Landmark Image Classification Using 3D Point Clouds

Xian Xiao\(^1,2\), Changsheng Xu\(^1,2\), Jinqiao Wang\(^1,2\)

\(^1\)National Lab of Pattern Recognition, Institute of Automation, CAS, Beijing 100190, China
\(^2\)China-Singapore Institute of Digital Media, Singapore, 119615, Singapore

Email: {xxiao, csxu, jqwang}@nlpr.ia.ac.cn

ABSTRACT

Most of the existing approaches for landmark image classification utilize either holistic features or interest of points in the whole image to train the classification model, which may lead to unsatisfactory result due to involvement of much information non-located on the landmark in the training process. In this paper, we propose a novel approach to improve landmark image classification result via a process of 2D to 3D reconstruction and 3D to 2D projection of iconic landmark images. Particularly, we first select iconic images from labeled landmark image collections to reconstruct a 3D landmark represented in point clouds. Then, 3D point clouds are projected back onto the same iconic images to obtain the landmark-region of each iconic image and subsequently extract SIFT features from the landmark-region to construct a k-dimensional tree (kd-tree) for each landmark. This process is able to filter out noise points corresponding to clutter background and non-landmark objects in the iconic images. Finally, the unlabeled images can be classified into predefined landmark categories based on the amount of matched feature points between the image features and the kd-trees. The experimental result and comparison with the state-of-the-art demonstrate the effectiveness of our approach.

Categories and Subject Descriptors

I.4.8 [Scene Analysis]: Object recognition

General Terms

Algorithms, Performance, Experimentation

Keywords

Landmark Image Classification, 3D Reconstruction, SIFT Matching

1. INTRODUCTION

The proliferation of photo-sharing websites such as Facebook and Flickr has led to enormous sightseeing pictures uploaded and spread. Among these sightseeing pictures, landmark pictures (Figure 1) are one of the most attractive contents for users. The landmark images are usually assigned some tags when they are uploaded. Due to the various circumstances (illumination, viewpoint, zoom in/out, occlusion, etc.) when a landmark image is taken, the same landmark image may have various presentation styles. How to make use of tagged landmark images to correctly classify an untagged landmark image taken under various circumstances is a challenging task.

Image classification has been extensively studied \([1]\). Compared with image classification, landmark image classification is more challenging due to uniqueness of the landmark and various presentation styles of the same landmark. The existing work on landmark image classification can be summarized into three categories: (1) Bag-of-Word (BoW) based method \([2]\), (2) Spatial Pyramid Matching (SPM) based method \([3],[4]\), and (3) iconic graph based method \([5]\). Since most of the previous work utilized the entire local features \((e.g.,\text{SIFT}\ [6])\) or a global feature to obtain a landmark representation to classify the unlabeled images, the precision of the classification result is limited due to the involved noise in the non-landmark-regions.

In order to improve the classification result by filtering out noise information not located on the landmark-region, we propose a novel approach to describe each landmark structure with 3D point clouds through 3D reconstruction and represent the landmark with SIFT features extracted from the landmark-regions by projection of 3D point clouds back onto iconic images. The landmark-region is the area of the landmark appearing in an image. Through accurate SIFT matching between the landmark images and the SIFT features extracted from the landmark-regions, our approach can classify the landmark images with occlusion, different illumination, or variations in viewpoints and scales (see each column of Figure 1).

The framework of our approach is illustrated in Figure 2, which consists of four steps: 1) iconic images selection and 3D reconstruction, 2) landmark-region identification, 3) k-dimensional tree (kd-tree) construction and 4) unlabeled landmark image classification. Firstly, we select iconic images from the labeled landmark image collections for each landmark using the global descriptor GIST and reconstruct 3D structure of each landmark with the structure-from-motion method \([7]\). Then, the 3D point clouds are projected to the corresponding iconic images and obtain the landmark-regions of each image. Next, we extract SIFT features from the landmark-regions and build a kd-tree for each landmark. Finally, the unlabeled images are classified into predefined landmark categories by matching with each kd-tree.

Compared with the existing approaches, the contribution of our work is two-fold: (1) we propose a novel framework for landmark image classification to improve the classification result; (2) we propose a landmark-region identification approach by projecting the 3D point clouds to the corresponding iconic images.

The rest of the paper is organized as follows. The details of landmark representation and image classification are described in Section 2 and 3 respectively. Experimental results are reported in Section 4. We conclude the paper with future work in Section 5.
2. LANDMARK REPRESENTATION
In order to achieve a good classification result, it is necessary to build a representation model for each landmark. Such a model should reflect the most representative features of the landmark. In our approach, firstly, a number of iconic images are selected from the labeled image collections and utilized to reconstruct the 3D point clouds for the landmarks. Then, landmark-regions in iconic images are detected by projecting the 3D point clouds to the corresponding iconic images. Finally, we extract SIFT features from the landmark-regions to represent each landmark. It is difficult to detect landmark-regions directly from the selected iconic images due to variation of the same landmark in the images and lack of global spatial context among these images. Through 2D to 3D reconstruction and 3D to 2D projection of the iconic images, we can detect the landmark-region from each iconic image, where we can obtain much more SIFT features compared with the features corresponding to the 3D point clouds.

2.1 Landmark 3D Reconstruction
Reconstructing a 3D landmark model from a number of images containing the same landmark taken under various circumstances is able to incorporate the landmark features from different images. Among various 3D reconstruction approaches, the structure-from-motion method [7] can reconstruct good 3D model and estimate accurate camera parameters, particularly when the camera focal length is known.

Our landmark images are collected from photo-sharing websites (Facebook and Flickr), where about a quarter of images are attached with camera focal length. Therefore, we select the representative images from these landmark images with camera focal length to reconstruct 3D landmark model. Here, we adopt the global GIST feature, which is impactful for grouping images by the perceptual similarity [11], to represent the image content and utilize k-means algorithm to cluster images. Within each cluster, the N most representative images whose GIST descriptors are closest to the cluster center are selected as its iconic images to be used for 3D reconstruction. It is worth noting that clusters with the number of images less than N are not considered for reconstruction.

The structure-from-motion method [7] is utilized to reconstruct the 3D landmark from iconic images. We start from the image pair with the largest number of correspondence to estimate the camera parameters using the five point algorithm [8]. Then we apply RANSAC [9] and the direct linear transform [10] to estimate the camera parameters of the rest images according to the largest number of correspondence between images without camera parameters and images with the estimated camera parameters. The 3D point clouds can be obtained by the projection matrix from the estimated camera parameters of images. Several iconic images may not be used for 3D reconstruction because the number of correspondence between them and other iconic images are too small to estimate the camera parameters of these images.

2.2 3D point Clouds Projection
In order to accurately detect the landmark-regions from iconic images, we map the 3D point clouds to the corresponding iconic images according to the projection matrices. The projection matrices can be obtained by the estimated camera parameters easily. The 2D coordinates of the 3D points in each iconic image are calculated as follows:

$$
\begin{align*}
\begin{bmatrix}
u \\
v \\
1
\end{bmatrix} &=
\begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & f
\end{bmatrix}
\begin{bmatrix}
R & t \\
o^T & 1
\end{bmatrix}
\begin{bmatrix}
x_w \\
y_w \\
z_w
\end{bmatrix}
\end{align*}
$$

where \(u\) and \(v\) are the coordinates in the image coordinate system, \(x_w, y_w, z_w\) are the coordinates in the world coordinate system, \(K\) is the camera intrinsic parameters matrix, \(R\) and \(t\) are the camera extrinsic parameters, \(f\) is the focal length and \(o^T\) is a 1×3 null matrix. The projected 2D point set can effectively describe the landmark-region in each iconic image as long as the 3D point clouds accurately describe the landmark in 3D space.

2.3 Landmark Region Identification
We identify the landmark region in each iconic image using the projected 2D point set. The 2D points with the maximum and minimum \(x\)-coordinates for each \(y\)-coordinate value are located at the boundary in the 2D point set. The boundary of the landmark region is formed by a convex polygon derived from the 2D boundary points.

We extract SIFT features from landmark-regions of the iconic images and use these features to build landmark representation. This exhibits two advantages: (1) since the amount of SIFT features in the landmark-region is much more than that of the 3D points, the features located in the landmark region are more detailed descriptor compared with the features corresponding to the 3D points; (2) since the features outside the landmark-regions are filtered out, the landmark features whose size is only about half of the amount of the SIFT features in the iconic images represent the landmark more accurately.
3. IMAGE CLASSIFICATION
Since we use features extracted only from the landmark regions of the images as training data while we do not know the landmark-region in an unlabeled image, using discriminative approaches (e.g. SVM) to train the classifier may not obtain good classification result due to different feature distribution in the training and test images. Here we propose a classification method to classify the unlabeled landmark images, which is able to handle different feature distribution in training and test images.

We classify the test image according to the number of correspondence between the features of the test image and the features extracted from landmark-regions of each landmark. A \textit{kd-tree} for each landmark is constructed using the SIFT features extracted from landmark-regions to achieve fast feature matching. Not all of the extracted features are used to build the \textit{kd-tree}. The SIFT features from different landmark images corresponding to the same 3D point generate a track. The track center represents the mean of the SIFT features. We only keep the feature with the minimum distance between the features and the track-center in the track. Therefore, the final \textit{kd-tree} of each landmark consists of the features located on the landmark-region and has no duplicate correspondence in 3D space. The approximate nearest neighbor search [12] is utilized on the \textit{kd-tree} to obtain accurate feature matching.

To directly classify the unlabeled image into the landmark category with the maximal correspondence is not an optimal way because the two maximal numbers of the correspondence maybe more or less the same which may lead to wrong classification result. Our proposed classification algorithm is shown in Table 1. We set a ratio threshold $\delta (\delta \geq 1)$ to decide whether the image can be successfully classified by determining whether the maximal correspondence is much more than other correspondences. The unsuccessfully classified images are tackled by sparse coding based liner SPM (ScSPM) [4] which has been demonstrated to have good performance among state-of-the-art image classification approaches.

4. EXPERIMENTAL RESULTS
We conduct landmark image classification experiments to validate the effectiveness of our proposed method. Comparison with BoW based method [2] and ScSPM [4] is also given.

4.1 Data Preparation
Our dataset is built by image collections of six landmarks downloaded from Flickr and Facebook using keyword search. The six landmarks are: the Notre Dame in Paris, the Statue of Liberty in New York, the Statue of Liberty in New York, the United States Capitol in Washington, D.C, the Leaning Tower of Pisa in Tuscany, the Potala Palace in Lhasa and the Himeji Castle in Himeji. Each landmark category consists of over 5,000 labeled landmark images.

For each landmark, we cluster the images attached with camera focal length using \textit{k-means} with $k = 50$ which is experimentally set. Then $N (N=5$ in our experiment) images are selected as iconic images from each cluster as described in Section 2.1. The rest images are the test landmark images.

4.2 3D Reconstruction Result
Two instances of the 3D point clouds are illustrated in Figure 3. They are the 3D reconstruction results of the Statue of Liberty and the Potala Palace. As shown in the 3D point clouds, the reconstructed points describe the landmark well and the 3D points are all located on the landmark region. Therefore, we can detect accurate landmark-regions by projecting the 3D point clouds to the corresponding iconic landmark images.

4.3 Evaluation of Image Classification
We conduct three groups of experiments to demonstrate the effectiveness of our proposed method. The first experiment proves the usefulness of our landmark-region detection scenario. The second proves the significance of the ratio threshold $\delta$. The third proves the effectiveness of our landmark images classification approach.

4.3.1 Comparison on different representations of landmark
We represent each landmark with three different \textit{kd-trees} (Figure 4) to demonstrate the effectiveness of the landmark-region detection scenario. The blue curve illustrates the \textit{kd-tree} constructed with the SIFT features extracted from the landmark-regions of selected iconic landmark images; the red curve shows the \textit{kd-tree} constructed with all the SIFT features extracted from the selected iconic landmark images; the green curve shows the \textit{kd-tree} constructed with the SIFT features corresponding to the reconstructed 3D points.
The x-axis and y-axis in Figure 4 denote ratio threshold $\delta$ (Section 3) and precision. The precision only considers the classification results of those successfully classified images decide by $\delta$. That is why the precision is growing with the increase of $\delta$. It is obvious that the blue curve has higher precision than the red and green curves. The $kd$-tree denoted as the red curve involves too much noise 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 precision. The performance shown by the blue curve demonstrates the usefulness of our landmark-region detection scenario.

4.3.3 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 and $ScSPM$, respectively. The worst, average and best performance of our approach are also provided.

Figure 5 illustrates that the performance of the $BoW$ based method is the worst, and the average and best performance of our approach are better than $ScSPM$. The $BoW$ based method discards degenerate data (QDEGSAC). In CVPR, 2006.

7. REFERENCES