Generating Representative Images for Landmark by Discovering High Frequency Shooting Locations from Community-Contributed Photos

Shuhui Jiang, Xueming Qian†, *Member IEEE*, Yao Xue, Fan Li, Xingsong Hou,
School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an China.

ABSTRACT

Representative images generation offers a comprehensive knowledge for landmark and is a hot research area recent years. This paper presents a representative images generation system by discovering high frequency shooting locations from geo-tagged community-contributed photos. We discover that the views (e.g. far and near, front, back and side) of the photos taken in the same location are usually similar and but different in different shooting locations. Our system is realized by three steps: 1) Landmark dataset is filtered from social media by the combination of tags and geo-tags. 2) High frequency shooting locations are mined by geo-tag cluster. 3) Visual feature is then used for removing irrelevant images and ranking by intra and inter SIFT matching. This work is the first attempt to generate representative images by high frequency shooting locations mining. Evaluating on ten landmarks shows its effectiveness .

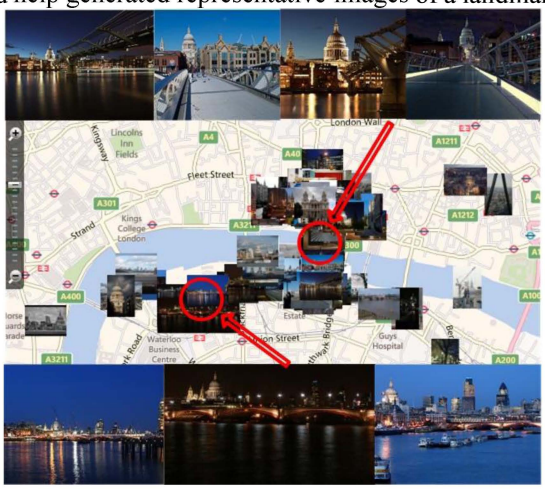
Index Terms—Geo-tagged photos, social media, Representative Image Selection

1. INTRODUCTION

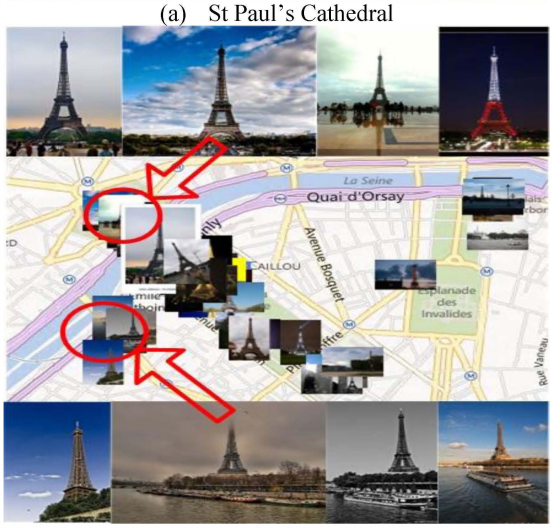
The birth of web 2.0 technologies brings volume of social media sites such as Flickr, Facebook, Twitter, and so on. Community – contributed photos are associated with a rich set of metadata such as timestamp, tags, GPS locations, etc. Recent years, it is a hot area to mining landmarks or events from social media [5]. Representative images generation is also interesting and emergent, which could offer a comprehensive knowledge for landmarks.

Most existing work on representative images generating view summarizing based on visual features [2-4],[6-7]. However, fewer efforts have been put into mining high-frequency shooting locations to discover representative views of landmarks.

As shown in Figure 1, we present the photos on the map according to their corresponding geo-tags, in other words, shooting locations. We could see that these photos do not stick to a small point but extend to an area. A landmark is Presented through different views (e.g. far and near, front, back and side) under different shooting locations. We discover that the views of the photos taken in the same location are usually similar but are different in different shooting locations. Finding frequency shooting locations could help generated representative images of a landmark.



(a) St Paul's Cathedral



(b) Eiffel Tower

Fig. 1. Two example landmarks photos' distribution on the map by their corresponding geo-tags. (a) St Paul's Cathedral. (b) Eiffel Tower. The photos at the same shooting location are similar. Two shooting location with zoomed photos are offered for each landmark.

†Corresponding author. This work is supported by NSFC No.60903121, No.61173109 and Microsoft Research Asia.

The two main challenges of this work are 1) to mine high-frequency shooting location from community-contributed photos and 2) to filter irrelevant images. Many efforts have done of mining landmarks with geo-tagged community-contributed photos [5] by mean-shift cluster, but less work could be referred of high frequency shooting locations mining. In this paper we discuss whether mean-shift cluster is also effective for shooting location mining. And as we know, the bandwidth should be given when clustering. We find the bandwidth is related to the geo-distance of photo distribution, so the bandwidth is always in a range.

To address the challenges of filtering irrelevant images, we use both tags and geo-tags when filter initial photos in the large dataset. And after finding higher frequency locations, the combination of intra and inter class SIFT matching could rank the intra cluster and inter cluster photos and move the irrelevant photos and even the irrelevant cluster.

Three main contributions in this paper are:

- We proposed a new way to generate representative images for landmark. Views are summarized by mining high frequency shooting locations from community- contributed photos.
- We offer an effective method to mine high- frequency shooting locations.
- We offer an effective way to move irrelevant images after finding views by intra and inter cluster local feature match.

The rest of the paper is organized as follows: Related works on representative images generation is introduced in Section 2. In Section 3, we describe the system by three steps. Experiment and discussion is shown in Section 4. In Section 5 the conclusions are drawn.

2. RELATED WORK

The main research efforts related to our work is representative images generation.

Most existing work focuses on view summarizing based on visual features[2],[4],[5-7]. Lyndon et al. cluster both global features and local features to find view point[2]. They use point-wise linking to detect whether two images belong to the same object. Xue and Qian [6] describe viewpoints in horizontal, vertical, scale and orientation aspects by modeling an image's viewpoint using a 4-D viewpoint vector. Identical Semantic Points are selected from SIFT points of the image to capture major and unique parts of a landmark. Zhao et al. make first attempt to generate representative views from scenic theme (e.g. sunny, night , view, snow) mining using Dirichlet Process Gaussian Mixture Model [7].

Community – contributed photos offers not only the visual feature. Rudinac et al. find representative image by modeling visual features, text associated with the photos as well as users and their social network using a multimodal graph. Four types of nodes constructed the graph are image, visual, text and user nodes.

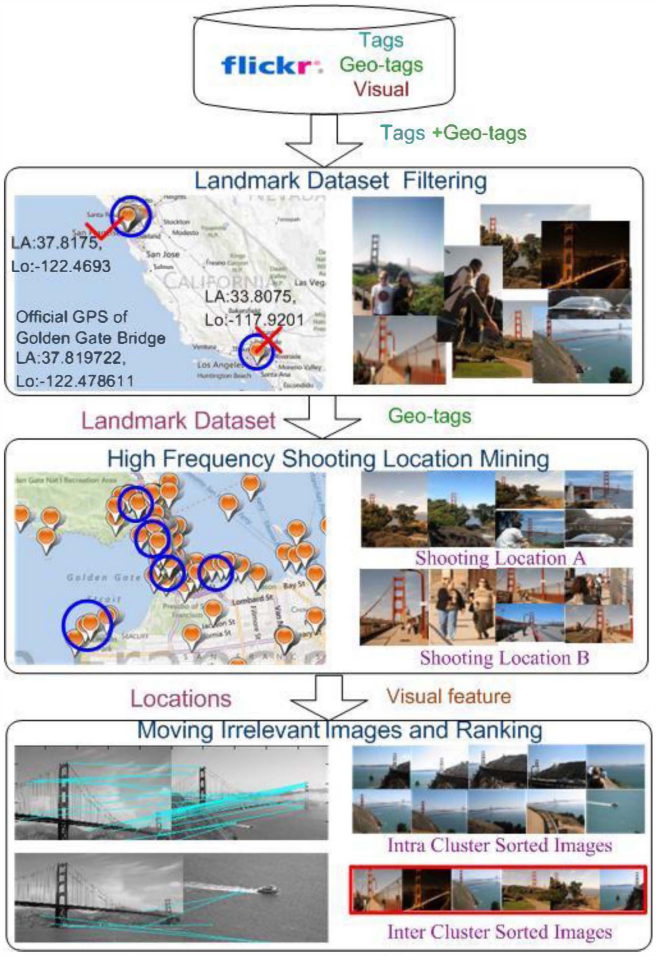


Fig. 2. The system overview. The figure presents the system by the example landmark, "Golden Gate Bridge". The input is community- contributed photos from Flickr, containing tags, visual and geo-tags. First, we use the combination of tags and geo-tags to filter initial photos to construct the Landmark Dataset. These photos are with the tag of landmark name and close to the location. Second, we cluster the geo-tags of the Landmark Dataset to mine high frequency shooting location. Photos in the same location are similar and various from location A to B. However there are still some noise photos. So third, we use intra and inter cluster sift matching to move irrelevant images and rank the photos. If the photos are of the same view, there are more blue lines of the match pairs. And the Inter Cluster Sorted Images are the final results.

Different to their work, we mine high-frequency shooting locations from geo-tagged community – contributed photos, and define the representative images of these locations as the representative images of the landmark.

3. APPROACH

In this section, we introduce the proposed approach detailed by three steps: 1) Landmark Dataset Filtering. 2) High

Frequency Shooting Location Mining. 3) Denoising and Ranking. The system overview is shown in Figure 2.

3.1 Landmark Dataset Filtering

Tags and geo-tags are complementary when filtering photos of a specific landmark from large data set. First, we match the tags of photos with the location name like "Golden Gate Bridge". These photos may include many irrelevant images.

Second, to solve the problem that different places may share the same name, we put the photos on the map by the corresponding geo-tags and circle the limitation of the specific landmark. As shown in the first step in Figure 2, there are two photo groups on the map whose tags are all include "golden gate bridge". We only choose the group whose GPS is closest to the official GPS as landmark dataset. It may still contain some noise, but content of most of the photos belong to the landmark.

3.2 High Frequency Shooting Location Mining

After getting Landmark Dataset, in this section, we introduce the approach to mine high frequency shooting locations by geo-tags. First, we cluster the geo-tags of the photos in the Landmark Dataset by **Mean-shift** cluster [1]. Mean-shift is the famous and effective method to mine landmark from GPS data and geo-tag data [5]. To $\forall X$, Mean shift is defined as follows:

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

where k is the number of observations falling within $S_h(X)$ region. As our geo-tag are 2-dimition, X is the shooting location of photo. $S_h(X)$ is the circle whose radius is h. So h is related to the geo-distance between photos. And, obviously, h is the "Bandwidth" in this paper.

When using Mean-shift, we need not to set the number of clusters like k-means but we have to set the bandwidth. It is difficult to define to best bandwidth, but we've already discuss the relationship between bandwidth and the geo-location of the photo distribution of the landmark. The bandwidth is always in a small range. If the bandwidth is too large, the circle of geo-distance is large. So photos of different views are in the same cluster. If the bandwidth is too small, the photos of the same view could not be cluster to the same cluster.

After clustering, not all the centers of clusters could present high frequency locations. In many clusters, there are only one or two photos. Then we only choose the clusters whose numbers of photos are higher than a threshold which is eight in this paper. The locations of these clusters are high frequency shooting location. We finally get H clusters C_1, …, C_H, the number of photos in C_i is N_i (i=1,2,…,H). Let X_l denote the l-th (l=1,2,…, N_i)photo in C_i.

Table 1. The number of high shooting locations in Different Bandwidth of "Big Ben" and "Gold Gate Bridge"

Bandwidth	Big Ben	Golden Gate Bridge
0.1	1	1
0.01	2	3
0.005	3	3
0.001	6	7
0.0008	8	7
0.0006	8	11
0.0005	9	11
0.0004	5	11
0.0001	2	7
0.00005	2	6

Because bandwidth is related to geo-distance on the map, and the geo-distance of the photos of a landmark is always in a certain range. Table 1 shows the number of high shooting locations in different bandwidth (BW) of "Big Ben" and "Tower Bridge". As shown in Table 1, 0.0005 is a best bandwidth to "Big Ben", because the number of views is largest. And to Tower Bridge, the best bandwidth is at 0.0004-0.0006. To different landmark, the best bandwidth is different, but always ranges from 0.001 to 0.0001. After experiment and evaluation, we offer a universal bandwidth 0.0005. For half of the landmarks, it is the best bandwidth. And for other landmarks, it is also acceptable.

3.3 Moving Irrelevant Images and Ranking

In this section we introduce the method to remove the irrelevant images and rank the photos of frequent shooting locations. The input of the section is the C_i (i=1,2,…,H) we get in Section 3.1.

Obviously, there are irrelevant photos in the clusters and even the situation that the whole cluster is irrelevant. We use **intra and inter cluster method** to remove the irrelevant images. **Intra cluster** step aims at ranking the photos in each cluster and find representative photo for the cluster. And **inter cluster** step aims at ranking the representative images of the cluster and generate the final representative images of the landmark.

We use visual similarity to determine whether these photos belong to the same view-scenic. The existing work uses both global visual feature (e.g. color and texture) and local visual feature (e,g. SIFT) to do the view summary [2]. In the paper, view summarization is carried out by mining high frequency shooting location. The scenic themes in the same shooting location could be different like day or night, sunshine and snow, summer or autumn, but the main building is the same

one. So, we use only local visual feature to do summarization. We represent the photos via local interest point descriptors given by SIFT. The feature of X_p is denoted by $X_p^f = \{S_{1,p}, S_{2,p}, ..., S_{np,p}\}$. np is the number of SIFT point in X_p.

3.3.1 Intra Cluster Moving Irrelevant Images and Ranking

Our method of intra cluster moving irrelevant images and ranking includes three main steps.

In the first step, we find match SIFT pairs using the basic method show in [2]. Given two images X_p and X_q in the C_i, each with a set of SIFT points, $X_p^f = \{S_{1,p}, S_{2,p}, ..., S_{np,p}\}$, and $X_q^f = \{S_{1,q}, S_{2,q}, ..., S_{nq,q}\}$. The Nearest Matching Point to two images is considered a match only if the Euclidean distance between the two descriptors is less than the distance between the first descriptor and all other points in the second image by a given threshold. To ensure symmetry, the match pair is the nearest of the first image against the second and the nearest of the second image against the first. We defined the number of match pairs of X_p and X_q as m_{pq}.

In the second step, we calculate the number of Match Pairs of one photo towards all the other photos and sum Match Pairs. Matrix M_i contains the number of match pairs to each two photos of C_i.

$$M_i = \begin{bmatrix} - & m_{12} & m_{1p} & \cdots & m_{1q} & m_{1N_i} \\ m_{21} & - & m_{2p} & \cdots & m_{2q} & m_{2N_i} \\ m_{p1} & m_{p2} & - & \cdots & m_{pq} & m_{pN_i} \\ \cdots & \cdots & \cdots & - & \cdots & \cdots \\ m_{q1} & m_{q2} & m_{qp} & \cdots & - & m_{qN_i} \\ m_{N_i1} & m_{N_i2} & m_{N_ip} & \cdots & m_{N_iq} & - \end{bmatrix}. \quad (2)$$

To X_p, the number of the matched pair of all the other photos is defined as m_p as follows:

$$m_p = \sum_{k \in \{1,\cdots,N_i\}, k \neq p} m_{pk} \quad (3)$$

Different to [2], we directly calculate the number of match pairs. In [2], they first calculate whether the Match Pairs of two images exceeds a threshold, which equals to three in their paper. Then they calculate the number of photos satisfy the requirement. In our paper, as the photos in one cluster are shot at the same location, the number is large enough and various from one scenic to the other. So we just calculate the number of Match Pairs avoiding setting the threshold.

Third we rank the photos according to the m_p $(p = 1, 2, ..., H)$. The fundamental idea of our method is that, the most appeared scenic in the cluster could represent the cluster, and the photos not similar to other photos are more likely to be irrelevant photos. And at the same time, a photo which could find more matching pairs is more likely to have more similar photos. Then we got the ranked cluster which is defined as $C_i^{'} = \{X_1^{'}, ..., X_{N_i}^{'}\}$. $X_1^{'}$ is also present as R_i, the representative image of C_i. Then the representative image of all the cluster is defined as $\{R_1, ...R_i, ...R_{N_i}\}$.

We show the example of intra cluster ranked results for landmarks: Tower Bridge and Big Ben in Fig. 3 (a) and Fig.3 (b) respectively. We find that the top-ranked photos are similar. Although the first photo may not be the best on subjective judgments, it could represent most of the photos in the cluster. In Fig.3 (b), we offer both relevant and irrelevant cluster. Whether the cluster itself is relevant or not, the top ranked photo could represent the cluster. The top ranked photos from each of cluster is the candidate set for us to find the final representative images of the landmark. The next section we'll introduce how to solve the influence of irrelevant clusters.

(a) an example ranked cluster of Tower Bridge

Relevant Cluster

Irrelevant Cluster

(b) two example ranked cluster of Big Ben

Fig. 3. Two examples ranked intra cluster. (a) Tower Bridge. (b) Big Ben. In (a), we offer 8 photos of the cluster in left-to-right and top-to-bottom order. In (b), we offer both relevant and irreverent cluster.

3.3.2 Inter Cluster Moving Irrelevant Images and Ranking

After getting representative images of the frequent shooting locations $R = \{R_1, ...R_i, ...R_{N_i}\}$ in intra cluster step, then we introduce the inter cluster step. We know that if all the photos or most of the photos are irrelevant of the landmark, the representative photo of the cluster is irrelevant of the landmark. As shown in Figure 3, a whole cluster "singing girl" is irrelevant. If we do not rank the representative

images of the clusters, the "singing girl" would appear in the final result of the system.

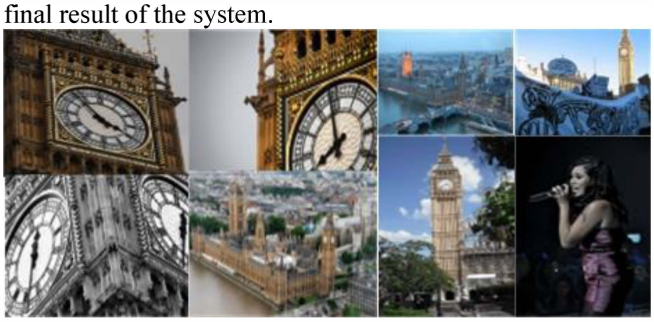


Fig. 4. Example ranked inter cluster of Big Ben. We could see the top 7 photos are relevant to the landmark and the photo irrelevant is at the end of array. And the 8 photos are given in left-to-right and top-to-bottom order.

To move the representative image of irrelevant clusters, we do the inter cluster ranking towards R. In the inter cluster ranking, visual feature helps to determine whether the photo belongs to the landmark. The fundamental idea is that most of the clusters are relevant to the landmark because they the high-frequency locations we mined. In most landmarks, only small numbers of clusters are irrelevant. The ranked R is defined as $R' = \{R'_1, ...R'_j, ...R_{N_c}\}$. So photos at the top of R' belong to the landmark and photos at the end of R' are more likely to be irrelevant.

We also use three-step SIFT match to rank the inter cluster photos. Though candidate set may contain far and near, front, back and side view, the scale-invariance, rotation - invariance and translation-invariance SIFT descriptor could solve the problem . As the method is similar to the method in section 3.3.1, we do not introduce it again.

Figure 4 shows the inter cluster ranked results of the landmark. The image of "singing girl" is in the end of the array. And the top seven photos are all belongs to the landmark.

4. EXPERIMENT

In this section, we first introduce the dataset and evaluation criterion. Then we show the evaluation of the proposed framework on ten landmarks.

4.1 Dataset and Evaluation

We collected 7 million Flickr images uploaded by 7,387 users and the heterogeneous metadata associated with the images with Flickr API.

We choose ten landmarks to evaluate our method. They are: 1)Tower Bridge, 2) St paul's Cathydral, 3) Eiffel Tower, 4) Angkor, 5) Big Ben, 6) Cologne Cathedral ,7) Colosseum, 8) Golden Gate Bridge, 9) Statue of Liberty and 10) Taj Mahal.

We invite 10 volunteers to evaluate top 8 targeted results for each of the 10 landmark. The volunteers have study the landmarks before and familiar with the landmarks. We use two-aspect criteria to evaluate the representative images for each of the ten evaluated landmarks [2]. The criteria of "representative" is also in that paper, but after "noise reduction" step to both our methods and the comparative method, almost all the images are related to the location. We mainly focus on these two aspects:

Unique. Whether the photos are view diversity? Are there many redundancies (0-10)?"

Comprehensive. Does this set of results offer a comprehensive view of the landmark (0-10)?

4.2 Evaluation with Repetitive Images

We compare our method using High Frequency Shooting Locations (denoted as HFSL) with the using identical semantic point based representative image selection approach (denoted as ISP) [6]. We show the performance by the criteria of "Unique" and "Comprehensive" in Figure 5. The top ranked representative images of the two methods are shown in Figure 6.

As we have showed some landmarks in the paper like "Big Ben", in Figure 6, we show 6 landmarks with each 5 images. Both HFSL and ISP could offer representative images, but there are differences between the two methods. The 5 results in landmark (2) St Paul's Cathedral of HFSL are shooting at five different shooting places, so the views are more diversity than the results of ISP. ISP failed to find some views shown in HFSL, but the roof of cathedral is with different viewpoints. In the landmark (6) and (10), HFSL successfully find different shooting locations, and ISP still lacks some views. ISP shows some different viewpoints of the same door (10) and window (6).

As there many redundancies in ISP, to most landmarks, performance of "Unique" and "Comprehensive" criteria are lower than HFSL. As show in Fig. 5 (b)and (c), in landmark (1) (8), both ISP and HFSL perform very well and the scores are all higher than 9.5. Different views of bridge are discovered, and there are less redundant images. To landmarks (2), (3), (5), (6) and (10), the performance of HFSL is 10 % higher than ISP. This is in accordance with the visual evaluation. And to Landmarks (4) and (7), both HFSL and ISP are not performing well. To (4), there are too "focuses" (interest things) in the landmark, instead of a main building like other landmark. Even in the same shooting place, the "focus" could be totally different. But the views HFSL discovered are more comprehensive of the landmark like "smiling face of Khmer" and "the Angkor Wat temple". To (7) "Colosseum", due to the column-shape of the landmark itself, the images shooting at different location looks similar. And to (9), the performance of ISP is 10% higher than HFSL, because that ISP is good at finding tiny

diversity of front-side and bottom-top viewpoints. And HFSL does not have this function.

Analyzing the causes, in [6], they mainly test their method on the Oxford building image set (Oxbuild). Different to Community–Contributed photo set, first, Oxbuild has less irrelevant images. Second, the shooting locations are almost the same to a building, but the viewpoints (e.g. front-side viewpoint, bottom-top viewpoint, close-distant viewpoint, etc.) are little different. So their method is good to find view points to the building at a same shooting location and our work is good at finding the views at different shooting locations.

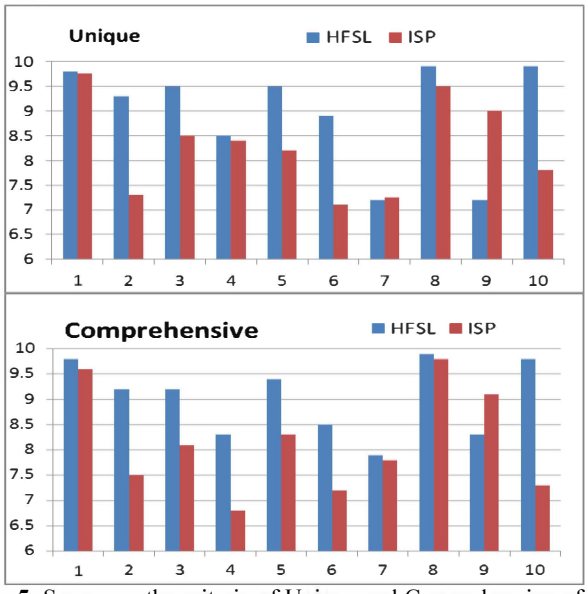


Fig. 5. Scores on the criteria of Unique and Comprehensive of the HFSL and ISP on the 10 Landmarks.

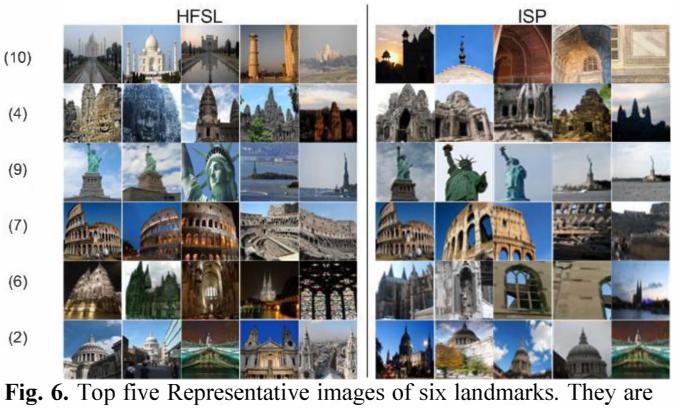


Fig. 6. Top five Representative images of six landmarks. They are (10) Taj Mahal, (4) Angkor, (9) Statue of Liberty , (7) Colosseum, (6) Cologne Cathedral and (2) St paul's Cathydral.

We analyze the performance of our method is related to the type of landmark. The scores of (1)(2)(3)(5)(8)are high, because there are sea, lake, mountain etc. around the main building. So photos present quite different views at different shooting locations. However, (6) Pisa Leaning Tower, 7) Colosseum are column structure. They look similar in different shooting locations. So the score of them are lower. The score of 9) Statue of Liberty is also not high, because most people take photos under and near the stature. The front or side view could not be distinguished by the cluster of location. However, we offer different views of the statue itself and the statue with the sea.

5. CONCLUSIONS

In this paper, we propose a new method to generate representative images from mining high-frequency geo-tagged community-contributed photosets. The advantages of our work are 1) We take advantage of geo-tags of photos and the computing time of two-dimension data is far less than only using visual features as most existing works. 2) We propose the method to mine high-frequency shooting locations of landmarks. 3) As there are many irrelevant photos in the community-contributed photosets, we use both intra cluster and inter cluster, and the final reprehensive photos are relevant to the location.

The disadvantage of our work is that for some types of landmarks, the method is not effective enough. And the intra cluster ranking could not find the best photos of the cluster.

Our future work is to add more landmarks for the evaluation. Landmarks not famous enough is also our focus. The ranking algorithm should be optimized to fine best photos of the cluster to raise the performance of final result.

6. REFERENCES

[1] K. Fukunaga, L.Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," IEEE Transactions on Information Theory, vol.21(1),pp.31-40,1975.

[2] L.Kennedy, M. Naaman, "Generating diverse and representative image search results for landmarks," WWW. pp. 297-306, 2008.

[3] S. Rudinac, A. Hanjalic, and M. Larson, "Finding representative and diverse community contributed images to create visual summaries of geographic areas", Proceedings of the 19th ACM international conference on Multimedia, pp.1109-1112, 2011.

[4] W. Chen, A. Battestini, N. Gelfand, and Setlur, V, "Visual summaries of popular landmarks from community photo collections", Signals, Systems and Computers, 2009 , pp. 1248-1255.

[5] X. Lu, C. Wang, J. Yang, Y. Pang and L. Zhang, "Photo2Trip: Generating Travel Routes from Geo-Tagged Photos for Trip Planning," ACM MM,2010.

[6] Y. Xue, X. Qian,"Visual Summarization of Landmarks via Viewpoint Modeling", in Proc. ICIP, 2012.

[7] Y. Zhao, Y. Zhang, X. Zhou, and T. Chua, "Generating Representative Views of Landmarks via Scenic Theme Detection," MMM. pp.392-402,2011.