Image Location Estimation by Salient Region Matching

Xueming Qian, *Member, IEEE*, Yisi Zhao, and Junwei Han, *Member, IEEE*

***Abstract*— Nowadays, locations of images have been widely used in many application scenarios for large geo-tagged image corpora. As to images which are not geographically tagged, we estimate their locations with the help of the large geo-tagged image set by content-based image retrieval. In this paper, we exploit spatial information of useful visual words to improve image location estimation (or content-based image retrieval performances). We proposed to generate visual word groups by mean-shift clustering. To improve the retrieval performance, spatial constraint is utilized to code the relative position of visual words. We proposed to generate a position descriptor for each visual word and build fast indexing structure for visual word groups. Experiments show the effectiveness of our proposed approach.**

***Index Terms*— Image retrieval, bag-of-words, spatial constraint, salient area detection, mean-shift.**

I. Introduction

IN RECENT years, estimating the locations of images has received a lot of attention. Large quantities of images taken by the users are shared in social media websites such as Facebook, and Flickr every day. Many of the images are associated with the locations when they were taken. As to images without geo-tags, automatic location estimation for them is possible with the help of the large scale geo-tagged photos shared by millions of worldwide users. In this paper, we estimate their locations utilizing content based image retrieval approach. Our task is to estimate the location of an input image by mining image content.

State-of-the-art large scale image retrieval systems have relied on the bag-of-words (BoW) model [37] and local descriptors. And the idea of hierarchical vocabulary tree [34] accelerates the speed of clustering and quantizing for large scale image retrieval. Traditionally, a visual vocabulary is trained by clustering a large number of local feature descriptors, such as SIFT [42], SURF [49]. The exemplar descriptor of each cluster is called a visual word, which is then indexed by an integer. However, Li et al. find that some images can be recognized well via global feature matching and local feature refinement [4]. Only using local feature may neglect the information that global feature can provide [4]. Therefore, in this paper, we further explore global feature clustering and local feature refinement based approach to carry out image geographic location estimation. In our work, firstly, we determine the refined locations of an input image using global features clustering. This step can speed up the image location estimation process by selecting some candidate locations. Secondly, we exploit spatial information relied on the bag-of-words to improve the image location estimation performance.

Existing works show that the commonly generated visual words are still not as expressive as the text words. Quantization [33] limits the discriminative power and ignores geometric relationships among visual words. Spatial verification enforces geometric consistent constraint on visual words that query and dataset image share, such as RANSAC and spatial coding [8]. Spatial information of visual words should be exploited for better image retrieval performance. Motivated by the problem of mismatching SIFT features, Wu *et al.* [35] employ the detector of Maximally Stable Extremal Regions (MSER) to bundle SIFT features into groups instead of taking all of them individually. The bundled feature based methods have more discriminative performance than the methods using the single SIFT feature, because the bundled feature based methods employ group feature matching instead of single feature matching.

In our work, visual words mining and spatial constraint based image geographical location estimation approach is exploited. Considering that the distribution of an image's visual words directly reflects the distribution of the image's main content, we mine the salient features for location estimation and exploit spatial information from the selected useful visual words. Firstly, we utilize term frequency-inverse document frequency (tf-idf) to select visual words with higher weight. Secondly, we divide an image's useful visual words into multiple groups by Mean-shift clustering [28]. In this process, the coordinates of BoWs are utilized to provide a geometric constraint. A visual word group is composed of visual words in the corresponding cluster. Thirdly, group

Manuscript received February 13, 2015; revised June 8, 2015; accepted July 20, 2015. Date of publication July 28, 2015; date of current version August 14, 2015. This work was supported in part by the 973 Program under Grant 2012CB316400, in part by the National Natural Science Foundation of China under Grant 60903121, Grant 61173109, Grant 61332018, and Grant 61473231, in part by Microsoft Research Asia, and in part by the Fundamental Research Funds for the Central Universities. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ling Shao.

X. Qian is with the Key Laboratory for Intelligent Networks and Network Security, Ministry of Education, Xi'an Jiaotong University, Xi'an 710049, China, and also with the SMILES Laboratory, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: qianxm@mail.xjtu.edu.cn).

Y. Zhao is with the SMILES Laboratory, School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: zys@stu.xjtu.edu.cn).

J. Han is with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: junweihan2010@gmail.com).

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/TIP.2015.2462131

based spatial coding is conducted. We generate a position descriptor for each visual word.

The main contributions of this paper are summarized as following: 1) we propose a salient region mining and representation based image location estimation (image retrieval) approach. This approach makes full use of the saliency that explored from a group of visual words rather the individual visual words. Thus the visual word group based approach is more discriminative than the traditional visual word based approaches. Moreover, useful visual words selection is utilized, which can maintain the performance and but reduce the time cost efficiently. The small number of robust salient visual words is suitable for mobile end based image location estimation.

2) We propose a mean-shift based clustering approach to group visual words with a sizable number. We propose to generate a position descriptor for each visual word, which describes the spatial distribution in its group. The position descriptor is fusing both the related area and relative distance of the visual word to the visual word group. Thus it can bear the variation of scale, and partial occlusions to some extent.

3) We build fast inverted file structure for all images in the offline dataset to improve the efficiency. The index file records the images, visual word groups, position descriptors and image geographical location information (in short GPS information). When utilizing the built fast inverted file structure, the computational cost of the online search for an input image is less than 0.5% of the approach without indexing.

The rest of the paper is organized as follows: Firstly, related works on image location estimation (or image retrieval) are reviewed in Section II. Secondly, we provide the system overview in Section III. Finally, we give detailed description for our approach in Section IV, V, VI and VII. Experiments and discussions are given in Section VIII. In Section IX, the conclusion is drawn.

II. Related Work

Many methods are intended to estimate the geographic location of images. Our image location estimation is purely based on image content. We can convert the image location estimation problem as content based image retrieval [25], [55]–[57], [64] or object recognition problem [58]–[61]. In social media community, the shared photos are always attached with tags, time stamps, user's comments, and geo-coordinates that images are taken. Thus, image GPS can be estimated by combining both the textual descriptions and image content. The main process is as follows: firstly finding the similar images for the input image, and then assigning the visual similar images' GPS to that of the input image. From this point of view, the existing image retrieval approach can be utilized in image GPS estimation.

A. Content Based Image Retrieval

Bag-of-words image representation has been utilized for many multimedia and vision problems. Li et al. utilize multi-class SVM classifiers using bag-of-words for large scale image location estimation [2]. The computational costs are extremely high for model parameters' training. Also when the dataset is extended, the models need to be trained again. Han et al. propose an object-based image retrieval algorithm. They generate a feature descriptor based on context-preserving bag-of-words, and utilize a two-stage re-ranking technique to measure the similarity between the query image and each image in the dataset [27]. A mixture of multi-scale deformable part-based model is trained for each object category by training a latent support vector machine [3].

Quack et al. propose an approach to estimate the location of an image by utilizing the method of local feature matching [23]. The feature matching based GPS estimations approach is also very computational intensive when the scale of dataset is very large. To speed up the estimation process, user interaction is required to confine the locations of the input images to rough geographic area [23]. If the rough geographic area that the user assigned is with large error, then both the image GPS estimation performances and the computational cost results will be affected. Chum et al. propose an approach for estimating the location of the image by matching local feature [10]. And user interaction is required to confine the locations of the input images to really small ranges. Han *et al.* [19] reports a framework for effective image retrieval by employing memory learning. It forms a knowledge memory model to store the semantic information by simply accumulating user-provided interactions. Li et al. make full use of the visual similarity of the photos in the same places in terms of global feature and local feature [4]. They adopt a fast file index structure and use representative images for each GPS location to guarantee the estimation speed and accuracy.

The excellence of SIFT feature and BoW model have been manifested in image retrieval. However, there still exists deficiency in BoW model. For example, owing to the quantization loss, the visual word is not discriminative enough. Thus, many improved approaches are proposed to enhance the discrimination, e.g. visual synonyms [7], [34], [39], [40], [55], [56], embed geometry constraint [1], [8], [36], [55], [56], etc. The visual synonym can be acquired based on geometric coherence estimation. In [7], Gavves et al. define visual synonyms as pairs of independent visual words that could be mapped to each other in similar images via a trained homographic matrix. Spatial information [26], [37] can reinforce the discriminative power of single word. A paradigm of co-occurrence model is the spatial visual phrase model which describes the geometric information such as relative scale, orientation, Euclidean distance and the frequency that other words appear in the neighborhood of the specified word [37]. Zhou et al. proposed a spatial coding based image retrieval approach [8]. The spatial coding encodes the relative positions between each pair of features in an image. By using inverting construction, computational cost is low but with good retrieval performances. Zhang et al. propose a spatial coding based image retrieval approach by building the contextual visual vocabulary [31].

Saliency detection [50]–[54], [62], [63] is also useful for image semantic analysis such as auto image retargeting, image retrieval. Fu et al. introduce a new cluster-based algorithm for co-saliency detection [50]. Global correspondence between

the multiple images is implicitly learned during the clustering process. Three visual attention cues: contrast, spatial, and corresponding, are devised to effectively measure the cluster saliency. The final co-saliency maps are generated by fusing the single image saliency and multi-image saliency. Cao *et al.* [53] rebuilds the images and provides the reconstruction error regarded as a negative correlational value in co-saliency measurement. Feng et al. propose an indexing method for approximate nearest neighbor search of binary features [51]. They construct the hash keys by an online learning process instead of pure randomness. They obtain uniform hash buckets and high collision rates, which makes the method more efficient on approximate nearest neighbor search than LSH. Donoser et al. propose a method of matching interest points detected in the query image to a sparse 3D point cloud [52]. They project features to fern-specific embedding spaces, which yields improved matching rates in short runtime. The obtained correspondences are then used to recover a precise camera pose.

Yang et al. proposed to explore the contextual saliency information that mined from extended queries to improve image retrieval performances [55], [56]. They further rank the saliency for scalable mobile image retrieval with geometric consistency checking. They show that the contextual saliency can not only improve the performance, but also reduce the quantity of the data that needs transmitted from mobile end to cloud/server end.

Sparse coding compresses the original BoW histogram of query by reconstructing the it with linear combination of some bases [38], [41], [46], [48]. Thus the high dimensional BoW histogram is projected into a low dimensional vector via transformation matrix or dictionary. And some new technologies continuously emerge, such as domain-adaptive global feature descriptor [43], re-ranking schemes for dataset images [44]. Moreover, the database can be constructed with a 3D model. [6], [13], [17], and [20] are related to GPS location estimation using constructed 3D models from large scale geo-tagged photos.

B. Multi-Source Based Image Retrieval

The visual information in combination with the textual information is helpful in predicting which landmark in a given city is represented in an image. Laere et al. propose a two-step to geo-referencing tagged resources [29]. They first use language models to find an area which is likely to contain the location of the resource. Then, the location is determined by choosing the most similar resources in the second step. In the method, tags are taken into consideration to measure the similarity between the input and offline images. Hauff and Houben look beyond the single platform and investigate if location estimation can be improved when considering traces of the image owner on other social Web platforms [5]. They focus on user traces across the micro-blogging platform Twitter. With the development of mobile phones, mobile image retrieval draws attention recently. By utilizing the user's photo album, mobile image retrieval helps to narrow the gap between user's intent and the description of query in the way of interaction [32], [33]. Recently, the multi-model is proposed to improve the visual researches, e.g. [45] selects crucial features by analyzing the shared information among multiple tasks, and [47] generates multimodal spatial-temporal theme to describe landmarks better.

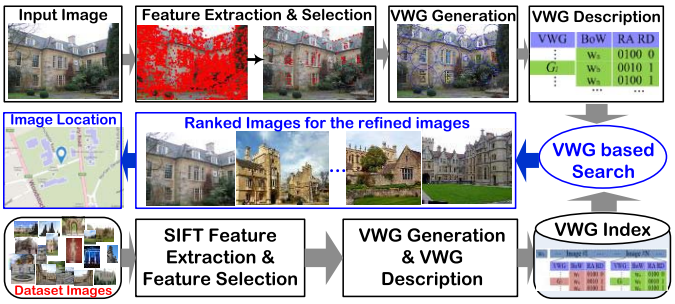

Fig. 1. Block diagram of the location estimation system.

Automatic annotation of video lacking of geographical data also gives us some insights. As for videos' visual content, they use the key frames of videos, and represent each frame by its visual features. Kelm et al. proposed a framework to geo-tag video using textual and visual information of shared media [30]. They make use of external resources like gazetteers to extract homonyms in the metadata. And visual and textual features are used to identify similar content. The videos' location is classified into possible regions by utilizing the method of fusion of visual and textual features. At the end, the Flickr videos are tagged with the geo-information of the most similar training image within the regions that is previously filtered by the probabilistic model for the test video.

III. System Overview

The system of our proposed image location approach is shown in Figure1. Firstly, we obtain refined locations of an input image. In this process, global feature clustering is utilized as that in our previous work [4], and refined locations are determined by cluster selection. Secondly, local feature of an image is in full use to refine the global feature clustering result. In our work, visual word groups (VWGs) are generated by mean-shift cluster. And we generate position descriptor for each visual word in the detected VWGs. Finally, we estimate the location of an input image by VWG and spatial consistency based image search.

IV. Refined Locations Generation

In the part, we introduce how to select the refined locations. In order to show the effectiveness of our proposed image location estimation approach, we utilize the same visual features in [4]. We also utilize the suggested parameters in [4] for global feature clustering.

A. Grouping Images Into Clusters

We cluster the image dataset using global features of the images. The global feature clustering is carried out on the 215D vector including 45D color moment feature and 170D texture feature [4]. Through global feature clustering,

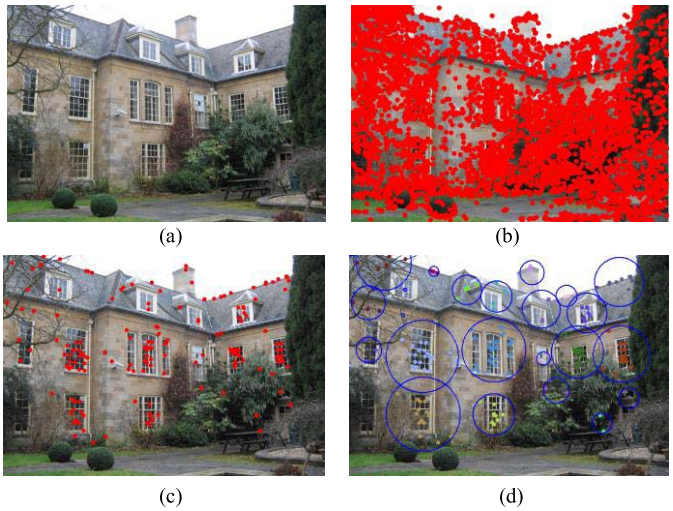

Fig. 2. For an input image (a), we extract raw SIFT points which are shown in (b). And useful visual words (c) are selected. The number of SIFT points of (a) is 4,002, while the number of useful words is only 169. Most of the salient visual words occur in the house, the main content of (a). In (d), we generate 23 VWGs, each with several useful visual words.

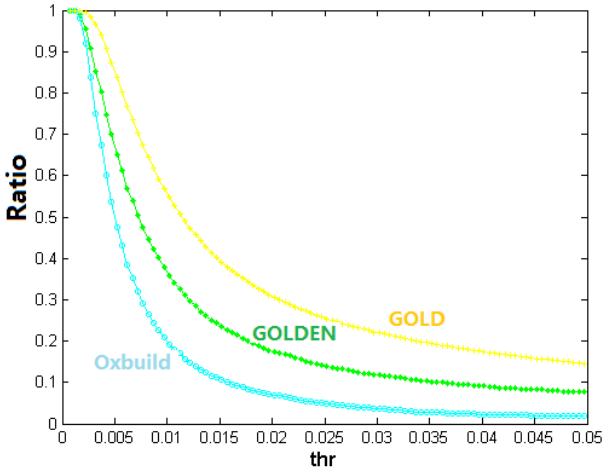

Fig. 3. The statistics on Oxbuild, GOLD and GOLDEN dataset. The distribution of scores between 0 and 0.05.

the whole dataset can be divided into several small scale groups. K-means clustering is utilized to divide the dataset into M small clusters, denoted as $C_n(n = 1, \ldots, M)$. In this paper, we set M to be 50 according to [4].

B. Cluster Selection

Let F_x denotes the 215D global features of the input image. Based on the obtained M clusters, we select candidate clusters for the input image according to the distance between F_x and M centers $C_{\mathrm{n}}(n = 1, \ldots, M)$, as that in [4].

The top ranked $S(S < M)$ clusters are selected. In this paper, we set S to be 15. We further obtain occurred locations of images in the selected clusters. The occurred locations are served as the refined locations of the input image.

V. Visual Word Group Building and Spatial Constraint

Based on the refined locations of the input image, we mine visual word group for refined images (selected images by global feature clustering), and enforce spatial constraint to improve image location estimation performance. The detailed process includes three steps: 1) SIFT feature extraction and useful feature selection, 2) visual word group building, and 3) position descriptor generation.

A. SIFT Feature Extraction and Useful Feature Selection

Here a BoW description is computed for the images. We represent each image by a set of visual words. For example, for the input image as shown in Figure2 (a), there are 4,002 SIFT points with their coordinates are shown in Figure2 (b).

However, not all the visual words are contributive to image location estimation. Different visual words have different weights of importance for identifying the query scene. Some visual words are non-distributive. As shown in Figure2 (b), the vast visual words often appear in the part of grass and trees, which are confusing for accurate location estimation. Thus it is rational to use some discriminative visual words rather than all the visual words. To mine useful features, we compute the score of each visual word while considering its frequency and the weight by employing a tf-idf weighting scheme. For an image, the score of a visual word w is computed as follows:

$$\mathrm{S_w} = \frac{\mathrm{f_w}}{\sum_{\mathrm{w}} \mathrm{f_w}} \times \log \frac{\mathrm{N}}{\mathrm{n_w}} \tag{1}$$

where f_w is the frequency of the w-th BoW in the image, n_w is the number of images containing the w-th BoW. The score reflects the important degree of BoW for its corresponding image and the entire image dataset. So we can select visual words based on the scores. We select the visual words whose scores are larger than thr as useful features.

In order to determine thr, a statistic based approach is utilized. The ratio curves of images which still have visual words left after the setting of different thresholds thr on three datasets: Oxbuild, GOLD, and GOLDEN are shown in Figure3. Oxbuild has 5K images, GOLD has 22.7K images selected from 3.3 M, and GOLDEN has 5.2M images [4]. The statistic on GOLDEN shows that when thr is equal to 0.001, all images in it have visual words left after selection. Of course, we can set thr arbitrary value which is less than 0.001. So in our experiments on OxBuild and GOLD, we set thr to be 0.001 considering the time cost. The selected visual words are called "useful features". The "useful" is embodied in maintaining the performance and saving the time cost. The comparison is shown in the latter discussion.

After useful feature selection, we pick out these high-frequency visual words for retrieval. As shown in Figure2(c), there are only 169 visual words left, which is far less than the raw SIFT features.

B. Visual Word Group Building

Bag-of-word model is often utilized in image retrieval. However it has limited discrimination power. To compensate

the shortcomings, geometric constrains are often adopted [1], [8], [36], [55], [56]. We aim at representing image by the salient area information rather than using single visual word. The main motivation of our approach is that the salient area information is far more robust than visual word. Thus, in this section, we group the BoW into visual word group (VWG) for each image.

For an image, we cluster the coordinates of its useful visual words by mean-shift clustering [28]. Usually, each SIFT point has a 128D descriptor vector and a 4D DoG key-point detector vector (x, y, scale, and orientation). Here the coordinates (x, y) of visual words are utilized. Let $v = \{(x_i, , y_i)\}_{i=1}^{h}$ denote the locations of the h SIFT points after useful feature selection. To $\forall v$, mean-shift is defined as follows:

$$\begin{cases} M_b(v) = \dfrac{1}{N_o} \sum_{v_i \in S_b(v)} v_i - v \\ S_b(v) = \{z : (z-v)^T (z-v) \leq b^2\} \end{cases} \tag{2}$$

where $S_b(v)$ is the region whose radius is b and whose centroid is v. N_o is the number of observations falling within $S_b(v)$ region. z represents the set of visual words falling within $S_b(v)$ region. b is the bandwidth parameter [28]. The large bandwidth means that the circle of geo-distance is large. So, if the bandwidth is large enough, we get only one VWG for an image. That is to say, all of the useful visual words are in the same VWG. If the bandwidth is too small, the method degenerates into conducting on single feature. At this circumstance, each BoW corresponds to a VWG, and then the VWG based search is degraded to the traditional BoW based approach.

After the clustering, we obtain several clusters and corresponding centers. A cluster is considered as a VWG, which is composed of the corresponding visual words in the cluster. So, the number of VWGs is equal to the number of clusters. Assuming that L VWGs are generated, we denote them as $G_l, l = 1, 2, \ldots, L$.

For the useful words shown in Figure2 (c), the corresponding VWGs after clustering are shown in Figure2 (d). Totally there are 23 VWGs. In order to visually display the VWGs, we mark a unique color for each VWG on Figure2 (d). For each VWG, we draw a blue circle. The center of the circle is the coordinates of the center of corresponding cluster. And the radius is the maximum of distances of center and words. After visual word group building, we represent an image by VWGs, each containing some visual words.

C. Position Descriptor Generation

Based on the obtained VWGs for an image, we further utilize the spatial information of visual words in each VWG to improve their discrimination power. We generate a position descriptor (PD) for each visual word to describe its distribution in the corresponding VWG. Assuming that a VWG G_l has n visual words which are denoted by $\{w_1, w_2, \ldots, w_n\}$), our position descriptor PD includes the following two aspects: 1) the relative area (RA), and 2) the relative distance (RD).

1) RA Representation: We set the center of a cluster as the center of the corresponding VWG. For a visual word, we record its relative area (RA) to the center of the VWG. We divide the VWG space into quadrants using its center as the origin of the quadrants. For each visual word w_i in the G_l, we record its relative spatial position against the origin. When the visual word is a bottom-right word, we define that its (RA) is [0 0 0 1]. For a visual word w_i, its position matrix is defined as follows.

$$RA_i = \begin{cases} [1\ 0\ 0\ 0], & \text{if } x_i > a_0, \quad y_i > b_0 \\ [0\ 1\ 0\ 0], & \text{if } x_i < a_0, \quad y_i > b_0 \\ [0\ 0\ 1\ 0], & \text{if } x_i < a_0, \quad y_i < b_0 \\ [0\ 0\ 0\ 1], & \text{if } x_i > a_0, \quad y_i < b_0 \end{cases} \tag{3}$$

where (x_i, y_i) is the coordinates of visual word w_i. (a_0, b_0) denotes the coordinates of the center of the VWG. Thus the RA of a visual word is a 4 bit descriptor, which shows its relative spatial distribution in a VWG. Actually, this 4 bit descriptor can be further compressed into a 2 bit vector.

2) RD Representation: We calculate the relative distance RD between the visual word and the center of the VWG. That is to say we want to know whether the distance between visual word and its corresponding center of ROI is large relatively.

We calculate the distance of each visual word w_i and the center by Euclidean distance, which is denoted by d_i. Meanwhile, we obtain the average distance of the VWG's visual words and the center as the denominator in Eq.(4). Then we compare a visual word's distance with the average distance. The relative distance of word w_i is calculated as follows:

$$\tilde{d}_i = \frac{d_i}{\frac{1}{n}\sum_{k=1}^{n} d_k} \tag{4}$$

For a visual word, if its distance d_i to the center is less than the average distance, i.e. the relative distance $\tilde{d}_i$ of a visual word is less than or equal to one, we think that the word is near to the center, otherwise, we think that the word is relatively far away from the center. Thus, we represent the quantized relative distance RD_i of the visual word w_i as follows.

$$RD_i = \begin{cases} 0, & if\ \tilde{d}_i \leq 1 \\ 1, & if\ \tilde{d}_i > 1 \end{cases} \tag{5}$$

According to the method described above, RD is a 1-bit spatial descriptor, which reflects the visual word's distance is relatively far from the center of the region or not. For a visual word w_i its position descriptor PD_i combines both relative area RA_i and relative distance RD_i. Thus PD_i is a five bits descriptor.

VI. Image Indexing

We build inverted file structure for all images in the offline dataset. Our BoW based image indexing is shown in Figure4. As for the BoW w_x, the images it belonging to are recorded. And correspondingly, the GPS location of image #I which is denoted by $Label_I$ is all recorded in another image-location inverted file list. Different image contains various number of VWGs $G_l, l = 1, 2, \ldots, L$. So, the VWG G_l visual word w_x belonging to is recorded too. Besides, we need to record the relative position descriptor of visual word. Therefore, position descriptor of visual word w_x including a four bits RA and one bit RD in the VWG G_l of the image #I is also recorded.

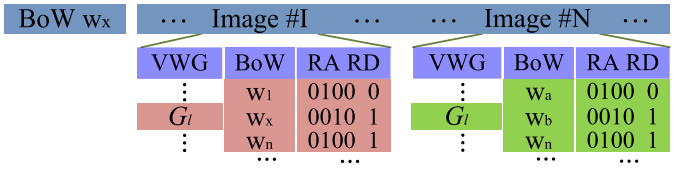

Fig. 4. Illustration of Inverted file structure for the offline image set. G_l is the l-th VWG that w_x belongs to. RA and RD are the position descriptors of w_x in the VWG G_l.

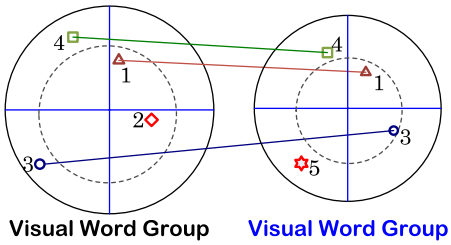

Fig. 5. VWG based image matching. The dotted line is a circle, whose radius is the average distance of all visual words in the VWG.

VII. VWG Based Image Search

An image retrieval method based on the VWG and the spatial geometric consistency is presented in this section. In the offline system, the SIFT points and VWGs are detected. And the PDs of visual words are calculated. In the online system, we extract SIFT features for the input image. The VWGs and the PD vectors are generated. Then we introduce how to calculate the similarity between the input image #q and the refined image #r. The process includes the following two steps.

A. Matched Group Pair Detection

In this section, we find the matched group pairs (MGPs) for the query image #q and the refined image #r. For each useful BoW occurred in the query image #q, we use the obtained inverted files to find refined images which contain the same BoW. The VWG that the visual word belonging to is also obtained. Let G_i^q denotes the i-th VWG of image #q. And G_j^r is denoted as the j-th VWG of image #r. We call the two VWGs as a MGP if they contain common visual words.

We iteratively search all visual words in the query image to obtain the total number of MGPs between the input image and refined image. Assuming that the input image #q and the refined image #r have m MGPs, we denote them as P_i, $i = 1, 2, \ldots, m$. However, this constraint is relatively weak. We further compare the position descriptors in each MGP to make constrain strong. The matching score of each MGP is calculated from their common visual words and their corresponding PDs. As shown in Figure5, the MGP has three common visual words, i.e. #1, #3 and #4. Correspondingly, we can get the PDs for the visual words in the VWGs. The solid line circle denotes the region by mean-shift clustering, while for the dotted line circle, its radius is the average distance of all visual words in the VWG. For example, the RAs of the #1 in the two VWG are 10000, the RDs of them are both 0, and their position descriptors are both 10000. While for the visual word #3, the RAs are respectively 0000 and 0001, and their corresponding RDs are 1 and 0. Then we compare their corresponding position descriptors to determine their matching score.

For a MGP P_i, let G_i^q and G_i^r denote the VWGs from the input image #q and the refined image #r. Assuming that G_i^q and G_i^r have m common visual words, their corresponding position descriptors are denoted as $PD_q^k, (k = 1, 2, \ldots, m)$ and $PD_r^k, (k = 1, 2, \ldots, m)$ respectively. We verify the spatial layout of the common visual words to determine their matching score (denoted by S_r^k) as the following:

$$S_r^k = 1 - \frac{1}{a} \sum_{k=1}^{m} PD_q^k \oplus PD_r^k \tag{6}$$

where $\oplus$ is Logical Exclusive (XOR) operation.

The larger S_r^k means that spatial consistent score of the MGP is higher. That is to say the two images are more similar. If some parts of two images match well, then we can find them by our approach. It has better performance than that way of considering the entire content of an image. Moreover, our VWG is unbounded with its position and shape.

B. Similarity Measurement

Based on that the input image #q and the refined image #r have m MGPs, we obtain m values, which are denoted by $S_r^j, (j = 1, 2, \ldots, m)$. In this paper, the maximum of S_r^j, $(j = 1, 2, \ldots, m)$ is selected as the score of the refined image to the input image. The fundamental thought is that the best matched MGPs in two images are more likely to be with high visual similarity. Thus in this paper, we utilize the best matched pairs to represent the similarity of query image #q and refined image #r as follows.

$$Score(r) = \max_j(S_r^j), \quad j = 1, 2, \ldots, m \tag{7}$$

This kind of similarity measurement approach is robust to the variations of image, such as rotation, scaling etc. In our experiments other kind of similarity measurement approaches are evaluated including using the average of the matching scores to show the effectiveness. Thus, we obtain the scores of all refined images to the input image. Then the refined images are ranked according to their spatial consistency with the input image.

In order to improve the final location estimation performance, k-nn based approach is also utilized in location estimation for the input image. It is very likely that images taken from one certain place can be distributed into different clusters due to the various appearances of the images [4], [15]. And images from different locations sometimes are similar. So the k-nn is necessary for improving the location estimation performance. The top ranked k images are selected. And then we count the number of images for each occurred location. The majority location in the k images is assigned for the input image. In this paper, we utilize $k = 50$ as suggested in [4] and [15].

VIII. Experiments and Discussions

In order to test the performance of the proposed location estimation approach, comparisons are made with IM2GPS [9],

hierarchical global feature clustering and local feature refinement under cosine similarity based measurement (denoted as CS) [4], spatial coding based approach (denoted as SC) [8], method of salient region mining using maximally stable extremal region (denoted as MSER) [35], method of adopting word spatial arrangement for an image (denoted as WSA) [24] and ours (denoted as VWG).

In the compared approach MSER, the input is the same as VWG, i.e. the useful features after selection. The only difference is that we utilize maximally stable extremal region [35] rather than that of our mean-shift based region detection approach (visual word group building approach as described in Part B of Section V) to bundle visual words into groups. The other parts are the same with our method. The goal is to show the effectiveness of our mean-shift based salient group detection description approach. Similarly, in the compared approach WSA, the other parts of WSA are also the same as these of VWG. The only difference is that, in WSA, we adopt word spatial arrangement [24] rather than the position descriptor in VWG. We do this is to show the effectiveness of our position description generation approach.

Experiments are carried out on a PC with 48G memory and Intel® Core(TM)2, Quad CP Q8400 with 2.26GHz on Matlab.

A. Datasets

Experiments are done on two datasets: OxBuild and GOLD. OxBuild is used for preliminary tests. In order to show its effectiveness for large scale dataset with more locations, our approach is tested on GOLD. All the experiments are performed on the same environment.

The categories of OxBuild are served as locations. Thus, the GPS numbers of OxBuild is 11. 100 images are selected randomly from the whole dataset as the test set, while the rest is served as training set in the offline system.

GOLD contains more than 3.3 million images together with their Geo-tags [4]. And it covers more than 65K places in the world. It is crawled from Flickr using its public API. 80 travel spots are selected for testing, i.e. the number of locations is 80. The test dataset for the 80 sites contains randomly selected 5000 images.

B. Performance Evaluation

For an input image, if the estimated location is exact with its ground-truth location, it is correctly estimated, otherwise falsely estimated. We utilize the average recognition rate (AR) to evaluate the performance of image location estimation performance which is given as follows:

$$\mathrm{AR} = \frac{1}{\mathrm{G}} \sum_{\mathrm{i}=1}^{\mathrm{G}} \mathrm{RR_i} \tag{8}$$

where G is the number of locations. It is 11 and 80 for OxBuild and GOLD respectively. RR_i is the recognition rate of the i-th location, which is defined as follows:

$$\mathrm{RR_i} = \frac{\mathrm{NC_i}}{\mathrm{NI_i}} \times 100\%, \quad \mathrm{i} \in \{1, 2, \ldots, G\} \tag{9}$$

where NC_i is the correct estimated image number, NI_i is the test image number.

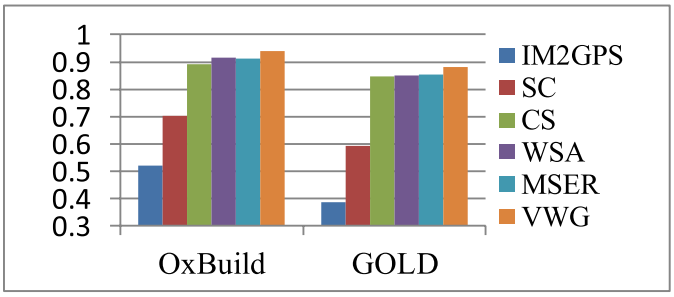

Fig. 6. AR values of IM2GPS, SC, CS, WSA, MSER, and VWG on the two dataset OxBuild and GOLD.

TABLE I
AVERAGE COMPUTATIONAL COSTS OF SC, IM2GPS, CS, WSA, MSER AND OUR APPROACH VWG ON OXBUILD AND GOLD

Dataset	SC	IM2GPS	CS	WSA	MSER	VWG
OxBuild	5.42 ms	33.74 ms	0.47 ms	0.40s	1.54s	0.38s
GOLD	47.0 ms	64927 ms	0.96 ms	0.64s	2.01s	0.62s

The location estimation performances of IM2GPS, SC, CS, MSER, WSA and VWG are shown in Figure6. The average computational costs of different methods on the two test sets are shown in Table1. We find that that our method VWG outperforms the other methods on the two datasets considering the performance and the time cost.

The AR values of our VWG on OxBuild and GOLD are 93.85% and 88.16% respectively. The results of IM2GPS in the two test datasets are 52.15% and 38.81% respectively, which are with lowest performance. The average recognition rates of SC are 70.39% and 59.48%. While the results of our previous approach CS are 89.27% and 84.86% respectively. The performance of CS is better than IM2GPS and SC. We can conclude that both image global and local visual features are beneficial in image location estimation.

The method MSER bundle visual words into groups. The average recognition rates of MSER are 91.12% and 85.47%. We divide the useful visual words of an image into VWGs by mean-shift clustering. This shows that our salient region mining approach is effective. In WSA, word spatial arrangement is utilized to encode the distribution of a useful visual word in an image. And the average recognition rates of WSA are 91.63% and 85.01%. Its results are not as good as our cluster center based spatial coding in VWG.

C. Discussion

In this section we give a comprehensive discussion of the impacts of the bandwidth b and k in k-nn to the final image location estimation performances. The parameters in our baseline algorithm are set as $k = 50$, and $b = 60$, and the size of BoW is set to be 60K [57]. We also discuss the impacts of using useful features or not, the selection of VWG's center and the different spatial constraint approaches.

1) The Impact of Using Useful Features: In the part of local feature refinement, we select useful features of images by tf-idf based approach, rather than the repeated visual words extracted from multiple queries [55]–[57]. For images, their

TABLE II
AR (%) OF USING ALL FEATURES AND USEFUL FEATURES OF IMAGES

Dataset	All features	Useful features
OxBuild	92.54	93.85
GOLD	86.97	88.16

TABLE III
THE COMPARISON OF AVERAGE COMPUTATIONAL COSTS (s)

Dataset	All features		Useful features	
	with indexing	without indexing	with indexing	without indexing
OxBuild	0.79	162.14	0.38	107.72
GOLD	1.02	236.04	0.62	183.06

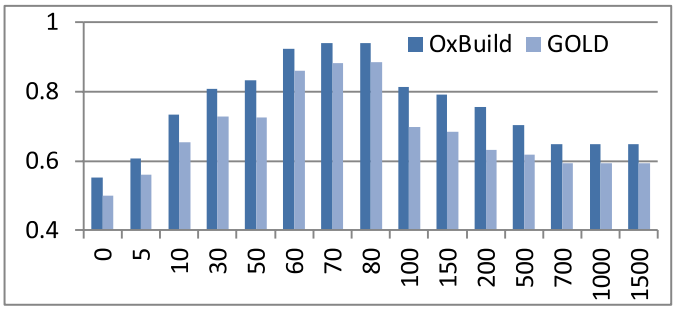

Fig. 7. Impact of bandwidth b (x-axis) to image location estimation performance.

visual words have different weights for location estimation. The comparison of using all features and useful features is shown in Table2, from which we find that the performance of using useful features is a little better than using the all the raw features.

In order to show the efficiency of the using useful feature and fast image indexing approach, we provide a complete comparison with using all features and using useful features under the cases with image indexing and without indexing. The corresponding computational costs (in second) are shown in Table3 respectively. We find that when utilizing useful features the computational cost is about 75% of that of using all features. Moreover, we find that when building fast indexing structure for dataset images, the algorithm is very efficient. The computational cost of utilizing fast indexing structure is less than 0.5% of that of without indexing.

2) The Impact of Bandwidth: In the section of VWGs generation, we cluster the useful words by Mean-shift cluster. After clustering, we obtain several clusters and corresponding centers. A cluster is considered as a VWG, which is composed of all visual words in the cluster. So, the multi-VWG generation is closely connected with the bandwidth b. Here, we discuss the impact of bandwidth b to image location estimation performance. If the bandwidth is too large, the circle of geo-distance is large. So, visual words with large distance are in the same VWG. If the bandwidth is too small, related visual words could not be cluster to the same VWG. Figure7 shows that with the increase of b, the AR is first increasing and then into decline. b is set 70 in our experiments.

TABLE IV
AR (%) UNDER DIFFERENT REFERENCE CENTER SETTING APPROACHES

Dataset	Center	Mean
OxBuild	93.85	94.02
GOLD	88.16	87.57

TABLE V
AR (%) OF USING RA OR RD RESPECTIVELY

Dataset	RA	RD	PD
OxBuild	91.52	90.65	93.85
GOLD	84.70	85.17	88.16

TABLE VI
AVERAGE RECOGNITION RATES (%) OF USING THE MAXIMUM AND AVERAGE OF MATCHING SCORES

Dataset	Max	Average
OxBuild	93.85	92.93
GOLD	88.16	87.74

Fig. 8. Four query exemplars.

From Figure7, we find that when $b > 700$, the performance doesn't change much. The slit differences are caused by the initial centers of mean-shift clustering. Because when the bandwidth is large enough, we get only one VWG for an image. All of its useful visual words are in the same VWG. Only the center may be different. So the performances are not the same. The bandwidth $b = 0$ is denotes that each visual word corresponds to a VWG. In this case, the VWG based image search is identical to the traditional BoW based image search.

3) The Impact of Different Reference Center Setting Approach: When calculating the position descriptor for a visual word, we set the center of a cluster (in short Center), which is determined by mean-shift clustering, as the center of the corresponding VWG. Based on the reference center, then we record the relative area (RA) in relation to the center of the VWG and calculate the relative distance (RD) between the visual word and the center of the VWG. We also conduct an experiment by setting the mean of coordinates of a VWG's visual words as the VWG's center (in short Mean). The corresponding comparisons are shown in Table4 respectively. From Table4 we find that different center determination approaches do not influence the final performance very much.

4) The Impact of Different Spatial Constrains: In our experiments, to improve the discrimination power of BoW, we further utilize the spatial information of BoW in each VWG. We generate a position descriptor PD for each visual word

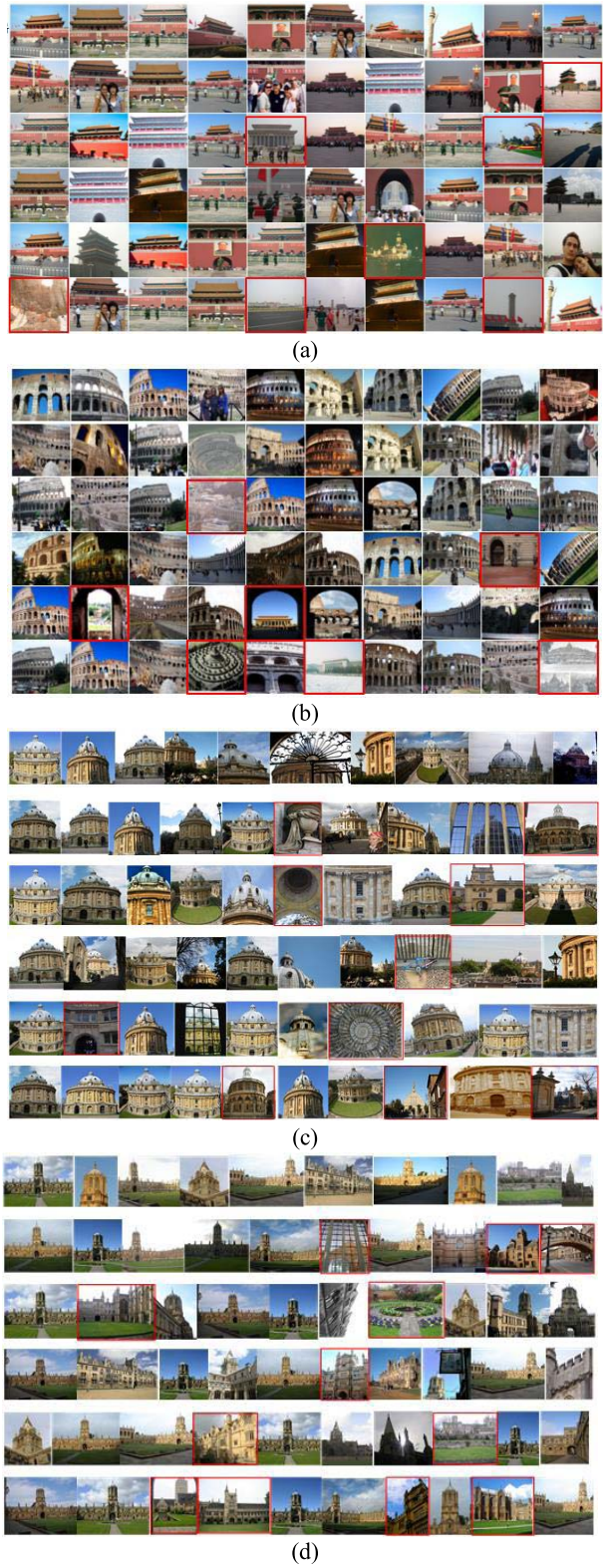

Fig. 9. Top 10 Ranking results of the six different methods for four queries as shown in Fig.8. Methods of the first raw to the sixth raw are 1) VWG, 2) SC, 3) CS, 4) WSA, 5) MSER, 6) IM2GPS. Images in red frames are the error results. (a) The results of the first query. (b) The results of the second query. (c) The results of the third query. (d) The results of the fourth query.

to describe its distribution in the corresponding VWG. Our position descriptor includes two aspects: the relative area RA and the relative distance RD. In the later image searching, we compare their position descriptors in each MGP. Here we discuss the impact of using RA or RD respectively to image location estimation performance. The corresponding results are shown in Table5 respectively. We find that combining both RA and RD better performances are achieved.

5) The Impact of Different Similarity Measurement Approach: In our similarity measurement, we utilize the best matched pairs to represent the similarity of query image #q and refined image #r by Eq.(7). The fundamental thought is that the best matched MGP is more likely to represent the similarity of the two images. The advantage of this similarity measurement approach can improve the partial overlapping content matching. In order to show the effectiveness of this approach, we compare it with the average based approach (in short Average). In this approach, we determine the similarity by averaging the total matched visual word groups as follows:

$$Score(r) = \frac{1}{m}\sum_{j=1}^{m} S_r^j \tag{10}$$

From Eq.(10), we find that the average value relatively weakens the good matching performance of some groups but not too serious. The corresponding results are shown in Table6 respectively. The reason that performances of Max as shown in Eq.(7) and Average based similarity measurement approaches are very close is due to follow two aspects: 1) matched group pair guarantees content overlapping. The matched group pair at least sharing one common visual word; 2) effective region representation approach. Our proposed position descriptor models the relative geometric information and relative distribution information.

D. Subjective Retrieval Results

In order to show the effectiveness of the proposed VWG based image location estimation performances, we provide the top ranked 10 results of four example images as shown in Figure8. The corresponding retrieval results of each query of the six compared approaches: 1) VWG, 2) SC, 3) CS, 4) WSA, 5) MSER, and 6) IM2GPS are shown from the first raw to the sixth raw respectively in Figure9 (a)∼(d). The irrelevant images to the query are marked out by red frames. From the above comparisons we find that the salient region based approaches VWG, WSA and MSER achieve better performances.

IX. Conclusion

In this paper, we propose a salient region mining and representation based image location estimation approach. The saliency that explored from a group of visual words is far more discriminative than the individual visual words in image retrieval. Mean-shift based clustering approach is proposed to group visual words with a sizable number. The proposed visual word group mining based image search is robust to find similar images even with partial occlusion. Useful feature representation does not improve the image location estimation performances but can save a quarter of computational costs.

We propose to generate a position descriptor for each visual word in each visual word group by fusing both the related area

and relative distance information. We build fast inverted file structure for dataset images by recording the images ID, visual word groups, position descriptors and image geographical location information. The fast indexing structure can reduce the computational cost dramatically. When utilizing the built fast inverted file structure, the computational cost of the online search for an input image is less than 0.5% of the approach without indexing.

REFERENCES

[1] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li, "Descriptive visual words and visual phrases for image applications," in *Proc. 17th ACM MM*, 2009, pp. 75–84.

[2] Y. Li, D. J. Crandall, and D. P. Huttenlocher, "Landmark classification in large-scale image collections," in *Proc. IEEE 12th ICCV*, Sep./Oct. 2009, pp. 1957–1964.

[3] G. Cheng *et al.*, "Object detection in remote sensing imagery using a discriminatively trained mixture model," *ISPRS J. Photogram. Remote Sens.*, vol. 85, pp. 32–43, Nov. 2013.

[4] J. Li, X. Qian, Y. Y. Tang, L. Yang, and T. Mei, "GPS estimation for places of interest from social users' uploaded photos," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 2058–2071, Dec. 2013.

[5] C. Hauff and G.-J. Houben, "Placing images on the world map: A microblog-based enrichment approach," in *Proc. 35th ACM SIGIR*, 2012, pp. 691–700.

[6] 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. 20th ACM Int. Conf. Multimedia*, 2012, pp. 9–18.

[7] E. Gavves, C. G. M. Snoek, and A. W. M. Smeulders, "Visual synonyms for landmark image retrieval," *Comput. Vis. Image Understand.*, vol. 116, no. 2, pp. 238–249, 2012.

[8] W. Zhou, Y. Lu, H. Li, Y. Song, and Q. Tian, "Spatial coding for large scale partial-duplicate Web image search," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 511–520.

[9] J. Hays and A. A. Efros, "IM2GPS: Estimating geographic information from a single image," in *Proc. IEEE Conf. CVPR*, Jun. 2008, pp. 1–8.

[10] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *Proc. IEEE 11th ICCV*, Oct. 2007, pp. 1–8.

[11] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 2006, pp. 2169–2178.

[12] X. Qian *et al.*, "HWVP: Hierarchical wavelet packet descriptors and their applications in scene categorization and semantic concept retrieval," *Multimedia Tools Appl.*, vol. 69, no. 3, pp. 897–920, 2014.

[13] G. Baatz, K. Köser, D. Chen, R. Grzeszczuk, and M. Pollefeys, "Leveraging 3D city models for rotation invariant place-of-interest recognition," *Int. J. Comput. Vis.*, vol. 96, no. 3, pp. 315–334, Feb. 2012.

[14] Y. Xue and X. Qian, "Visual summarization of landmarks via viewpoint modeling," in *Proc. 19th IEEE ICIP*, Sep./Oct. 2012, pp. 2873–2876.

[15] J. Li, X. Qian, Y. Y. Tang, L. Yang, and C. Liu, "GPS estimation from users' photos," in *Proc. 19th Int. Conf. MMM*, 2013, pp. 118–129.

[16] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.

[17] C. Wu, F. Fraundorfer, J.-M. Frahm, and M. Pollefeys, "3D model search and pose estimation from single images using VIP features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2008, pp. 1–8.

[18] Y.-T. Zheng *et al.*, "Tour the world: Building a Web-scale landmark recognition engine," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 1085–1092.

[19] J. Han, K. N. Ngan, M. Li, and H.-J. Zhang, "A memory learning framework for effective image retrieval," *IEEE Trans. Image Process.*, vol. 14, no. 4, pp. 511–524, Apr. 2005.

[20] M. Park, J. Luo, R. T. Collins, and Y. Liu, "Beyond GPS: Determining the camera viewing direction of a geotagged image," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 631–634.

[21] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *Proc. 18th Int. Conf. WWW*, 2009, pp. 761–770.

[22] E. Kalogerakis, O. Vesselova, J. Hays, A. A. Efros, and A. Hertzmann, "Image sequence geolocation with human travel priors," in *Proc. IEEE 12th Int. Conf. CVPR*, Sep./Oct. 2009, pp. 253–260.

[23] T. Quack, B. Leibe, and L. Van Gool, "World-scale mining of objects and events from community photo collections," in *Proc. Int. Conf. CIVR*, 2008, pp. 47–56.

[24] O. A. B. Penatti, F. B. Silva, E. Valle, V. Gouet-Brunet, and R. da Silva Torres, "Visual word spatial arrangement for image retrieval and classification," *Pattern Recognit.*, vol. 47, no. 2, pp. 705–720, 2014.

[25] Y. Zhao, X. Qian, and T. Mu, "Image taken place estimation via geometric constrained spatial layer matching," in *Proc. 21st Int. Conf. MMM*, 2015, pp. 436–446.

[26] J. Yuan, Y. Wu, and M. Yang, "Discovery of collocation patterns: From visual words to visual phrases," in *Proc. IEEE Conf. CVPR*, Jun. 2007, pp. 1–8.

[27] J. Han, M. Xu, X. Li, L. Guo, and T. Liu, "Interactive object-based image retrieval and annotation on iPad," *Multimedia Tools Appl.*, vol. 72, no. 3, pp. 2275–2297, 2014.

[28] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Trans. Inf. Theory*, vol. 21, no. 1, pp. 32–40, Jan. 1975.

[29] O. Van Laere, S. Schockaert, and B. Dhoedt, "Ghent University at the 2010 placing task," in *Proc. MediaEval Workshop*, 2010, pp. 1–2.

[30] P. Kelm, S. Schmiedeke, and T. Sikora, "How spatial segmentation improves the multimodal geo-tagging," in *Proc. MediaEval Workshop*, 2012, pp. 1–2.

[31] S. Zhang, Q. Huang, G. Hua, S. Jiang, W. Gao, and Q. Tian, "Building contextual visual vocabulary for large-scale image applications," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 501–510.

[32] H. Li, Y. Wang, T. Mei, J. Wang, and S. Li, "Interactive multimodal visual search on mobile device," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 594–607, Apr. 2013.

[33] J. Sang, T. Mei, Y.-Q. Xu, C. Zhao, C. Xu, and S. Li, "Interaction design for mobile visual search," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1665–1676, Nov. 2013.

[34] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 2006, pp. 2161–2168.

[35] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate Web image search," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 25–32.

[36] J. Chen, B. Feng, L. Zhu, P. Ding, and B. Xu, "Effective near-duplicate image retrieval with image-specific visual phrase selection," in *Proc. 19th IEEE ICIP*, Sep./Oct. 2010, pp. 1909–1912.

[37] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE ICCV*, Oct. 2003, pp. 1470–1477.

[38] R. Ji, L.-Y. Duan, J. Chen, and W. Gao, "Towards compact topical descriptors," in *Proc. IEEE Conf. CVPR*, Jun. 2012, pp. 2925–2932.

[39] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. IEEE Conf. CVPR*, Jun. 2008, pp. 1–8.

[40] W. Tang, R. Cai, Z. Li, and L. Zhang, "Contextual synonym dictionary for visual object retrieval," in *Proc. 19th ACM MM*, 2011, pp. 503–512.

[41] X. Yang, L. Liu, X. Qian, T. Mei, J. Shen, and Q. Tian, "Mobile visual search via hierarchical sparse coding," in *Proc. IEEE ICME*, Jul. 2014, pp. 1–6.

[42] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[43] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 723–742, Apr. 2012.

[44] L. Shao, L. Liu, and X. Li, "Feature learning for image classification via multiobjective genetic programming," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 7, pp. 1359–1371, Jul. 2014.

[45] Y. Yang, Z. Ma, A. G. Hauptmann, and N. Sebe, "Feature selection for multimedia analysis by sharing information among multiple tasks," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 661–669, Apr. 2013.

[46] C. Yang, J. Shen, J. Peng, and J. Fan, "Image collection summarization via dictionary learning for sparse representation," *Pattern Recognit.*, vol. 46, no. 3, pp. 948–961, 2013.

[47] W. Min, B.-K. Bao, and C. Xu, "Multimodal spatio-temporal theme modeling for landmark analysis," *IEEE Multimedia*, vol. 21, no. 3, pp. 20–29, Jul./Sep. 2014.

[48] J. Huang, H. Liu, J. Shen, and S. Yan, "Towards efficient sparse coding for scalable image annotation," in *Proc. 21st ACM MM*, 2013, pp. 947–956.

[49] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. 9th ECCV*, 2006, pp. 404–417.

[50] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3766–3778, Oct. 2013.
[51] Y. Feng, Y. Wu, and L. Fan, "Online learning of binary feature indexing for real-time SLAM relocalization," in *Proc. ACCV Workshop Big Data 3D Comput. Vis.*, 2015, pp. 206–217.
[52] M. Donoser and D. Schmalstieg, "Discriminative feature-to-point matching in image-based localization," in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 516–523.
[53] X. Cao, Y. Cheng, Z. Tao, and H. Fu, "Co-saliency detection via base reconstruction," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 997–1000.
[54] C. Li, J. Xue, N. Zheng, and Z. Tian, "Nonparametric bottom-up saliency detection using hypercomplex spectral contrast," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 1157–1160.
[55] X. Yang, X. Qian, and Y. Xue, "Scalable mobile image retrieval by exploring contextual saliency," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1709–1721, Jun. 2015.
[56] X. Yang, X. Qian, and T. Mei, "Learning salient visual word for scalable mobile image retrieval," *Pattern Recognit.*, vol. 48, no. 10, pp. 3093–3101, 2015.
[57] X. Qian, Y. Xue, X. Yang, Y. Y. Tang, X. Hou, and T. Mei, "Landmark summarization with diverse viewpoints," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.
[58] L. Liu, M. Yu, and L. Shao, "Multiview alignment hashing for efficient image search," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 956–966, Mar. 2015.
[59] F. Zhu and L. Shao, "Weakly-supervised cross-domain dictionary learning for visual recognition," *Int. J. Comput. Vis.*, vol. 109, nos. 1–2, pp. 42–59, Aug. 2014.
[60] L. Zhang, X. Zhen, and L. Shao, "Learning object-to-class kernels for scene classification," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3241–3253, Aug. 2014.
[61] F. Zhu, Z. Jiang, and L. Shao, "Submodular object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 2457–2464.
[62] J. Han, S. He, X. Qian, D. Wang, L. Guo, and T. Liu, "An object-oriented visual saliency detection framework based on sparse coding representations," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 12, pp. 2009–2021, Dec. 2013.
[63] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.
[64] Y. Zhao and X. Qian, "Spatial constraint for image location estimation," in *Proc. ICMR*, 2015, pp. 515–518.

Xueming Qian (M'10) received the B.S. and M.S. degrees from the Xi'an University of Technology, Xi'an, China, in 1999 and 2004, respectively, and the Ph.D. degree from the School of Electronics and Information Engineering, Xi'an Jiaotong University, in 2008. He was an Assistant Professor. He was an Associate Professor from 2011 to 2014, and then a Full Professor. He is the Director of the SMILES Laboratory. He was a Visiting Scholar with Microsoft research Asia from 2010 to 2011. His research interests include social media big data mining and search. His research is supported by NSFC, the Microsoft Research, and MOST. He received the Microsoft Fellowship in 2006, and the Outstanding Doctoral Dissertations of Xi'an Jiaotong University and Shaanxi Province in 2010 and 2011, respectively.

Yisi Zhao is currently pursuing the M.S.D. degree with the SMILES Laboratory, Xi'an Jiaotong University, Xi'an, China. Her research interests include large-scale image retrieval and image content understanding.

Junwei Han received the B.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 1999 and 2003, respectively. He was a Visiting Student with Microsoft Research Asia and a Visiting Researcher with the University of Surrey. He has been a Research Fellow with Nanyang Technological University, the Chinese University of Hong Kong, Dublin City University, and the University of Dundee. He is currently a Professor with Northwestern Polytechnical University. His research interests include computer vision and multimedia processing.