Enhanced 3D Modeling for Landmark Image Classification

Xian Xiao, Changsheng Xu, Senior Member, IEEE, Jinqiao Wang, Member, IEEE and Min Xu, Member, IEEE

Abstract—Landmark image classification is a challenging task due to the various circumstances, e.g. illumination, viewpoint, zoom in/out and occlusion under which landmark images are taken. Most existing approaches utilize features extracted from the whole image including both landmark and non-landmark areas. However, non-landmark areas introduce redundant and noisy information. In this paper, we propose a novel approach to improve landmark image classification consisting of three steps. First, an attention-based 3D reconstruction method is proposed to reconstruct sparse 3D landmark models. Second, the sparse 3D models are projected onto iconic images in order to identify images of the hot regions. For a landmark, hot regions are parts of a landmark which attract photographers’ attention and are popularly captured in photos. These hot region images are later used to enhance reconstructed sparse 3D models. Third, the landmark regions are obtained through mapping the enhanced 3D models to landmark images. A k-dimensional tree (kd-tree) is then constructed for each landmark based on scale invariant feature transform (SIFT) features [5] extracted from the landmark area to classify unlabeled images into pre-defined landmark categories. The proposed method is evaluated using 291,661 images of 51 landmarks. Experiments of comparison indicate that our method outperforms Bag-of-Words (BoW) based approach [25] 18.5% and method of Spatial-Pyramid-Matching using Sparse-Coding (ScSPM) [3] 8.4%.

Index Terms—Landmark Image Classification, Attention Analysis, Attention Based 3D Reconstruction, 3D model enhancement

I. INTRODUCTION

MORE and more enormous sightseeing pictures are uploaded and spread with the proliferation of photo-sharing websites such as Facebook and Flickr. Among these sightseeing pictures, landmark pictures are one of the most attractive contents for users. Some landmark images are uploaded together with assigned tags. However, those assigned tags might not be accurate. Moreover, much more images are uploaded without any tags. Web users need to spend a lot of time to find landmark images of their interests. Properly classifying and labeling landmark images helps users easily search for their interested landmarks. In different images, a landmark could be presented in different styles due to various circumstances of illumination, viewpoint, zoom in/out and occlusion, when photographing (shown in Fig. 1). Traditional image classification approaches of similarity matching cannot successfully classify landmark images with different styles. Therefore, to correctly classify landmark images is a challenging task.

Although existing approaches [2-4] [15-16] achieved acceptable classification accuracy for landmark images, features extracted from non-landmark region brought much noisy and redundant information to classification. In order to handle the problems described above and improve existing work, it is vital to identify the landmark region and combine information from various presentation styles of a landmark. Recently, popular interesting region detection methods are based on image segmentation [24] and visual attention detection [19-22]. Although image segmentation approaches [24] can generate satisfactory segmentation results, it is still difficult to determine which region is a landmark region. Moreover, the region of interest (ROI) resulted by visual attention analysis always contains regions that attract people’s attention but not belongs to any parts of a landmark (e.g. people in front of the landmark, trees on the side of the landmark, sky behind the landmark, etc). Moreover, the detected boundaries of the saliency regions are also not very accurate. Therefore, we introduced a 3D model in [14]. Firstly, a 3D landmark model is reconstructed utilizing iconic landmark images. Then, landmark regions are identified by projecting the 3D model back to the iconic images. Finally, a kd-tree for each landmark is constructed with SIFT features extracted from the landmark regions to classify the unlabeled landmark images. Our approach [14] has two advantages. Firstly, it avoids
mis-classification introduced by noisy and redundant information. Secondly, it combines information related to the same landmark in various presentation styles for consideration.

The framework of our approach is illustrated in Fig. 2, consisting of three steps, where Step 1), Step 2) and KD-tree construction in Step 3) are training parts and classification in Step 3) is testing part.

1) Iconic image selection and attention based 3D reconstruction: for each landmark class, we applied \textit{k-means} with the global descriptor \textit{GIST} [26] on the landmark images. Landmark images are clustered into different clusters. A group of images, which are close to the cluster center, are selected as iconic image candidates. Then, visual attention region is detected for each image candidate. Geometry constraints are applied on visual attention region in order to select the iconic images.

The visual attention regions of the selected iconic images are further used to reconstruct a sparse 3D model of each landmark with the structure-from-motion method [6].

2) 3D model enhancement: for each landmark class, the constructed 3D model is projected to the corresponding iconic images in order to identify the hot region images. By calculating the similarity between hot region images and its corresponding cluster members, a number of hot region images are selected to enhance the sparse 3D model. Same 3D model reconstruction method as in Step 1) is applied here to add selected hot region images into reconstruction to enhance the 3D model. Finally, the enhanced 3D model is projected onto all the images used for 3D reconstruction and enhancement to identify the landmark region.
3) *k*-dimensional tree (*kd*-tree) construction and unlabeled landmark image classification: a *kd*-tree for each landmark is constructed based on *SIFT* features extracted from the landmark regions to classify unlabeled images into pre-defined landmark categories.

This work improves our previous work [14] by the following three points: 1) compared with our previous method, increased hot region images are selected to enhance 3D model. Normally, hot regions are components of a landmark which are photographed most often. By this way, the hot regions can be used to enhance the 3D model. 2) 3D model reconstruction is a time-consuming task. Improving the previous work, we utilize visual attention region on the iconic images to reconstruct 3D model in order to reduce the computational cost. 3) In this research, we significantly enlarge experimental data set ten times compared with our previous work.

In this paper, we propose a framework for landmark image classification utilizing enhanced 3D model which is able to improve the landmark image classification accuracy. Compared with the existing approaches, the contributions of our work are summarized as follows:

1) The proposed framework avoids mis-classification introduced by noisy and redundant information and combines information related to the same landmark in various presentation styles.
2) An attention based landmark 3D reconstruction is proposed to reduce computational cost in 3D reconstruction.
3) Through 3D to 2D projection, the landmark region is detected accurately.

The rest of the paper is organized as follows. The related work is reviewed in Sec II. The details of attention based landmark 3D reconstruction, enhancement, and image classification are described in Sec III, IV and V respectively. Experimental results are reported in Sec VI. We conclude the paper with future work in Sec VII.

II. RELATED WORK

Image classification has been extensively studied [1] recently. Compared with general image classification, landmark image classification is a challenging task due to uniqueness of the landmark and various presentation styles of the same landmark. Existing work on landmark image classification can be summarized into four categories: (1) Bag-of-Word (*BoW*) based method [2][15], (2) Spatial Pyramid Matching (*SPM*) based method [3], (3) iconic graph based method [4] and (4) Geo tag based method [6]. Since most of the previous research utilized the entire local features (e.g. *SIFT* [5]) or a global feature to train classifiers, the classification accuracy is limited due to the involved noisy and redundant information. Our previous research [14] utilized 3D models to classify the landmark images and improve the classification accuracy. However, the 3D reconstruction process wastes much computational power on matching the features located on the landmark region with the features located on the non-landmark regions. Moreover, *hot region* attracts more attention from photographers. *Hot region* images occupy large portion of the landmark image collections. The previous model treats each region in landmark equally without giving extra consideration to *hot region*.

Generally, 3D reconstruction approaches can be classified into two categories: sparse 3D reconstruction [6][17] and dense 3D reconstruction[18]. Both kinds of approaches need to know the camera parameters. The previous approaches always utilized local features (e.g. *SIFT* [5]) extracted from whole images to estimate the camera parameters and then generated 3D model of a whole scene. Our previous work [12] analyzed the spatial-temporal attention to obtain the region of interest (ROI) in video sequence and then reconstructed 3D model of ROI. 3D model reconstruction on ROI within an image instead of the whole image saved the computational cost. However, different from our previous work, we are not able to utilize the temporal attention because the landmark images are almost independent between each other.

In order to identify the region of interest (ROI), visual attention has been studied and widely used in computer vision, artificial intelligence and multimedia processing. Most of the pioneer work which studied the static information on the still images [19] can be classified into contrast based method and information theory based method according to the physiological basis. Contrast based attention analysis [19][20] takes the notion that the center-surround structure of receptive field provides human visual system (HVS) sensitivity to feature contrast. Information theory based methods [21] adopt the premise that visual attention proceeds entirely by maximizing the information sampled from an image. Contrast and information sampling are two factors used to evaluate saliency in computational visual attention. In [22], the contrast and information are combined to obtain visual attention. Motivated by [22], we integrate the contrast and information to calculate saliency map.

Image segmentation [24] and visual attention analysis [19-22] can be utilized to identify the landmark regions. Although some image segmentation approaches [24] are able to generate satisfactory segmentation results, it is difficult to determine which region is the landmark region and the landmark may be segmented into several separate regions. Moreover, visual attention analysis can obtain the visual attention regions which are not always overlapped with the landmark region.

Iconic images/Scene summarization for landmark image collections has been discussed by Simon et al.[23], where images are clustered based on exhaustive pair wise feature matching. However, such summary generation approach tried to present the most interesting and important aspects of the scene with minimal redundancy while considered nothing for the 3D reconstruction work.

III. ATTENTION BASED LANDMARK 3D RECONSTRUCTION

The traditional image classification methods may not appropriate for landmark classification due to the problems of various presentation styles of each landmark, noise from non-landmark regions and redundant information. In our previous work [14], the structure-from-motion (SFM) method
[6] is utilized to reconstruct 3D landmark models and handle the problems above. Although our previous work is able to improve the classification result, 3D landmark model reconstruction wasted much computational power. Our previous work [12] introduced a spatial-temporal attention based on static, location and motion analysis of video sequence to find a visual attention region within a video frame. In order to reduce the computational cost, we only consider the visual attention region for 3D reconstruction. However, different from the video sequence analyzed in our previous method [12], the landmark images in this work do not have temporal information.

Many collections of landmark images consist of a small number of popular viewpoints, from which photos are taken. The images taken from same popular viewpoints always have the similar geometrical characteristic. On the other hand, Snavely et al. [7] has found that reconstruction using several representative images from the collection is able to provide a good approximation to a reconstruction using all the images. Therefore, we reconstruct a 3D model for each landmark with a group of representative landmark images from different popular viewpoints.

Our landmark images are collected from photo-sharing websites (Facebook and Flickr). About 1/10 of the images are attached with camera focal length. It is well known that the structure-from-motion (SFM) [6] method may make more accurate 3D reconstruction results when the camera focal length is known [6]. Therefore, we select the representative images from these landmark images with camera focal length to reconstruct 3D landmark model.

In this paper, the attention based landmark 3D reconstruction consists of five modules: 1) GIST clustering, 2) iconic image candidate selection, 3) attention analysis (static), 4) geometric constraints for iconic images selection and 5) attention based 3D reconstruction.

A. GIST Clustering

Our goal is to represent the landmark collections by identifying a set of iconic views corresponding to dominant aspects in a 3D scene. If there are many frames belonging to very similar viewpoints, some of them will at least have a similar image appearance, which can be efficiently matched using a low-dimensional global description of their pixel patterns.

We adopt the global GIST feature which is impactful for grouping images by the perceptual similarity [10] to represent the image content and utilize k-means algorithm with \( k = 100 \) to cluster images. In particular, we can expect images with very similar viewpoints to end up in different GIST clusters because of the different circumstances (illumination, viewpoint, zoom in/out, occlusion, etc) when they are taken. This will not cause a problem for our approach, because our 3D model is able to combine different representation styles of the same landmark.

B. Iconic Image Candidate Selection

In order to select a group of iconic images which are efficient for 3D reconstruction, within each cluster, a predetermined rate of the most representative images, whose GIST descriptors are close to the cluster center, are selected as its iconic image candidates. At least one image is selected for each cluster. The final group of iconic images comes from the iconic image candidates. We calculate the rate as follows:

\[
\eta = \frac{k}{n} 
\]

where \( \eta \) is the rate, \( n \) is the total number of the images in the landmark image collection and \( k \) is the number of clusters.

For each cluster, we calculate the number of selected images from each cluster as follows:

\[
S_i = \left\lceil n_i \times \eta \right\rceil 
\]

where \( S_i \) is the number of selected images for the \( i \)th cluster, \( n_i \) is the total number of the images in the \( i \)th cluster. The larger the cluster is the more iconic image candidates are selected from this cluster. The large clusters typically contain more diverse information than small clusters.

Therefore, the iconic image candidates can be denoted as follows:

\[
C = \{C_{ij} \mid i = 1, 2, \ldots, k, j = 1, 2, \ldots, S_i \} 
\]

where \( C \) is the iconic image candidates, \( C_{ij} \) is the \( j \)th image selected from the \( i \)th cluster, \( k \) is the number of the cluster. Iconic images are later selected from the above candidates by the following process.

C. Attention Analysis

In our research, we assume that iconic images are taken by different persons. Therefore, we only consider static attention...
as there is no any temporal information in iconic image candidates. Through attention analysis, we can obtain a saliency map which describes the visual attention region in a landmark image. There are two important factors which are utilized to evaluate saliency in computational visual attention. One is the contrast of the region compared with its neighborhood; the other is the information importance of the region in a global image. Motivated by [22], we integrate the contrast and information importance to calculate saliency map as follows:

\[ Saliency(x, y) = Con(x, y) \times ID(x, y) \]  

where \( Saliency(x, y) \) is the attention analysis result of \( point(x, y) \) in the landmark images, \( Con(x, y) \) and \( ID(x, y) \) are contrast and information density of \( point(x, y) \) and normalized to \([0, 1]\). In our experiments, we calculate the mean value of \( Saliency(x, y) \) of a whole image and define the region consisting of points whose values of \( Saliency(x, y) \) are larger than the mean value to be attention region. There are some attention analysis results shown in Fig. 3. Although most of the attention regions are located on the landmark, the boundaries are not very accurate.

D. Iconic Image Selection

In order to select a group of representative landmark images from the iconic image candidates, we introduce iconic image group and geometric constraints. One iconic image candidate randomly selected from each cluster forms an iconic image group. Therefore, the iconic image groups are denoted as follows:

\[ Ig_n = \{ Ig_{ni} | i = 1, 2, ... k \} \]
\[ n = 1, ..., \prod S_i \ (i = 1, 2, ..., k) \]  

where \( Ig_n \) is the \( nth \) iconic image group, \( Ig_{ni} \) is the image that belongs to the \( nth \) group and is selected from the \( ith \) cluster, \( \prod S_i \) is the number of the iconic image groups, \( k \) is the number of the cluster.

To select the best iconic image group for 3D reconstruction, we sort all iconic image groups using geometric constraints. The geometric constraints perform verification of each iconic image group to confirm whether the iconic images in each group share a common 3D structure. We extract SIFT features [5] from the visual attention region of each iconic image, use RANSAC algorithm [8] to estimate a fundamental matrix and obtain the accurate SIFT matches between different iconic images in the same iconic image group. For an iconic image group, each image has a number of inliers to the others. The sum of inliers of an iconic image group is named as a geometric constraint score. The score is calculated as follows:

\[ GCScore_n = \sum_{s \neq t}^{k} \text{Inlier}_{st}^n \]  

where \( GCScore_n \) is the score of the \( nth \) iconic image group, \( \text{Inlier}_{st}^n \) is the number of SIFT matches between the \( sth \) and the \( ith \) image in the \( nth \) iconic image group.

We rank the iconic image groups with the geometric constraint scores \( GCScore_n \) by descending order and the first group with the highest score is selected as the best iconic image group. If several iconic image groups have the same max geometric constraint score, any of them can be selected as the iconic image group.

E. Attention Based 3D Reconstruction

The attention-based 3D reconstruction only considers visual attention regions and consists of four steps: 1) estimate the camera parameters of two initial pair of images; 2) estimate the camera parameters of a newly added image; 3) add points to 3D model; 4) repeat the step 2 and step 3) until the camera parameters have been estimated for all images.

Step 1: We start with estimating the camera parameters of an initial pair of iconic images. In order to obtain an accurate 3D reconstruction model of the initial two images, the initial two images should have a large number of matching points. Therefore, we choose two images with the largest number of correspondence as the initial two images. The numbers of correspondence between different iconic images have been obtained in Sec III.D by using features extracted only from visual attention regions. The camera parameters for the initial pair are estimated using the five point algorithm [7] and then the points visible in the two images are triangulated. Bundle adjustment is always invariably used as the last step to obtain optimal 3D structure and camera parameter estimation. Therefore, we apply a two-image sparse bundle adjustment [13] to refine estimated camera parameters in order to complete the reconstruction by the initial pair.

Step 2: We continue estimating the camera parameters of a newly added iconic image. A newly added iconic image is selected as the image with the largest number of correspondence to the images used for camera parameter estimation. Same as step 1, only visual attention regions have been considered. Camera parameters are initialized using the direct linear transform (DLT) method [9] inside a RANSAC [8] procedure. The DLT approach returns an upper-triangular matrix \( K \) which can be used as an initial estimation of camera intrinsic parameters. A sparse bundle adjustment [13] is further used to optimize camera parameters.

Step 3: We add points observed by the newly added iconic
image into the camera parameter optimization. A point which is observed by at least one other added iconic image is added if triangulating it gives a well-conditioned estimate of its location. The condition of all pairs of rays which could be used to triangulate this point is estimated to obtain the pair of rays with the maximum angle of separation. If this maximum angle is larger than a threshold (we use 2.0 degrees in our experiments), then the point is triangulated. Once the new points have been added, we run a global bundle adjustment [13] to refine the entire model and find the minimum error solution.

Step 4: The 3D reconstruction is completed once all iconic images have been added and the optimized camera parameters have been obtained. Iconic images with less than five correspondences are ignored for 3D reconstruction.

After reconstruction, the generated 3D model consists of 3D points corresponding to those matching points among iconic images.

IV. 3D MODEL ENHANCEMENT

In our previous work [14], we projected the 3D model back to the corresponding iconic images to obtain landmark region in each image. We utilized SIFT features extracted from the landmark regions to construct a *kd-tree* for each landmark. The *kd-tree* is later used to classify the unlabeled landmark images. Although this approach improved the accuracy of landmark image classification, it did not consider the hot region (see Fig. 4), which is a significant part of the landmark image.

Since hot regions attract much attention from photographers, hot region images occupy a large portion of landmark image collections. 3D point distribution on a 3D model constructed by using the hot region is significantly uneven. The 3D points within a hot region are much more than in other regions. For various landmarks, the hot region occupies different location within the landmark. Therefore, the distribution of 3D points somehow contributes to landmark image classification.

We consider applying hot region images to enhance the existing 3D model. Firstly, we need to find iconic images which are hot region images. Constructed 3D model is projected back to the iconic images which are used to reconstruct sparse 3D model according to the corresponding projection matrices. Many points will be projected out of an image, if the image is a hot region image. By this way, we are able to identify hot region images from the iconic image group easily. In Sec III.A, we have already obtained 100 clusters for each landmark. The cluster whose iconic image is identified as hot region image is regarded as hot region cluster. Some hot region images are selected from hot region cluster with high similarity to hot region iconic images. Finally, visual attention region of the selected hot region images are added to reconstructed 3D landmark model for 3D model enhancement by the attention based 3D reconstruction method described in Sec III.E, step 2, 3 and 4.

A. 3D Model Projection

As discussed above, in order to identify hot region images, we need to project 3D model to 2D images by using the following projection matrices with estimated camera parameters.

\[
P = \begin{bmatrix} K & R \end{bmatrix}^T \begin{bmatrix} 1 & 0 & 0 & 0 R & t \ 0 & 1 & 0 & 0 \ 0 & 0 & 1 & o_3^T \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix}
\]  

(7)

where \( P \) is the projection matrix, \( K \) is the camera intrinsic parameters matrix, \( R \) and \( t \) are the camera extrinsic parameters, \( f \) is the focal length and \( o_3 \) is a 1×3 null matrix.

The coordinates of the 2D projective point in each image are calculated as follows:

\[
[u \ v \ 1]^T = P \times [x_\omega \ y_\omega \ z_\omega \ 1]^T
\]  

(8)

where \( P \) is the projection matrix, \( u \) and \( v \) are the coordinates in the image coordinate system, \( x_\omega \), \( y_\omega \), \( z_\omega \) are the coordinates in the world coordinate system.

B. Hot Region Image Selection

There are three key problems to be resolved in the hot region image selection process: 1) how to identify hot region clusters; 2) how many images should be selected from each hot region cluster; and 3) which image should be selected from each hot region cluster.

Through projecting the 3D model back to the iconic images, we obtain the distribution of the 2D projective points corresponding to the 3D model in the iconic images. The distribution of the 2D projective points can be used to distinguish between the hot region images and the whole scene images. The sufficient and necessary condition of an iconic image to be a whole scene image is that most of the 2D projective points should be distributed in the iconic image. Therefore, if there are many 2D projective points distributed out of the iconic image, the iconic image should be a hot region image. In our experiments, an iconic image with more than half of the 2D projective points located out of the image is a hot region image. Several iconic images may not estimate the camera parameters and their corresponding cluster will be skipped for the hot region image selection.

Selected hot region images should be evenly distributed in each hot region cluster. For the hot region cluster with the smallest number of images, only one image is selected. The number of the images selected from other hot region clusters can be calculated as follows:

\[
HR_i = \frac{CN_i}{CN_{\text{min}}}
\]  

(9)

where \( HR_i \) is the number of the images selected from the \( i \)th hot region cluster, \( CN_i \) is the number of the images in the \( i \)th hot region cluster, \( CN_{\text{min}} \) is the number of the images in the smallest hot region cluster.

The larger numbers of the correspondence points between newly selected hot region images and images that have been used for reconstruction exist, the more accurate reconstruction is achieved. The newly selected hot region images of each
cluster should have more correspondence points than unselected images. Considering computational power, we match images in hot region clusters to the iconic image selected from the same cluster. The images in each hot region cluster are ranked with the number of correspondence points by descending order. Then, the top HR images are selected from the ith hot region cluster.

It is noted that hot region is different from ROI. We identify an iconic image with more than half of the 2D projective points located out of the image as a hot region image. In this paper, ROI is equivalent to visual attention region which is obtained by visual attention analysis.

C. 3D Model Enhancement

The newly selected hot region images are utilized to enhance 3D model. Firstly, visual attention regions are detected for the newly selected hot region images by using the same approach described in Sec III.C. Then, SIFT features [5] are extracted from the detected visual attention regions. RANSAC algorithm [8] is used to estimate a fundamental matrix and further obtain the accurate SIFT matches between the newly selected hot region images and the iconic images. Finally, the newly selected hot region images are added as extra inputs to the reconstructed 3D model. The 3D model is enhanced by using the same method described in Sec III.E, step 2, 3 and 4.

V. IMAGE CLASSIFICATION

In order to avoid the noise from non-landmark region, we only use information from landmark regions to train our landmark classifier. Traditional discriminative classification approaches, such as SVM, are generally applied for the classification problem with consistent distribution of feature points within training images and testing images. In our work, the distribution of feature points contains 1) features extracted from the landmark region within the image and 2) features extracted from the whole image without considering landmark region. Therefore, traditional discriminative classification approaches are not suitable for our work as we do not know the landmark locations in unlabeled images.

We propose a SIFT feature based classification method to classify unlabeled landmark images using k-dimensional tree (kd-tree).

A. Landmark Region Identification

To find the landmark region within images used in Sec III.E and Sec IV.C for 3D reconstruction and enhancement, we project enhanced 3D model into 2D images. The region in 2D image with 2D projective points is identified as landmark region.

Although the 2D projected points describe the landmark region in 2D images, we still need to decide the boundary of landmark regions. The coordinates of the 2D projected points are scanned to identify the landmark region. The 2D projected points with the maximum and minimum x-coordinate and y-coordinate value are located as points on the 2D boundary of the landmark regions. The boundary of the landmark region is formed by a convex polygon derived from the 2D boundary points.

SIFT features are extracted from landmark regions within each image selected in Sec III.D and Sec IV.B and further used to build landmark representation. Compared with SIFT features corresponding to the back projected 3D points, SIFT features extracted from landmark regions within images are more detailed landmark representation. The images mentioned above are the images which have been used for 3D model reconstruction and enhancement. 3D models are reconstructed by the matching points between selected images. For each individual image, un-matching points but within landmark region are actually useful to represent landmark. Moreover, since feature points outside the landmark regions are ignored, the landmark representation is accurate.

B. Landmark Image Classification

By finding the SIFT matching points between a test image and landmark regions of each landmark as described in Sec V.B, the test image is classified as the landmark with the largest number of SIFT matching points. A kd-tree for each landmark is constructed using the SIFT features from landmark regions of training images to achieve fast feature matching. Points, belonging to different images used to generate a same 3D point in 3D model, generate a track. We calculate the mean of the SIFT features of every point in a track as the track center. A point which is the closest to the track center is selected to represent the track. By this way, the kd-tree of each landmark is constructed by considering all possible features located on the landmark region and avoids duplicate information. The approximate nearest neighbor search [11] is utilized for kd-trees to obtain accurate feature matching.

Possibly, the numbers of matching points of a test image compared with two k-d trees are very close, i.e. difference of 1 or 2 points. The number of matching points between a test image and kd-tree i is donated as KD_i. Only if existing a KD_i that is much more than other KD_x ( x ∈ 1, 2, ..., k, x ≠ i ), the test image will be labeled as landmark i. Otherwise, sparse coding based linear SPM (SC-SPM) [4] is further performed to identify the landmark class. SC-SPM has been demonstrated to have competitive performance among state-of-the-art image classification approaches. We set a ratio threshold δ (δ≥1) to make above decision as follows:

\[
\text{Image}_{c} = \begin{cases} 
    i & \exists i, \forall x \in 1, 2, ..., k, x \neq i \quad \frac{KD_i}{KD_x} \geq \delta \quad (\delta \geq 1) \\
    C_{\text{S-SPM}} & \text{Otherwise}
\end{cases}
\]

where Image_c is the category of the current test landmark image, i is the ith kd-tree, x is the xth kd-tree, δ is the ratio threshold, k is the number of the cluster, KD_i is the number of matching points between the current test landmark image and kd-tree i, KD_x is the number of matching points between the current test landmark image and kd-tree x.
Fig. 5. Examples of 3D models. From left to right, from top to bottom, the 3D models are Potala Palace, Himeji Jo Castle, Statue of Liberty, London Tower Bridge, Brandenburg Tor, St. Martin's Dome Church, Christ the Redeemer, Trevi Fountain, Colosseum, Hagia Sofia, Notre Dame, Qi Nian Dian, Taj Mahal, Saint Basil the Blessed and Pantheon.
### Table I

<table>
<thead>
<tr>
<th>Landmark Name</th>
<th>Time Cost 1</th>
<th>Time Cost 2</th>
<th>Time Cost Reduction</th>
<th>Landmark Name</th>
<th>Time Cost 1</th>
<th>Time Cost 2</th>
<th>Time Cost Reduction</th>
<th>Landmark Name</th>
<th>Time Cost 1</th>
<th>Time Cost 2</th>
<th>Time Cost Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saint Giorgio Maggiore</td>
<td>4 h</td>
<td>3 h</td>
<td>25%</td>
<td>Basílica de São Paulo</td>
<td>4.5 h</td>
<td>4 h</td>
<td>11%</td>
<td>Monument to Vittorio Emanuele II</td>
<td>5.4 h</td>
<td>3.1 h</td>
<td>43%</td>
</tr>
<tr>
<td>Neuschwanstein Castle</td>
<td>4.5 h</td>
<td>3.2 h</td>
<td>29%</td>
<td>São Paulo Novo Museum</td>
<td>3 h</td>
<td>2.5 h</td>
<td>17%</td>
<td>Saint Basil the Blessed</td>
<td>6.4 h</td>
<td>4.4 h</td>
<td>31%</td>
</tr>
<tr>
<td>Brandenburg Tor</td>
<td>4.3 h</td>
<td>4 h</td>
<td>11%</td>
<td>Hagia Sofia</td>
<td>5 h</td>
<td>4 h</td>
<td>70%</td>
<td>Saint Maria Maggore</td>
<td>7.4 h</td>
<td>4.4 h</td>
<td>37%</td>
</tr>
<tr>
<td>Trevi Fountain</td>
<td>5.5 h</td>
<td>3.5 h</td>
<td>36%</td>
<td>Partenon</td>
<td>7 h</td>
<td>5 h</td>
<td>29%</td>
<td>Piazza Del Popolo</td>
<td>8.5 h</td>
<td>5.8 h</td>
<td>35%</td>
</tr>
<tr>
<td>Eiffel Tower</td>
<td>4 h</td>
<td>3 h</td>
<td>25%</td>
<td>Leaning Tower Pisa</td>
<td>6 h</td>
<td>4.5 h</td>
<td>25%</td>
<td>Basílica de San Marco</td>
<td>9.2 h</td>
<td>6.3 h</td>
<td>32%</td>
</tr>
<tr>
<td>Notre Dame</td>
<td>6 h</td>
<td>4 h</td>
<td>33%</td>
<td>Himeji Jo</td>
<td>5 h</td>
<td>3.5 h</td>
<td>30%</td>
<td>Ponte di Rialto</td>
<td>6.6 h</td>
<td>4.7 h</td>
<td>29%</td>
</tr>
<tr>
<td>Kinkaku Ji</td>
<td>4.2 h</td>
<td>3 h</td>
<td>29%</td>
<td>David</td>
<td>4.2 h</td>
<td>3 h</td>
<td>29%</td>
<td>Christ the Redeemer</td>
<td>4.4 h</td>
<td>3.1 h</td>
<td>30%</td>
</tr>
<tr>
<td>Piazza della Reppublica</td>
<td>4 h</td>
<td>7 h</td>
<td>40%</td>
<td>Washington Monument</td>
<td>4 h</td>
<td>3.1 h</td>
<td>73%</td>
<td>Saint Maria di Nazareth</td>
<td>6.4 h</td>
<td>4.7 h</td>
<td>27%</td>
</tr>
<tr>
<td>Tower Bridge, London</td>
<td>4 h</td>
<td>2.8 h</td>
<td>30%</td>
<td>Place de la Concorde</td>
<td>4 h</td>
<td>2.4 h</td>
<td>40%</td>
<td>Statue of Venice</td>
<td>4 h</td>
<td>2.3 h</td>
<td>43%</td>
</tr>
<tr>
<td>Piazzale of Saint Pietro</td>
<td>4.5 h</td>
<td>4 h</td>
<td>11%</td>
<td>Saint Maria dei Fiore</td>
<td>4.5 h</td>
<td>4 h</td>
<td>11%</td>
<td>Taj Mahal</td>
<td>4.8 h</td>
<td>4.1 h</td>
<td>12%</td>
</tr>
<tr>
<td>Prague Square</td>
<td>4 h</td>
<td>4.8 h</td>
<td>18%</td>
<td>Arc de Cortant</td>
<td>4.8 h</td>
<td>3.1 h</td>
<td>23%</td>
<td>Lotus Temple</td>
<td>8.6 h</td>
<td>6.8 h</td>
<td>27%</td>
</tr>
<tr>
<td>Saint Martin</td>
<td>7 h</td>
<td>4.8 h</td>
<td>31%</td>
<td>Campidoglio</td>
<td>8.3 h</td>
<td>5.3 h</td>
<td>36%</td>
<td>Westminster Abbey</td>
<td>7.4 h</td>
<td>5.3 h</td>
<td>28%</td>
</tr>
<tr>
<td>Capitol</td>
<td>4.4 h</td>
<td>3.1 h</td>
<td>30%</td>
<td>Castel Saint Angelo</td>
<td>4.2 h</td>
<td>3.1 h</td>
<td>26%</td>
<td>Big Ben</td>
<td>5.1 h</td>
<td>3.2 h</td>
<td>37%</td>
</tr>
<tr>
<td>Piazza della Signoria</td>
<td>4.9 h</td>
<td>3.3 h</td>
<td>33%</td>
<td>Colosseum</td>
<td>7.6 h</td>
<td>5.3 h</td>
<td>28%</td>
<td>Saint Paul's Cathedral</td>
<td>4 h</td>
<td>3 h</td>
<td>25%</td>
</tr>
<tr>
<td>Mount Rushmore</td>
<td>5.3 h</td>
<td>3.9 h</td>
<td>26%</td>
<td>Pantheon</td>
<td>4.4 h</td>
<td>3.8 h</td>
<td>14%</td>
<td>Qi Nian Dian</td>
<td>4.3 h</td>
<td>3.1 h</td>
<td>31%</td>
</tr>
<tr>
<td>Statue of Liberty</td>
<td>4.3 h</td>
<td>3 h</td>
<td>.8%</td>
<td>Pizza Navona</td>
<td>12 h</td>
<td>7.3 h</td>
<td>39%</td>
<td>Potala Palace</td>
<td>8.4 h</td>
<td>6.9 h</td>
<td>18%</td>
</tr>
<tr>
<td>Buckingham Palace</td>
<td>4.6 h</td>
<td>3.3 h</td>
<td>28%</td>
<td>Osaka Castle</td>
<td>4.3 h</td>
<td>3.2 h</td>
<td>26%</td>
<td>Huang He Lou</td>
<td>5.3 h</td>
<td>4.2 h</td>
<td>21%</td>
</tr>
</tbody>
</table>

Fig. 6. Examples of 3D to 2D projection. (Images with mourning border are hot region images.)
VI. Experimental Results

We conduct landmark image classification experiments to validate the effectiveness of our proposed method. The proposed method is compared with BoW based method [2], ScSPM [3] and our previous method [14] respectively. All our experiments use the same computer with the configuration as follows: Win7 system, Intel core 2 Duo CPU E4700 @ 2.60GHz, 4G Memory.

A. Data Preparation

Our experimental dataset consists of 51 landmarks selected from 27 famous cities. Most of them are downloaded from Flickr and Facebook website using keyword searching. Some images are taken by us in Italy. The list of the landmarks can be found in Table I. There are a total of 298,384 landmark images. Each landmark class consists of over 5,000 landmark images. In each landmark class about 1/10 of the images are attached with camera focal length. For each landmark, we cluster the images attached with camera focal length using $k$-means with $k = 100$ which is experimentally set. 6,723 landmark images including both iconic images and hot region images are used to reconstruct and enhance 3D models for 51 landmarks. The above 6,723 landmark images are also used as training images. The rest 291,661 images are used for testing.

B. 3D Model Reconstruction

Some instances of the 3D models are illustrated in Fig. 5. As shown in Fig. 5, most of the 3D points are located on the landmark region. This makes accurate detection of landmark regions possible.

Compared with traditional structure-from-motion (SFM) approach, our attention based approach significantly saves the computational time for 3D reconstruction. Details can be found in Table I. This is because the attention based approach avoids the computational cost from feature matching for non-landmark region.

C. 3D To 2D Projection

Examples of 3D to 2D projection are shown in Fig. 6. From the distribution of the 2D projected points, we can easily distinguish the hot region images from the whole scene images. As shown in Fig. 6, the landmark images with more than half of the 2D projected points located out of the images are recognized as hot region images (shown as images with mourning border in Fig. 6). It is obvious that the 2D projected points located inside 2D images well describe the landmark region.

D. Landmark Image Classification

We conduct four groups of experiments to demonstrate the effectiveness of our proposed method. The first group proves the effectiveness of image selection method for 3D model construction. The second group proves the usefulness of our landmark region detection scenario. The third group proves the significance of the ratio threshold $\delta$. The fourth group proves the effectiveness of our landmark image classification approach.

Comparison on different image groups for 3D reconstruction

We select four different landmark image groups and compare their landmark image classification accuracy to prove the effectiveness of image selection method for 3D model reconstruction, including iconic image selection and the hot region image selection (Fig. 7). The four groups are (1) 100 iconic images and N newly added images; (2) 100 iconic images and N randomly selected images; (3) 100+N iconic images (cluster the landmark images into 100+N clusters); (4) 100+N randomly selected images. N is the number of the newly added images for 3D model enhancement calculated in Sec IV.B.

By adding hot region images for 3D model enhancement, we take special consideration to the most representative parts of the landmark. As shown in Fig. 7, the performance of our image selection approach is better than other three. We find that the third group is better than the second group. It might be because the third group considers more representation styles than the second group. The fourth group is the worst among the four groups due to the random selection.

Comparison on different representations of landmark

We represent each landmark with three different $kd$-trees (Fig. 8) to demonstrate the effectiveness of the landmark region detection scenario. In Fig.8, The blue curve illustrates the
kd-tree constructed with the SIFT features extracted from the landmark regions of selected iconic landmark images; the red curve shows the kd-tree constructed with all the SIFT features extracted from the selected iconic landmark images; the green curve shows the kd-tree constructed with the SIFT features corresponding to the reconstructed 3D points. The x-axis and y-axis in Fig. 8 denote ratio threshold $\delta$ and classification accuracy. The classification accuracy only considers the classification results of those confidently classified images decided by $\delta$. This is why the classification accuracy is growing with the increase of $\delta$. It is obvious that the blue curve has higher classification accuracy than the red and green curves. The kd-tree denoted as the red curve involves too many noisy features located on the non-landmark regions and the kd-tree denoted as the green curve involves too few features, which lead to the lower classification accuracy. The performance

<table>
<thead>
<tr>
<th>CLASSIFICATION ACCURACY OF CLASSIFICATION WITH DIFFERENT $\delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
</tr>
<tr>
<td>----------------------------------------------------------</td>
</tr>
<tr>
<td>Brandenburg Tor</td>
</tr>
<tr>
<td>Buckingham Palace</td>
</tr>
<tr>
<td>United States Capitol</td>
</tr>
<tr>
<td>Christ the Redeemer</td>
</tr>
<tr>
<td>Colosseum</td>
</tr>
<tr>
<td>Eiffel Tower</td>
</tr>
<tr>
<td>Hagia Sofia</td>
</tr>
<tr>
<td>Huang Ji</td>
</tr>
<tr>
<td>Huang He Lou</td>
</tr>
<tr>
<td>Leaning Tower Pisa</td>
</tr>
<tr>
<td>Mount Rushmore</td>
</tr>
<tr>
<td>Notre Dame</td>
</tr>
<tr>
<td>Pantheon</td>
</tr>
<tr>
<td>Potala Palace</td>
</tr>
<tr>
<td>Qi Nian Dia</td>
</tr>
<tr>
<td>Saint Basil’s Cathedral</td>
</tr>
<tr>
<td>Saint Martin’s Church</td>
</tr>
<tr>
<td>Statue of Liberty</td>
</tr>
<tr>
<td>Taj Mahal</td>
</tr>
<tr>
<td>Tower Bridge London</td>
</tr>
<tr>
<td>Trevi Fountain</td>
</tr>
<tr>
<td>Average Accuracy</td>
</tr>
<tr>
<td>Classifier Confidence</td>
</tr>
</tbody>
</table>

Fig. 9. Performance comparison with BoW, ScSPM and our previous method [14]. The training images utilized in ScSPM* are randomly selected from the landmark image collection, the training images utilized in our previous method [14] are selected using the representative image selection approach in [14] and the training images utilized in other approaches are selected using our iconic image selection and hot region image selection approach.
Comparison on different $\delta$

The classification accuracy of landmark image classification with different $\delta$ is shown in Table II where the highest classification accuracy for each landmark is marked bold. The classifier confidence is calculated by the percentage of the testing images classified by our method. The average classification accuracy has the highest value when $\delta = 1.3$. Meanwhile, the classification accuracy has no much change when $1.2 \leq \delta \leq 1.5$. The number of images that are classified by our method decreases as $\delta$ increases. When $\delta < 1.4$, with $\delta$ increasing, the classification accuracy increases. When $\delta \geq 1.4$, the number of images classified by ScSPM increases dramatically. For example, when $\delta$ is tuned from 1.4 to 1.5, the number of images classified by ScSPM increases 10.2%. High percentage of images classified by ScSPM makes the average classification accuracy decrease.

Comparison on classification

To further investigate the effectiveness of our landmark image classification approach, we compare the performance of our method with the BoW based method, ScSPM and our previous method [14] as shown in Fig. 9, respectively. Images used for BoW, ScSPM and our method are selected by the method described in Sec III.D and Sec IV.B. The worst, average and best performance of our approach are also provided with different value of $\delta$ as shown in Fig. 9. The best experimental results of our previous method [14] are provided in Fig. 9 to show the improvement of our method. In order to show the comparisons not affected by training data selection, randomly selected images are used for ScSPM classification as an extra. The classification accuracy of ScSPM* is shown in Fig. 9. A total of 6,723 training images utilized in ScSPM* are randomly selected from the landmark image collection, a total of 5,100 training images utilized in our previous method [14] are selected using the representative image selection approach in [14] and a total of 6,723 training images utilized in other approaches are selected using our iconic image selection and hot region image selection approach. The worst and best performances of our approach are achieved when the values of $\delta$ are 1.0 and 1.3 respectively. We use the mean of performance of $\delta$ from 1.0 to 2.0 with an interval of 0.1.

Fig. 9 illustrates that the performance of the BoW based method is the worst, the average and best performance of our approach are better than other methods. The BoW based method discards the spatial order of local descriptors while ScSPM does not, thus the performance of ScSPM is better than that of BoW. ScSPM* utilized randomly selected images to train classifiers and ScSPM utilized images selected by our proposed method to train classifiers. The images selected by our proposed method cover more representation styles than the randomly selected images. This might be the reason why ScSPM outperforms ScSPM*. Our approach trains classifier using the features located in the landmark regions, while the BoW based method and ScSPM use all the features extracted from training images. The training features of our approach are more accurate descriptor compared with the training features of the BoW based method and ScSPM. Therefore, our approach performs better than them. However, the classification accuracy of the worst result of our approach is lower than ScSPM because the ratio threshold $\delta = 1$ is too weak. Our average performance is better than ScSPM demonstrating the effectiveness of our method. Moreover, our average performance outperforms our previous method [14] 4.7% which proves the usefulness of the hot region.

In order to further demonstrate the effectiveness of our method, the ROC curves of our method, BoW and ScSPM are shown in Fig. 10. The training images and testing images are same as those used in Sec VI.A. Fig. 10 illustrates that the performance of our method is better than the other two methods. Different from BoW and ScSPM which utilize features extracted from the whole image, the features utilized in our method are extracted from the landmark regions of training images. Therefore, the training features of our approach describe the landmark more accurately than BoW and ScSPM. ScSPM adds the spatial order of local descriptors in image description while the BoW based method discards them, thus the performance of ScSPM is better than the BoW based method.

VII. CONCLUSIONS

In this paper, we have presented a novel approach for landmark image classification using enhanced 3D model. In our approach, through visual attention analysis, a 3D landmark model can be reconstructed quickly for each landmark. We project the 3D model onto iconic images used in 3D model reconstruction to give additional consideration to hot regions, and then obtain enhanced 3D models. Enhanced 3D models are projected to iconic images and newly selected hot region images to identify landmark regions accurately. Through representing each landmark using SIFT feature points extracted from the landmark regions, we avoid mis-classification due to noisy and redundant information and combine information related to the same landmark in various presentation styles for consideration. In order to obtain the number of matching points
between an unlabeled landmark image and a landmark image representation quickly, a *kd-tree* is constructed for each landmark representation. By this way, we improve landmark image classification result. Experimental results on 3D model reconstruction, 3D to 2D projection, and landmark image classification have demonstrated the effectiveness of the proposed approach.

In the future, we will investigate landmark image retrieval and personalized landmark image ranking using 3D model of landmarks.

### References


