Mobile Landmark Search with 3D Models

Weiqing Min, Changsheng Xu, *Fellow, IEEE*, Min Xu, Xian Xiao, and Bing-Kun Bao

***Abstract*—Landmark search is crucial to improve the quality of travel experience. Smart phones make it possible to search landmarks anytime and anywhere. Most of the existing work computes image features on smart phones locally after taking a landmark image. Compared with sending original image to the remote server, sending computed features saves network bandwidth and consequently makes sending process fast. However, this scheme would be restricted by the limitations of phone battery power and computational ability. In this paper, we propose to send compressed (low resolution) images to remote server instead of computing image features locally for landmark recognition and search. To this end, a robust 3D model based method is proposed to recognize query images with corresponding landmarks. Using the proposed method, images with low resolution can be recognized accurately, even though images only contain a small part of the landmark or are taken under various conditions of lighting, zoom, occlusions and different viewpoints. In order to provide an attractive landmark search result, a 3D texture model is generated to respond to a landmark query. The proposed search approach, which opens up a new direction, starts from a 2D compressed image query input and ends with a 3D model search result.**

***Index Terms*—3D reconstruction, 3D to 2D matching, content-based image retrieval, mobile landmark search.**

I. Introduction

With the rapid development of smart phones, more and more people search travel information pervasively. Landmarks are not only the tour destinations but also used to find the way to the travel destinations. Especially, mobile landmark search becomes more crucial to improve the quality of travel experience, through which users only need to take a picture of a landmark to know what they are looking at or download a picture of a landmark from Internet during destination search or planning.

Manuscript received February 01, 2013; revised June 23, 2013 and September 21, 2013; accepted September 27, 2013. Date of publication January 27, 2014; date of current version March 13, 2014. This work was supported in part by National Basic Research Program of China (No. 2012CB316304), National Natural Science Foundation of China (No. 61225009), and Beijing Natural Science Foundation (No. 4131004). This work is also supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tao Mei.

W. Min and X. Xiao are with the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: wqmin@nlpr.ia.ac.cn)

C. Xu and B.-K. Bao are with China-Singapore Institute of Digital Media, Singapore, 139951, and also with the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: csxu@nlpr.ia.ac.cn; bkbao@nlpr.ia.ac.cn).

M. Xu is with iNEXT, School of Computing and Communications, University of Technology Sydney, Sydney, Australia (e-mail: Min.Xu@uts.edu.au).

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

Digital Object Identifier 10.1109/TMM.2014.2302744

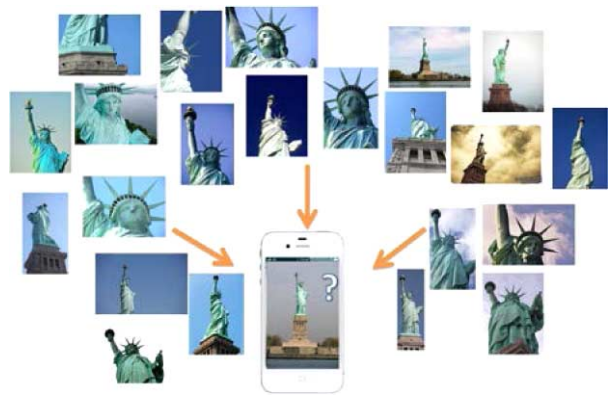

Fig. 1. Examples of various landmark representation styles.

Nowadays, mobile visual search makes a wide range of applications possible. Meanwhile, replacing text-based search, content-based visual search becomes one of the major mobile applications. In these applications, the reference database containing millions of images is stored in a remote server. In searching process, an image captured by a smart phone is firstly sent to the remote server over the mobile network as a query image. Then, the query image is matched with the reference images. The best matching one/ones is/are returned as the search result. Such mobile search functionalities have been shown in several commercial systems, such as the Point and Find [30], Snaptell [31] and Goggles [32]. However, transmitting an original query image occupies much resource of mobile network, which increases the waiting time of mobile visual search.

More recently, with the increasing of mobile computational ability, mobile visual search locally extracts image features on the mobile and sends a feature vector as a query to the reference database [24]–[29]. Compared with sending a query image, sending a feature vector reduces transmission data size, therefore saves network bandwidth and further reduces the transmission cost. However, computing features will consume the power of battery significantly. Obviously, this challenges the tolerant attitudes of users to a short battery running time, since recharging is usually inconvenient for users, especially when they are travelling.

Mobile visual search requires a comprehensive solution which can efficiently save the mobile network resource while effectively decrease the computational cost on mobile side. In this paper, we intend to use a compressed query image with low-resolution to replace the original image or a feature vector extracted from the original image. Compared with the existing methods, query by a low-resolution image brings difficulty to mobile visual search. In addition, compared with mobile image search, mobile landmark search has its own characteristics.

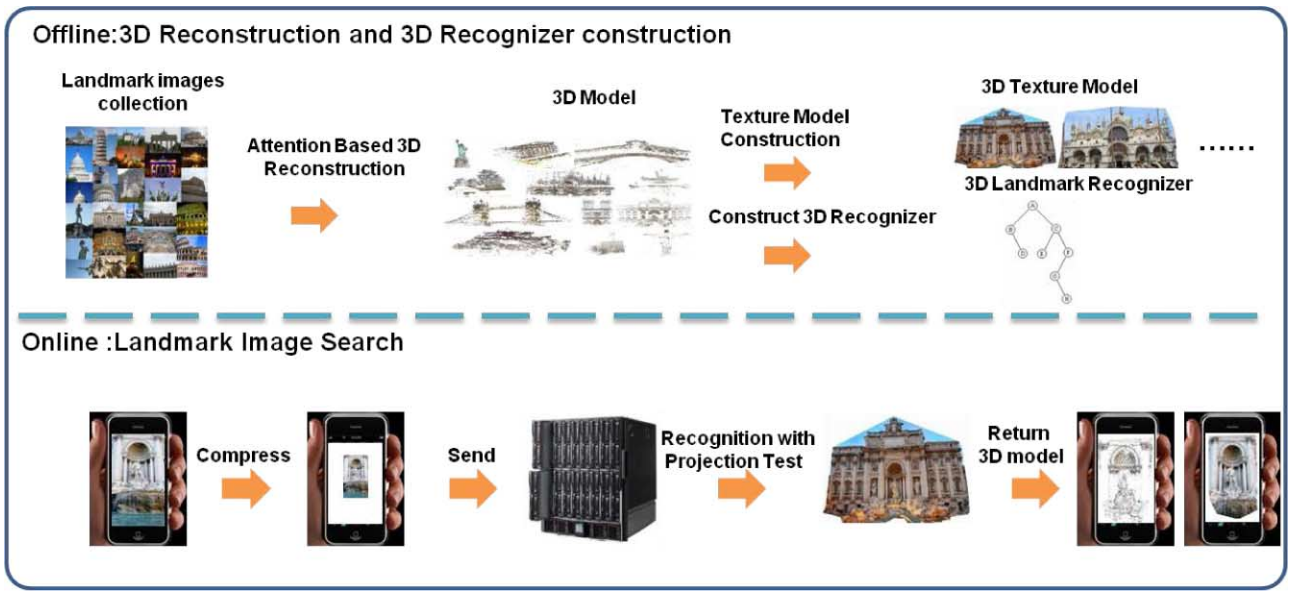

Fig. 2. Our framework for mobile landmark search.

Landmark can be captured with different presentation styles (Fig. 1). Images of the same landmark can appear differently due to various circumstances during picture capturing, including lighting, viewpoint, zoom and occlusion as shown in Fig. 1. The above characteristics make mobile landmark search a challenging problem. The state-of-the-art content-based methods of landmark search [1]–[3] can only return images which are visually similar to the query image. Images of the same landmark with different presentation styles cannot be found. Besides, images which only contain a part of a landmark makes mobile landmark search much more difficult.

To deal with the above difficulties, we propose to utilize 3D models to recognize landmarks from the query images and respond to landmark queries. A 3D landmark model is able to describe a landmark from any scales and any viewpoints. This makes it possible to handle problems coming from various landmark presentation styles, especially when the image only involves a small part of the landmark, in real-world usage scenarios. Moreover, different from existing landmark search, we attempt to return a 3D model to enrich the experience of mobile landmark search. Compared with text based search and image based search, a 3D model is vivid and attractive for users.

Fig. 2 illustrates the framework of our proposed mobile landmark search approach, which consists of two independent modules: offline 3D model reconstruction and online search.

In the offline module, we firstly collect landmark images from *Google* image and *Flickr*. Then, an attention based 3D landmark reconstruction method is employed to construct 3D model by using a number of selected iconic images from the landmark image collection. For each landmark, we select a landmark image which contains all the parts of a landmark to produce a 3D texture model. Finally, all the 2D SIFT feature points corresponding to 3D points in a 3D model are extracted from different iconic images to construct a landmark recognizer. In order to extract more SIFT features for 3D points, we introduce Affine-SIFT [18]. By using all SIFT features (the coordinate of each 3D point is given to these SIFT features as an extra characteristic) corresponding to 3D points, a k-dimensional tree (KD-tree) is constructed to achieve fast landmark image recognition. The KD-tree is actually a 3D landmark recognizer.

In the online search module, we first deliver a compressed query image captured by a mobile device to the remote server. Then, the landmark within the query image is recognized by a 3D landmark recognizer. Finally, a 3D texture model is returned as the query result. SIFT features are extracted from a query image and matching points between the query image and each 3D landmark recognizer are obtained with a projection test. We select a landmark with the largest number of matching points as the recognition result. Our 3D landmark recognizers are designed to recognize low-resolution images. For each recognized query image, a 3D model and a 3D texture model corresponding to the landmark regions in the query image are returned to the user. Note that our work is geo-tag independent. We would like not to constrain our approach on the on-site users, but benefit the users who are not near the landmark. Then, the users, who are not near the landmark, can send any landmark image either downloaded from Internet or photocopied from a magazine, to see the panoramic view of this landmark. This would make our system more helpful on tour destination decision, tour planning, and landmark browsing.

Compared with existing approaches, the contributions of our work are summarized as follows:

1. By offering different compressed rates and sending the compressed images instead of the original images or image feature vectors for search, the proposed method provides users the flexibility to choose on recognition precision vs. the transmission time based on their preference. It not only reduces the query transmitting time but also saves the computational cost on mobile side.

2. A 3D landmark recognizer is generated to recognize various landmark images with different representation styles. With 3D models, images captured under different circumstances, from different viewpoints, containing different parts of landmarks, can be accurately recognized as the corresponding landmark.
3. The proposed method enriches the experience of landmark search by returning the 3D texture model corresponding to the landmark region in the query image. A 3D texture model is vivid and attractive to users, which is able to present a landmark in various viewpoints and scales.

The rest of the paper is organized as follows. The related work is reviewed in Section II. The details of 3D landmark reconstruction, landmark recognizer construction and query image recognition are described in Section III, IV and V respectively. Experimental results are reported in Section VI. We conclude the paper with future work in Section VII.

II. Related Work

A. Mobile Visual Search

With the growing wireless Internet services, mobile visual search has provided a wide range of applications. Meanwhile, compared with text-based search, content based-search in mobile is becoming one of the major mobile applications in mobile visual search. Such mobile search functionalities have been shown in several commercial systems, such as the Point and Find [30], Snaptell [31] and Goggles [32].

A lot of recent work on mobile visual search focused on location recognition [35], landmark retrieval [43], [45], [46], [47], [49] and CD/book cover search [28], [36], [38]. As the computational power on mobiles is growing, several recent work [24]–[29] proposed to directly extract image features on the mobile and represent a query image as a vector, and then send the vector to the remote server over the wireless network. In several mobile visual search systems, approximate similarity matching techniques such as bag-of-words models [36], [39], [40] were utilized to generate vectors to represent the query image. In order to significantly reduce the delay in computation and communication, several compact descriptors were proposed [25], [26], [28], [29], [38], [40]. For instance, Tsai *et al.* [28] proposed to transmit the spatial layouts of descriptors. From each image, about 1,000 points were extracted by local feature detectors and the overall transmission was about 8 KB. Compared with sending the query image (typically over 50 KB) to the remote server, the network flow consumption in [28] was much less. Meanwhile, local features such as Speeded Up Robust Feature (SURF) [33], MPEG-7 image signature, and Scale-Invariant Feature Transform (SIFT) feature [15] were devised to handle luminance and geometry variances. However, although these methods can save much network flow, the power of battery is consumed quickly. In the real-world usage scenarios, users cannot stand an application consuming too much battery power, since it is inconvenient to charge the battery outside.

As the 3G network and free Wi-Fi network become pervasive, in our opinion, most of the computation on the query image in mobile visual search can be done on the remote server.

B. Landmark Image Recognition

Landmark image recognition is very challenging due to various presentation styles of the same landmark. The existing work on landmark image recognition can be summarized into three categories: 1) threshold based method [7], [8], [48] 2) classifier based method [6], [9], [13] and 3) Geo-tag based method [12]. Threshold based method builds a recognizer by selecting the most representative features from landmark images and measures the similarity between the recognizer and the query image. The similarity is compared with a pre-defined threshold to determine whether the query image contains the landmark. In [8], the authors used a visual graph clustering to discover different views of each landmark and then built their recognizers. Besides the visual clustering, in [7] the authors introduced 3D geometric constraints to construct the iconic scene graph of the landmark as its recognizer. The difficulty of threshold based methods is to decide pre-defined thresholds. Classifier based method trains a recognizer for each landmark and recognizes the query image using the trained recognizers. In [9] and [13], the authors detected popular locations and then mined landmark names from tags in these locations. Relevant images were used for training a classifier. The performance of [9] was limited due to the involved noise in the non-landmark region of landmark images. In [6], a 3D point cloud projection was applied to select landmark regions of each image as the training data for improving the performance. However, classifier based method may result in wrong classification result, if the query image, which is actually none of the trained classes, has to be classified to one of the other classes. Geo-tag based method estimates location of an image using training images with location information. For example, Chen *et al.* incorporated GPS information to improve the precision of landmark identification [12]. It is difficult to use geo-tag base method to recognize landmarks from those images without geo-tags. Existing work mainly utilized information on 2D images. In our work, we will make use of 3D information in landmark image recognition to improve recognition accuracy.

III. 3D Landmark Reconstruction

Utilizing all the images of a landmark to reconstruct the 3D model is time-consuming and not necessary. Generally speaking, the images taken from the same viewpoints are always very similar. On the other hand, the work in [14] found that 3D reconstruction using several representative images is able to provide a better approximation to a 3D reconstruction than using all the images. Therefore, for each landmark, we select a group of representative landmark images from different viewpoints to reconstruct a 3D model. These representative images are called as iconic images in our paper. Then, the attention based 3D reconstruction method [5] is utilized to achieve a quick 3D reconstruction. Finally, a whole scene image is selected from iconic images through projecting the 3D model to each iconic image using the corresponding projection matrices. The whole scene image is also utilized to produce a complete 3D texture model. The details of 3D landmark reconstruction will be presented in the rest of this section.

A. Iconic Image Selection

Our goal is to represent each landmark image collection by identifying a set of iconic views corresponding to dominant aspects in the 3D scene. For each landmark, we apply k-means with the global descriptor GIST on the landmark images since GIST is widely used for grouping images using the perceptual similarity [17]. Different clusters indicate various landmark presentation styles. Images which are taken from very similar viewpoints, but different circumstances including illumination, zoom in/out, occlusion, etc., will be clustered into different GIST clusters. This will not cause a problem for our approach, because 3D landmark reconstruction is able to combine different representation styles of the same landmark.

Within each cluster, the most representative landmark image whose GIST descriptor is the closest to the cluster center is selected as the iconic image and further utilized for 3D reconstruction.

B. 3D Model Reconstruction With Static Attention

Different from traditional structure-from-motion (SFM) 3D reconstruction methods [14], the attention based 3D reconstruction method [5] analyzes the spatial-temporal attention to obtain the region of interest (ROI) in video sequence, and then reconstructs 3D model of ROI. 3D reconstruction on ROI within an image instead of a whole image can reduce the computational power. However, compared with the video sequence analyzed in [5], the landmark images in this work do not have much temporal information. We utilize the static attention instead of the spatial-temporal attention to obtain the ROI. In our work, the attention based 3D reconstruction method consists of five steps: 1) analyze attention region of each iconic image; 2) estimate the camera parameters of an initial pair of images; 3) estimate the camera parameters of a newly added image; 4) add points to 3D model; 5) repeat the step 3) and step 4) until the camera parameters have been estimated for all iconic images.

Step 1: A static attention analysis method is used to obtain the attention region. Motivated by [16], we integrate the contrast and information importance to calculate the saliency map as follows:

$$Saliency(x,y) = Con(x,y) \times ID(x,y) \tag{1}$$

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) respectively and normalized to [0, 1]. Contrast Con (x,y) and information density ID (x,y) are calculated by the following Eqn. (2) and Eqn. (3) respectively [16].

$$Con(x,y) = -\sum\nolimits_{(u,v)\in I} d(f(x,y), f(u,v)) \times DoG(u-x, v-y) \tag{2}$$

$$ID(x,y) = \sum\nolimits_{(u,v)\in center} I(u,v) \times DoG(u-x, v-y) \tag{3}$$

where I denotes one image and (u,v) denotes one pixel of this image. $d(f(x,y), f(u,v))$ is the distance between two features $f(x,y)$ and $f(u,v)$. $DoG(u-x, v-y)$ is the Difference of Gaussian [50] and used to model the structure of receptive field. In Eqn. (3), $center$ denotes the receptive field's center. $I(u,v) = -\log(H(r,g,b))$ is the amount of information contained in pixel (u,v), where H is the normalized color histogram of the image.

Several static attention analysis results are shown in Fig. 3.

Step 2: We estimate the camera parameters of an initial pair of iconic images. In order to obtain an accurate 3D reconstruction model of the initial two images, we choose two images with the largest number of correspondence as the initial two images. The camera parameters for the initial pair are estimated using the five point algorithm [19]. We then apply a two-image sparse bundle adjustment [22] to refine estimated camera parameters in order to complete the reconstruction by the initial pair.

Step 3: We continually estimate 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. Camera parameters are initialized using the direct linear transform (DLT) [21] inside a RANSAC [20] procedure. A sparse bundle adjustment [22] is further used to optimize camera parameters and complete the reconstruction.

Step 4: 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. Once the new points have been added, we run a global bundle adjustment [22] to refine the entire model and find the minimum error solution.

Step 5: The 3D reconstruction is completed once all iconic images have been added and the optimized camera parameters are obtained. Iconic images with less than five correspondences are ignored for 3D reconstruction.

C. 3D Texture Modal Construction

A 3D model consisting of 3D points is obtained for each landmark, while a 3D texture model is more vivid and attractive than a 3D model. In order to select an iconic image to construct a texture landmark model, we project 3D points to iconic images and denote the 2D point (in the landmark iconic image) corresponding to a 3D point as a 2D projective point. The iconic image with most of 2D projective points distributed in the image region is an integrated landmark image. We can select each of them to construct a 3D texture model.

Fig. 3. Examples of static attention analysis.

As discussed above, we project 3D points in a 3D model to 2D iconic images by using the following projection matrix with estimated camera parameters:

$$P = [K] \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1/f & 0 \end{bmatrix} \begin{bmatrix} R & t \\ {o_3}^T & 1 \end{bmatrix} \tag{4}$$

where P is the projection matrix, K is 3×3 camera intrinsic parameter 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 points in each image are calculated as follows:

$$s \times [uv1]^T = P \times [x_w y_w z_w 1]^T \tag{5}$$

where P is the projection matrix, 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, s is the scale.

After projection, an iconic image with the most number of 2D projective points distributed in the image region is selected as a whole scene as there are many whole scene images among iconic images. Several examples of 3D to 2D projection are shown in Fig. 4. The left image of each landmark in Fig. 4 is a whole scene image. Then, 2D projective points inside the selected whole scene of the iconic image are triangulated. Each triangle is formed by three 2D projective points and corresponds to a 2D texture in the whole scene iconic image. The 2D texture is utilized to texture the region formed by three 3D points whose 2D projective points form the triangle of the 2D texture. By this way, a 3D texture model is constructed.

IV. 3D Landmark Recognizer Construction

We describe the method of 3D landmark recognizer construction through 3D model in this section. Although existing approaches [7]–[13] have achieved acceptable landmark recognition results, features extracted from non-landmark region brought noisy and redundant information. The method in [6] obtained landmark regions through a 3D to 2D projection and extracted landmark information from the landmark regions. However, noisy information is also extracted from regions which are occluded in landmark regions. Furthermore, since the method in [6] utilized local feature (SIFT) in landmark image recognition without considering geometric constraints, the recognition results are trustless. Generally, geometric constraint refers to a homography matrix that constrains matching points of two images in geometry. In order to handle the aforementioned problems, in this section, a 3D landmark recognizer is built by SIFT features corresponding to 3D points in a 3D model.

A. 3D Feature Structure Definition

In order to exclude noisy and redundant information, we utilize the 3D models to build landmark recognizers. It is obvious that each 3D point in a 3D model corresponds to several 2D points extracted from different iconic images. Each of these 2D points has a SIFT feature. In our work, we define a 3D feature structure to store the 3D coordinates and SIFT features of a 3D point. The 3D feature structure is denoted as follows:

$$P_{3i} = \{(x_i, y_i, z_i) : SFi_1, SFi_2, \ldots, SFi_{mi}\} \tag{6}$$

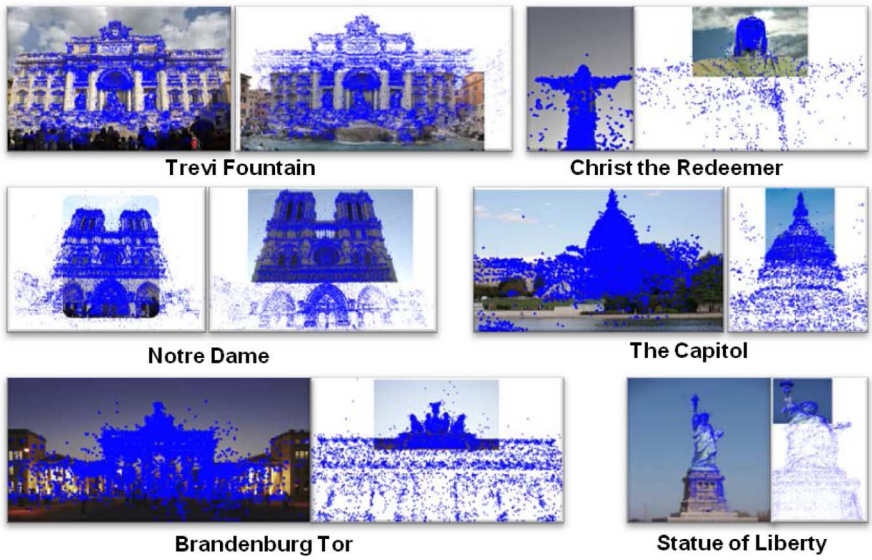

Fig. 4. Examples of 3D model to 2D image projection.

where $P_{3\mathrm{i}}$ is the 3D feature structure of the i-th 3D point, $(x_{\mathrm{i}}, y_{\mathrm{i}}, z_{\mathrm{i}})$ is the 3D coordinate in the world coordinate system, SFi_{mi} is the mi-th SIFT feature, mi is the number of 2D correspondences of the i-th 3D point.

B. ASIFT Addition

In the real world images, the landmark may appear with a viewpoint significantly different from the iconic images used to generate the 3D models. It is impossible to reconstruct a 3D model by incorporating images of the landmark from all possible viewpoints. The state-of-the-art method to handle this problem is Affine-SIFT [18], which extracted features from a finite set of affine transformations of two original images and then matched all the extracted features. The number of matching points is significantly increased. Motivated by [18], we incorporate features extracted from affine warped images into 3D feature structures. An affine transformation A can be decomposed as follows:

$$\begin{aligned} A &= H_\lambda R(\psi) T_t R(\phi) \\ &= \lambda \begin{bmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \end{aligned} \quad (7)$$

where $H_\lambda > 0$, $R(\psi)$ and $R(\Phi)$ are rotation matrices with ψ and $\Phi \in [0, \Omega)$, $T_t (t \geq 1)$ is a tilt. In the decomposition, λ corresponds to the zoom and $R(\psi)$ corresponds to the planar rotation of the camera, Φ is the longitude angle and $t = 1/\cos(\theta)$.

We utilize tilts of $t = \{1, 2/\sqrt{3}, \sqrt{2}, 2\}$ corresponding to latitude angles of $\theta = \{0, 30, 45, 60\}$ degrees in our approach. For each t, we follow θ and sample the longitude angles Φ by an arithmetic series $\Phi = \{0, b/t, \ldots, kb/t\}$ for $b = 72$ degrees and $k = [180 * t/b]$. Each pair $\{t, \Phi\}$ specifies an affine transformation $A_{\mathrm{t},\Phi}$, which is utilized to transform an iconic image as follows:

$$Ic_{t,\phi}(x, y) = Ic(A_{t,\phi}(x, y)) \quad (8)$$

where $Ic_{\mathrm{t},\Phi}$ is an affine transformed iconic image, Ic is an original iconic image and $A_{\mathrm{t},\Phi}$ is an affine transformation matrix.

We extract SIFT features from each $I_{\mathrm{ct},\Phi}$ and compute the location of each key point on the original iconic image. We refer to the new extracted SIFT features as affine features (AF). By using the estimated camera parameters of each iconic image in 3D landmark reconstruction, we search for correspondence of each AF along the epipolar lines in iconic images. If an AF corresponds to a 2D projective point which is obtained by projecting 3D points to iconic images, the AF is added to the 3D feature structure corresponding to the 2D projective point.

C. Recognizer Construction

Since SIFT features in each 3D feature structure are utilized to recognize query images, all the SIFT features in 3D feature structures are utilized to construct a k-dimensional tree (KD-tree) to achieve fast feature matching between recognizers and query images. Each KD-tree is a 3D landmark recognizer.

The number of features in each KD-tree can be calculated as follows:

$$Feature^{i}_{num} = \sum_{j=1}^{P^{i}_{num}} m_j \quad (9)$$

where $Feature^{\mathrm{i}}_{num}$ is the number of features in the i-th KD-tree, which is the 3D landmark recognizer of the i-th landmark, P^{i}_{num} is the number of 3D points in the i-th 3D model, m_j is the number of features corresponding to the j-th 3D point.

V. Image Recognition Using 3D Landmark Recognition

In online module, we compress the query image according to the quality and resolution. Considering the time cost in network transmission, we compress the query image taken by mobile phone to low quality and resolution. Here, we do not care about

which method is utilized to compress the query image on the mobile devices. The simpler, the better.

In our work, the number of matching points between a query image and each 3D landmark recognizer is utilized to recognize a query image. We obtain the number of matching points by direct 2D to 3D matching. The initial matching points are obtained by using the approximate nearest neighbor search (ANN) [23]. Since SIFT feature is shape-based and very ambiguous, the one-to-one matching is infeasible. Therefore, a projection matrix is estimated to verify whether the matching points are correct. The landmark with the largest number of correct matching points is selected as the search result. A 3D model and a 3D texture model are reconstructed for the search result.

ANN's priority search algorithm is applied to limit each query feature for visiting a maximum of 200 bins in the KD-tree. We estimate a projection matrix by using all the matching points in ANN. We do not estimate the projection matrix for query image whose number of matching points to the 3D recognizer is smaller than the lower bound of the number of matching points for projection matrix estimation. As described in (2), a projection matrix is a 3×4 matrix with 12 unknown numbers. Therefore, the lower bound of the number of matching points is straightforwardly set to 12. Since the projection matrix is calculated by intrinsic parameter matrix K, the focal length of a camera and extrinsic parameters including camera rotation (R) and camera translation (t), the projection matrix estimation process is also a camera calibration process. We initialize the query image's camera extrinsic parameters using the direct linear transform (DLT) [21] inside a RANSAC procedure. We utilize an outlier threshold of 0.4% of max (image width, image height) in the RANSAC step. In addition to providing an estimate of R and t, the DLT returns an upper-triangular matrix K' which can be used as an estimate of the camera intrinsic parameter matrix K. We use K' to initialize the focal length of the new camera. Starting from these initial parameters, a sparse bundle adjustment [22] is used to optimize camera parameters.

In order to test whether the matching points are correctly matched, we project 3D points of the matching points to the query image according to the projection matrix estimated above and set an outlier threshold to 0.4% of maximum image dimension.

After projection test, we calculate the number of matching points for each landmark. If there are more than one landmark having more than a pre-defined number of correct matching points (denoted as CMP_{thr}), we select the landmark with the biggest number of correct matching points as the category of the query image. Otherwise, the query image is a non-landmark image. The pre-defined CMP_{thr} will influence the precision and recall of query image recognition. If the number of matching points between a query image and none of 3D landmark recognizers is larger than the lower bound of the matching points, the query image is a non-landmark image (or the query image may be a landmark out of our 3D model dataset). For each recognized query image, in order to obtain the 3D model corresponding to the landmark region in the query image, we project all the 3D points of the selected landmark to the query image according to the estimated projection matrix. Each 3D point has a 2D projective point. The 3D points with 2D projective points inside the query image are selected to constitute a new 3D model corresponding to the landmark region in the query image. Finally, the same method described in Section III-C is utilized to construct a new 3D texture model corresponding to the landmark region in the query image. The information about new constructed models is sent to the mobile device as the landmark search result. The user can also select to receive the integrated 3D texture model as the landmark search result.

VI. Experiment

We conduct various experiments to validate the effectiveness of our proposed approach. The first experiment illustrates several reconstructed 3D models and 3D texture models. The second experiment shows comparison on time cost of 3D model reconstruction by using attention based method and original SFM method. The third experiment indicates the influence of ASIFT addition on the number of features utilized to construct 3D landmark recognizers. In the fourth experiment, the power consumption of our proposed landmark search approach is compared with Google Goggles. In the fifth experiment, the proposed recognition approach is compared with a classifier based method [6] and a threshold based method [7]. The sixth experiment illustrates several landmark search results using landmark images as query. The last experiment gives user experience comparison between our method and Google Goggles.

A. Experimental Setup

To construct our dataset, we download 502,185 images of 217 landmarks to select two subsets, i.e., one is iconic image set for 3D model construction, and the other is the test set to evaluate the performance of our method. We first extract all the landmark names from Wiki page of "category: landmark by country",[1] and manually select 217 famous landmarks. All the images are then crawled from Google Image and Flickr by querying the landmark name. For iconic image set construction, we cluster all the images of the same landmark into k groups,[2] and the image closest to the cluster center is selected as the iconic image. In this way, we finally collect 10,850 ($= 50 \times 217$) images for the iconic image set to construct 3D model. For test set, we first randomly choose 10,000 landmark images from the rest images which exclude the iconic images, and then crawl 20,000 non-landmark images from Flickr. Therefore, the whole test dataset consists of 30,000 images, where the biggest resolution is 1024×1024. We compress test images to three different resolutions (500×500, 300×300 and 200×200) with the quality factor of 70%. Besides the constructed whole test dataset, we also select two subsets of it for evaluation. For the first subset, we manually choose 1,000 landmark images which only contain part of the landmark. This subset is used to evaluate the recognition performance on the images of partial landmarks. For the second subset, 80 landmark images and 20 non-landmark images are randomly selected from the test dataset to constitute a subset named Image100. Image100 is utilized in the experiment of comparing the power consumption and performance of our method with Google Goggles.

[1] http://commons.wikimedia.org/wiki/Category:Landmarks_by_country

[2] In practical, k is set as 50.

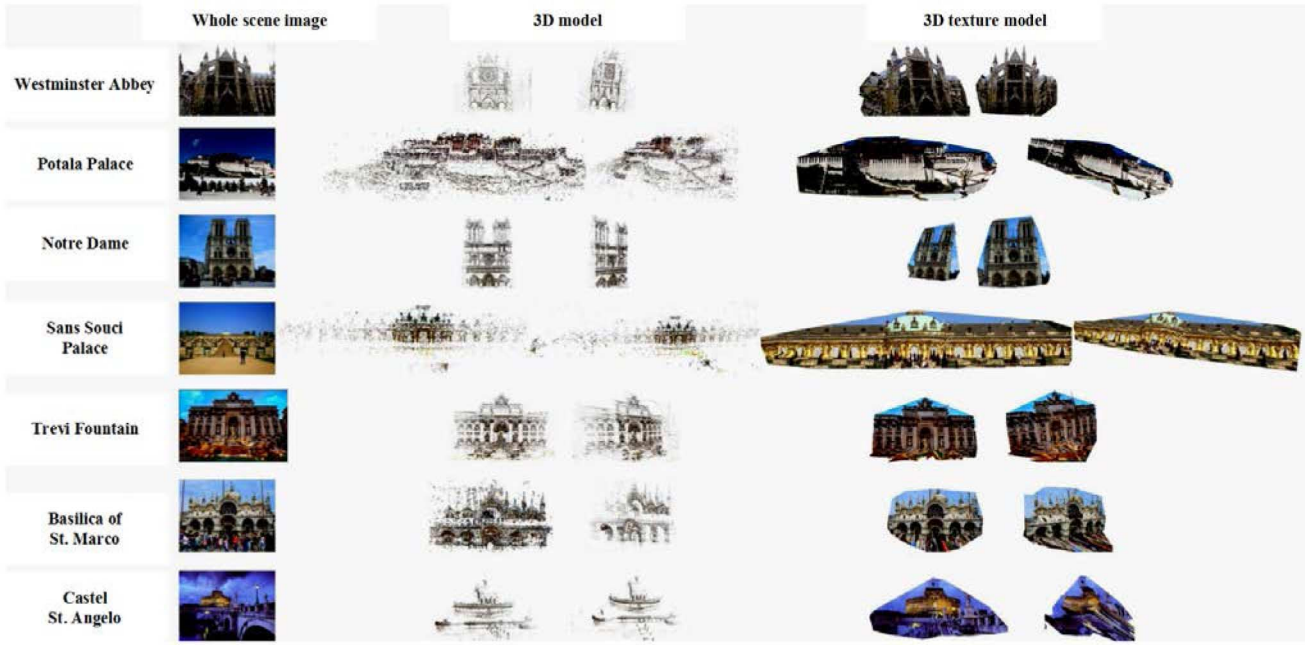

Fig. 5. Examples of 3D models and 3D texture model.

B. 3D Model and 3D Texture Model Reconstruction

Several instances of reconstructed 3D models and 3D texture models are illustrated in Fig. 5. We can see that the 3D points of 3D models, referred to the third column in Fig. 5, are located on the landmark region, and 3D texture models, referred to the last column in Fig. 5, provide different viewpoints and scales for users to view landmarks, which are more attractive than 2D images. This demonstrates that the proposed method of using SIFT features corresponding to 3D points to build landmark recognizer is able to avoid noisy and redundant information in training.

C. Time Cost of 3D Reconstruction

Table I shows the comparison on time cost of 3D landmark reconstruction by using attention based method and traditional structure-from-motion (SFM) method. We utilize the traditional SFM as our baseline, and its process is followed by [14]. Particularly, the camera parameters are first estimated by a pair of iconic images using the five point algorithm [19], and a two-image sparse bundle adjustment [22] is applied to refine the camera parameters. Then, the camera parameters are continually estimated by a newly added iconic image. At last, points observed by the newly added iconic image are added into the camera parameter optimization to further refine the entire model and find the minimum error solution. This process is repeated until the camera parameters have been estimated for all iconic images. As shown in Table I, the sum time cost of the two methods is 5,357 minutes (attention based method) and 7,260 minutes (traditional SFM method), respectively. From the sum time cost at the last row in Table I, we can see that our attention based method reduces 35.52% time cost of the traditional SFM method. This is because the attention based approach avoids the computational cost from feature matching for non-landmark regions. The percentage of time cost reduction might vary for different landmarks. For example, the time cost reduction for Monument to Vittorio Emanuele II reaches 56.33% compared with 6.32% for Basilica of St. Macro. The reason is that the proportions of non-landmark regions in different landmark images are different.

D. 3D Landmark Features

Numbers of features corresponding to 3D feature structures are also shown in Table I. The numbers of features before and after ASIFT addition are provided. It is obvious that the numbers of features increase 28.28% by using ASIFT addition. This is because many SIFT features are extracted from the affine transformed iconic images and many of them correspond to 3D points in the 3D model. As shown in Table I, the percentages of increased features might vary for different landmarks. We can also see that the percentages of increased features vary with the number of features before ASIFT addition in most cases. This is because more features provide more chances for affine features (AF) to find correspondences. However, the positive correlation is not strict. For example, Neuschwanstein Castle has the least percentage of increased features (17.83%), while its number of features before ASIFT is not the least. The AF, which can find correspondences, is extracted from the landmark region. Therefore, one possible reason is that the landmark regions in Neuschwanstein Castle images are too small to be used for extracting AF.

E. Comparison on Power Consumption

In our experiments, Iphone4 is utilized to run our method and Goggles on the Image100 dataset. Iphone4 is equipped with an embedded camera with maximal 2592×1936 resolution and an A4 processor with 1 GHz frequency. Its battery is 3.7 V with a capacity of $\sim$ 1400 mAh ($\sim$ 18 KJ). The power consumption is measured by the energy measurement application

TABLE I
TIME COST OF 3D LANDMARK RECONSTRUCTION AND NUMBER OF FEATURES UTILIZED IN 3D LANDMARK RECOGNIZER CONSTRUCTION

Landmark Name	Time cost of 3D Reconstruction (minutes)			#Number of Features		
	Attention Based	Traditional SFM	Time cost Reduction	After ASIFT	Before ASIFT	Feature Addition
Basilica of St. Pietro	118	173	31.79%	72753	56561	28.24%
Monument to Vittorio Emanuele II	138	316	**56.33%**	51184	40911	25.11%
Christ the REdeemer	104	145	28.28%	32923	27636	19.13%
Piazza Della Signoria	217	267	18.73%	58620	46277	26.67%
Potala Palace	189	262	27.86%	61614	47076	30.88%
Prague Square	286	337	15.13%	114509	86392	32.55%
Tower Bridge	133	192	30.73%	39483	32236	22.48%
Neuschwanstein Castle	239	271	11.81%	45776	38850	**17.83%**
Trevi Fountain	225	282	20.21%	97274	74320	30.89%
Notre Dame	303	386	21.50%	159632	119177	33.95%
Capitol	131	187	29.95%	31378	26025	20.57%
Statue of Liberty	177	216	18.06%	25789	21742	18.61%
Basilica of St. Marco	326	348	**6.32%**	231192	170994	**35.20%**
Castel St. Angelo	129	162	20.37%	93343	72802	28.21%
Himeji Jo	156	228	31.58%	50957	41144	23.85%
Sum	**2871**	**3772**	**23.89%**	**1166206**	**902143**	**29.27%**

TABLE II
POWER CONSUMPTION FOR A SINGLE QUERY

	Our Method	Goggles
Energy(J)	21.36	28.00

in iOS platform.[3] A testing process on power consumption for each query starts from capturing an image and ends at receiving the returned result from the remote server. Table II shows the power consumption comparison on our system and Goggle. The reported power consumption is for a single query and is calculated from the total power consumption of 200 queries. Among 200 queries, the number of wrongly returned results for Goggle is 101 while 85 for our method. Obviously, the energy consumed in our method is less than that in Goggles. This is mainly because compressing images may have a much less computational cost occurring on the mobile side.

F. Landmark Recognition

We compared the proposed recognition approach with a classifier based method [6] and a threshold based method [7] on the whole test dataset and the subset respectively. The method in [6] detected landmark regions of training images and extracted local features from landmark regions to train classifiers. This method only utilized SIFT features extracted from landmark region and avoided noisy and redundant information in training classifiers. The method in [6] did not give additional consideration to geometric constraints when landmark image classification results are obtained by matching test images with classifiers. The method in [7] constructed an iconic graph to recognize query images with global features GIST. Therefore, comparisons with the methods in [6] and [7] also reflect comparisons with the local feature based methods and global feature based methods.

We also compared the recognition performance of the proposed method with Goggles on the Image100 dataset.

1) Experiments on the Whole Test Dataset: The performance of landmark recognition is evaluated by plotting a recall/precision curve of the test images ordered from the highest to lowest score. The recall of our proposed method is decreasing when CMP_{thr} (defined in Section VI) is increasing. Comparisons on the performance of five different landmark recognition methods are shown in Fig. 6, Fig. 7 and Fig. 8 where the query images (with original resolutions over 1024×1024) are compressed to the resolution of 500×500, 300×300 and 200×200. The five methods are: 1) our proposed approach, 2) our proposed 3D recognizer without ASIFT, 3) our proposed 3D recognizer without projection test, 4) a classifier based method [6], [5]) a threshold based method [7].

As shown in Fig. 6, Fig. 7 and Fig. 8, our proposed method achieves the best performance among all the methods. Comparisons on the first three methods demonstrate the effectiveness of ASIFT addition and projection test. Since the ASIFT addition adds not only more features to 3D landmark recognizer but also more wrong matching points to landmark image recognition. In many cases, the projection test has more influential than ASIFT addition on landmark image recognition. This also proves that the projection test is necessary. Since ASIFT adds more matching points between the 3D landmark recognizer and the query image and projection test filters out outliers, the performance of our method is better than the second and third methods. Compared with the classifier based method and threshold based method, our method demonstrates the advantages of the 3D landmark recognizer. Since our method avoids

[3]http://developer.apple.com/library/ios/#recipes/instruments_help-energy-measurement-help/Logging_Energy_Usage_in_an_iOS_Device/10_Logging_Energy_Usage_in_an_iOS_Device.html#//apple_ref/doc/uid/TP40011372-CH17-SW1

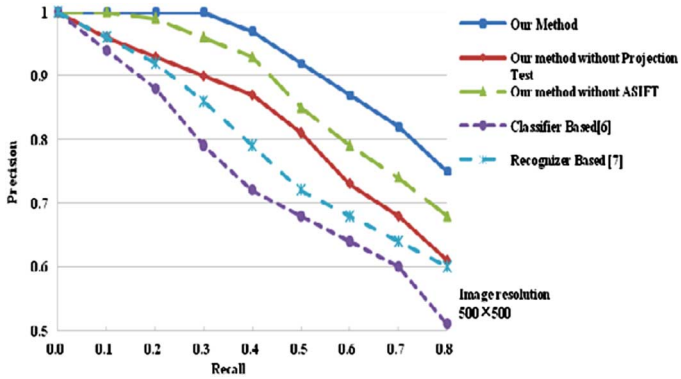

Fig. 6. Performance of landmark recognition (500×500).

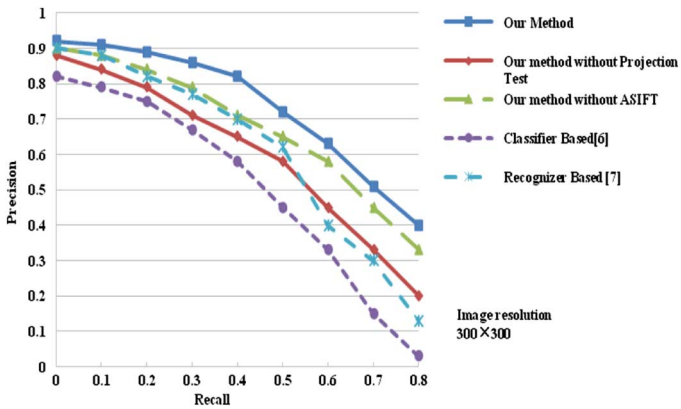

Fig. 7. Performance of landmark recognition (300×300).

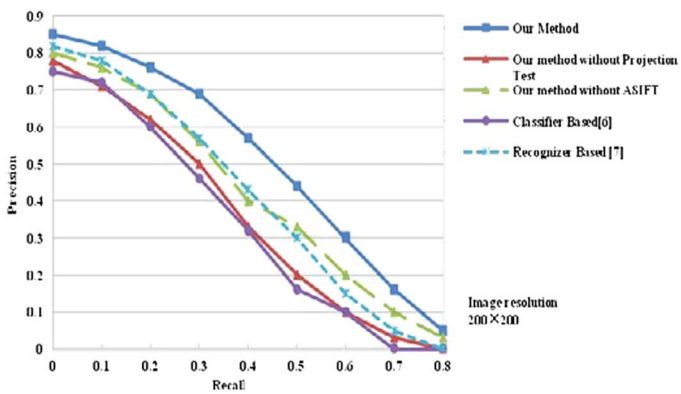

Fig. 8. Performance of landmark recognition (200×200).

the involvement of noisy and redundant information and utilizes ASIFT addition and projection test to find more correct matching points, the performance of our method is the best.

Besides the evaluation on recognition performances under different algorithms, we also evaluate recognition precision vs. transmission time under various resolutions, shown in Table III. Considering that the upload speed is not good in some areas such as countryside, we assume this speed as 60kbps as [44] suggested. We test four kinds of resolutions, that is, the original one 1024×1024, 500×500, 300×300, and 200×200 respectively. The experimental results in Table III showed that when the original image (1024×1024) is compressed into 500×500, the precision decreases about 0.7% (from 98.7% to 98%) while

TABLE III
COMPARISON ON PERFORMANCE OF OUR METHOD FOR THE TEST IMAGE UNDER DIFFERENT RESOLUTIONS (Recall $= 40\%$)

Resolution	200×200	300×300	500×500	Original(1024 × 1024)
Data (KB)	18.7	28.2	56.7	>180
Precision	59%	82%	98%	98.7%
Time (second)	2.50	3.76	7.56	>24

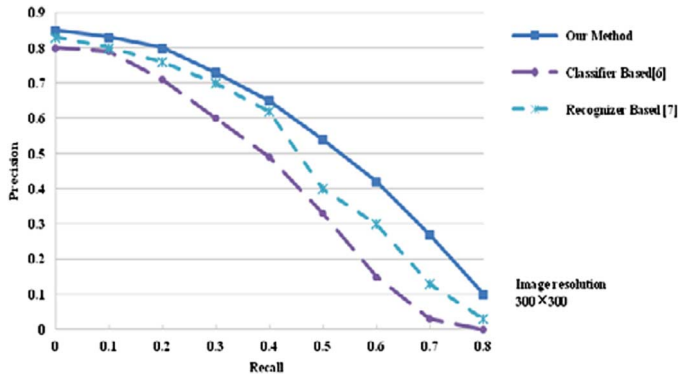

Fig. 9. Performance of landmark recognition on the subset (300×300).

TABLE IV
COMPARISON ON PERFORMANCE OF OUR METHOD AND GOGGLES

	Precision	Recall	True negative rate	Accuracy
Our Method	52.5%	42%	85%	59%
Goggles	47.5%	38%	35%	45%

the transmission time decreases to 1/3 (from 24 s to 7.56 s) of the original one. If the original image is compressed to 300×300, the transmission time is further reduced to 1/6 (from 24 s to 3.76 s) while there is only 16% loss (from 98.7% to 82%) on precision. Our method allows user to flexibly choose on recognition precision vs. the transmission time by setting different compressed rate based on their preference. With small sacrifice of the recognition precision, the transmission time can be reduced significantly. If the user needs higher precision, they can choose the lower compression, with the cost of longer transmission time.

2) Experiment on the Subset of the Test Dataset: The performance of our method, the classifier based method [6] and the threshold based method [7] on the subset are shown in Fig. 9. As shown in Fig. 9, our method has more superiority on recognizing images that only contain a part of the landmark. This is because our 3D model based method puts all the features into a whole 3D recognizer and the similarity between the query image and the 3D recognizer can be found easier than the other two methods.

3) Experiment on the Image100 Dataset: For recognition task, the terms true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are the statistical data on the recognition result. The precision $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, recall $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, true negative rate $\mathrm{TN}/(\mathrm{TN}+\mathrm{FP})$ and accuracy $(\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})$ are usually utilized to compare the performance of different recognizers.

TABLE V
THE QUERY TIME COMPARISON UNDER DIFFERENT RESOLUTIONS

	200 × 200	300 × 300	500 × 500	Original size (1024 × 1024)
Image Uploading	2.5	3.76	7.56	>24
Feature Extraction	0.3	0.5	1.0	3.1
Landmark Recognition	1.45	2.2	2.42	2.6
3D Texture Model Reconstruction	0.75	0.90	1.08	2.0
3D Texture Model Downloading	3.3	5.1	9.2	16.6
Total	8.9	12.46	21.26	48.3

The performance of our method and Goggles is shown in Table IV. Iphone4 is utilized to run our method and Goggles on the Image100 dataset. The images in Image100 dataset are compressed to 200×200 resolution before the experiment. From Table IV, our method outperforms Goggles, especially for true negative rate and overall accuracy. The experiment on the Image100 dataset demonstrates the efficiency of our proposed method.

G. 3D Landmark Recognition

We compare our 3D landmark search approach with the content-based method on landmark search [1] and Goggles [32]. Some search results using same query images are illustrated in Fig. 10 and Fig. 11.

In Fig. 10, the search result is very similar to the query image. Therefore, we have to use many different query images to search landmark images if we are interested in various presentation styles of the landmark.

In Fig. 11, Goggles returns the landmark name and URL (for quickly search) while our method provides a 3D model and a 3D texture model for each recognized query image. In the offline constructed 3D models, compared with Fig. 10, we can view the landmark from various viewpoints and different scales by rotating and zoom in/out the model in the 3D space. The online constructed 3D model is obtained by projecting 3D points to the query image according to the estimated projection matrix in Section III-C. The 3D points with 2D projective points inside the query image are selected to constitute a 3D model corresponding to the landmark region in the query image. Then, a 3D texture model is obtained by using the same method described in Section III-C. We compare online query time under different resolutions in Table V. For the compressed query image, time cost for constructing 3D models for each query image is at most 1.08 seconds, which is acceptable for online use. Although data size of a returned 3D texture model is about 4 times more than the size of a returned image, the time spending on transmitting a 3D texture model from a server to a mobile phone is still small because of high download speed. From a mobile phone point of view, a 3D texture model only occupies relatively small percentage of the memory of a mobile phone. Also, only 40% of the mobile phone CPU is utilized by the relevant operations. These are within the tolerance range for a proper running system. Therefore, users can easily use mobile devices to interact with a 3D texture model. The online constructed 3D models correspond to the landmark region in the query image, which shows more detailed and attractive information of the query image from various viewpoints and scales. Furthermore, the landmark images with different angles can be also returned for the query landmark, which are shown at the last two rows of our method in Fig. 11. This is an attractive landmark search experience.

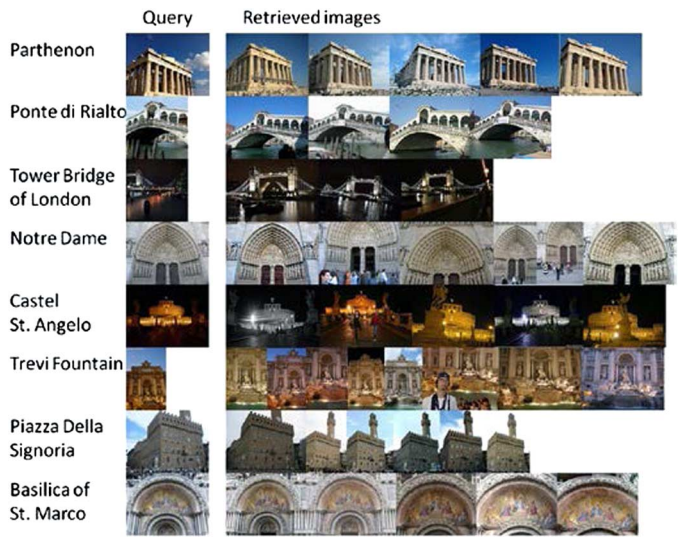

Fig. 10. Content-based landmark search results [1].

H. User Experience: Our System vs. GOGGLES

As the interface design is not the focus of this work, we focus on the quality of the returned images/3D texture models of our system and Goggles for user experience study. Note that the resolution of query images is 300x300. For comparison, we asked 12 participants to randomly take photos of 3 different landmark images to test our system and Goggles respectively.

All the participants are asked to evaluate these two systems from four aspects: 1) effectiveness, namely whether the returned results are correct; 2) attractiveness, namely whether the returned results are vivid and attractive; 3) time efficiency, namely whether the time cost on one complete query is accepted for users; 4) satisfaction, namely whether users are satisfied with the system. All the evaluations are categorized into five levels, i.e. $\{1, 2, 3, 4, 5\}$, indicating "very bad", "bad", "average", "good" and "very good", respectively. The average statistics from 12 participants are shown in Table VI. From the results, we can see that, overall, our system outperforms Goggles, and the superiority is very obvious in terms of attractiveness while the time

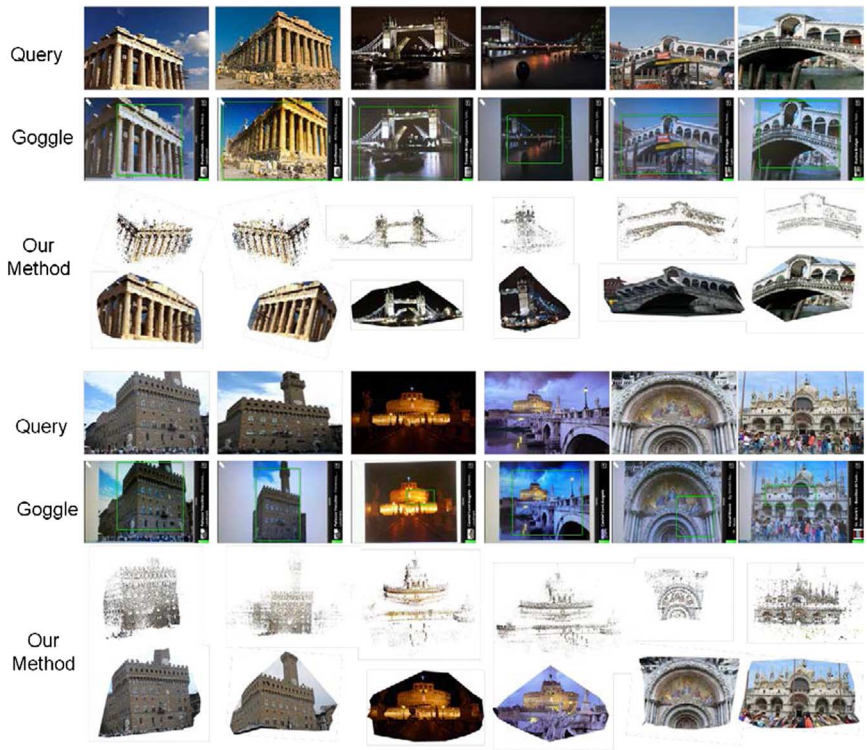

Fig. 11. Search result of Google Goggles and our method.

TABLE VI
THE USER EXPERIENCE COMPARISON BETWEEN OUR SYSTEM AND GOGGLES

Aspects	Our system	Goggles
effectiveness	3.9	3.5
attractiveness	4.2	3.2
time efficiency	3.5	3.9
satisfaction	4.0	3.6

efficiency is a little bit lower than Goggle. Possible explanations include 1) our 3D model based method can robustly handle query images under various conditions of lighting, zoom, occlusions or ones even containing only a small part of the landmark, thus the high performance of recognition can be guaranteed; 2) for low time efficiency, this is mainly caused by low configuration of our server and long transmission time for returned 3D texture model; and 3) our system utilizes 3D texture model to provide users various viewpoints, scales for each landmark, which are more vivid and attractive than the images provided by Goggles. In addition, we conduct another user study to compare the following three different aspects of returned results: (1) 3D texture model (2) 3D model and (3) 2D images. Note that 2D images are the iconic images used to construct the 3D model, and 3D texture model is our result. All the evaluations are also categorized into five levels, i.e. $\{1, 2, 3, 4, 5\}$, indicating "very bad", "bad", "average", "good" and "very good", respectively. The average statistics from 12 participants are shown in Table VII. We can see that the attractiveness score of 3D texture model is the highest. This further verifies that the returned 3D texture model is more vivid and attractive than other forms.

TABLE VII
THE USER EXPERIENCE COMPARISON ON DIFFERENT FORMS OF RETURNED RESULTS FROM OUR METHOD

Aspects	3D texture model	3D model	2D Images
attractiveness	4.2	3.7	3.4

VII. CONCLUSION

In this paper, we have presented a novel approach for robust landmark search on mobile devices. In our work, 3D models are obtained offline. In order to achieve high performance of landmark search, we build 3D landmark recognizers to recognize unlabeled landmark images by a direct 2D to 3D matching. The ASIFT addition in offline module and the projection test in online module improve the performance of landmark recognition. In the online module, we compress the query images to achieve fast image transmission and save network bandwidth. We construct a 3D model and a 3D texture model corresponding to the landmark region in query images and return the 3D model and 3D texture model as landmark search results. This enriches the user experience of landmark search. Experimental results on landmark recognition and landmark search have demonstrated the effectiveness of our method.

In the future, we will investigate landmark rendering and interactively landmark touring in virtual reality to further complement and enhance the proposed approach.

REFERENCES

[1] Y. Avrithis, Y. Kalantidis, G. Tolias, and E. Spyrou, "Retrieving landmark and non-landmark images from community photo collections," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 153–162.

[2] Y. Kalantidis, G. Tolias, Y. Avrithis, M. Phinikettos, E. Spyrou, P. Mylonas, and S. Kollias, "VIRaL: Visual image retrieval and localization," *Multimedia Tools Applicat.*, vol. 51, no. 2, pp. 555–592, Jan. 2011.
[3] E. Gawes and C. G. M. Snoek, "Landmark image retrieval using visual synonyms," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1123–1126.
[4] [Online]. Available: http://en.wikipedia.org/wiki/List_of_landmarks/
[5] X. Xiao, C. S. Xu, and Y. Rui, "Video based 3D reconstruction using spatio-temporal attention analysis," in *Proc. ICME*, 2010, pp. 1091–1096.
[6] X. Xiao, C. S. Xu, and J. Q. Wang, "Landmark image classification using 3D point clouds," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 719–722.
[7] X. Li, C. Wu, C. Zach, S. Lazebnik, and J. M. Frahm, "Modeling and recognition of landmark image collections using iconic scene graphs," in *Proc. ECCV*, 2008, pp. 427–440.
[8] Y. T. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T. S. Chua, and H. Neven, "Tour the world: Building a web-scale landmark recognition engine," in *Proc. Comput. Vis. Pattern Recognit.*, 2009, pp. 1085–1092.
[9] Y. P. Li, D. J. Crandall, and D. P. Huttenlocher, "Landmark classification in large-scale image collections," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 1957–1964.
[10] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. Comput. Vis. Pattern Recognit.*, 2006, pp. 2169–2178.
[11] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. Comput. Vis. Pattern Recognit.*, 2009, pp. 1794–1801.
[12] D. Chen, G. Baatz, K. Koser, S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk, "City-scale landmark identification on mobile devices," in *Proc. Comput. Vis. Pattern Recognit.*, 2011, pp. 737–744.
[13] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *Proc. ACM Int. Conf. World Wide Web*, 2009, pp. 761–770.
[14] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3d," in *Proc. ACM SIGGRAPH*, 2006, pp. 835–846.
[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[16] H. Y. Liu, S. Q. Jiang, Q. M. Huang, and C. S. Xu, "A generic virtual content insertion system based on visual attention analysis," in *Proc. ACM Int. Conf. Multimedia*, 2008, pp. 379–388.
[17] J. Hays and A. A. Efros, "Scene completion using millions of photographs," in *Proc. ACM SIGGRAPH*, 2007.
[18] J. Morel and G. Yu, "ASIFT. A new framework for fully affine invariant image comparison," *SIAM J. Imag. Sci.*, 2009.
[19] D. Nistér, "An efficient solution to the five-point relative pose problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 6, pp. 756–770, 2004.
[20] J. M. Frahm and M. Pollefeys, "RANSAC for (quasi-) degenerate data (QDEGSAC)," in *Proc. Comput. Vis. Pattern Recognit.*, 2006, pp. 453–460.
[21] R. I. Hartley and A. Zisserman, *Multiple View Geometry*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[22] M. I. A. Lourakis and A. A. Argyros, "SBA: A software package for generic sparse bundle adjustment," *ACM Trans. Math. Softw.*, vol. 36, no. 1, Mar. 2009.
[23] J. S. Beis and D. G. Lowe, "Shape indexing using approximate nearest-neighbor search in high-dimensional spaces," in *Proc. Comput. Vis. Pattern Recognit.*, 2003, pp. 1000–1006.
[24] D. Chen, S. Tsai, V. Chandrasekhar, G. Takacs, J. Singh, and B. Girod, "Tree histogram coding for mobile image matching," in *Proc. DCC*, 2009, pp. 143–152.
[25] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk, and B. Girod, "Chog: Compressed histogram of gradients a low bit-rate feature descriptor," in *Proc. Comput. Vis. Pattern Recognit.*, 2009, pp. 2504–2511.
[26] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, J. Singh, and B. Girod, "Transform coding of image feature descriptors," in *Proc. SPIE VCIP*, 2009.
[27] M. Makar, C. Chang, D. Chen, S. Tsai, and B. Girod, "Compression of image patches for local feature extraction," in *Proc. ICASSP*, 2009, pp. 821–824.
[28] B. Girod, V. Chandrasekhar, D. Chen, N. M. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. Tsai, and R. Vedantham, "Mobile visual search," *IEEE Signal Process. Mag.*, vol. 28, no. 4, pp. 61–67, 2011.
[29] R. Ji, L. Y. Duan, H. X. Yao, Y. Rui, S. F. Chang, and W. Gao, "Towards low bit rate mobile visual search with multiple-channel coding," in *Proc. ACM Int. Conf. Multimedia*, 2011, pp. 573–582.
[30] "NOKIA point and find," [Online]. Available: http://pointandfind.nokia.com/.
[31] "SnapTell," [Online]. Available: http://www.snaptell.com/.
[32] "Google goggles," [Online]. Available: http://www.google.com/mobile/goggles/.
[33] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded-up robust features," in *Proc. ECCV*, 2008, pp. 346–359.
[34] X. Xiao, C. S. Xu, J. Q. Wang, and M. Xu, "Landmark recognition and retrieval: From 2D to 3D," in *Proc. ACM, J-HGBC*, 2011, pp. 77–78.
[35] J. A. Lee, K. C. Yow, and A. Sluzek, "Image based information guide on mobile devices," *Adv. Vis. Comput.*, 2008.
[36] J. Philipin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabulary and fast spatial matching," in *Proc. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
[37] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. Comput. Vis. Pattern Recognit.*, 2006, pp. 1000–1006.
[38] S. Tsai, D. Chen, G. Takacs, V. Chandrasekhar, J. Singh, and B. Girod, "Location coding for mobile image retrieval," *ICST MobileMedia*, 2009.
[39] A. Irschara, C. Zach, J. Frahm, and H. Bischof, "From structure-from-motion point clouds to fast location recognition," in *Proc. Comput. Vis. Pattern Recognit.*, 2009, pp. 2599–2606.
[40] G. Schindler and M. Brown, "City-scale location recognition," in *Proc. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–7.
[41] V. Chandrasekhar, D. Chen, A. Lin, G. Takacs, S. Tsai, N. Cheung, Y. Reznik, R. Grzeszczuk, and B. Girod, "Comparison of local feature descriptors for mobile visual search," in *Proc. ICIP*, 2010, pp. 3885–3888.
[42] W. Zhang and J. Kosecka, "Image based localization in urban environments," in *Proc. 3DVT*, 2006, pp. 33–40.
[43] R. Ji, L. Y. Duan, J. Chen, H. Yao, J. Yuan, Y. Rui, and W. Gao, "Location discriminative vocabulary coding for mobile landmark search," *Int. J. Comput. Vis.*, vol. 96, no. 3, pp. 290–314, 2011.
[44] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk, and B. Girod, "CHoG: Compressed histogram of gradients a low bit-rate feature descriptor," *Int. J. Comput. Vis.*, vol. 96, pp. 384–399, 2012.
[45] M. Wang, K. Yang, X. S. Hua, and H. J. Zhang, "Towards a relevant and diverse search of social images," *IEEE Trans. Multimedia*, vol. 12, no. 8, pp. 829–842, 2010.
[46] H. Liu, T. Mei, J. Luo, H. Li, and S. Li, "Finding perfect rendezvous on the go: Accurate mobile visual localization and its applications to routing," in *Proc. ACM Int. Conf. Multimedia*, 2012, pp. 9–18.
[47] M. Wang, H. Li, D. C. Tao, and X. Wu, "Multimodal reranking for web image search," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4649–4661, 2012.
[48] M. Wang, Y. Gao, K. Lu, and Y. Rui, "View-based discriminative probabilistic modeling for 3D object retrieval and recognition," *IEEE Trans. Image Process.*, vol. 22, no. 4, 2013.
[49] H. Liu, T. Mei, J. Luo, H. Li, and S. Li, "Robust and accurate mobile visual localization and its applications," *ACM Trans. Multimedia Comput., Commun., Applicat.*, to be published.
[50] R. E. Soodak, "Two-dimensional modeling of visual receptive fields using Gaussian subunits," in *Proc. Nat. Acad. Sci. USA*, 1986, vol. 83, pp. 9259–9263.

Weiqing Min received the B.E. degree from Shandong Normal University, Jinan, China, in 2008 and M.E. degree from Wuhan University, Wuhan, China, in 2010, respectively. He is currently pursuing the Ph.D. degree at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing.

In 2012, he was an intern student in the China-Singapore Institute of Digital Media (CSIDM). His current research interests include landmark based multimedia search and mining, computer vision and pattern recognition.

Changsheng Xu (M'97–SM'99–F'14) is a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, and an Executive Director of the China-Singapore Institute of Digital Media, Singapore. His current research interests include multimedia content analysis/indexing/retrieval, pattern recognition, and computer vision. He holds 30 granted/pending patents and has published over 200 refereed research papers.

Dr. Xu is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA and ACM *Transactions on Multimedia Computing, Communications*. He served as a Program Chair of ACM Multimedia in 2009. He has served as an Associate Editor, Guest Editor, General Chair, Program Chair, Area/Track Chair, Special Session Organizer, Session Chair and TPC Member for over 20 prestigious IEEE and ACM multimedia journals, conferences, and workshops. He is an ACM Distinguished Scientist.

Min Xu received the B.E. degree from University of Science and Technology of China, in 2000, M.S. degree from National University of Singapore in 2004 and Ph.D. degree from University of Newcastle, Australia in 2010.

Currently, she is a lecturer in School of Computing and Communications, Faculty of Engineering and IT, University of Technology, Sydney. Her research interests include multimedia content analysis, video adaptation, interactive multimedia, pattern recognition and computer vision.

Xian Xiao received the B.E. degree from Jilin University, Changchun, China, in 2007, and the Ph.D. degree from Institute of Automation, Chinese Academy of Science, Beijing.

In 2010, he was an intern student in the China-Singapore Institute of Digital Media (CSIDM). His current research interests include 3D reconstruction, landmark image classification and recognition, multimedia content analysis, pattern recognition and computer vision.

Bing-Kun Bao received the Ph.D. degree in Control Theory and Control Application, Department of Automation, University of Science and Technology of China (USTC), China, in 2009.

Dr. Bao is currently an assistant researcher at Institute of Automation, Chinese Academy of Sciences, and a researcher at China-Singapore Institute of Digital Media. Her research interests include cross-media cross-modal image search, social event detection, image classification and annotation, and sparse/low rank representation. She received the Best Paper Award from ICIMCS'09. She served as a technical program committee member of several international conferences (MMM2013, ICIMCS2013, etc.), and a guest editor in Multimedia System Journal.