Scalable Mobile Image Retrieval by Exploring Contextual Saliency

Xiyu Yang, Xueming Qian, *Member, IEEE*, and Yao Xue

***Abstract*— Nowadays, it is very convenient to capture photos by a smart phone. As using, the smart phone is a convenient way to share what users experienced anytime and anywhere through social networks, it is very possible that we capture multiple photos to make sure the content is well photographed. In this paper, an effective scalable mobile image retrieval approach is proposed by exploring contextual salient information for the input query image. Our goal is to explore the high-level semantic information of an image by finding the contextual saliency from multiple relevant photos rather than solely using the input image. Thus, the proposed mobile image retrieval approach first determines the relevant photos according to visual similarity, then mines salient features by exploring contextual saliency from multiple relevant images, and finally determines contributions of salient features for scalable retrieval. Compared with the existing mobile-based image retrieval approaches, our approach requires less bandwidth and has better retrieval performance. We can carry out retrieval with <200-B data, which is <5% of existing approaches. Most importantly, when the bandwidth is limited, we can rank the transmitted features according to their contributions to retrieval. Experimental results show the effectiveness of the proposed approach.**

***Index Terms*— Mobile image retrieval, salient visual vocabulary pair, spatial layout descriptor, scalable image retrieval, multiple queries.**

I. Introduction

NOWADAYS, mobile phones have been immensely pervasive. According to the statistics, there are 4.5 billion mobile phones and 1.7 billion smart phone users in the world in 2014. Mobile phones have been indispensable for most people, especially the young. They tend to use smart phones for doing many things, such as sharing photos, inquiring bus route, surfing the Internet and so on. People are so dependent on smart phones that they hope to use smart phones to handle as more things as possible. Hence, image retrieval has to be applied to mobile end as well, e.g. to look at a restaurant's environment or to search scenic spots. Mobile image retrieval is particularly effective to help users search unknown objects. For example, when a user sees an unknown building and wants to know more about it, he can take photos of the building, and searches it on the internet. Thus the user can acquire information about the building from the descriptions of its similar images.

For mobile image retrieval, the poor condition of wireless channel is a big challenge [20], [22]. Both the limited bandwidth and instability of channel need to be considered [21]–[23]. Typically, mobile image retrieval is based on text. Some search engines, like Google, could provide thousands of relevant images successfully when a user inputs a textual query. Text based image retrieval depends on the images' tags, labels or related text to a large extent. Nevertheless, numerous images may not have tags, and the text depiction about images is not always accurate. With the mobile cameras at hand, it is convenient to adopt content based image retrieval. Recently, researchers managed to achieve visual search in mobile end. The state of art focuses on extracting more compact descriptor [17], [18] or compressing the BoW (bag-of-word) histogram [19]–[22]. In most cases, the BoW histogram is transmitted in compact form to reduce the volume of data. However, BoW representation has its innate deficiencies. Firstly, it contains quantization loss that results in synonym and polysemy phenomenon. Secondly, the BoW representation neglects the rich spatial relationship among the visual words as well. Thirdly, it does not have enough descriptive power. Due to the quantization loss, a single visual word cannot describe local image region exactly and discriminatingly.

The existing retrieval systems usually require a single query image [17]–[22]. For this case, on one hand, the system rarely achieves good performance if the object in the query cannot be seen clearly. On the other hand, there are too many local interest points (such as SIFT feature points) detected in an image. A large part of the interest points are noise or irrelevant to the crucial object of the image, which increases the computational complexity. And unstable noisy local features have indeed negative effect on retrieval performance. Query expansion (QE) [15] updates the original query by combining it with retrieval results to extend the query and weaken the disturbance of noise. The goal of QE is to explore more relevant photos iteratively to eliminate the deficiency of single image based retrieval.

Usually, we may take multiple photos at a scene to make sure at least one photo is satisfying. That is to say, in mobile image retrieval, a user may take many relevant photos before sending a desired one to the server terminal to carry out

Manuscript received July 22, 2014; revised December 13, 2014; accepted February 16, 2015. Date of publication March 9, 2015; date of current version March 23, 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 61173109, Grant 60903121, and Grant 61332018, and in part by Microsoft Research. The associate editor coordinating the review of this manuscript and approving it for publication was Mr. P.-M. Jodoin. *(Correspond author: Xueming Qian.)*

The authors are with the SMILES Laboratory, Xi'an Jiaotong Univeristy, Xi'an 710049, China (e-mail: yangxiyu@stu.xjtu.edu.cn; qianxm@mail.xjtu.edu.cn; xy_again@sina.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.2015.2411433

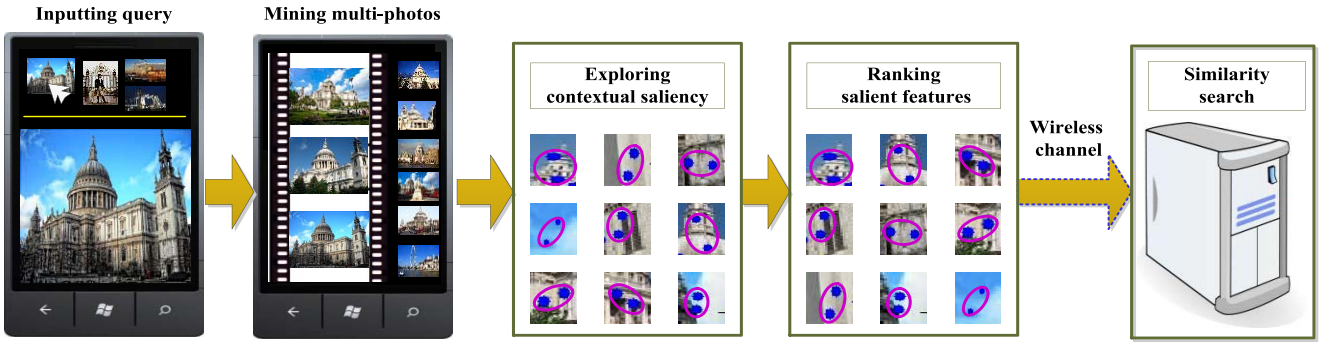


Fig. 1. The flow chart of whole system.

retrieval. It is rational to mine contextual salient features from multiple photos to improve image retrieval performances and reduce the bandwidth requirement. Thus a scalable mobile visual search algorithm is proposed via exploring contextual saliency from multiple relevant photos.

Our approach consists of the following 3 steps as shown in Fig.1: 1) Multi-relevant photo mining. In a user's album at his mobile end, there may be multiple photos relevant to the image that he submitted to carry out image retrieval. And at the same time there may be some irrelevant photos to the query images. So, it is necessary to determine the relevant multiple photos that contain the contextual information of the current input image, as shown in the left most two parts of Fig.1. 2) Contextual saliency exploring from the multiple relevant photos. With the determined multiple relevant photos, we mine salient features for image retrieval, because the main content is usually repeated in the multiple relevant images, as shown in the third component of Fig.1. The saliency explored from multiple relevant photos is more robust, stable and significant. Moreover, by further enforcing spatial and geometric constrains on the detected salient local features, better retrieval performance can be achieved. 3) Salient features ranking for scalable mobile image retrieval. According to the mined contextual saliency from multi-relevant images, the contribution of each feature to the retrieval can be measured. As the available bandwidth of mobile network is time variant, determining the contributions of features to the retrieval is helpful for setting the priority of feature transmission. This is also the foundation of our scalable mobile image retrieval. When the bandwidth is limited, we can transmit features according to their priorities (i.e. contributions to image retrieval).

The main contributions of this paper are summarized as follows: 1) we propose a novel mobile image retrieval approach by exploring saliency from the contextual information in multiple images rather than using single query image. 2) We propose an effective salient feature mining algorithm from multiple relevant images. The features extracted from multiple images are more robust and stable than that extracted from single query image. 3) We propose a scalable mobile image retrieval approach by determining the contribution of feature to retrieval. We can rank the features for bandwidth limited transmission according to their contributions.

Compared with our preliminaries [8], [14], [44], [46], several enhancements have been made in this paper. We summarize them as the follows: 1) we propose an adaptive context exploring approach to mine multiple relevant photos, rather than requiring users to input multi-relevant photos [8] in mobile image retrieval. 2) We enhance salient feature for mobile image retrieval by reinforce the spatial and geometric constrains to improve their robustness in image retrieval; 3) We propose an effective approach to rank the features, based on which we can carry out scalable mobile image retrieval. And 4) more experiments and comparisons are made.

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 summarizes the proposed mobile retrieval system. Section 4 depicts the method of exploring saliency from multiple relevant photos. The approach of achieving scalable retrieval is described in Section 5. Comparing experiments and discussion about some parameters are given in Section 6. Conclusions are drawn in Section 7.

II. Related Work

In recent years, content based image retrieval has experienced a rapid development due to the BoW representation [1] and local features, like SIFT [2], its variants SURF [3], PCA-SIFT [31] and so on. The idea of hierarchical vocabulary tree [4] accelerates the speed of clustering and quantizing for large scale image retrieval, and makes it feasible to realize scalable recognition. Despite of its notable advantages, its deficiency attracts attention as well. Many works make contributions to remedy these defects, such as introducing spatial verification [11], [34], using multiple queries [8], [14]–[16], [40], [44], [46] and compact descriptors [33]–[35].

A. Spatial Verification

The spatial relationship within the visual words is often ignored, but it plays a great role in the retrieval, such as weighting the features [26]–[28] and re-ranking the results [23], [24]. Some works [5]–[8] explore the application of synonyms by verifying the geometry relations between visual words in the image retrieval task. The visual synonyms can extend the visual words in the query image and boost the recall rate. In [5] and [6], Gavves et al. defined visual synonyms as pairs of independent visual words which meet the geometry coherence estimation. Tang et al. [7] constructed a contextual dictionary to reveal the visual word's synonyms

which have the similar contextual distribution. The contextual distribution of a visual word is the statistical aggregation of the spatial distribution of its neighboring visual words in the dataset. In our previous work [8], we introduced geometry difference into the local features extracted from multi-photos input to detect the visual synonyms for retrieval, which captures the saliently important visual words.

The spatial information is usually embedded in bundled visual words [9], [10], [12], [29], [30]. In most cases, the relative geometry information is applied to near-duplicate image retrieval (NDIR) as spatial verification. Chen et al. [9] introduced a spatial visual phrase model describing the relative scale and orientation between the two visual words. Zhang [10] et al. exacted local feature group from images, and measure the spatial contextual similarity between groups to find a best matcher order which is used to calculate the group distance for NDIR and topic based image re-ranking. And [11] encodes the spatial relationship among local features into binary matrix based on coordinate.

The approach of visual synonym extends the visual words to improve the recall, but still does not use geometry relation in retrieval stage. The application of spatial verification like [11] usually demands restrict spatial consistency, which may misjudge the matched visual phrase as discrepant. Recently, geometry relation is also used to search similar images directly. In [12], visual phrases are constructed to be embedded spatial layout constraints in image retrieval. And in [13], descriptive visual words (DVWs) and visual phrases (DVPs) are generated and selected for each image category, and then the images are indexed by DVW and DVP.

B. Using Multiple Queries

The works aforementioned except for [8], all utilize the spatial information in single query. In fact, there are too many local features extracted in the query, and a portion of them are noise and unstable. In our previous work [14], [44], [46], we detect identical salient point (ISP) from the topic album which contains a set of relevant images of the same landmark. An ISP is a subset of similar SIFT points occurs in most of the images in the album, which can capture the major and unique part of landmark. Similar to [14], in this paper, we detect salient visual words (SVWs) from multiple photos which are mined from the user's personal photos. Differently, the SVW requires not only the ISPs in multiple queries are similar in descriptor space but also the features in an ISP are assigned to same visual words.

The famous application of multi-queries is query expansion proposed by Chum et al. [15], [16], which refines the query by combining it with new results returned every time to yield a better query model. The retrieval results can be combined in other ways as well. As in [41], the top candidates retrieved by different methods are jointly ranked to enhance the precision. Recently, multiple queries are exploited in content based image retrieval. In [42], the textual query is input, then the results are regarded as multiple queries to perform visual search. And Fernando and Tuytelaars [40] proposed a pattern based image retrieval approach by mining multiple queries which are given directly by the user. The visual patterns in [40] are actually sets of neighboring visual words that frequently co-occur in the queries. With the development of digital cameras, users' personal photos contain meaningful information. Liu et al. [43] train the classifiers with the retrieved web images to rank personal photos. Actually, we can learn personal photos to search web images in turn. In this paper, we mine multiple queries automatically to make sure that they are visually relevant to the query. In addition, by contrast to [40], our approach describes the specific geometric relationship besides the co-occurrence relationship. Different from query expansion, our approach mines multiple relevant photos from the mobile end. The better representative model of query in our approach is generated by analyzing the multiple photos instead of enriching the query model by doing search over and over again as in query expansion. Besides, query expansion uses RANSAC [25] to perform spatial verification, while our approach uses salient visual pair (SVP) and spatial layout descriptor (SLD) for searching directly [39], which improves the mobile image retrieval performances.

C. Compact Descriptors for Mobile Image Retrieval

With the development of the smart phone, mobile image retrieval catches attention recently. Most works fall in two frames: extracting more compact descriptor, such as CHoG [17], [18], PCA-SIFT [31] and CEDD [32] and compressing the BoW histogram, such as [19]–[22], [33]–[35]. The BoW histogram is usually compressed in three ways: 1) transmitting the intact BoW without redundancy. For example, BoW histogram is encoded as intervals between positive-count nodes of scalable vocabulary tree in [19]; 2) shrinking the scale of vocabulary tree. As in [20], some trivial branches of vocabulary tree are pruned to decrease the dimension of BoW; 3) projecting the high dimensional BoW into a low dimensional vector via transformation matrix or dictionary, such as [21], [22], and [33]–[35]. The learning of the dictionary resorts to solve optimization problem like sparse coding [21], [22], [34]. Sparse coding takes a post processing operation on BoW histogram by representing it as a linear combination of dictionary elements. Sparse coding schemes, such as Lasso [36] can learn the dictionary from original BoW codebook. To control the data size, the geometry information is disregarded in many works. In [23], the orientation of visual word is transmitted along with the frequency for re-ranking. But just a portion of visual words occur once in the query, so many points' geometry information is abandoned. In our approach, we explore salient features from multiple photos. The salient feature is suitable for mobile image retrieval because the feature is discriminative and with small size. We embed spatial information into the feature to improve the performance. Furthermore, we propose scalable transmission for salient features to adapt to the various channel condition.

III. System Overview

As shown in Fig. 1, the proposed mobile image retrieval approach by exploiting contextual saliency consists of three steps: 1) multiple relevant photos mining;

2) contextual saliency exploring; 3) salient features ranking for scalable search. Once a user inputs a query image, our system mines multiple most relevant photos automatically instead of asking user to input them directly. Then, with the multi-photos, we explore saliency from them to eliminate noise, improve precision and reduce computational complexity. For the multiple photos are relevant to the same object, there is an intersection of them, i.e. the repeated content about the object. By exploring the contextual information, we mine robust and stable salient features. Finally, to adapt to the variant wireless channel, we rank the salient features in the light of their contributions to retrieval. According to the current condition of channel, we can transmit only a fit number of salient features to the server end for retrieval.

IV. Multiple Relevant Photos Mining

It is possible that there are many relevant photos to the image that the user submits to retrieval. Our aim is to mine multiple relevant photos from user's mobile end, and to learn contextual salient features to carry out retrieval. Mining relevant photos to carry out retrieval can not only reduce the number of transmitted features but also can improve the image retrieval performances. It consists of the following two steps: 1) feature extraction and quantization; and 2) multiple relevant photos mining.

A. Feature Extraction and Quantization

We describe each image with a set of local features. An image represented through local features can be more powerful than global features [37]. SIFT (scale invariant feature transform) feature is robust against illumination, affine change, scale and other local distortions [2]. A SIFT feature consists of a 128D descriptor vector and a 4D DoG key-point detector vector (x, y, scale, and orientation). Each of the 128-dimension SIFT descriptors of an image is quantized to a visual vocabulary with W code words by hierarchical quantization [14], [44], [46]. In this paper, W is 61,724.

B. Multiple Relevant Photos Mining

Mining multiple relevant photos actually aims to find visually similar images in the user's mobile terminal. As is known that the contextual information for the photos user took, includes time, location, and visual information. Usually, we take multiple photos at a place at a certain time range, and sometimes the geographical information of each photo can also be available. This temporal and geographical information are also valuable if available. However, sometimes, the location information is not available under the circumstances that the GPS devices of mobiles are not ready or open. So, we focus on the contextual saliency extraction from the visual information.

In this paper, we propose to adopt a feature based approach to find the relevant images. For the query image selected by the user, we take the following two steps to mine the multiple relevant photos: 1) searching candidate relevant photos; 2) removing the noisy images.

1) Searching Candidate Relevant Photos: To mine the most relevant multiple photos, we measure the similarity between the query and other images in mobile end by a simple histogram based approach. Assuming that the normalized BoW histograms of the input image and the images in mobile end are respectively denoted as h_q *and* $h_m(k)$, the similarity score of k-th image in smart phone to query, $\boldsymbol{D}(k)$, can be calculated using the city block distance as following:

$$D(k) = \exp(-\left\| h_q - h_m(k) \right\|_1) \tag{1}$$

where $\|v\|_1$ denotes L1 norm of vector v, and $k = 1, \cdots, P$, P is the number of images in mobile end, which are primarily from user's photo album.

2) Removing Noisy Images: We sort the distances in ascending order. The top ranked M-1 results along with the original query image form candidate multiple photos.

Although the candidate multiple photos are the most relevant to the input, there still exist noisy images among them. As the noisy images degenerate the performance and the number of multiple photos is tightly related to the calculating cost, it is necessary to remove the noisy. In this paper, a simple threshold based approach is adopted [8]. If the similarity score of one candidate photo is too small, we eliminate it. In addition, the remaining photos may be duplicates, so we set another threshold to remove the duplicates with too high similarity score. The remnant X candidates are final multiple relevant photos which are used for exploring saliency. It is possible that the mined multiple images are all eliminated and only the input is remained. In this case, the image retrieval degrades into visual searching framework with single input, such as BoW [3] or sparse coding [34] based image retrieval. In this paper, BoW model [3] is applied. Judging the relevance of a photo to the input based on its similarity score is insufficient. We further reduce the impact of noise in contextual saliency exploring stage.

V. Contextual Saliency Exploring

After finding multiple relevant photos for the query image at user's mobile end, we explore the robust and stable contextual saliency in multiple photos. In this paper, the contextual saliency exploring approach comprises the following two parts: 1) contextual information in multiple relevant photos; 2) contextual geometric relationship between salient features. We take advantage of contextual information in multiple photos to mine salient features, and contextual geometric relationship between salient features to enhance the features.

A. Exploring Contextual Information in Multiple Photos

The multi-photos contain the same crucial content, i.e. they are contextually relevant. Generally, the crucial content occurs more frequently than disturbance in these photos, i.e. the frequency of visual words occurring in crucial content is higher than that in background. As shown in Fig. 2, the house is the crucial content, which occurs more frequently than the trees and pedestrians. Our purpose is to pick out these high-frequency salient visual words for retrieval. We take the following two steps to explore the contextual information

Fig. 2. The first row displays the SIFT points in 3 relevant images. The second row sets out the ISPs that occur in 3 images. For ISP_l, if $d_l^i \neq \emptyset$, $\forall 1 \leq i \leq 3$, it occurs in 3 images. The third row shows the ISPs with CS=3, i.e. SVWs that are accordant in 3 images. The average amount of SIFT in three images is 5418, ISP is 263, and SVW is 27.

in multi-photos to mine salient features: 1) identical salient points (ISPs) detection; 2) salient visual word ranking.

1) Identical Salient Points Detection: Following [14], [46], we perform optimal matching pair determination between every two images in multi-images to capture repeated content. During each image-image match, we record all the optimal matched SIFT points pairs (u, q) and their matching scores $MS(u, q)$. The similarity score of two optimal matched SIFT points (u, q) are measured as follows:

$$MS(u, q) = (u \cdot q) \big/ (|u| \cdot |q|) \tag{2}$$

where u and q denote 128D SIFT descriptor vector and $|\boldsymbol{v}|$ denotes the norm of the vector $\boldsymbol{v}$, and "$\cdot$" denotes dot product.

Identical salient points are determined based on the matching score. An ISP is a set of matched SIFT points, denoted as:

$$ISP_l = \{d_l^1, \ldots, d_l^i, \ldots, d_l^X\} \tag{3}$$

where ISP_l denotes the l-th ISP, X denotes the number of the multiple images, d_l^i is the SIFT descriptor of the l-th ISP in the i-th image, which implies the occurrence of the l-th ISP in the i-th image. $d_l^i = \emptyset$, if no feature in the i-th image matches with d_l^1. d_l^1 is the SIFT feature extracted from original query. A speed-up approach is discussed in Part D of our experiment.

2) Salient Visual Words Ranking: The ISPs are extracted based on the similarity in descriptor space. However, the basis of our approach is visual word, and an ISP is different when it occurs in multiple photos. Hence, the SIFT features in an ISP have to meet the consistency of visual words, i.e. the ISP in multi-queries must be assigned to the same visual word. We mark consistency of an ISP as C:

$$C_l = \{c_l^1, \ldots, c_l^i, \ldots, c_l^X\} \tag{4}$$

where c_l^i stands for the consistency of l-th ISP in i-th image. $c_l^i = 1$, if the l-th ISP is appeared in the i-th image and d_l^i is quantified to the same visual word as d_l^1, otherwise $c_l^i = 0$.

The significance of the l-th ISP is measured based on its consistency score (CS) in the multiple images F_l as follows:

$$CS_l = \sum_{i=1}^{X} c_l^i \tag{5}$$

Thus by ranking the consistency score CS_l for all the identical salient points, we select the top ranked ISPs. The ISPs with equal frequency are put on an equal footing in this step. They will further be ranked according to their stabilities, which we will describe in Section VI. In this paper, we select the ISPs with $CS_l \geq 2$, i.e. we remain the ISPs which is consistent in at least two images. The corresponding visual word of the selected ISP is defined as salient visual word (SVW). The number of the consistent ISPs is much less than that of initial ISPs. As shown in Fig.2, the average SIFT point number of the three images in the first row is 5418, the average ISP number of them in the second row is only 263, which is about 5% of raw SIFT points. While the corresponding SVW number is only 27 (as shown in the last row), which is only about 10% of the ISP and 0.5% of raw SIFT points. Hence, exploring contextual saliency is very effective to reduce the size of transmitted features and computational cost. Moreover, SVWs are more stable and pertinent to the crucial content, which manifests that exploring contextual saliency can eliminate the noise effectively.

B. Exploring Contextual Geometric Relationship

By exploring contextual information in multiple relevant photos, we can mine salient visual words. Owing to the quantization loss and noise, single visual word cannot stand for a visual region accurately. In general, the visual words derived from similar visual regions should be the same. Hence, it is rational to infer that if a visual word represents similar visual content in two images, its neighboring features respectively in the two images should be assigned to the same visual words. That is, exploring contextual geometric relationship, i.e. the relationship between neighboring features, can help to distinguish the essentially different features which are quantified to same visual word. We explore the contextual geometric relationship by three steps: 1) bundling neighboring SVWs as salient visual pairs (SVPs) to strengthen the discriminative power of visual words; 2) describing each SVP with a spatial layout descriptor (SLD) to enforce geometric constraint on SVP; 3) merging the SVPs that contain the same SVWs in multiple images to reduce the transmitted data, and measuring the stability of merged SVPs.

1) SVP Generation: As in [39], we construct Salient Visual Pairs (SVPs) by combining the neighboring feature points. For an ISP, it is bound with its nearest and second nearest neighbor ISPs respectively, constructing two SVPs. As illustrated in Fig. 3, the point Pc is bundled with Pn1 and Pn2 separately; Pn1 and Pn2 are the nearest two ISPs around Pc. An ISP is combined with its next nearest ISP besides the nearest ISP for its nearest ISP is not constant in

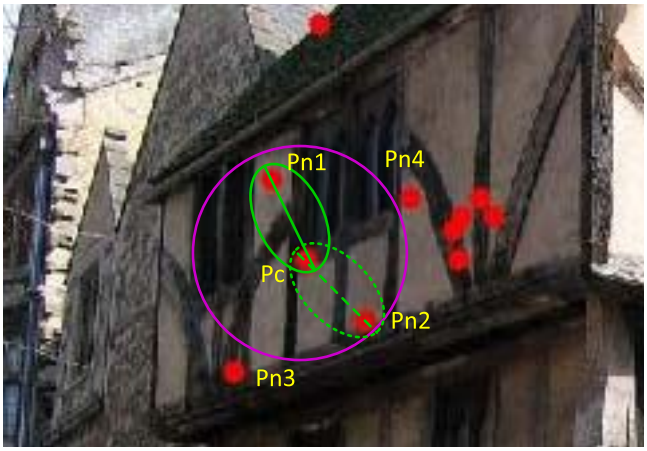


Fig. 3. Construction of SVP.

Fig. 4. Blue circles mark the false visual pairs and green circles mark the accordant visual pairs.

multiple relevant images. For example, in Fig. 3 Pn2 may be closer than Pn1 to Pc, or the Pn1 may be covered by branches of tree, if the cameraman takes the picture in other viewpoints.

Considering that the spatial layout between SVWs in each image may be different, the SVPs generated in each image are not always the same. As shown in Fig. 4, the SVPs in green ellipses are matched in all the three images, and the SVPs in blue ellipses just occur in one image. To reduce the computational cost and data volume to be transmitted, we reserve the stable and accordant SVPs that containing the same SVWs according to (6),

$$\begin{cases} SVW_{l1}^{i} = SVW_{l1}^{j} \\ SVW_{l2}^{i} = SVW_{l2}^{j} \end{cases} \tag{6}$$

where SVW_{l1}^{i} and SVW_{l2}^{i} are the corresponding SVWs of the two ISPs which construct SVP_{i}^{l}, and SVP_{i}^{l} is the l-th SVP in the i-th photo. The accordant SVP must meet (6), $\forall 1 \leq i \neq j \leq X$. X denotes the real number of multiple relevant photos.

2) SVP Description: Each SVP is a pair of visual words. If we only use SVP for retrieval, it is difficult to measure the similarity between the SVP in query and the SVP in matched image. Actually, even though the two visual regions contain the same SVP, they may be different owing to the difference in spatial layout between the two visual words in each region. For the two SVPs that constructed by the same visual words, we measure the similarity between them according to the spatial layout of the two visual words in each image. Following [39], for each accordant SVP, we describe it with a Spatial Layout Descriptor (SLD), which utilize the scale and distance information between the two visual words, defined as Scaled Distance (SD):

$$SD_{l}^{i} = d\left(SVW_{l1}^{i}, SVW_{l2}^{i}\right) / \left(s\left(SVW_{l1}^{i}\right) + s\left(SVW_{l2}^{i}\right)\right) \tag{7}$$

where SD_{l}^{i} denotes the SD of the l-th SVP in the i-th image. $s\left(SVW_{l1}^{i}\right)$ and $s\left(SVW_{l2}^{i}\right)$ are the scale of the SVW_{l1}^{i} and SVW_{l2}^{i} respectively. $d\left(SVW_{l1}^{i}, SVW_{l2}^{i}\right)$ denotes the Euclidean distance between the two salient features that are assigned to SVW_{l1}^{i} and SVW_{l2}^{i} in the i-th image. It is defined as:

$$d\left(SVW_{l1}^{i}, SVW_{l2}^{i}\right) = \sqrt{\left(X_{l1}^{i} - X_{l2}^{i}\right)^{2} + \left(Y_{l1}^{i} - Y_{l2}^{i}\right)^{2}} \tag{8}$$

where X_{l1}^{i} and Y_{l1}^{i} denote the abscissa and ordinate value of SVW_{l1}^{i} in the i-th image.

The SVP enforces co-occurrence restriction on visual words, and SLD reflects the contextual geometric relationship between two local features. They enhance the salient visual words together.

3) SVPs Merging: By exploring contextual geometric relationship, each photo is represented as a series of SVPs and corresponding SLDs as follows:

$$\mathrm{I}_{i} = \{(SVP_{1}^{i}, SLD_{1}^{i}), \cdots, (SVP_{l}^{i}, SLD_{l}^{i}), \cdots, (SVP_{L}^{i}, SLD_{L}^{i})\} \tag{9}$$

where $\mathbf{I}_{i}$ denotes the i-th image, SVP_{l}^{i} is the l-th SVP in the i-th image and SLD_{l}^{i} is the SLD for SVP_{l}^{i}, L is the number of accordant SVPs in multiple relevant images.

Since we have required the SVPs in multiple images to be accordant, i.e. $SVP_{l}^{i} = SVP_{m}^{i}$, $l \neq m$, it is a waste of bandwidth resource to transmit all the SVPs generated from the multiple photos. In addition, the spatial layout descriptor of the accordant SVP should be similar, for the multiple images are relevant. Therefore, it is feasible and necessary to merge the accordant SVPs in multiple images as one SVP. The average SD of the multiple accordant SVPs is regarded as the SLD for the new pooled l-th SVP denoted as:

$$ASD_{l} = \sum_{i=1}^{X} SD_{l}^{i} / X \tag{10}$$

where X denotes the number of multiple images.

It is possible that slight difference in spatial layout of accordant SVPs still exists. Taking into account the difference of spatial layout, we calculate the stability weight for each pooled SVP using the standard deviation of the SLDs in multiple photos as (11):

$$w_{SD_{l}} = \exp\left(-\sqrt{\sum_{i=1}^{X}\left(SD_{l}^{i} - ASD_{l}\right)^{2} / X}\right) \tag{11}$$

where $w_{SD_{l}}$ denotes the stability weight in Scale Distance of l-th pooled SVP.

We define a Salient Feature Group (SFG) as the set of SVP, ASD, and w_{SD} like the following:

$$SFG_{l} = \left(SVP_{l},\ ASD_{l}, w_{SD_{l}}\right) \tag{12}$$

Algorithm 1 SVP Ranking Algorithm

Input: SVW{ s }, s =1, 2, ⋯, X
Output: ASVP
Initialization:, s =X;
While s >0
  Step 1: generating SVP with SVW{ s }
  Step 2: merging the accordant SVP, denoted as ASVP{ s }
  Step 3: SVW{ s -1}= SVW{ s -1}∪ RSVW{ s }
  Step 4: s = s -1;
end

Thus the multiple images can be jointly represented as a series of SFGs as (13):

$$Inp = \{SFG_1, \cdots, SFG_l, \cdots, SFG_L\} \tag{13}$$

where ***Inp*** denotes the final sorted features groups in mobile terminal. The number of the salient feature groups to be transmitted depends on the channel condition.

VI. Salient Feature Ranking for Scalable Retrieval

Wireless channel is vulnerable to interference. There exists serious latency when mobile devices suffer from weak signals. To adapt to the variant wireless channel, we propose scalable mobile image retrieval. We rank the salient feature groups according to their contributions to the retrieval, so that we can adjust the data volume to the channel condition. In addition, we propose a salient feature group based similarity measurement approach.

A. Salient Feature Ranking

In our approach, SFGs are transmitted instead of compact BoW histogram to cope with the problem of small capacity of the channel. In the case of poor channel condition, it is effective to transmit a part of data to carry out retrieval to reduce the latency. Whether a salient feature group is transmitted is determined based on its contributions to retrieval. We rank the SFGs in two sequential levels: 1) frequency of occurrence of SVP to rank the feature group on the whole, and 2) stability in the multi-photos to rank the feature group in detail.

1) Ranking Based on Frequency in Multi-Photos: It is rational to rank the SFGs according to the frequency of occurrence of their SVPs in multiple photos. For example, if *SVPi* is accordant in s photos, and *SVPj* is accordant in s-1 photos, then *SFGi* should be prior over *SFGj* to be transmitted. We denote SVWs with consistency score CS=s as $SVW\{s\}$, and SVP co-occur in s images as $ASVP\{s\}$. RSVW{s} denotes the remnant SVWs of SVW{s}, which cannot build the accordant SVPs in s images. The SVP ranking algorithm based on frequency is given in **Algorithm 1**.

2) Ranking Based on Stability in Multi-Photos: By ranking based on frequency, the SFGs whose SVP appears in s images are denoted as SFG{s}. For SFGs in SFG{s}, we further rank them based on the scale-stability weights as shown in (11). The SVP with bigger stability weight ranks higher. For example, if $w_{SD_i} > w_{SDj}$, SFG_i should be transmitted prior to SFG_j. After ranking based on stability, all the SFGs are ranked. Therefore, the system can send suitable size of data automatically according to the channel condition. When bandwidth is narrow, a small number of Salient Feature Groups which ranked higher are preferentially sent to the server end for retrieval. If the channel condition improves, more SFGs which rank lower will be transmitted. In our system, each SFG is independent and can be used for visual search solely. The similarity between a dataset image and the query is calculated by adding up the similarity score computed based on each SFG.

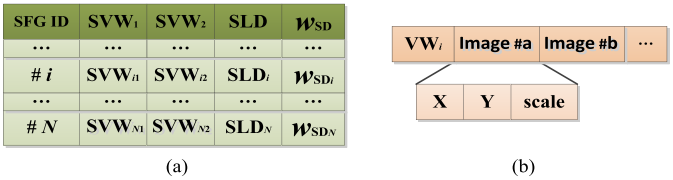


Fig. 5. Structure of (a) SFG and (b) IFIS.

B. Similarity Search

Dataset images also undergo the process of SIFT extraction and feature quantization. Each dataset image is represented as bag of words. Then we build Inverted-File Indexing Structure (IFIS) to speed up the retrieval. However, compared with traditional IFIS, our IFIS records the coordinates and scale information besides the images' ID for each visual word as shown in Fig. 5 (b). In Fig.5 (a) the N SFGs extracted from multiple relevant photos at mobile end with their corresponding salient visual word pairs (SVW_1, SVW_2), SLD and w_{SD} are shown. In Fig. 5(b), VW_i of IFIS denotes the i-th visual word of IFIS. Image #a and image #b are the images that contain VW_i. X, Y, and scale denote the horizonal and vertical coordinates and scale information of the corresponding feature that is assigned to SVW_{i1} in image #a.

Given that N SFGs are extracted from multiple photos at the mobile end, part of which will be selected and send from mobile to server terminal to compute the similarity score in turn. Images similarity calculation consists of the following two steps: 1) SVP consistency verification, 2) SLD similarity measurement. For SVP_i *of* SFG_i, we first search the SVW_{i1} and SVW_{i2} through the inverted-file, only the pictures containing both SVW_{i1} and SVW_{i2} can pass the SVP consistency verification. In other words, we don't store the SVP as the index in inverted file, but construct the SVP only if the database image conforms to the consistency verification for SVP. Suppose that there are T pairs of qualified SVP constructed in database image I, we calculate the similarity score for the eligible database image I according to SLD as follows:

$$Sim(Q, I) = \sum_{j=1}^{T} \exp\left(-\left|SD_{Qj} - SD_{Ij}\right|\right) \times w_{SD_{Qj}} \tag{14}$$

where SD_{Qj} and SD_{Ij} denote the scale distances of the j-th matched SVP respectively in query image Q and database image I. $w_{SD_{Qj}}$ is the stability weight of the j-th matched SVP in Q. $Sim(Q, I)$ is the similarity measurement between

Q and I, and higher score means higher similarity. Finally, we rank all the retrieved images according to their similarity score.

VII. Experiments

We test the proposed approach on two datasets: Oxford Buildings Dataset and GOLD [38], [45]. To show the effectiveness of our approach, we compare our method with query expansion [16], spatial coding [11], and sparse coding [34]. Some main factors that influence the performance are discussed as well. In addition, we speed up the process of mining salient visual words and show the performance. Experiments are all carried out on a PC with 48G memory and Intel® Core(TM)2, Quad CP Q8400 with 2.26GHz on Matlab.

A. Datasets

The Oxford Buildings Dataset consists of 5,062 images collected from Flickr by searching for particular Oxford landmarks, 11 landmarks in total. 55 query images given by Oxford Buildings Dataset as test set. We suppose that the Oxford Buildings Dataset is pre-saved in mobile end, so the first step of our approach, obtaining multiple relevant photos, is run on Oxford Buildings set. If the system is applied in reality, the searching in the first step is conducted on photos stored in mobile end.

GOLD is a geo-tagged large scale web image set associated with their geographic coordinates [38], [45], which is crawled from Flickr through its API. GOLD contains more than 227 thousand images together with 80 places-of-interests which are selected from 60 world-wide cities with about 3.3 million images. We take it as disturbance when carrying out experiments on Oxford Building dataset.

B. Baselines

To be convenient, our approach is called MP in the following experiments. We compare our method with three representative algorithms: (1) Query expansion (QE) [16]. In paper [16], a query region is given as input. To be fair with our method, we carry out retrieval with the whole image instead of query region. (2) Spatial coding (SP) [11]. SP encodes the relative position between matched features. The similarity score between two images is calculated based on spatial verification as in [11]. The spatial verification proposed in paper [11] is applied in near duplicate image retrieval (NDIR). Actually, for universal dataset, spatial coding is too rigid, which requires the spatial map of matched features to be identical. Hence, to make fair comparisons we slacken the spatial constraint. If 90 percent of features around the matched feature are consistent in spatial map, the two features are truly matched. The methods QE, SP, and the proposed approach, all use the same codebook with 61,724 visual words. (3) Sparse coding (SC) [34]. Due to the extremely high computational cost of compressing 61,724D BoW histogram, we train a codebook which contains 8,623 visual words, and compress the 8,623D BoW histogram into 1500 dimensions by Lasso regression. Our approach uses the smaller cookbook to perform experiments to be compared with SC as well.

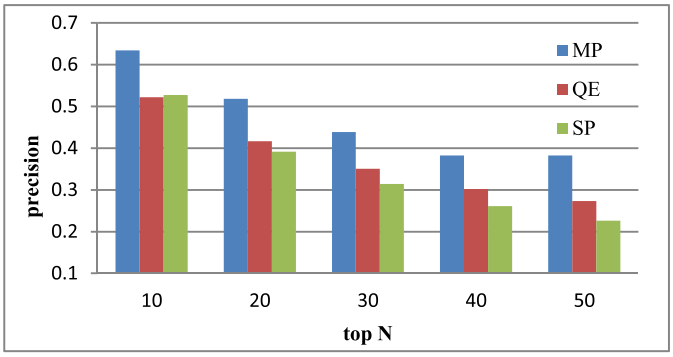


Fig. 6. Comparison between our method MP, query expansion QE and spatial coding SP.

Average precision at the top N (AP@N) is the evaluation criterion to measure the mean percent of relevant images in the top N retrieved results. It is defined as:

$$AP@N = (1/K) \times \sum_{i=1}^{K} (R_i/N) \tag{15}$$

where K is the size of test set, and R_i denotes the number of retrieved relevant images up to N for the i-th query image.

C. Performance Comparison

Experiments are conducted on Oxford Buildings Dataset. M in the experiment is set to be 3 by pursuing tradeoff between precision and consuming time. In the situation that no multiple relevant photo is mined, image retrieval is carried out based on BoW similarity. QE, SP, and SC are the baselines to be compared with our approach. We use top ranked 20 SFGs to execute retrieval.

Fig. 6 shows the average precision @ top N of MP, SP, and QE in the condition of using 61,724 visual words. The AP@10 of MP is 0.6127, which is higher than 0.5273 of SP and 0.5218 of QE. Query expansion cannot achieve good result when the query image is complicated, i.e. the object region is not clear enough. Spatial coding does not perform well as in NDIR owing to its over strict spatial constraint. Our approach measures the difference of spatial layout in a soft way, and explores the contextual information in multiple relevant photos, so our proposed algorithm outperforms other methods.

Furthermore, each SFG consists of one pair of SVP along with its SLD and stability weight w_{SD}. Each SVP contains two SVWs. We store each SVW as unsign short integer, i.e. two bytes is needed. SLD and w_{SD} are set as floating type (4 bytes). That is, every SFG needs 12 bytes memory. We only use 20 SFGs for retrieval, so our method only transmits 240 bytes in total. The small amount of data is particularly suitable for mobile image retrieval. In extreme condition, we can use less than 10 group SFGs to seach images as well, i.e the transmitted data is less than 120 bytes. To highlight the remarkable small amount of data of our method, we list the required amount of data of other approaches in Table 1. It demonstrates that our approach needs the least bandwidth source. We find that our approach only requires 4% of the data of sparse coding [34]. Even though sparse coding

TABLE I
THE REQUIRED DATA SIZE OF MP, QE, SC, SP AND JPEG IMAGE

approaches	MP	QE	SP	SC	JPEG
Data (bytes)	240	316K	18K	6K	385.8K

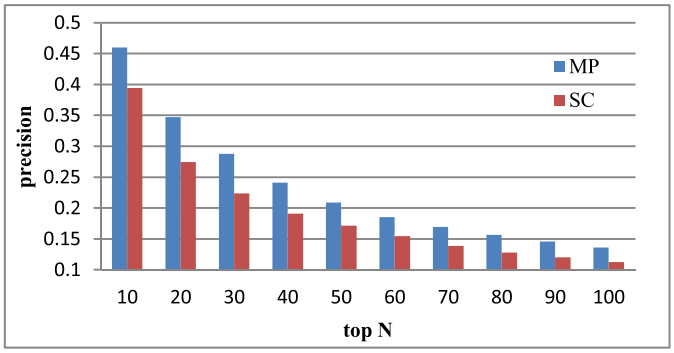


Fig. 7. Comparison between our method MP nd sparse coding SC.

compresses BoW histogram into 100 dimensions, 400 bytes is still needed. And the performance will deteriorate if the dimension of compact descriptor in sparse coding is reduced.

In addition, we use a codebook that consists of 8,623 visual words to test our approach and SC [34]. The small codebook makes the visual words not distinctive enough. In this case, using 20 SFGs cannot perform well. Therefore, 60 SFGs are selected for retrieval in our method, and 1500D compressed BoW histogram is used in SC. The result is shown in Fig. 7. Our approach performs better because we introduce the spatial layout between salient features. Moreover, the data volume sent to the server end in our approach (720 bytes) is much smaller than that in sparse coding (at least 3-4 KB).

D. Discussion

In our approach, the retrieval performance is influenced by two main parameters: (1) M, i.e. the number of the candidate multiple relevant photos; (2) L, i.e. the number of the SFGs sent to the server end. We discuss the impacts of the two parameters in our experiments. Besides, to confirm the justifiability of the major parts of the proposed approach, we add the following comparison experements: 1) comparison about adopting SFG ranking or not; 2) comparison about mining salient visual words from multiple relevant photos or from single image; 3) measureing the influence of distance and scale to spatial layout descriptor; 4) discussing the effect of noisy image in multiple photos. Furthermore, we speed up the process of exploring the saliency from multiple photos and show the consuming time and precision. Finally, we test our approach on GOLD dataset [38], [45].

1) Impact of M: For we have removed noisy images from candidate multiple relevant photos in the stage of mining multiple photos, the number of the multiple photos that are actually used to explore the saliency is not definite. We use M to discuss the impact of the number of the multiple images. The parameter M, i.e. the number of the

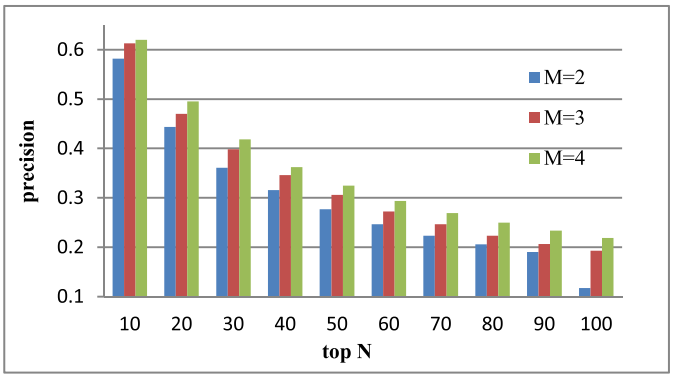


Fig. 8. Result of using different M.

candidate multi-relevant photos, has impact on both precision and computational complexity. In the experiments, at most 20 groups of SFGs are sent to the server end no matter how M is. As shown in Fig. 8, bigger M leads to better performance, for the system can detect more stable SVPs and give these SVPs bigger weights. In our approach, the SVPs are ranked according to their frequency of occurrence and stability in multiple photos. If a SVP occurs in more images of the multiple queries, then it is likely more relevant to the images about the scene in query. That is to say, if we use bigger number of multiple photos, we can select the more important SVPs. Therefore, the SVPs can be ranked more exact with big number of multiple photos. The SVWs which are more relevant to the theme of query play an important part in retrieval. However, M is tightly related to the computational cost, because the procedure of detecting ISP needs matching features between multiple photos. Supposing that there are $S1$ SIFT features extracted from the first photo and $S2$ from the second, matching features between the two images needs $S1 \times S2$ times of matches. Each matching algorithm involves 128 times of multiplication and one add operation. If we use 3 photos, the matching operation will add extra $S3 \times S2 + S3 \times S1$ times given that there are $S3$ features in the third image. The cost of increasing M is considerable calculation. We must consider the tradeoff between performance and complexity.

2) Impact of L: The number of the salient visual pairs L sent to server end influences both the expended data volume and the retrieval precision. We transmit 10, 20, 40 and 60 SFGs seperately and test the performance. The average precision @ top 10 is 0.5818, 0.6127, 0.6236, 0.6272 respectively. The results in Fig.9 demonstrate that the number of the uploaded SFGs has influence on the retrieval precision, but the impact is not remarkable. With the increasing of data, the searching performs better. But when the number of the SFGs reaches 40, the rising tendency of MP turns particularly slow. According to the results, it is feasible to transmit few SFGs when the channel condition is poor. In addtion, even the bandwidth of wireless channel is wide enough, users can send a little data to save online traffic. The above experiments show that our method achieves scalabel retrieval successfully.

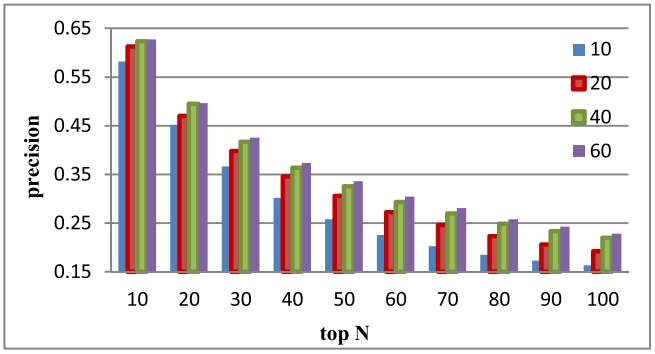


Fig. 9. Result of changing the parameter L.

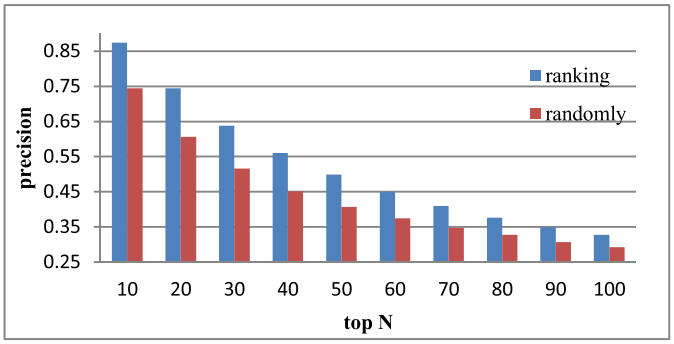


Fig. 10. Comparison ranking the features with transmitting them randomly.

3) Impact of Ranking Features or Not: To confirm that our ranking scheme for SFGs is rational and effective, we select 5 SFGs randomly (denoted by randomly) for retrieval in contrast to 5 groups selected according to ranking order (denoted by ranking). To highlight the effect of ranking, we pick out 97 test images from Oxford Buildings Dataset which are all able to mine multiple images. The correspoinding precisions under top 10 to top 100 are shown in Fig.10 respectively. The average precision @ top 10 of ranking is 0.8742 which is higher than 0.7443 of selecting randomly. The result in Fig. 10 shows that ranking scheme plays an important role in improving performance, because ranking scheme selects the features that contribute most to retrieval.

4) Impact of Using Multi-Photos or Single Photo: The advantages of multi-relevant photos help to remove noisy features and catch salient points. If we don't use the multiple queries, we wonder how the performance will be. For a single image, we perform local feature refinement on it by resizing it as in [10]. In our experiment, the original image is resized to 40 percent of original size in vertical and horizonal direction. The feature that repeats in original and resized image is defined as refined feature. Then the refined features are used for building visual pairs and visual search. One visual pair in single query is ranked according to the sum of TF-IDF values of its two visual words. We compare the performances of using 20, 50, and 80 visual pairs extracted from single photos and multi-photos respectively. The result is shown in Fig. 11. In Fig.11, MP and SQ denote using multi-photos and single query. For image retrieval by single query, the refined feature

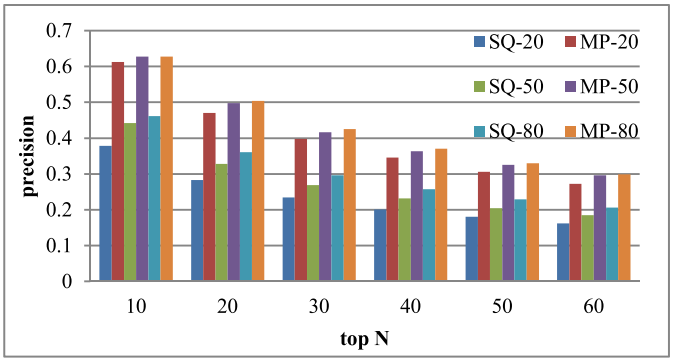


Fig. 11. Result of using 20, 50, and 80 visual pairs generated in 3 relevant photos and constructed in single photos respectively.

TABLE II
THE PERFORMANCE FOR THE 44 QUERIES

AP@N	10	20	30	40	50
Noise	0.6864	0.5409	0.4575	0.3920	0.3578
No noise	0.6818	0.5409	0.4530	0.3949	0.3532

is the stable one that survives from affine transformation. But it is likely that not the salient feature is closely relevant to the key content of query image. Due to the impact of the irrelevant features, a small number of viual pairs will not lead to good retrieval result. Similar to the situation in MP, more visual pairs lead to better performance. It is obvious that using multi-relevant photos outperforms using single photo. Local feature refinement can eleminate noisy and unstable ones which are sensitive to deformation. But the refinement cannot catch salient points which are relevant to the key content of the image, while MP works remarkably in this aspect.

5) Impact of Noisy Image in Multiple Photos: In our experiments, the threshold which is used to remove noisy images in multiple photos is set as 1.77 empirically based on vast observations. Actually, simply setting threshold cannot filter out all the noisy images. However, the noisy images in multiple photos influence the retrieval results little. To study the effect of noisy image, we intentionally add a noisy image to the mined multiple relevant photos. We analyze the effect of noisy image in two cases: 1) multiple relevant photos are mined; 2) no multiple photos are mined. When the threshold is set as 1.77, for 44 of 55 queries of Oxford Buildings Dataset, multiple relevant photos can be mined. We computed the average precision for the 44 queries in Table 2. We also test the performance for the other 11 queries in the case of combining a noisy image with original query as multiple photos. The result is shown in Table 3. In Table 2 and 3, the term Noise means adding noisy image to the multiple photos, similarly, No noise means no noisy image added deliberately.

By comparing the performances shown in Table 2 and Table 3, we find that the noisy image effect much on the second case and little on the first case. In second case, the noisy image is irrelevant to the query. So, few SVWs are mined and fewer SVPs remained. And the remaining SVPs likely do not represent salient visual content.

TABLE III
THE PERFORMANCE FOR THE 11 QUERIES

AP@N	10	20	30	40	50
Noise	0.1273	0.0909	0.0727	0.0614	0.0509
No noise	0.4091	0.3227	0.2697	0.2386	0.2127

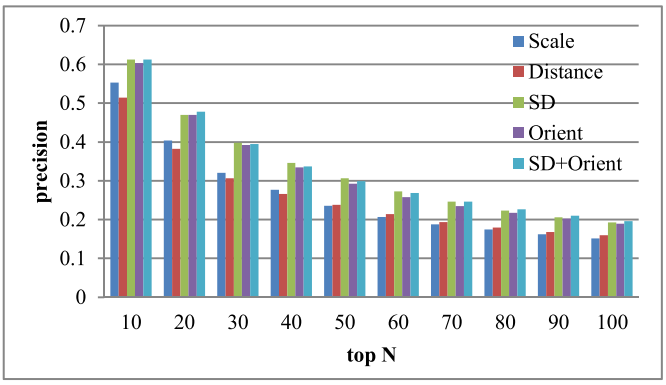


Fig. 12. Comparing the performance when we use scale, distance, orientation and combine them to describe the SVP.

Therefore, the noisy image deteriorates the performance in this case. Actually, in this case the similarity score between the noisy image and the query is very low, so the noisy image can be removed by the threshold in reality. That is to say, noisy images can be further removed during contextual saliency exploring.

However, in first case, besides the noisy image, there still exist other relevant multiple photos which make our approach robust to noise. The reasons are as follows. First, identical semantic points detection matches SIFT descriptors between every two of multiple photos. Only a small part of the SIFT features in noisy image are optimally matched with the features in other photos. That is to say, few SIFT features of noisy image are remained as SVW. Second, we construct salient visual pairs with the SVWs. Each SVP consists of two neighboring SVWs. And the SVPs are required to be accordant in the multiple photos. Even though the noisy photo contains SVWs, the SVPs in the noisy image are mostly not accordant with the SVPs in other relevant photos. Hence, most of the SVPs in the noisy images are removed. Thus, we can conclude our approach is robust to the noisy images.

6) Impact of Different Spatial Constraints to Performance: In this paper, SD is used as the spatial layout descriptor, which utilizes the scales and spatial distance between two SVWs. Actually, either the scale or the distance can be utilized to describe the spatial layout solely. And the orientation of the SVW is also usable. To convince that integrating the scale and distance is better, we test the performance in four conditions: 1) using total scale $s\left(SVW_{l1}^{i}\right)+s\left(SVW_{l2}^{i}\right)$ as SLD to describe the spatial layout for the l-th SVP in the i-th image; 2) using distance $d\left(SVW_{l1}^{i}, SVW_{l2}^{i}\right)$ as SLD; 3) the difference of the orientations $o\left(SVW_{l1}^{i}\right)+o\left(SVW_{l2}^{i}\right)$, where o denotes the orientation; 4) combining orientation and SD. In the fourth case, the final similarity score is the sum of the similarity score of SD and orientation. The results are shown in Fig. 12. In Fig. 12, Orient denotes the third case, and the term SD+Orient stands for the fourth case. The results demonstrate that the Scale Distance performs the best.

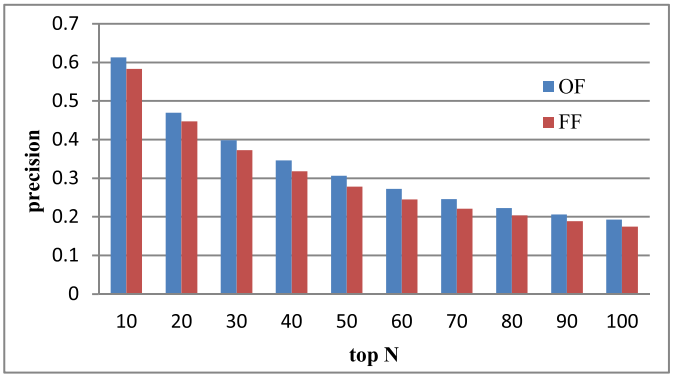


Fig. 13. Performance comparison between original algorithm of mining SVW and improved algorithm.

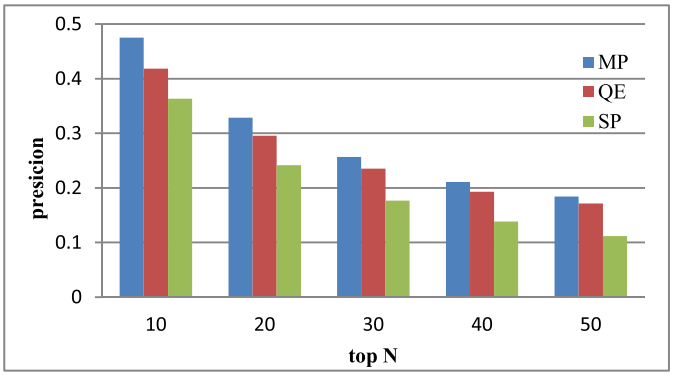


Fig. 14. Performance of 3 methods on GOLD dataset.

7) Speeding Up the Process of Detecting SVW: The proposed method of mining salient visual word is based on ISP detection [14]. Detecting ISP needs to match SIFT features between every two images. For one local feature in an image, it is matched with all the features in other images to detect the optimal matched pair. To speed up the process of mining salient visual word, we perform feature matching on features that are assigned to the same visual word. Thus the scope of features that one SIFT feature is matched with is shrunk tremendously. It takes an average of 4.27 seconds to mine SVWs before. After we speed up the algorithm, it takes an average of 0.41 seconds. The performance comparison is shown in Fig. 13. OF denotes original algorithm, and FF denotes fast algorithm. The retrieval precision of FF is slightly inferior to OF, because some unstable points are regarded as ISP in the case that few features are quantized to the same words. In OF, average 440 ISPs are detected, whereas in FF, average 838 ISPs are detected. A portion of them are unstable actually.

8) Test on GOLD Dataset: The above experiments and discussions are conducted on Oxford Buildings Dataset with GOLD as disturbance. It is essential to test our method on a bigger dataset. GOLD dataset contains about 270 thousand images [38]. About 80 pairs of SVPs are used for retrieval. We compare our approach MP with QE and SP. The hierarchical vocabulary used to quantify the two datasets has 8 levels, and the branch factor of that is 10. The result in Fig. 14 shows

that our approach is suitable for big dataset and outperforms query expansion and spatial coding.

VIII. Conclusion

In this paper, a novel mobile visual searching algorithm is proposed by exploring saliency from multi-relevant photos. The proposed method can transmit less than 100 bytes data when wireless channel suffers from bad situation, and send more data to server end if bandwidth turns wide to improve retrieval results. The performance of our approach can be further improved by introducing GPS and time information in the stage of mining multiple relevant photos. Our future work focuses on further reducing the computational complexity, especially in the procedure of capturing SVWs, and exploring more excellent salient feature.

References

[1] 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.

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

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

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

[5] E. Gavves and C. G. M. Snoek, "Landmark image retrieval using visual synonyms," in *Proc. Int. Conf. MM*, 2010, pp. 1123–1126.

[6] E. Gavves, C. G. M. Snoek, and A. W. M. Smeulders, "Visual synonyms for landmark image retrieval," *Comput. Vis. Image Understand.*, vol. 166, no. 2, pp. 238–249, Feb. 2012.

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

[8] Y. Xue, X. Qian, and B. Zhang, "Mobile image retrieval using multi-photos as query," in *Proc. IEEE ICMEW*, Jul. 2013, pp. 1–4.

[9] J. Chen, B. Feng, L. Zhu, P. Ding, and B. Xu, "Effective near-duplicate image retrieval with image-specific visual phrase selection," in *Proc. IEEE ICIP*, Sep./Oct. 2012, pp. 1909–1912.

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

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

[12] Y. Zhang, Z. Jia, and T. Chen, "Image retrieval with geometry-preserving visual phrases," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 809–816.

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

[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] 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.

[16] O. Chum, A. Mikulík, M. Perdoch, and J. Matas, "Total recall II: Query expansion revisited," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 889–896.

[17] V. Chandrasekhar, G. Takacs, D. Chen, and S. Tsai, "CHoG: Compressed histogram of gradients—A low bit-rate feature descriptor," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 2504–2511.

[18] B. Girod *et al.*, "Mobile visual search," *IEEE Signal Process. Mag.*, vol. 28, no. 4, pp. 61–76, Jun. 2011.

[19] D. Chen, S. 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.

[20] J. Chen, L.-Y. Duan, R. Ji, and W. Gao, "Pruning tree-structured vector quantizer towards low bit rate mobile visual search," in *Proc. IEEE ICASSP*, Mar. 2012, pp. 965–968.

[21] J. Lin, L.-Y. Duan, J. Chen, R. Ji, S. Luo, and W. Gao, "Learning multiple codebooks for low bit rate mobile visual search," in *Proc. IEEE ICASSP*, Mar. 2012, pp. 933–936.

[22] R. Ji, L.-Y. Duan, J. Chen, H. Yao, Y. Rui, and W. Gao, "Location discriminative vocabulary coding for mobile landmark search," *Int. J. Comput. Vis.*, vol. 96, no. 3, pp. 290–314, Feb. 2012.

[23] Y. Wu, S. Lu, T. Mei, J. Zhang, and S. Li, "Local visual words coding for low bit rate mobile visual search," in *Proc. 20th ACM Int. Conf. MM*, 2012, pp. 989–992.

[24] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. CVPR*, Jun. 2007, pp. 1–8.

[25] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[26] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, and T. X. Han, "Contextual weighting for vocabulary tree based image retrieval," in *Proc. IEEE ICCV*, Nov. 2011, pp. 209–216.

[27] M. Perd'och, O. Chum, and J. Matas, "Efficient representation of local geometry for large scale object retrieval," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 9–16.

[28] M. Marszalek and C. Schmid, "Spatial weighting for bag-of-features," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 2006, pp. 2118–2125.

[29] 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.

[30] 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.

[31] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun./Jul. 2004, pp. II-506–II-513.

[32] S. A. Chatzichristofis and Y. S. Boutalis, "CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval," in *Proc. ICVS*, May 2008, pp. 312–322.

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

[34] R. Ji, H. Yao, W. Liu, X. Sun, and Q. Tian, "Task-dependent visual-codebook compression," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2282–2293, Apr. 2012.

[35] R. Ji, L. Duan, J. Chen, H. Yao, and W. Gao, "A lowbit rate vocabulary coding scheme for mobile landmark search," in *Proc. IEEE ICASSP*, May 2011, pp. 2316–2319.

[36] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc., Ser. B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[37] A. Qamra and E. Y. Chang, "Scalable landmark recognition using EXTENT," *Multimedia Tools Appl.*, vol. 38, no. 2, pp. 187–208, Jun. 2008.

[38] 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.

[39] Q. Luo, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Scalable mobile search with binary phrase," in *Proc. 5th ICIMCS*, 2013, pp. 66–70.

[40] B. Fernando and T. Tuytelaars, "Mining multiple queries for image retrieval: On-the-fly learning of an object-specific mid-level representation," in *Proc. IEEE ICCV*, Dec. 2013, pp. 2544–2551.

[41] S. Zhang, M. Yang, T. Cour, K. Yu, and D. N. Metaxas, "Query specific rank fusion for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 803–815, Apr. 2015.

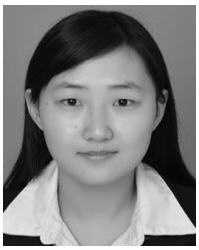
[42] R. Arandjelović and A. Zisserman, "Multiple queries for large scale specific object retrieval," in *Proc. BMVC*, 2012, pp. 1–11.

[43] Y. Liu, D. Xu, I. W. Tsang, and J. Luo, "Textual query of personal photos facilitated by large-scale web data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 1022–1036, May 2011.

[44] 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. [Online]. Available: http://dx.doi.org/10.1109/TCSVT.2014.2369731

[45] 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.

[46] X. Yang, X. Qian, and T. Mei, "Learning salient visual word for scalable mobile image retrieval," *Pattern Recognit.*, to be published. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2014.12.017



Xiyu Yang is currently pursuing the M.S.D. degree with the Smiles Laboratory, Xi'an Jiaotong University, Xi'an, China. Her research interests include mobile end image retrieval.

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, Xi'an, in 2008, where he was an Assistant Professor. He was a Visiting Scholar with Microsoft Research Asia from 2010 to 2011. He was an Associate Professor from 2011 to 2014, and he is currently a Full Professor and the Director of the Smiles Laboratory. His research interests include social media big data mining and search. His research was supported by the National Natural Science Foundation of China, Microsoft Research Asia, and the Ministry of Science and Technology. He was a recipient of the Microsoft Fellowship in 2006. He was also a recipient of the outstanding doctoral dissertations from Xi'an Jiaotong University and Shaanxi Province, in 2010 and 2011, respectively.

Yao Xue received the M.Eng. degree in electronic and information engineering from Xi'an Jiaotong University, China, under the supervision of Dr. X. Qian, in 2013. He is currently pursuing the Ph.D. degree with the Department of Computing Science, University of Alberta, Canada. He is also with the Centre for Intelligent Mining Systems, under the supervision of Dr. N. Ray. His research interests cover computer vision related areas, like image retrieval, visual summary, object detection, and recognition.