Pattern Recognition 48 (2015) 3093–3101

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Learning salient visual word for scalable mobile image retrieval ☆

Xiyu Yang [a], Xueming Qian [a,*], Tao Mei [b]

[a] SMILES LAB, Xi'an Jiaotong University, Xi'an 710049, China
[b] Microsoft Research Asia, Beijing 100080, China

ARTICLE INFO

Article history:
Received 11 August 2014
Received in revised form
17 December 2014
Accepted 20 December 2014
Available online 2 January 2015

Keywords:
Mobile image retrieval
Scalable retrieval
Salient visual word (SVW)
Multiple relevant photos
Spatial verification

ABSTRACT

Owing to the portable and excellent phone camera, people now prefer to take photos and share them in social networks with their friends. If a user wants to obtain relevant information about an image, content based image retrieval method can be utilized. Taking the limited bandwidth and instability of wireless channel into account, in this paper we propose an effective scalable mobile image retrieval approach by exploiting the advantage of mobile end that people usually take multiple photos of an object in different viewpoints and focuses. The proposed algorithm first determines the truly relevant photos according to visual similarity in mobile end, then learns salient visual words by exploring saliency from these relevant images, and finally determines the contribution order of salient visual words to carry out scalable retrieval. Moreover, to improve the retrieval performance, soft spatial verification is proposed to re-rank the results. Compared to the existing approaches of mobile image retrieval, our approach transmits less data and reduces the computational cost of spatial verification. Most importantly, when the bandwidth is limited, we can transmit only a part of features according their contributions to retrieval. Experimental results show the effectiveness of the proposed approach.

1. Introduction

The bag of word (BoW) [1] model and local features, such as SIFT [2] and SURF [3] make significant breakthroughs in content based image retrieval (CBIR). And 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. Recent years, new technologies continuously emerge, which facilitate the development of visual search and recognition, such as domain-adaptive global feature descriptor [46], cross-domain dictionary learning [53], re-ranking schemes for database images [21,47], query expansion [5–7], visual synonym [8–13], co-occurrence pattern [14–15] and geometric verification [16–17]. Query expansion enriches the query model over and over again by combining it with top returned retrieval results. Co-occurrence pattern constructs visual phrase or group and represent image as bag of visual groups [14–15]. Visual synonyms are defined as the visual words that correspond to similar visual patch. The goal of visual synonym is to expand a visual word with its synonyms to narrow down the semantic gap in visual word quantization [9]. Besides visual synonyms, discriminative features are proposed [15,57]. Shao et al. [57] learn relative feature by max-margin criterion between the input and its dissimilarity with the prototype images. Spatial verification enforces geometric consistent constraint on common words that query and dataset image share, such as RANSAC [16] and spatial coding [17]. Spatial coding performs well in partial duplicate image retrieval. However, due to the rapid development of digital camera, photos usually have high definition, which results in that too many local features are extracted from one photo. Thus spatial coding will be time-consuming.

Recently, the multi-model is exploited to improve the visual researches, e.g. [47] learns global feature via multi-objective genetic programming, [48] selects crucial features by analyzing the shared information among multiple tasks, and [50] generates multimodal spatio-temporal theme to describe landmarks better. Multiple models are correlated and complementary to each other. Therefore using multiple models helps to make up the deficient of single model. For the images about same object, the views of them are usually various. It is effective to mine the view invariant features to represent the object. The core methodology that tackles visual problems with changes in viewpoint is to discover the shared knowledge irrespective to such viewpoint changes [52]. Thus it is feasible to explore shared information from multiple relevant photos to represent the query better.

Smartphone is experiencing booming development recently. According to the statistics, there are 4.5 billion mobile phones and 1.7 billion smart phone users in the world in 2014. And the number

☆This work is supported in part by the Program 973 No. 2012CB316400, by NSFC No. 60903121, 61173109, 61332018, and Microsoft Research Asia.
* Corresponding author.
E-mail addresses: yangxiyu@stu.xjtu.edu.cn (X. Yang), qianxm@mail.xjtu.edu.cn (X. Qian), tmei@microsoft.com (T. Mei).

http://dx.doi.org/10.1016/j.patcog.2014.12.017
0031-3203/

of smart phone users will mushroom in the future. Mobile phone has been an indispensable part of people's lives. With the powerful phone camera, most people now prefer to take pictures with mobile phone. Usually, people are accustomed to taking many photos about same object or view to ensure that at least one of them is satisfying and to fully present the object from different viewpoints. Now that many photos refer to the same scene, it is rational to comprehensively analyze multiple relevant photos to acquire salient visual words. The salient visual words should be stable and significant, which capture the repeated crucial content from multiple photos.

For the mobile image retrieval, two factors need to be considered: (1) the limited bandwidth and instability of wireless channel [22–24] and (2) the electric quantity of the battery. Hence, the mature approach for web image retrieval is not suitable to be applied in mobile end. The mobile image retrieval requires that: (1) the less data are transmitted; (2) the computational cost of the algorithm that performed in mobile end is low; and (3) the transmitted data volume is changeable according to the condition of the wireless channel. The state-of-the-art mobile image retrieval approaches focus on extracting more compact descriptor, such as CHOG [25–26] and PCA-SIFT [27], or compressing the BoW histogram [22–24,28,29]. In most cases, the BoW histogram is transmitted in a compact form to reduce the volume of data. As to compact descriptors, the total amount of data is up to the number of features extracted from an image. And BoW histogram compressing approaches, like sparse coding, occupy much memory and are very complex. Both of the two kinds of methods do not take the instability of channel into account. Actually, the instable condition of wireless channel requires the algorithm to be scalable, i.e. the transmitted data can be adjusted according to the variant channel capacity.

The idea of scalability is successfully applied in Scalable Video Coding (SVC), such as spatial and quality scalability in H.262|MPEG-2 Video, H.263 and MPEG-4 Visual. H.264 includes temporal scalability besides. The SVC technology makes the length of video stream variable to satisfy the users' need in the condition of current channel condition. Inspired by SVC, we proposed a scalable mobile visual search method by adjusting the number of salient visual words that are sent to server end.

In this paper, a novel spatial verification algorithm is proposed based on salient visual word for mobile image retrieval. Our approach consists of 3 steps: (1) mining multiple relevant photos. Once a user inputs a query, our approach automatically mines some relevant photos from mobile end; (2) extracting salient visual word (SVW) and ranking them for scalable image retrieval. With the relevant photos, we extract the stable, robust and distinctive visual words; (3) re-ranking the retrieval results based on spatial verification to improve the performance.

The main contributions of this paper are summarized as follows: (1) we learn salient visual words, which eliminates the effect of noisy, unstable and irrelevant features; (2) the small number of robust salient visual words is suitable for mobile retrieval; (3) we change the restrict spatial consistent constraint into a soft type of accumulating consistent score, which makes spatial coding applicable to universal image retrieval task besides duplicate image retrieval, and achieve notable performance; and (4) considering the instability of invariance of wireless channel, we propose a selection scheme for salient visual words, which is the fundamental of scalable mobile image retrieval.

The remainder of this paper is organized as follows. In Section 2, related work is reviewed. Section 3 overviews the system. Section 4 describes the method of mining multiple relevant photos. Section 5 details the strategy of extracting salient visual word from multiple relevant photos and the re-ranking scheme. In Section 6, we introduce the spatial verification model. Experimental results and discussion are represented in Section 7. Conclusions are drawn in Section 8.

2. Related work

The CBIR thrives in recent years. The excellence of the SIFT feature and BoW model have been manifested in computer vision. However, there still exists deficiency in BoW model. For example, SIFT is sensitive to little disturbance like viewpoints, which makes the SIFT features extracted from similar visual patches not identical. And owing to the quantization loss, the visual word is not discriminative enough. Recently, plenty of papers make contribution to remedy these defects, such as learning synonyms [8-13], introducing spatial verification [14,15, 17, 30–33] , using multiple queries [34,35] and learning compact descriptors [24-29, 37,44].

2.1. Learning Synonyms

The visual vocabulary usually has to be rather large to successfully distinguish one image from dissimilar ones in large scale image retrieval. Nevertheless, the over large codebook may contain many synonyms owing to over-splitting in the process of clustering. The methods to address synonym phenomenon resorts to assign a SIFT feature to more than one visual words such as soft quantization [8]. The idea of soft quantization focuses on assigning features according to Euclidean distance in descriptor space. Actually, synonyms can be defined in visual level and semantic level. For instance, in [9] visual synonym is aimed at mining visual words that correspond to same semantic meaning. The synonyms are the words with similar contextual distribution which is the statistics of both co-occurrence and spatial information of surrounding words. The visual synonym can also be acquired based on geometric coherence estimation. In [10,11], 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. And in [12] synonyms are explored by counting the frequency that two visual words are coherent in training set. In our previous work [13], 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.

2.2. Spatial Verification

The spatial relationship within the visual words attracts much attention recently. It plays a great role in the retrieval, such as weighting the features [18–20] and learning visual synonyms [21]. And spatial information can reinforce the discriminative power of single word. A direct method to distinguish the same word in two images is to compare the orientation information or neighboring visual words in two images. Usually the spatial information is embedded in bundled visual words which are near to each other and co-occur frequently [14–15,30–32]. 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 of other words' appearing in the neighborhood of the specified word [14]. Spatial verification is introduced to verify whether the retrieved image in truly matched with the query image, it performs well in near-duplicate image retrieval (NDIR). For example, [21] exacts local feature groups from images, and measures the spatial contextual similarity between groups to find a best matcher order which is used to calculate the group distance for NDIR and [17] encodes the relative position among local features into binary spatial maps based on coordinate. The spatial verification in NDIR requires restrict geometric consistency, e.g. in [17] the spatial maps of query image and dataset image must be same if they are truly matched. However, in semantic image retrieval, spatial consistency should be in a soft type. Ref. [33] presents the word spatial arrangement to describe the rough distribution of the visual words in an image and compute the similarity between images. In [15], the images are indexed by descriptive visual

words (DVWs) and visual phrases (DVPs). In [32], visual phrases are constructed to embed spatial layout constraints in image retrieval.

2.3. Using multiple queries

Typically, the input of the image retrieval system is one query image. For one query image, the following shortcomings catch attention: (1) too many local features extracted in the query; (2) it is difficult to remove the noise and unstable features: (3) the significant visual words cannot be selected. To address the above problems, multiple queries is proposed. The multiple queries are achieved in two ways: by asking the user to input directly [34] and utilizing the feedback of retrieval result such as query expansion [5–6]. On one hand, the multiple queries can be used to select key points, as in [34] the words that appear at least two among queries are regarded as key points. And in our previous work [35], identical salient point (ISP) is detected from the topic album which contains a set of relevant images of the same landmark to measure the viewpoint of each image. Similar to [35], in this paper, we learn salient visual word (SVW) from multiple photos which are mined from mobile end. The SVW requires not only the ISPs are similar in descriptor space but also the features in an ISP are assigned to the same visual words. And in this paper we rank the SVWs according to their significance. On the other hand, multiple queries are useful to expand the synonyms [13] or enrich the query model [5–6]. Query expansion [5–6] improves the representative model of query by combining it with new results returned every time and adopts the complex RANSAC to perform geometric verification. Contrast to query expansion, multiple relevant photos in our approach derive from the initial result, and we represent the query model as a set of concise salient visual words.

2.4. Compact descriptor for mobile image retrieval

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 [42,43]. To deal with the challenges of low bit rate, compact descriptors are proposed to replace SIFT, such as CHoG [25,26], PCA-SIFT [27] and CEDD [36]. And some words focus on compressing the BoW histogram [22–24,28,29]. One way to compress the BoW histogram is removing the redundancy, e.g. BoW histogram is encoded as intervals between positive-count nodes of scalable vocabulary tree in [28]. Another way is reducing the scale of vocabulary tree. As in [29], some trivial branches of vocabulary tree are pruned to decrease the dimension of BoW. To further reduce the transmitted data and meanwhile maintain the performance, sparse coding is introduced. Sparse coding compresses the original BoW histogram of query by reconstructing the it with a linear combination of some bases [22,23,37,41,44,49,51]. Thus the high dimensional BoW histogram is projected into a low dimensional vector via a transformation matrix or dictionary. 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 [38] can learn the dictionary from original BoW codebook. Considering the loss of information from dimension reduction, Fu et al. [56] propose locally adaptive subspace and similarity metric learning based on locally embedded analysis which preserves the local nearest neighbor affinity. To control the data size, the geometry information is disregarded in many works. In [24], the orientation of visual word is transmitted along with 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 mine salient visual words from multiple relevant photos. The salient visual words are discriminative, and their number is small which is suitable for mobile image retrieval. Furthermore, we rank them according to their stability and significance to achieve scalable transmission.

3. System overview

As shown in Fig. 1, the proposed mobile image retrieval approach consists of the following three steps: (1) multiple relevant photos mining; (2) salient visual words learning and re-ranking; and (3) performing spatial verification to re-rank the initial retrieval results. After a user appoints one of the images in his mobile as the query image, our system mines multiple most relevant photos automatically according to the visual similarity. Then, with the multi-photos, we extract salient visual words from them. The salient visual words are the most stable and prominent words that represent the pivotal content of the multiple photos. To make our algorithm adaptive to labile wireless channel, we rank the salient visual words according to their stability in multiple relevant images. Thus in the circumstance that bandwidth is narrow, we transmit part of the salient visual words to server end. In the server end, we perform spatial verification to re-rank the initial results that are retrieved by SVWs. Because the corresponding visual word of the noisy feature in dataset image may be same with salient visual word, spatial verification can judge whether the features matched in word level are truly matched in spatial level.

4. Mining multiple photos

As a user usually takes many photos of the same object, it is possible that there are many photos relevant to the appointed query image. Our aim is to find visually similar images in the user's mobile end and extract salient visual words from them for retrieval.

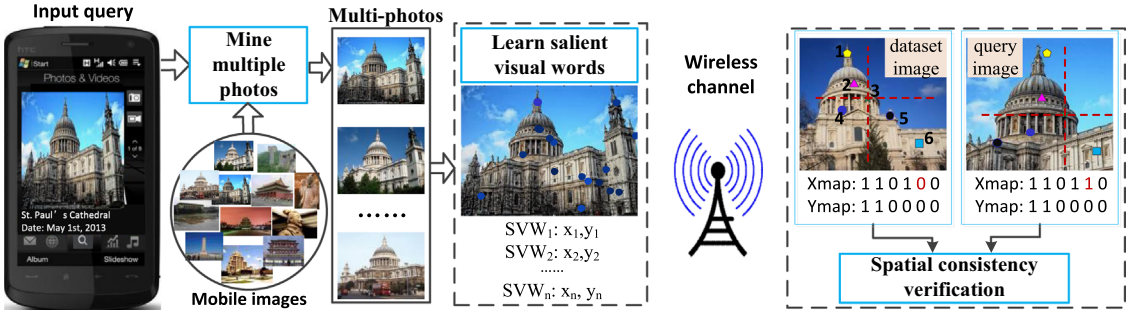

Fig. 1. The flowchart of the whole system. After a user inputs the query, our approach first mines multiple most relevant photos. Then the salient visual words (SVWs) are learned and transmitted to server end along with the corresponding coordinate positions. In server end, SVWs are used to search candidate similar images. The position information of SVWs is used to re-rank the similar images by spatial consistency verification. A patch of the query is amplified to illustrate the spatial verification in the right.

In mobile end image retrieval, some valuable contextual information can be utilized to mine multiple relevant photos such as the temporal information and GPS information. With the information, we can preliminarily remove most of images which are much different with the query. Thus we just need to perform visual similarity on fewer images in mobile end to reduce the computational cost. In this paper, we focus on utilizing visual information to mine multiple photos.

We describe each image with a set of local features. An image represented through local features can be more powerful than global features [39]. SIFT (scale invariant feature transform) feature is robust against illumination, affine change, scale and other local distortions [2]. A SIFT feature consists of a 128-D descriptor vector and a 4-dimensional DoG key-point detector vector (x, y, scale, and orientation). Each of the 128-dimension SIFT descriptors of an image is quantized to a bag-of-words visual vocabulary with W codebooks by hierarchical quantization [13].

To mine the most relevant multiple photos, we measure the similarity between the query and other images in mobile end. 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 the k-th image in smart phone to query, $D(k)$, can be calculated using the city block distance as following:

$$D(k) = \exp(-|h_q - h_m(k)|) \quad (1)$$

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

We sort the similarity scores in descending order. The top ranked $M-1$ results along with the original query 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. If the similarity score of one candidate photo is too small, we eliminate it. And if the similarity score of one is too big e.g. it is approximately equal to 1, then the image may be a duplicate of the query and should be removed as well. The remnant X candidates are final multiple relevant photos which are used for exploring saliency. If no multiple photos are mined finally, the retrieval system degrades into mobile visual search with single query. In this paper, the typical BoW model is used to search similar images in this case.

5. Mining and ranking SVW

After finding multiple relevant photos for the query image at a user's mobile end, we learn the robust and distinctive salient visual words from these relevant photos. Since people focus on their object when they are taking pictures, the object will exist in most of their photos. Thus the object occurs more frequently than disturbance in these photos, i.e. the frequency of visual words corresponding to crucial content is higher than that to background as the background is always changed if user takes photo in different viewpoints. As shown in Fig. 2, the tower is the object, which occurs more frequently than the trees and other buildings. Our purpose is to pick out these high-frequency salient visual words for retrieval. Then, to achieve scalable mobile image retrieval, we rank the salient visual words before transmission.

5.1. Detecting identical semantic point

We mine salient visual word based on identical semantic point (ISP) detection in our previous work [35,54]. 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 a landmark. As in [35], 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. ISP detection is based on the idea that one feature has its optimal

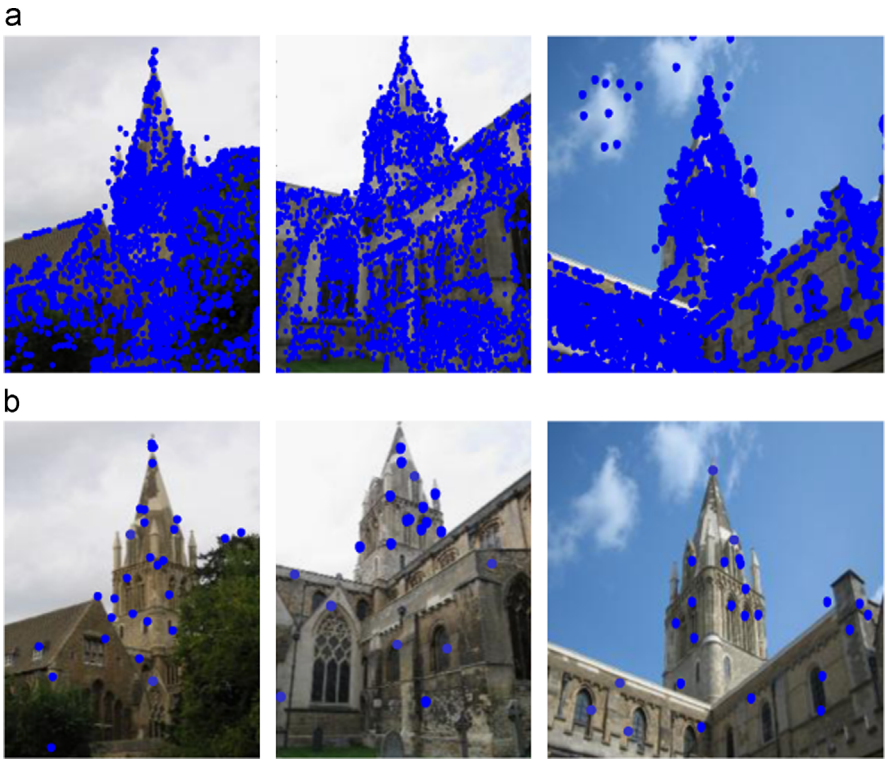

Fig. 2. The comparison between raw SIFT features and extracted ISPs. The average number of SIFT points is 3697, while the average number of SVWs is 42. (b) Salient visual words mainly occur in the tower, the common object of the multiple photos.

matched feature in another similar image. In theory, the optimal matched features should represent identical visual content. 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 which one SIFT is matched with is shrunk tremendously.

Firstly, we find common words that at least two of the mined multiple relevant photos share. Given that w is a visual word that occurs in the i-th and the j-th image, we denote the local features that are assigned to w in the two images as S^i and S^j respectively.

Following [35], then 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) is measured as follows:

$$MS(u,q) = (u \times q^T)/(|u| \times |q|) \tag{2}$$

where u denotes 128-D SIFT descriptor vector from S^i and, q is from S^j. $|x|$ denotes the norm of vector x.

Identical Salient Points (ISP) is determined based on the matching score. An ISP is a set of matched SIFT points, denoted as

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

where ISP_l denotes the l-th ISP, X denotes the number of multiple images, d_l^i is the SIFT ID of the l-th ISP in the i-th image, which implies the occurrence of the l-th ISP in the i-th image. $d_l^i = 0$, if no feature in the i-th image matches with other features in ISP_l.

The corresponding visual word of the ISP is defined as salient visual word (SVW). SVWs are pertinent to the crucial content, and the number of SVWs is very small. As shown in Fig. 2, the average SIFT point number of the three images is 3697, while the average SVW number is only 42, which is about 1% of raw SIFT features. And the SVW rarely occurs in the trees or the lower house, which manifests that extracting SVW eliminates the noise effectively. Owing to the small number of SVWS, we can cut down the time cost on searching remarkably.

5.2. Ranking the salient visual word

Wireless channel is vulnerable to interference. There exists serious latency when mobile devices suffer from weak signal. To adapt to the variant wireless channel, we propose scalable retrieval. We rank the salient visual words according to their contribution to the retrieval, so that we can adjust the data volume to the channel condition. We rank the SVWs in two levels: frequency of occurrence of SVW to rank them on the whole and stability in the multi-photos to rank them in detail.

We denote occurrence of an ISP in multiple relevant images as C

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

where, c_l^i stands for the occurrence of the l-th ISP in the i-th image. $c_l^i = 1$, if $d_l^i \neq 0$, otherwise $c_l^i = 0$.

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

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

Thus by ranking the consistency score CS for all the identical salient points, we rank the SVWs on the whole. The ISPs with equal frequency are put on an equal footing.

Then we rank the SVWs in detail. We accumulate the total matched score of the descriptors in an ISP to measure the stability (Sta) of the ISP as follows:

$$Sta_l = \sum_{i,j,i \neq j} MS(d_l^i, d_l^j) \tag{6}$$

In general, the SVWs are firstly ranked in light of CS in descending sort. Then for the SVWs with same occurring frequency in multiple photos, they are ranked according to Sta. After the SVWs are ranked, a fit number of SVWs that ranked highly are transmitted according to the available bandwidth of wireless channel.

6. Spatial verification on SVW

In mobile end, we have finished mining multiple relevant photos, learning salient visual words from multiple relevant photos and ranking the salient visual words. The salient visual words along with their coordinate information in the query image are sent over to the server end. In server end, we first search the candidate similar images as the initial results through an inverted file indexing structure. The candidate similar image should contain at least one visual word that is the same with the salient visual word transmitted from the mobile end.

For the candidate similar images, we perform spatial verification to re-rank the initial retrieval results. Spatial coding [17] is adopted to describe the relative position among SVWs. It is possible that the mined multiple images are all eliminated and only the input is remained. In this case, we refine the features extracted from the query image as in [21].

Firstly, SIFT features assigned to the same visual word will be considered as valid match when its orientation difference with the query feature is less than π/t. t is set as 4 in this paper. More discussions are given in Section 7.4.4.

Spatial coding encodes the spatial relationship among visual words in an image into two binary maps: X-map and Y-map. The two maps describe the relative position of each valid feature pairs.

Each element in X-map and Y-map is defined as following:

$$X\ \mathrm{map}_{i,j} = \begin{cases} 1 & \text{if } x_i < x_j \\ 0 & \text{if } x_i > x_j \end{cases} \tag{7}$$

$$Y\ \mathrm{map}_{i,j} = \begin{cases} 1 & \text{if } y_i < y_j \\ 0 & \text{if } y_i > y_j \end{cases} \tag{8}$$

where x_i and x_j denote the horizontal coordinate of the i-th feature and the j-th feature, respectively, and y_i and y_j denote the vertical coordinate.

For query image Iq and matched image Im, X-map and Y-map are generated for each, denoted as (Xq, Yq) and (Xm, Ym), which encode the spatial relationship among the salient visual words which occur in database image. Hence, to verify the spatial layout of common visual words is to compare the X-map and Y-map. Logical Exclusive OR (XOR) operation $\oplus$ is performed on the spatial maps as following:

$$SV_X = X_q \oplus X_m \tag{9}$$

$$SV_Y = Y_q \oplus Y_m \tag{10}$$

where SV_X and SV_Y denote the difference in X-map and Y-map.

Thus the spatial difference of matched features in two images can be denoted as

$$SP_X(i) = \sum_{j=1}^{N} SV_X(i,j) \tag{11}$$

$$SP_Y(i) = \sum_{j=1}^{N} SV_Y(i,j) \tag{12}$$

where N denotes the number of common visual words. $SP_X(i)$ and $SP_Y(i)$ denote the spatial consistency of the i-th common visual word.

For partial duplicate image retrieval, $SP_X(i)$ and $SP_Y(i)$ are required to be zero strictly if the i-th common visual word are truly matched in Zhou's paper [17]. However, for universal image retrieval, too rigorous spatial constraint may regards the true matched features as false. To address this problem, we change the absolute way of judgment into a soft way, i.e. calculating the consistency score as follow:

$$Score = \sum_{i=1}^{N} \exp(-(SP_X(i)+SP_Y(i))/N) \times R(i) \quad (13)$$

where Score denotes the spatial consistency score of two images. $R(i)$ is a binary function. $R(i)=1$, if $(SP_X(i)+SP_Y(i))/N < thr$, otherwise $R(i)=0$. thr is the threshold. More discussions are given in Section 7.4.1.

After computing the spatial consistency score for each initial retrieved image, the initial results are re-ranked according to their spatial consistency with query image.

7. Experimentation

Most of our experiments are conducted on the Oxford Buildings Dataset. The scalable vocabulary tree (SVT) is learned on the dataset, including 61,724 leaf nodes in total. To show the effectiveness of our approach, we compare our method with Query Expansion (QE) [5] and the original spatial coding (SP) [17]. We denote our method as SSV. Some main factors that influence the performance are discussed as well. In addition, a bigger dataset, GOLD [40,45], is used to test our approach. The depth of the visual vocabulary used to quantize the GOLD is 8 levels, and the branch factor is 10. The further testify our approach, we create a real-world mobile image collection and perform our approach on it.

7.1. Datasets

The main dataset we tested our approach on was the Oxford Buildings Dataset which affords the test collection and ground truth [55]. The Oxford Buildings Dataset consists of 5062 images collected from Flickr by searching for particular Oxford landmarks, 11 landmarks in total. For each landmark, 5 possible queries are given. Our test set consists of the given 55 query images. For each query, the similar images are given in three types: good, OK and junk. We carried out retrieval with each query. If the result is one of the good or OK collections, it is regarded as a right result. The first step of our approach, obtaining multiple relevant photos, is run on Oxford Buildings set. If the system is applied in reality, the first step should be performed on photos stored in mobile end.

The other testing dataset is GOLD dataset which is a geo-tagged large scale web image set [40], which is crawled from Flickr. 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.

7.2. Evaluation criterion

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

$$P@K = (1/T) \times \sum_{i=1}^{T} (R_i/K) \quad (14)$$

where T is the size of test set, $T=55$ in this paper. R_i denotes the number of retrieved relevant images up to K for the i-th query image.

7.3. Performance comparison

We compare our approach with four typical methods: (1) the query expansion (QE) [5]; (2) original spatial coding (SP) [17]; (3) BoW model (BoW) [1]; and (4) Lasso based sparse coding (SCL) [37]. To be fair with our approach, no query region is specified in QE. The input of QE is the whole image. In SP, all the features extracted from the query image are used for retrieval. The Lasso based sparse coding compresses the 61,724-dimensional BoW histogram to 1501 dimensions. Our approach is denoted as SSV. The results shown in Fig. 3 demonstrate the effectiveness of our approach. Owing to the too strict requirement in spatial consistency, SP performs inferior in universal image retrieval to that in duplicate retrieval. When the object is not clear or occupies a small region of query, QE cannot perform well. The sparse coding method compresses the BoW histogram and reconstructs it in server end, which brings about loss. So the performance of sparse coding is inferior to BoW model. Our approach performs best with the least data because our approach can mine the salient visual word which is closely relevant to the crucial content of the query image.

In addition, to show the less necessary data volume of our approach, we estimate the data size of different methods. In our approach, the salient visual words along with their corresponding

Table 1
The comparison of necessary data size.

Approaches	SSV	SP	QE	BoW	SCL	JPEG
Data (bytes) (K)	600	18	316	60.3	5	385.8
Percent (%)	0.16	4.67	81.97	15.64	1.30	100

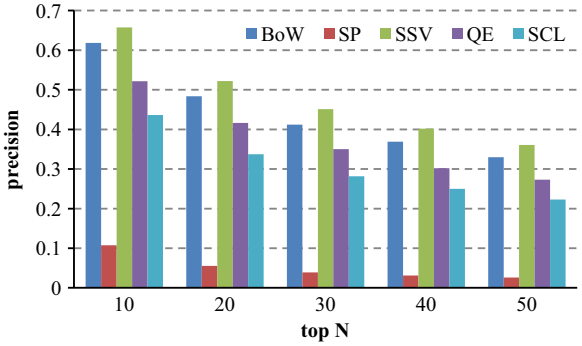

Fig. 3. The mean precision of the five different methods.

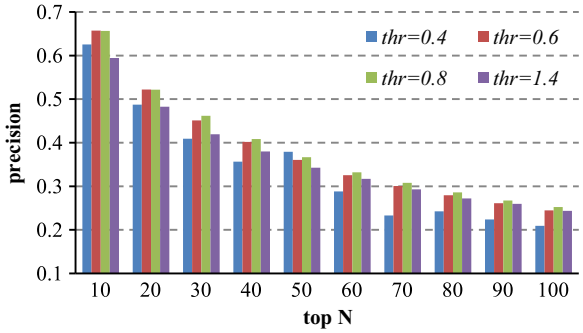

Fig. 4. The performance for different thr value.

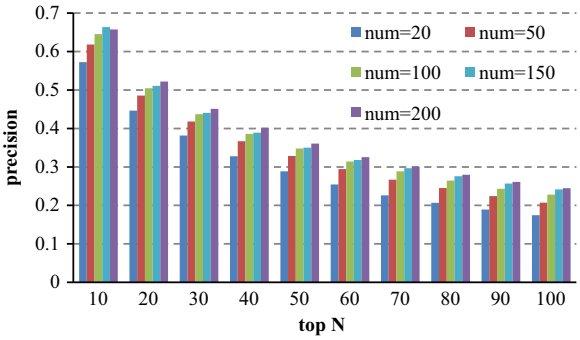

Fig. 5. The comparison for different data volume.

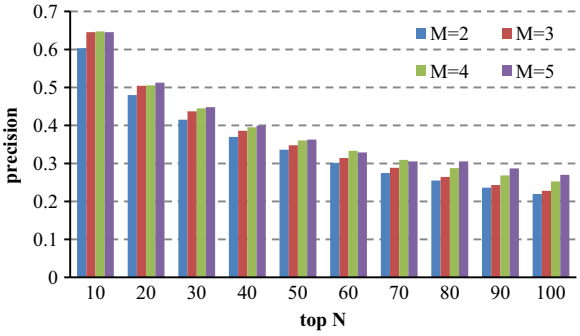

Fig. 6. The comparison for different values of *M*.

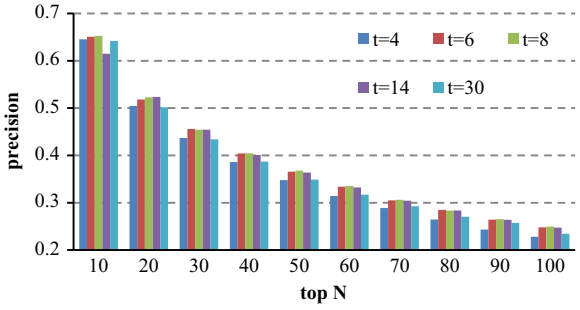

Fig. 7. The comparison for different values of *t*.

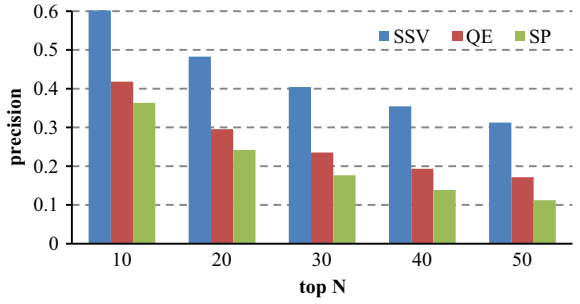

Fig. 8. The performance on GOLD dataset.

horizontal and vertical coordinates are transmitted. Considering the sparse distribution of SVWs, their coordinates can be rounded to short integer (2 bytes) memory. And each SVW needs 2 bytes. So each SVW along with its horizontal and vertical coordinate needs 6 bytes. Supposing that 100 SVWs are transmitted, 600 bytes are needed, while the compact descriptor like [37] needs about 5 K bytes. Table 1 show the data size of different methods. It demonstrates that our approach needs the least bandwidth source. We find that our approach only requires 12% of the data of compact descriptor [37].

From Table 1, we find that the average amount of data JPEG image is about 385.8 K bytes. When we repesent the image by features, the total amount of data can be decreased dramatically. In BoW, QE, SP and SSV, only15.64%, 81.97%, 4.67% and 0.16% of raw JPEG are required. It is interesting to find that our SSV is only required 3.33% data of SP while with the best performances.

7.4. Discussion

The performance of our approach is influence by following factors: (1) *thr* which judges whether a matched pair is spatial consistent; (2)the number of SVWs that are transmitted to server end; (3) *M*, i.e. the number of candidate multiple photos; and (4) *t* which controls the orientation difference between the matched visual words. We discuss their impact in this section. And finally we test our approach on GOLD.

7.4.1. The impact of thr

The parameter *thr* determines whether a matched feature pair is regarded as truly matched. Fig. 4 shows the performance with different *thr* values. The results show that the performance is the best when *thr* is around 0.8. Bigger *thr* will not lead a better performance, because some actually false matching will be taken as right matching. And over small values exclude a part of truly matched pairs.

7.4.2. The impact of data volume transmitted

Another main factor that influences the retrieval performance is the number of salient visual words that are sent to the server terminal. We use 20, 50,100, 150, and 200 SVWs to carry out retrieval respectively. Fig. 5 shows that more SVWs result in a better performance. However, when the data volume reaches 100 SVWs, the rising trend of precision decelerates. And we find that 20 SVWs are enough for retrieval, for SVWs are pertinent to the crucial content of the query image.

7.4.3. The impact of M

For we have removed noisy images from candidate multiple relevant photos, the number of the multiple photos that are actually used to mine salient visual words is not definite. We use *M* to discuss the impact of number of multiple images. The parameter *M*, i.e. the number of candidate multi-relevant photos, has impact on both precision and computational complexity. Fig. 6 shows that bigger *M* produces a better result. However, bigger *M* will expand the computational cost since mining SVW needs to match SIFT features between every two images. And the result presents that rising tendency of the performance turns slow from $M=4$. Considering the performance and complexity comprehensively, we set $M=3$ in our other experiments.

7.4.4. The impact of t

The parameter *t* influences the number of matched visual words in database image that are qualified to construct spatial map. Generally, bigger *t* filters out more visual words. Thus our approach performs spatial verification on less matched pairs, which reduces the necessary time to finish spatial verification. Actually, as shown in Fig. 7, *t* effects little to the precision, since the falsely matched visual words will not get through spatial consistency verification. However, the performance deteriorates if t is set too large, because the over strict constraint in orientation may remove some truly matched visual words.

Fig. 9. The query and the mine two relevant photos in mobile end.

Fig. 10. The top 10 results of test on mobile collection.

7.4.5. The performance on GOLD

To show the effectiveness of our method on large dataset, we test our approach on GOLD. 100 SVWs are used for retrieval in our approach (SSV). Fig. 8 shows the result. Our approach still performs superior to QE and SP. Moreover, the precision at top 10 is 0.6018 on GOLD dataset, and 0.6527 on oxford buildings dataset in the condition that all the parameters are set the same. Therefore, our approach is capable to be applied to the large scale image retrieval.

7.4.6. The performance on mobile image collection

To further testify our approach, we create a mobile photos collection to perform the first step. The real mobile photo collection contains 186 images from my mobile phone. A photo of the library of Xi'an Jiaotong University is set as the query. And some images of the library are put into Oxford building dataset as the database in server end. These photos are shared by 6 volunteers and downloaded from the Internet.

First, we select the query which was circled by black frame in Fig. 9. Then two more similar images were searched in mobile end. The other two images in Fig. 9 are the relevant photos.

We learned semantic features from above three images, and used 150 SVWs for retrieval on the extended Oxford building dataset. The top 10 retrieval results are shown in Fig. 10. In Fig. 10, there are 6 of 10 images relevant to the query, which demonstrates the effectiveness in mobile platform of our approach.

8. Conclusion

In this paper, we propose a novel mobile image retrieval scheme based on learning salient visual words from multiple relevant photos. The salient visual word is more robust and really pertinent to the theme of the query. Our approach achieves the better performance with less data. Generally our method requires less than 1 KB bandwidth to transmit data. In extreme condition, e.g. the user is in remote mountain area, we can transmit only hundreds of bytes to carry out the retrieval. Our future work will focus on mining salient visual words from single query image to make our method available in the case that multiple relevant images cannot be mined in mobile end. And we are trying to improve our approach to further speed up the process of learning salient word.

Conflict of interest

None declared.

References

[1] J. Sivic, A. Zisserman, Video google: a text retrieval approach to object matching in videos, in: Proceedings of ICCV, 2003.
[2] D.G. Lowe, Distinctive image features from scale-invariant keypoints, IJCV 60 (2) (2004) 91–110.
[3] H. Bay, T. Tuytelaars, L.V. Gool, Surf: speeded up robust features, in: Proceedings of ECCV, 2006.
[4] D. Nistér, H. Stewénius, Scalable recognition with a vocabulary tree, in: Proceedings of CVPR, 2006.
[5] O. Chum, J. Philbin, J. Sivic, M. Isard, A. Zisserman, Total recall: automatic query expansion with a generative feature model for object retrieval, in: Proceedings of ICCV, 2007.
[6] O. Chum, A. Mikulík, M. Perdoch, J. Matas, Total recall II: query expansion revisited, in: Proceedings of CVPR, 2011.
[7] R. Arandjelovic, A. Zisserman, Three things everyone should know to improve object retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012.
[8] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Lost in quantization: improving particular object retrieval in large scale image datasets, in: Proceedings of CVPR, 2008.
[9] W. Tang, R. Cai, Z. Li, L. Zhang, Contextual synonym dictionary for visual object retrieval, in: Proceedings of ACM MM, 2011.

[10] E. Gavves, Cees G.M. Snoek, Landmark image retrieval using visual synonyms, in: Proceedings of ACM, MM, 2010.
[11] E. Gavves, C. Snoek, and A. Smeulders, Visual synonyms for landmark image retrieval, in: Proceedings of CVIU, 2011.
[12] A. Mikulik, M. perdoch, O.Chum, J. Matas, Learning a fine vocabulary, in: Proceedings of ECCV, 2010, pp. 1–14.
[13] Y. Xue, X. Qian, B. Zhang, Mobile image retrieval using multi-photo as query, in: Proceedings of ICMEW, 2013.
[14] J. Chen, B. Feng, L. Zhu, P. Ding, B. Xu, Effective near-duplicate image retrieval with image-specific visual phrase selection, in: Proceedings of ICIP, 2010.
[15] S. Zhang, Q. Tian, G. Hua, Q. Huang, S. Li, Descriptive visual words and visual phrases for image, in: Proceedings of ACM MM, 2009.
[16] M.A. Fischler, R.C. Bolles, Random sample consensus, Commun. ACM 24 (6) (1981) 381–395.
[17] W. Zhou, Y. Lu, H. Li, Y. Song, Q. Tian, Spatial coding for large scale partial-duplicate web image search, in: Proceedings of ACM MM, 2010.
[18] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, T. Han, Contextual weighting vocabulary tree based image retrieval, in: Proceedings of ICCV, 2011.
[19] M. Perdoch, O. Chum, J. Matas, Efficient representation of local geometry for large scale object retrieval, in: Proceedings of CVPR, 2009.
[20] M. Marszałek, C. Schmid, Spatial weighting for bag-of-features, in: Proceedings of CVPR, 2006.
[21] S. Zhang, Q. Huang, G. Hua, S. Jiang, W. Gao, Q. Tian, Building contextual visual vocabulary for large-scale image application, in: Proceedings of ACM MM, October 2010.
[22] J. Lin, L.Y. Duan, J. Chen, R. Ji, Learning multiple codebooks for low bit rate mobile visual search, in: Proceedings of ICASSP, 2012, pp. 933–936.
[23] R. Ji, L.Y. Duan, J. Chen, H. Yao, Y. Rui, W. Gao, Location discriminative vocabulary coding for mobile landmark search 96 (2012) 290–314Int. J. Comput. Vis. 96 (2012) 290–314.
[24] Y. Wu, S. Lu, T. Mei, J. Zhang, S. Li, Local visual words coding for low bit rate mobile visual search, in: Proceedings of ACM Multimedia, 2012, pp. 989–992.
[25] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, CHoG: Compressed histogram of gradients A low bit-rate feature descriptor, in: Proceedings of CVPR, 2009.
[26] B. Girod, V. Chandrasekhar, D. Chen, N. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. Tsai, R. Vedantham, Mobile visual search, IEEEE Signal Process. Mag. 28 (2011) 61–76.
[27] Y. Ke, R. Sukthankar, PCA-SIFT: a more distinctive representation for local image descriptors, in: Proceedings of CVPR, 2004.
[28] D. Chen, S. Tsai, V. Chandrasekhar, G. Takacs, Tree histogram coding for mobile image matching, in: Proceedings of DCC, 2009, pp. 143–152.
[29] J. Chen, L.Y. Duan, R. Ji, W. Gao, Pruning tree-structured vector quantizer towards low bit rate mobile visual search, in: Proceedings of ICASSP, 2012, pp. 965–968.
[30] Z. Wu, Q. Ke, M. Isard, J. Sun, Bundling features for large scale partial-duplicate web image search, in: Proceedings of CVPR, 2009.
[31] J. Yuan, Y. Wu, M. Yang, Discovery of collocation patterns: from visual words to visual phrases, in: Proceedings of CVPR, June 2007, pp. 1–8.
[32] Y. Zhang, Z. Jia, T. Chen, Image retrieval with geometry-preserving visual phrases, in: Proceedings of CVPR, 2011.
[33] O.A.B. Penatti, F.B. Silva, E. Valle, V. Gouet-Brunet, R. da, S. Torres, Visual word spatial arrangement for image retrieval and classification, Pattern Recognit. 47 (2) (2014) 705–720.
[34] B. Fernando, T. Tuytelaars, Mining multiple queries for image retrieval: on-the-fly learning of an object-specific mid-level representation, in: Proceedings of ICCV, 2013.
[35] Y. Xue, X. Qian, Visual summarization of landmarks via viewpoint modeling, in: Proceedings of IEEE International Conference on Image Processing, 2012.
[36] S. Chatzichristofis, Y. Boutalis, CEDD: Color and edge directivity descriptor: a compact descriptor for image indexing and retrieval, Comput. Vis. Syst. 5008 (2008) 312–322.
[37] R. Ji, H. Yao, W. Liu, X. Sun, Q. Tian, Task-dependent visual codebook compression, IEEE Trans. Image Process. 21 (4) (2012) 2282–2293.
[38] R. Tibshirani, Regression shrinkage and selection via the Lasso., J. R. Stat. Soc. 58 (1) (1996) 267–288.
[39] A. Qamra, E. Chang, Scalable landmark recognition using EXTENT, Multimed. Tools Appl. 38 (2) (2008) 187–208.
[40] J. Li, X. Qian, Y. Tang, L. Yang, GPS estimation from users' photos, in: Proceedings of MMM, 2013, pp. 118–129.
[41] R. Ji, L. Duan, J. Chen, W. Gao, Towards compact topical descriptor, in: Proceedings of CVPR, 2012.
[42] H. Li, Y. Wang, T. Mei, J. Wang, S. Li, Interactive multimodal visual search on mobile device, IEEE Trans. Multimed. 15 (3) (2013) 594–607.
[43] J. Sang, T. Mei, Y. Xu, C. Zhao, C. Xu, S. Li, Interaction design for mobile visual search, IEEE Trans. Multimed. 15 (7) (2013) 1665–1676.
[44] X. Yang, L. Liu, X. Qian, T. Mei, J. Shen, Q. Tian, Mobile visual search via hierarchical sparse coding, in: Proceedings of ICME, 2014.
[45] J. Li, X. Qian, Y. Tang, L. Yang, T. Mei., GPS estimation for places of interest from social users' uploaded photos, IEEE Trans. Multimed. 15 (8) (2013) 2058–2071.
[46] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, Y. Pan., A multimedia retrieval framework based on semi-supervised ranking and relevance feedback, IEEE Trans. Pattern Anal. Mach. Intell. 34 (4) (2012) 723–742.
[47] L. Shao, L. Liu, X. Li, Feature learning for image classification via multi-objective genetic programming, IEEE Trans. Pattern Anal. Mach. Intell. 25 (7) (2014) 1359–1371.
[48] Y. Yang, Z. Ma, A.G. Hauptmann, N. Sebe, Feature selection for multimedia analysis by sharing information among multiple tasks, IEEE Trans. Multimed. 15 (3) (2013) 661–669.
[49] C. Yang, J. Shen, J. Peng, J. Fan, Image collection summarization via dictionary learning for sparse representation, Pattern Recognit. 46 (3) (2013) 948–961.
[50] W. Min, B. Bao, C. Xu, Multimodal spatio-temporal theme modeling landmark analysis, IEEE Multimed. 21 (3) (2014) 20–29.
[51] J. Huang, H. Liu, J. Shen, S. Yan, Towards efficient sparse coding for scalable image annotation, in: Proceedings of ACM MM, 2013.
[52] L. Shao, F. Zhu, X. Li., Transfer learning for visual categorization: a survey, IEEE Trans. Neural Netw. Learn. Syst. (2014), http://dx.doi.org/10.1109/TNNLS.2014.2330900.
[53] F. Zhu, L. Shao., Weakly-supervised cross-domain dictionary learning for visual recognition, Int. J. Comput. Vis. 109 (1–2) (2014) 42–59.
[54] X. Qian, Y. Xue, Y. Tang, X. Hou, T. Mei, Landmark summarization with diverse viewpoints, IEEE Trans. Circuits Syst. Vid. Technol. (2015), http://dx.doi.org/10.1109/TCSVT.2014.2369731.
[55] X. Li, C. Wu, C. Zach, S. Lazebnik, J. Frahm, Modeling and recognition of landmark image collections using iconic scene graphs, in: Proceedings of ECCV, 2008, pp. 427–440.
[56] Y. Fu, Z. Li, T.S. Huang, A.K. Katsaggelos, Locally adaptive subspace and similarity metric learning for visual clustering and retrieval, Comput. Vis. Image Underst. (CVIU), Spec. Issue Similarity Matching Comput. Vis. Multimed. 110 (3) (2008) 390–402.
[57] M. Shao, S. Li, T. Liu, D. Tao, T.S. Huang, Y. Fu. Learning relative features through adaptive pooling for image classification, in: Proceedings of IEEE International Conference on Multimedia and Expo (ICME), 2014, pp. 1–6.

Xiyu Yang received the B.A. degree from Lanzhou University in 2012, and now is a Master student in SMILES lab, Xi'an Jiaotong University.
Her research interests include computer vision, large scale image retrieval & recognition and data mining and knowledge discovery from social multimedia.

Xueming Qian (M'10) received the B.S. and M.S. degrees in Xi'an University of Technology, Xi'an, China, in 1999 and 2004, respectively, and the Ph.D. degree in the School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, China, in 2008. From 1999 to 2001, he was an Assistant Engineer at Shannxi Daily. From 2008 to 2011, he was an assistant professor, from 2014 till now, he was an full professor of the School of Electronics and Information Engineering, Xi'an Jiaotong University. His research interests include Social mobile multimedia mining learning and search. He is the director of SMILES LAB. He has authored or co-authored over 70 papers in journals and conferences. His research is supported by Microsoft, NSFC, etc.
He was awarded Microsoft fellowship in 2006. He was a visit scholar at Microsoft research Asia from Aug. 2010 to March 2011. He was TPC member of ICME, Multimedia Modeling, ICIMCS, and he is the session chairs/organizers of VIE08, ICME14,MMM14. He is a member of IEEE, ACM and Senior member of CCF.

Tao Mei (M'07-SM'11) is a Researcher with Microsoft Research Asia, Beijing, China. He received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively. His current research interests include multimedia information retrieval and computer vision. He has authored or co-authored over 140 papers in journals and conferences, eight book chapters, and edited two books. He holds six U.S. granted patents and more than 30 in pending.
Dr. Mei was the recipient of several paper awards from prestigious multimedia conferences, including the Best Paper Award and the Best Demonstration Award at ACM Multimedia in 2007, the Best Poster Paper Award at the IEEE MMSP in 2008, the Best Paper Award at ACM Multimedia in 2009, the Top 10% Paper Award at the IEEE MMSP in 2012, the Best Paper Award at ACM ICIMCS in 2012, the Best Student Paper Award at the IEEE VCIP in 2012, and the Best Paper Finalist at ACM Multimedia in 2012. He was the principle designer of the automatic video search system that achieved the best performance in the worldwide TRECVID evaluation in 2007. He received Microsoft Gold Star Award in 2010, and Microsoft Technology Transfer Awards in 2010 and 2012. He is an Associate Editor of Neurocomputing and the Journal of Multimedia, a Guest Editor of the IEEE Transactions on Multimedia, the IEEE Multimedia Magazine, the ACM/Springer Multimedia Systems, and the Journal of Visual Communication and Image Representation. He is the Program Co-Chair of MMM 2013, and the General Co-Chair of ACM ICIMCS 2013. He is a Senior Member of ACM.