POI SUMMARIZATION BY COMBINING AESTHETICS AND DIVERSITY USING 3D RECONSTRUCTION

Cheng Li, Xueming Qian†, IEEE Member, Guoshuai Zhao
†Xi'an Jiaotong University
Emails: ChengLi3435@yeah.net, qianxm@mail.xjtu.edu.cn

ABSTRACT

Place-of-Interest (POI) summarization by aesthetics evaluation is significant in image retrieve. In this paper, we propose a system that summarizes a collection of POI images regarding aesthetics. First, we generate 3D models for POIs. Second, we select candidate photos in terms of the proposed crowd source saliency model based on the 3D to 2D projection relationship. Third, we propose a crowd-sourced salient map detection approach by exploring the distribution of salient region in the 3D model. Then, we measure the composition aesthetics of each image and we explore crowd source salient feature to get salient map. Finally, we combine the diversity and the aesthetics to recommend aesthetic pictures. Experimental results show that the proposed POI summarization approach is with diverse camera distribution and aesthetics.

Index Terms—landmark summarization, 3D reconstruction, salient map, aesthetic measurement, crowd source salient feature

1. INTRODUCTION

In recent years, huge number of photos of POI shared on internet by people during their travels. It is difficult to get a satisfactory result when a user intends to search a landmark. There are many irrelevant and unpleasing pictures in the top ranked results. Estimating aesthetics of images has been attracting much attention

Some kinds of information such as the geo-tags, views, comments, and faves can be exploited to assist to image summarization. Usually, the representative and beautiful photos can be viewed, commented and forwarded by world users frequently [1]. 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.

Jiang et al summarize landmarks by exploring high-frequency shooting locations based on the geo-tag information of photos posted to social network. The system realized by three steps: 1) Filtering landmark dataset from social media by combination of tags and geo-tags. 2) Mining high frequency shooting locations by geo-tag cluster. 3) Using visual feature verification to remove irrelevant images and rank images through intra and inter high frequency shooting locations by SIFT (scale and invariant feature transform) feature matching [2], [3].

Simon et al proposed an unsupervised method for finding canonical views to form the POI summary [4]. Their basic idea is that an image selected as a representative image is similar to many other images in image dataset. They construct a feature-image incidence matrix to represent the image set. Then they decompose images into groups. Qian et al modelled an image's viewpoint in horizontal, vertical, scale and orientation aspects, and then, they used a 4-D vector to construct the viewpoint for each image. They selected identical semantic points (ISPs) from the raw SIFT points of the images to capture major and unique parts of a landmark [5].

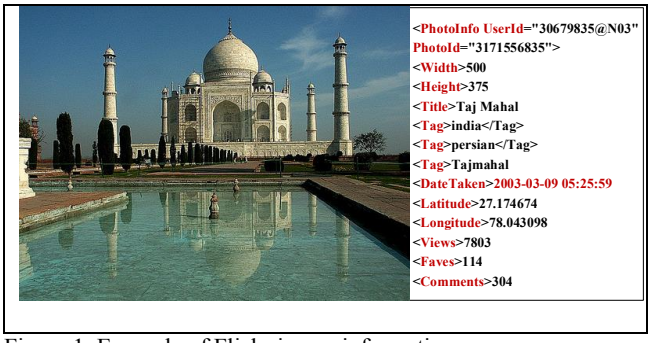

Figure 1. Example of Flickr image information.

Ren et al employ visual and views verification to select images from LOIs (location-of-interests) to summarize the POI [6]. 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. Lan at al proposed an optimal viewpoint selection system for robots to search for the optimal viewpoint [7]. They constructed evaluation functions based on certain known composition rules using three factors, namely, target size, visual balance, and composition fitting value. Then a score, which is a reflection of human evaluation, was obtained using these functions. The optimal viewpoint was selected from a number of candidates around the target group

Zhu et al propose a novel unsupervised hashing scheme, called topic hypergraph hashing, to address the limitations [11]. Firstly, relations between images and semantic topics are discovered. And then a unified topic hypergraph is constructed to model inherent high-order semantic

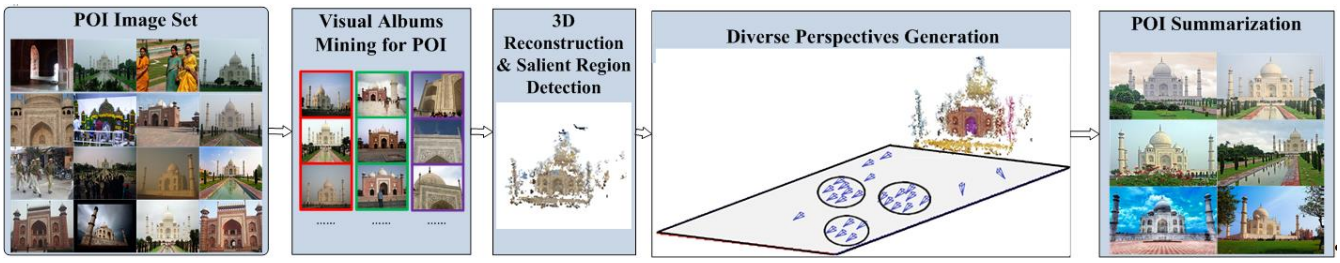

Figure 2. Flow chart of the proposed POI summarization with aesthetics evaluation from crowd-source social media

correlations of images. Finally, hashing codes and functions are learned. Zhu et al propose a novel unsupervised visual hashing approach called semantic-assisted visual hashing [12]. They propose a unified framework to learn hash codes by simultaneously preserving visual similarities of images, integrating the semantic assistance from auxiliary texts on modeling high-order relationships of inter-images, and characterizing the correlations between images and shared topics. Lei et al proposes a canonical view based compact visual representation to handle these problems via novel three-stage learning [13]. Xie et al propose Online Cross-modal Hashing which can effectively address the above two problems by learning the shared latent codes (SLC) [14] .Hash codes can be represented by the permanent SLC and dynamic transfer matrix.

From above, we find that existing methods carry out image summarization using image content clustering and exploring multimodality information of web images. In the content clustering-based approaches, representative images are selected in terms of image feature diversity. In our approach, we consider both the aesthetics of each images and diversity by exploring the crowd source social media information.

The rest of this paper is organized as follows: Section 2 describes the details of our approach. In Section 3, we give the experimental results. Finally, in Section 4, we make a conclusion on our system.

2. THE PROPOSED APPROACH

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 three steps: 1) Data Mining, 2) Salient Region Detection, 3) Image Ranking for POI Summarization. The following section will explain each part in detail.

2.1. Data Mining

As for a POI, there might exist several positions where photographers can take aesthetic images. Thus we utilize visual album generation approach to classify those miscellaneous images into different albums [5].

We utilize mean-shift algorithm to cluster images of a POI using GPS information to generate a set of coarse POI clusters is generated. Considering that images in a cluster are the same geographical location but with variance appearances, we propose a fine POI clustering approach to refine the coarse clustering results. It consists of the following three steps: a) Feature extraction; b) Similarity measurement: The similarity of two photos is measured by SIFT point matching [5]. 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 [5]. Actually, some albums are with very little number of images. These albums can be removed from final POI summarization.

Then we generate a 3D model for each POI album based on the SFM algorithm [8]. We get a group of 3D models, including their camera information C and geometric point's information G. 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 point, and a list of views the point is in. Camera information C contains includes 3-vector camera position, focal length F, 3-vector translation T, 3x3 matrix format of rotation R and the parameters of radial distortion k_1 and k_2 .

After above, this is effective to remove irrelevant images for POI summarization and mine useful data.

2.2. 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 in 3D space which are derived from the crowd-source social media information. If more people likely to take more photos at the region, then its "heat" level of it will be higher.

2.2.1. Salient region generation and representation

In order to eliminate some noise points in 3D models, an improved mean-shift algorithm is employed on the point cloud as follows:

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

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 points in $S_h(X)$, X denotes a cluster center in 3D space.

Each cluster corresponds to a region that containing a set of matched points. The regions with high saliency always indicate their attractiveness to people to take photos. 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:

$$D_r = \frac{1}{n}\sum_{i=0}^{n} f_i \quad (3)$$

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

$$f_i = N_i/N_a \quad (4)$$

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

We sort regions by D_r, and we remove the small regions which contain less points. In this paper, we keep the top ranked 80% of regions as salient regions to generate salient map and the rest 20% regions are removed as noise region. The kept regions are utilized in image aesthetic measurement for POI summarization.

2.2.2. Crowd source salient map detection

In order to get more accurate description of aesthetics, we utilize crowd source salient feature to select candidate images for POI summarization. And then we project all the salient regions in 3D into each 2D image to get the corresponding salient map.

The selected salient regions collaboratively describe photo local aesthetics. We select some representative pictures as candidate images from the total 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 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.

Following the process above, we can select images which contain more of salient region as candidate images, and the images containing less regions are removed.

2.2.3. Aesthetics Measurement

For candidate images, we propose an aesthetic measurement approach by exploring the distribution of salient regions. We build a computational model called dynamic balance which discloses distribution of salient region.

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

$$P_{sc} = \left(\frac{1}{N}\sum_i^N A_i x_i, \frac{1}{N}\sum_i^N A_i y_i\right) \quad (5)$$

where N is the number of points that mapped into an image; A_i is the weight of 2D point p_i=(x_i, y_i). In our experience, we use Dr of the salient region to represent A_i.

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

$$D_c = \left|P_{sc} - P_{pc}\right| \quad (6)$$

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

$$S_{DB} = 1 - 2arctan\frac{D_c}{\pi} \quad (7)$$

where S_{DB} ranges from 0 to 1. Larger S_{DB} means the better visual balance the salient regions in the picture.

2.3. 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. We recommend images from each visual album. Secondly, we combine aesthetics and diversity for summarization.

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

Images in the same album have much content overlap but they have different perspectives. Mean-shift algorithm is applied here to cluster cameras (the coordinates in 3D space) in a visual album into different perspectives.

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.

3. EXPERIMENTAL RESULTS

In order to evaluate the effectiveness of the proposed method, some comprehensive comparisons are made with existing approaches, including Canonical Views (denoted as CV) [4], Clustering, Ranking and Ranking (denoted as CRR) [9], Identical Semantic Points (denoted as ISP) [5], High Frequency Shooting Location (denoted as HFSL) [2] and Social-Contextual Constrained Geo-clustering (denoted as SCCG) [6].

3.1. Dataset

Experiments are conducted on the collected 7 million Flickr images uploaded by 7,387 users and the heterogeneous metadata 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.

3.2. Performances Evaluation

We utilized a user-driven approach that twenty volunteers were invited to evaluate the POI summarization performances.

We evaluate the results by assigning the aesthetics scores aes_i and diversity scores div for different summarization approaches. aes_i is aesthetics scores of i-th images, $aes_i \in \{0,1,2,3\}$, where the four discrete values are on behalf of inelegant, ordinary, good, and perfect. Let div denotes the diversity of the POI summarization result. We classify it into four categories: 3-excellent, 2-good, 1-normal, 0-irrelevant.

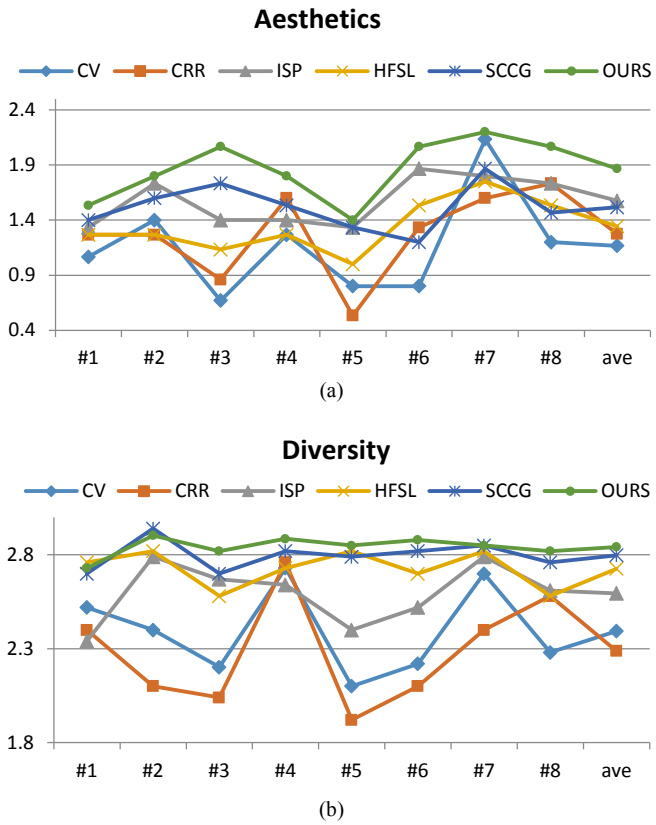

Figure 3. Average aesthetic scores and diverse scores CV, CRR, ISP, HFSL, SCCG and OURS on 8 POIs. The y-axis is the score and the x-axis is the POI index. (a) the average aesthetic scores. (b) the average diverse scores.

We utilize the average precision (AP) [10] for performance evaluation. Once we get the value of aes_i and div, the AP of the top-n images is determined as follows :

$$AP@n = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{i}\frac{aes_i}{i}\right) \quad (8)$$

Whether POI summarization is good or not is depend on whether it provides aesthetic images and make users fully u nderstand the POI.

3.3. Performances

The average aesthetic scores and diverse scores assigned by the 20 volunteers to eight POIs are shown in Fig.3 respectively. The average (denoted by ave) scores are also given in the last columns. From Fig.3, we find that ours method is with better aesthetics for all the eight POIs. The diversity of our approach is also outperforms the 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 need large enough accurate data. The others consider the representativeness of results, but it does not consider salient region using crowd source information.

From Fig.3 (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 makes sure the top ranked images selected from diverse perspectives.

4. CONCLUSIONS

In this paper, we propose 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 new method to build the salient map by calculating the frequency of points in 3D model appearing on the lens is proposed here. Crowd source salient feature is presented to gain more precise aesthetics evaluation which is 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.

REFERENCES

[1] D. Lu, X. Liu, X. Qian. "Tag based image search by social re-ranking". IEEE Transactions on Multimedia, 2016.

[2] S. Jiang et al., "Author Topic Model-based Collaborative Filtering for Personalized POI Recommendations", IEEE Trans. Multimedia, 2015, vol.17, no.6, pp.907-918..

[3] S. Jiang, X. Qian, "Generating representative images for landmark by discovering high frequency shooting locations from community-contributed photos," IEEE International Conference on Multimedia &Expo, pp.1-6, 2013.

[4] I. Simon, N. Snavely, and S. M. Seitz. "Scene Summarization

for Online Image Collections." IEEE International Conference on Computer Vision 2007:1-8.

[5] X. Qian, Y. Xue, X. Yang, Y.Y.Tang, X. Hou, and T. Mei, "Visual summarization of landmarks via viewpoint modeling," IEEE Trans. On Circuits and Systems for Video Technology, vol.25, no.11, 2015, pp.1857-1869.

[6] Y. Ren, X, Qian, S. Jiang, "Visual Summarization for Place-of-Interest by Social-Contextual Constrained Geo-clustering" IEEE, International Workshop on Multimedia Signal Processing IEEE, 2015.

[7] K. Lan, K. Sekiyama. "Autonomous Viewpoint Selection of Robot Based on Aesthetic Evaluation of a Scene" Journal of Artificial Intelligence and Soft Computing Research, 2016, 6(4).

[8] C. Wu, "Towards Linear-time Incremental Structure from Motion", 3DV 2013.

[9] L. Kennedy, M. Naaman, "Generating diverse and representative image search results for landmarks," in Proceedings of the 17th int. conf. on World Wide Web, pp. 297-306, 2008.

[10] X. Qian, D. Lu, and X. Liu. "Image Retrieval by User-oriented Ranking." Proceedings of the 5th ACM on International Conference on Multimedia Retrieval ACM, 2015.

[11] Z. Lei, J. Shen, X. Liang, et al. "Unsupervised Topic Hypergraph Hashing for Efficient Mobile Image Retrieval". IEEE Transactions on Cybernetics, 2016, PP(99):1-14.

[12] L. Zhu, J. Shen, L. Xie, et al. "Unsupervised Visual Hashing with Semantic Assistant for Content-Based Image Retrieval". IEEE Transactions on Knowledge & Data Engineering, 2017, 29(2):472-486.

[13] Z. Lei, J. Shen, X. Liu, et al. "Learning Compact Visual Representation with Canonical Views for Robust Mobile Landmark Search". Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, {IJCAI} 2016

[14] L. Xie, J. Shen, L. Zhu. "Online Cross-Modal Hashing for Web Image Retrieval". Proceedings of the Thirtieth {AAAI} Conference on Artificial Intelligence, February 12-17, 2016.