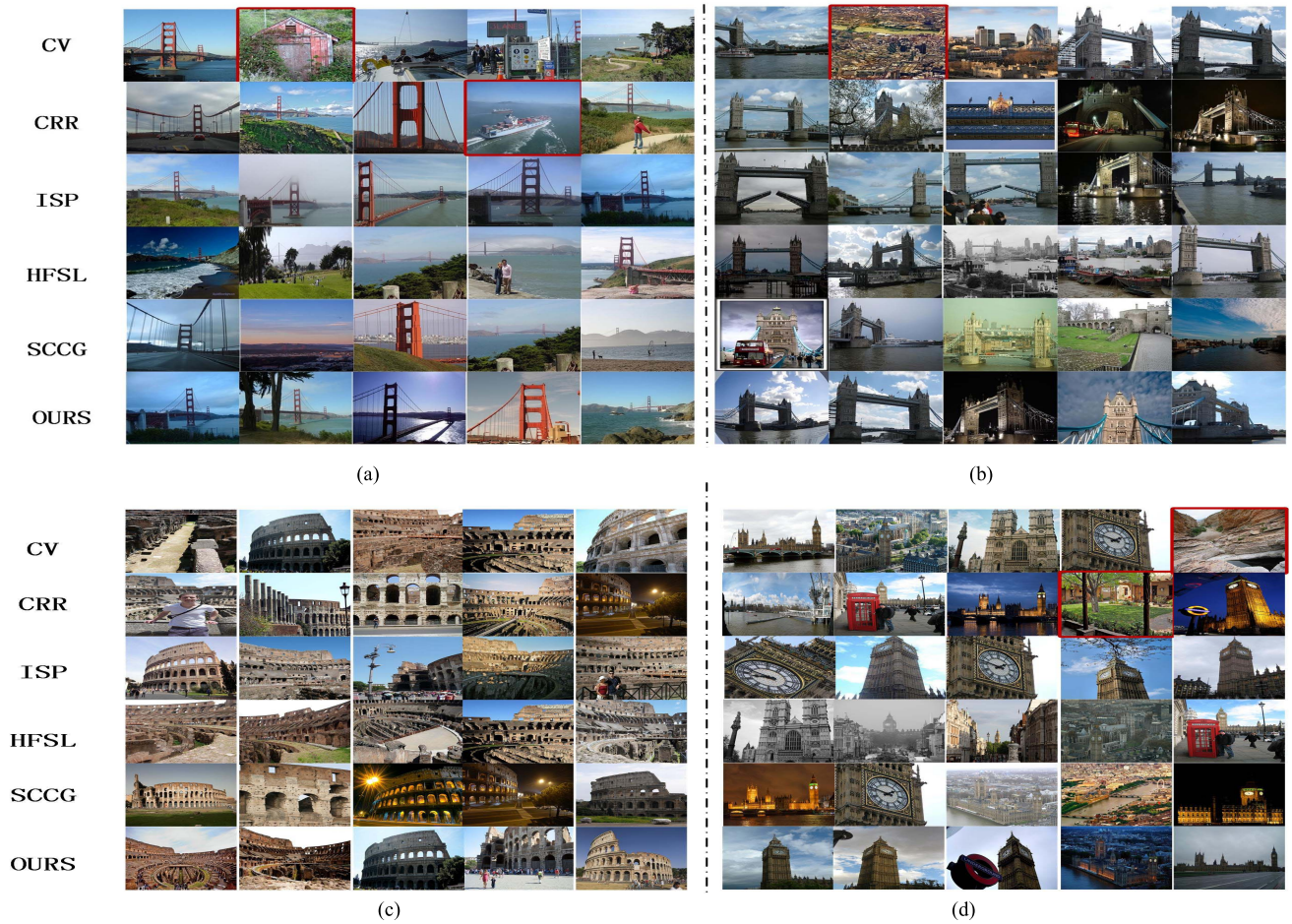


Fig. 14. The results of CV, CRR, ISP, HFSL, SCCG and OURS for four POIs: (a) Golden Gate Bridge, (b)Tower Bridge, (c) Colosseum, (d) Big Ban. Images in red frames are irrelevant to the POIs.

approach can be utilized in the mobile devices with small screen, such as pad and smart phone.

E. Subjective Performance Comparison

To show the POI summarization performances intuitively, we give four POIs and their corresponding summarization results. The detailed comparisons with CR, CRR, ISP, HFSL, and SCCG are shown in Fig.13. Please turn to the end of this paper for details. We find that the CV and CRR based POI approaches sometimes select irrelevant images in the top ranked image list, which are labelled by red frames. And diversity of CV is remarkable, since it decreases similar canonical views. However, it is hard to filter the irrelevant images.

We find that SCCG achieves good aesthetics performance because it takes the views into account, which indicate the attractiveness of the photos in websites. ISP works well with diversity, which generates an image's viewpoint by a 4D vector. Comparing with CRR, HSFL and SCCG, our approach achieves better diversity and aesthetics. Our approach considers the diversity using different position of cameras estimated by 3D reconstruction. As for the aesthetics, salient regions are used in POI summarization. Especially, when fusing crowd source salient feature, the selected images are with many attractive regions.

Other approaches such as ISP and SCCG contain local area for the POI as shown in Fig.14 (b) and (d). We find that our approach is effective to summarize the global view of the POI rather than the local views. This is caused by the fact that our saliency model and aesthetics are obtained from the global salient regions in 3D cloud points rather than the local salient regions. As for CV and CRR, there exit some irrelevant images as shown in Fig.14(a). In our POI summarization results, no irrelevant image exits. In Fig.14(d), we can see that the results of HFSL cannot describe the POI clearly. We also can see in our approach that the first two pictures are more similar to each other, because they belong to a 3D model built by images from the same visual album. But they are in different perspectives, and they are captured from different shooting angles. Other three pictures in our approach belong to different 3D models. If you want to return more pictures as recommend pictures, you can select more pictures from different 3D models.

VI. Conclusion

In this paper, we have proposed a new POI summarization by aesthetics evaluation from crowd source social media information. From the 3D models of the POI, irrelevant images can be well removed from the recommended image list. The density of the cloud points in 3D space embodies the heat

levels of the region in the POI. A novel method to build the saliency map by calculating the frequency of points in 3D model appearing on the lens was proposed here. Crowd source salient feature was presented to gain more precise aesthetics evaluation. Crowd source salient feature was used to guarantee that images in results have more salient regions. From the results, our method performs better than other methods, especially in aspect of aesthetics. By exploring the crowd source social media information, we cannot only model the aesthetics of the photos taken at the POI, and estimate the camera poses at the POI that takes the photos, which enables us to select aesthetic images with diverse viewpoints for POI summarization. Moreover, the crowd source saliency map can guide the image content adoption.

REFERENCES

[1] L. Zhang, Y. Gao, R. Zimmermann, Q. Tian, and X. Li, "Fusion of multichannel local and global structural cues for photo aesthetics evaluation," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1419–1429, Mar. 2014.

[2] B. Ni, M. Xu, B. Cheng, M. Wang, S. Yan, and Q. Tian, "Learning to photograph: A compositional perspective," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1138–1151, Aug. 2013.

[3] Y. Gao, M. Wang, Z.-J. Zha, Q. Tian, Q. Dai, and N. Zhang, "Less is more: Efficient 3-D object retrieval with query view selection," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 1018–1071, Oct. 2011.

[4] L. Zhang, Y. Han, Y. Yang, M. Song, S. Yan, and Q. Tian, "Discovering discriminative graphlets for aerial image categories recognition," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5071–5084, Dec. 2013.

[5] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *Proc. Int. Conf. Multimedia*, 2010, pp. 271–280.

[6] B. Cheng, B. Ni, S. Yan, and Q. Tian, "Learning to photograph," in *Proc. Int. Conf. Multimedia*, 2010, pp. 291–300.

[7] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato, "Aesthetic quality classification of photographs based on color harmony," in *Proc. CVPR*, Jun. 2011, pp. 33–40.

[8] L. Zhang, M. Song, Q. Zhao, X. Liu, J. Bu, and C. Chen, "Probabilistic graphlet transfer for photo cropping," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 802–815, Feb. 2013.

[9] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato, "Sensation-based photo cropping," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 669–672.

[10] K. Kataoka, K. Sudo, and M. Morimoto, "Region of Interest detection using indoor structure and saliency map," in *Proc. ICPR*, Nov. 2012, pp. 3329–3332.

[11] J. Liu, F. Meng, F. Mu, and Y. Zhang, "An improved image retrieval method based on SIFT algorithm and saliency map," in *Proc. Int. Conf. Fuzzy Syst. Knowl. Discovery*, Aug. 2014, pp. 766–770.

[12] T. Grill and M. Scanlon, *Photographic Composition*. New York, NY, USA: Amphoto Books, 1990.

[13] T. Liu *et al.*, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[14] A. Borji, M.-M. Cheng, H. Jiang, and Jia Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, 2015.

[15] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[16] H. L. Teuber, "Physiological psychology," *Annu. Rev. Psychol.*, vol. 6, no. 1, pp. 267–296, 1955.

[17] J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?" *Nature Rev., Neurosci.*, vol. 5, no. 6, pp. 495–501, 2004.

[18] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Ann. Rev. Neurosci.*, vol. 18, no. 1, pp. 193–222, 1995.

[19] S. K. Mannan, C. Kennard, and M. Husain, "The role of visual salience in directing eye movements in visual object agnosia," *Current Biol.*, vol. 19, no. 6, pp. R247–R248, 2009.

[20] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[21] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *Proc. IEEE Int. Conf. Comput. Vis*, Dec. 2013, pp. 1529–1536.

[22] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 194–201, Jan. 2012.

[23] M. Hayashi, "A study on dynamic balance of pictures (research of aesthetic apperception II)," *Jpn. J. Edu. Psychol.*, vol. 3, pp. 11–17, Mar. 1955.

[24] X. Qian, D. Lu, and X. Liu, "Image retrieval by user-oriented ranking," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, 2015, pp. 511–514.

[25] M. Taylor, J. Guiver, S. Robertson, and T. Minka, "SoftRank: Optimizing non-smooth rank metrics," in *Proc. WSDM*, 2008, pp. 77–86.

[26] A. Moffat and J. Zobel, "Rank-biased precision for measurement of retrieval effectiveness," *ACM Trans. Inf. Syst.*, vol. 27, no. 1, 2008, Art. no. 2.

[27] B. Yan, K. Sun, and L. Liu, "Matching-area-based seam carving for video retargeting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 2, pp. 302–310, Feb. 2013.

[28] X. Sun, H. Yao, R. Ji, and S. Liu, "Photo assessment based on computational visual attention model," in *Proc. ACM MM*, 2009, pp. 541–544.

[29] L. Wolf, M. Guttmann, and D. Cohen-or, "Non-homogeneous content-driven video-retargeting," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–6.

[30] B. Suh, H. Ling, B. B. Bederson, and D. W. Jaobs, "Automatic thumbnail cropping and its effectiveness," in *Proc. Annu. ACM Symp. Conf User Interface Softw. Technol.*, 2003, pp. 95–104.

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

[32] S. Lok, S. Feiner, and G. Ngai, "Evaluation of visual balance for automated layout," in *Proc. Intell. User Interfaces*, 2004, pp. 101–108.

[33] Z. Byers, M. Dixon, W. Smart, and C. Grimm, "Say cheese! Experiences with a robot photographer," *AI Mag.*, vol. 25, no. 3, pp. 37–46, Sep. 2004.

[34] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or, "Optimizing photo composition," *Comput. Graph. Forum*, vol. 29, no. 2, pp. 469–478, 2010.

[35] H. Liu, T. Mei, J. Luo, H. Li, and S. Li, "Finding perfect rendezvous on the go: Accurate mobile visual localization and its applications to routing," in *Proc. MM*, 2012, pp. 9–18.

[36] R. Datta and J. Z. Wang, "ACQUINE: Aesthetic quality inference engine-real-time automatic rating of photo aesthetics," in *Proc. ACM*, 2010, pp. 421–424.

[37] Y. Zhang, X. Qian, X. Tan, J. Han, and Y. Tang, "Sketch-based image retrieval by salient contour reinforcement," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1604–1615, Aug. 2016.

[38] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz, "Multicore bundle adjustment," in *Proc. CVPR*, Jun. 2011, pp. 3057–3064.

[39] N. Snavely, S. M. Sertz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3D," in *Proc. ACM MM*, 2006, pp. 835–846.

[40] S. Almaadeed, A. Bouridane, D. Crookes, and O. Nibouche, "Partial shoeprint retrieval using multiple point-of-interest detectors and SIFT descriptors," *Integr. Comput.-Aided Eng.*, vol. 22, no. 1, pp. 41–58, 2015.

[41] 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.

[42] B. Peterson, *Learning to See Creatively: Design, Color, and Composition in Photography*. New York, NY, USA: Amphoto Books. 2003.

[43] M. Freeman, *The Photographer's Eye*. Waltham, MA, USA: Focal Press, 2007.

[44] 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.

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

[46] X. Li, Q. Lv, and W. Huang, "Learning similarity with probabilistic latent semantic analysis for image retrieval," *KSII Trans. Internet Inf. Syst.*, vol. 9, no. 4, pp. 1424–1440, 2015.

[47] M. K. Kundu, M. Chowdhury, and S. R. Bulò, "A graph-based relevance feedback mechanism in content-based image retrieval," *Knowl.-Based Syst*, vol. 73, pp. 254–264, Jan. 2015.

[48] R. C. Bolles and M. A. Fischler, "A RANSAC-based approach to model fitting and its application to finding cylinders in range data," in *Proc. Int. Joint Conf. Artif. Intell.*, 1981, pp. 637–647.

[49] N. Snavely, S. M. Seitz, and R. Szeliski, "Modeling the world from Internet photo collections," *Int. J. Comput. Vis.*, vol. 80, no. 2, pp. 189–210, 2007.

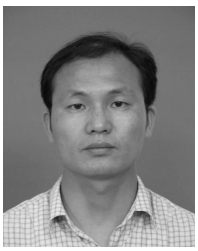
[50] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Netw.*, vol. 19, no. 9, pp. 1395–1407, 2006.
[51] J. Li, X. Qian, Y. Y. Tang, L. Yang, and T. Mei, "GPS estimation for places of interest from social users' uploaded photos," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 2058–2071, Dec. 2013.
[52] F. Dellaert, S. M. Seitz, C. E. Thorpe, and S. Thrun, "Structure from Motion without Correspondence," in *Proc. CVPR*, vol. 2. Jun. 2000, pp. 557–564.
[53] B. Krages, *Photography: The Art of Composition*. New York City, NY, USA: Allworth Press, 2005.
[54] A. Qamra and E. Y. Chang, "Scalable landmark recognition using EXTENT," *Multimedia Tools Appl.*, vol. 38, no. 2, pp. 187–208, 2008.
[55] 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*, Jul. 2013, pp. 1–6.
[56] I. Simon, N. Snavely, and S. M. Seitz, "Scene summarization for online image collections," in *Proc. ICCV*, Oct. 2007, pp. 1–8.
[57] L. S. Kennedy and M. Naaman, "Generating diverse and representative image search results for landmarks," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 297–306.
[58] 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. 2015.
[59] 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.
[60] 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.
[61] R. Hartley and A. Zisserman, "Multiple view geometry," *Encyclopedia Biometrics*, vol. 2, nos. 9–10, pp. 181–186, 2000.
[62] M. Freeman, *The Photographer's Mind: Creative Thinking for Better Digital Photos*. Library Journal, 2011.
[63] 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.
[64] 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.
[65] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. CVPR*, Jun. 2011, pp. 409–416.
[66] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. CVPR*, Jun. 2013, pp. 3166–3173.
[67] W. Wang, D. Cai, L. Wang, Q. Huang, X. Xu, and X. Li, "Synthesized computational aesthetic evaluation of photos," *Neurocomputing*, vol. 172, pp. 244–252, Jan. 2016.
[68] W.-Y. Chang and Y.-C. F. Wang, "Style-centric image summarization from photographic views of a city," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2016, pp. 2787–2791.
[69] 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.
[70] D. Zhang, J. Han, J. Han, and L. Shao, "Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1163–1176, Jun. 2016.
[71] D. Zhang, J. Han, L. Jiang, S. Ye, and X. Chang, "Revealing event saliency in unconstrained video collection," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1746–1758, Apr. 2017.
[72] C. Li, X. Qian, and G. Zhao, "POI summarization by combining aesthetics and diversity using 3D reconstruction," in *Proc. ICIP*, 2017, pp. 620–624.
[73] K. Lan and X. Qian, "Social image aesthetic measurement based on 3D reconstruction," in *Proc. ICIMCS*, 2014, p. 350.
[74] X. Qian, D. Lu, Y. Wang, L. Zhu, Y. 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.
[75] L. Stefan, M. Chli, and R. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proc. ICCV*, Nov. 2011, pp. 2548–2555.
[76] D. Weng, Y. Wang, M. Gong, D. Tao, H. Wei, and D. Huang, "DERF: Distinctive efficient robust features from the biological modeling of the P ganglion cells," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2287–2302, Aug. 2015.
[77] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. CVPR*, Jun. 2015, pp. 4353–4361.
[78] G. Ding, Y. Guo, J. Zhou, and Y. Gao, "Large-scale cross-modality search via collective matrix factorization hashing," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5427–5440, Nov. 2016.

Xueming Qian (M'10) 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 in 2008. He received outstanding doctoral dissertations of Xi'an Jiaotong University and Shaanxi Province, in 2010 and 2011, respectively. He received the Microsoft Fellowship in 2006. 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 the Director of the Smiles Laboratory, Xi'an Jiaotong University. His research interests include social media big data mining and search.

Cheng Li received the B.S. degree from Xi'an Jiaotong University, Xi'an, China, in 2014, where he received the M.S. degree from the School of Information and Communication Engineering in 2017.

Ke Lan received the B.S. degree from Xi'an Jiaotong University, Xi'an, China, in 2015.

Xingsong Hou received the Ph.D. degree from the School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, China, in 2005. From 1995 to 1997, he was an Engineer with the Xi'an Electronic Engineering Institute, where he was involved in the field of radar signal processing. He is currently a Professor with the School of Electronics and Information Engineering, Xi'an Jiaotong University. His research interests include video/image coding, wavelet analysis, sparse representation, sparse representation and compressive sensing, and radar signal processing.



Zhetao Li received the B.Eng. degree in electrical information engineering from Xiangtan University in 2002, the M.Eng. degree in pattern recognition and intelligent system from Beihang University in 2005, and the Ph.D. degree in computer application technology from Hunan University in 2010. He is currently a Professor with the College of Information Engineering, Xiangtan University. From 2013 to 2014, he was a Post-Doctoral Researcher in wireless network with Stony Brook University. From 2014 to 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 received the Second Prize of Fok Ying Tung Education Foundation Fourteenth Young Teachers Award in 2014.

Junwei Han received the B.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 1999 and 2003, respectively. He was a Research Fellow with Nanyang Technological University, The Chinese University of Hong Kong, Dublin City University, and the University of Dundee. He was a Visiting Student with Microsoft Research Asia and a Visiting Researcher with the University of Surrey. He is currently a Professor with Northwestern Polytechnical University. His research interests include computer vision and multimedia processing.