Spatial Verification for Scalable Mobile Image Retrieval

Xiyu Yang and Xueming Qian*
SMILES LAB, Xi'an Jiaotong University, Xi'an China, 710049
yangxiyu@stu.xjtu.edu.cn; qianxm@mail.xjtu.edu.cn

ABSTRACT
Owing to the portable and excellent phone camera, people now prefer to take photos and upload them by mobile phone. Content based image retrieval is effective for users to obtain relevant information about a photo. Taking the limited bandwidth and instability into account, we propose an effective scalable mobile image retrieval approach in this paper. The proposed mobile image retrieval algorithm first determines the relevant photos according to visual similarity in mobile end, then mines salient visual words by exploring saliency from multiple relevant images, and finally we determine the contribution order of salient visual words for scalable retrieval. In server terminal, spatial verification is performed 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 a part of features according their contributions to retrieval. Experimental results show the effectiveness of the proposed approach.

Categories and Subject Descriptors
I.2.10 [**Vision and Scene Understanding**]: VISION

General Terms
Algorithms, Measurement, Performance, Experimentation, Verification

Keywords
Mobile Image Retrieval; Scalable Retrieval; Salient Visual Word (SVW); Multiple Relevant Photos; Spatial Verification

1. INTRODUCTION
Recent years, the research on content based image retrieval flourished owing to the BoW [1] model and local features, such as SIFT [2]. And Chum et al. [10] proposed to update the query by combining it with retrieval results every time to learn better representation of query and improve the retrieval performance. Usually single visual word is not distinctive and stable enough. To improve the BoW model, co-occurrence pattern and spatial verification are introduced. Co-occurrence pattern constructs visual phrase or group and represents image as bag of visual groups [3]. Spatial verification enforces geometric consistent constraint on common words that query and dataset image share, such as RANSAC [4] and spatial coding [5]. Spatial coding performs well in partial duplicate image retrieval. 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.

Smartphone is experiencing booming development recently. It has been an indispensable part of people's lives. With the pervasiveness of digital image-capture devices such as mobile phone, it is likely that user take many photos about same object or scene. Hence, it is rational to acquire salient visual words from multiple relevant photos. These salient visual words should be stable and significant, which capture the repeated crucial content from multiple photos.

In this paper, a novel spatial verification algorithm is proposed based on salient visual word. Our approach consists of 3 steps: 1) mining multiple relevant photos. Once user inputs a query, our approach automatically mines some relevant photos; 2) extracting salient visual words (SVWs) and ranking them for scalable image retrieval. With the relevant photos, we extract the stable, robust and distinctive visual words from them for image retrieval; 3) re-ranking the retrieval results based on spatial verification to improve the performance. Figure1 shows the flowchart.

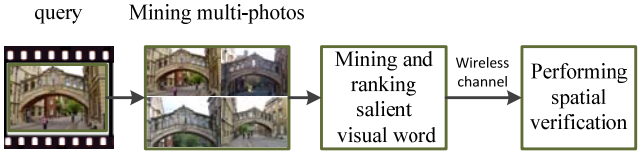

Figure 1. The system flowchart.

The main contributions of this paper are summarized as following: 1) we extract salient visual words, which eliminates the effect of noisy, unstable and irrelevant features; 2) the small number of robust salient visual words reduces the computational complexity of spatial verification and 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 achieves notable performance; 4) considering the instability of invariance of wireless channel, we propose selection scheme for salient visual words, which achieves scalable retrieval.

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

2. SYSTEM OVERVIEW
As shown in Fig. 1, the proposed mobile image retrieval approach consists of three steps: 1) multiple relevant photos mining; 2) salient visual words mining and re-ranking; 3) performing spatial

*Corresponding author.

CIKM'14, November 3–7, 2014, Shanghai, China.

http://dx.doi.org/10.1145/2661829.2661971

verification to re-rank the initial retrieval results. Once the user inputs a query image, our system mines multiple most relevant photos automatically in mobile end. Then, with the multi-photos, we extract salient visual word from them to eliminate noise, improve precision and reduce computational complexity. To make our algorithm adaptive to labile wireless channel, we rank the SVWs according to their stability in multiple relevant images. Thus in the circumstance than bandwidth is narrow, we transmit part of the salient visual words to server terminal. In the server end, we perform spatial verification to re-rank the initial results that are retrieved by SVWs. For the noisy feature in matched dataset image may assigned to same visual word with salient visual word, spatial verification can judge whether the matched feature in truly matched.

3. MINING MULTIPLE PHOTOS

It is possible that, there are many relevant photos to the image that user submitted to retrieval. Our aim is to find visually similar images in the user's mobile end, and to extract salient visual words from them for retrieval.

We describe each image with a set of local features. An image represented through local features can be more powerful than global features [6]. 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 (BOW) visual vocabulary with W codebooks by hierarchical quantization [7].

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 *hq* and *hm*(*k*), the similarity score of *k*-th image in smart phone to query, D(*k*), can be calculated using the city block distance as following:

$$D(k) = \exp(-\left|h_q - h_m(k)\right|) \qquad (1)$$

where $|\square|$ denotes L1 norm, and $k = 1, \cdots, P$, P is the number of images in mobile end, which are primarily from 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. The remnant X candidates are final multiple relevant photos which are used for exploring saliency.

4. MINING AND RANKING SVW

After finding multiple relevant photos for the query image at user's mobile end, we mine the robust and distinctive salient visual words from these relevant photos. 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. Then, to achieve scalable mobile image retrieval, we rank the salient visual words.

4.1 Detecting Identical Semantic Point

We mine salient visual word based on identical semantic point (ISP) detection in our previous work [8]. As in [8], 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 same visual word. Thus the scope of features that 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 i-th and j-th image, we denote the local features that are assigned to *w* in the two images as S^i and S^j respectively.

Figure 2. The comparison between raw SIFT features and extracted ISPs.

Following [8], 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) are measured as follows:

$$MS(u,q) = \left(u \times q^T\right) \Big/ \left(\left|u\right| \times \left|q\right|\right) \qquad (2)$$

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:

$$ISP_l = \{d_l^1, ..., d_l^i, ..., d_l^X\} \qquad (3)$$

where ISP_l denotes the *l*-th ISP, X denotes the number of multiple images, d_l^i is the SIFT ID of *l*-th ISP in *i*-th image, which implies the occurrence of *l*-th ISP in *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 5418, the average ISP number is only 263, which is about 5% of raw SIFT feature. And the ISP rarely occurs in pedestrian and trees, which manifest that extracting SVW eliminates the noise effectively.

4.2 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 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:

$$C_l = \{c_l^1, ..., c_l^i, ..., c_l^X\} \quad (4)$$

where, c_l^i stands for the occurrence of *l*-th ISP in *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 following:

$$CS_l = \sum_{i=1}^{X} c_l^i \quad (5)$$

Thus by ranking the consistency score CS for all the identical salient points, we rank the SVWs on the whole.

Then we rank the SVWs in detail. We accumulate the matched score of the descriptors in an ISP to measure the stability of this ISP. For the SVWs that occur in same number of multiple photos, they are ranked according to the total matched score of the corresponding ISP.

5. SPATIAL VERIFICATION ON SVW

The salient visual words along with their coordinate information in the query image are sent to server end. In server end, we first search the candidate similar images which should contain at least one of the salient visual words transmitted from the mobile end. For the candidate similar images, we perform spatial verification to re-rank them. Spatial coding [5] 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 [9].

Firstly, SIFT feature 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.

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:

$$Xmap_{i,j} = \begin{cases} 1 & if\ x_i < x_j \\ 0 & if\ x_i > x_j \end{cases} \quad (6)$$

$$Ymap_{i,j} = \begin{cases} 1 & if\ y_i < y_j \\ 0 & if\ y_i > y_j \end{cases} \quad (7)$$

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

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:

$$SV_X = X_q \oplus X_m \quad (8)$$

$$SV_Y = Y_q \oplus Y_m \quad (9)$$

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

Thus the spatial consistency of matched feature in two images can be denoted as:

$$SP_X(i) = \sum_{j=1}^{N} SV_X(i,j) \quad (10)$$

$$SP_Y(i) = \sum_{j=1}^{N} SV_Y(i,j) \quad (11)$$

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

For partial duplicate image retrieval, $SP_X(i)$ and $SP_Y(i)$ are required to be zero strictly if *i*-th common visual word is truly matched in Zhou's paper [5]. 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 following:

$$Score = \sum_{i=1}^{N} \left(SP_X(i) + SP_Y(i) \right) / N \times R(i) \quad (12)$$

where Score denotes the spatial consistency score of two images. $R(i)$ is a binary function. $R(i)$=1, if $\left(SP_X(i) + SP_Y(i) \right) / N < thr$, otherwise $R(i)$=0. *thr* is the threshold.

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.

6. EXPERIMENTATION

We conduct our experiments on the Oxford Buildings Dataset. The scalable vocabulary tree (SVT) is learned on the dataset. It includes 61724 leaf nodes in total. To show the effectiveness of our approach, we compare our method with BoW model [1], Query Expansion [10]. Some main factors that influence the performance are discussed as well.

6.1 Dataset

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. 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 conducted on photos stored in mobile end.

6.2 Evaluation Criterion

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

$$P@K = (1/T) \times \sum_{i=1}^{T} (R_i / K) \quad (13)$$

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 *i*-th query image.

6.3 Performance Comparison

We compare our approach with the BoW model and query expansion. In BoW model, the retrieval results are ranked based on their similarity of BoW histogram to the query. In query expansion, 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. In addition, our approach is also compared with the original spatial coding proposed in [5], denoted as SP, in which all

the local features extracted from the query image are used for spatial coding. The results shown in Fig. 3 demonstrate the effectiveness of our approach. SSV denotes our method. Owing to the too strict requirement in spatial consistency, SP performs inferior in universal image retrieval to in duplicate retrieval. When the object is not clear or occupies a small region of query, QE cannot perform well, whereas our approach can mine the salient visual word that is relevant to the object.

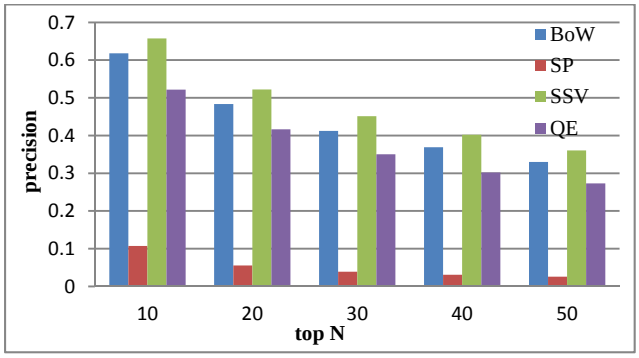

Figure 3. The mean precision of three different methods.

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 horizontal and vertical coordinates are transmitted. Considering the sparse distribution of SVWs, their coordinates are rounded to short integer. Supposing that 50 SVWs are transmitted, 300 bytes are needed. Table 1 show the data size of different methods.

Table 1. The comparison of necessary data size

Approaches	SSV	SP	BoW	JPEG
Data(bytes)	300	18K	60.3K	385.8K

6.4 Discussion

The performance of our approach is influence by two main factors: *thr* and the number of SVWs that are transmitted to server end. We discuss their impact in this subsection.

6.4.1 The impact of thr

The parameter *thr* determines whether a matched feature pair is regarded as truly matched. Figure 4 shows the performance with different *thr* value. The results show that the performance is best when *thr* is around 0.8. Bigger *thr* will not lead better performance, because some actually false matching will be taken for right matching.

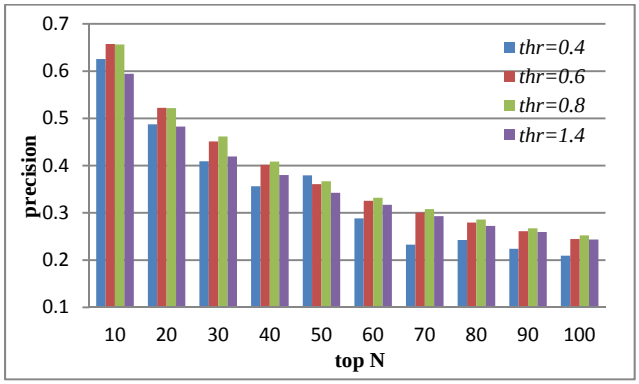

Figure 4. The performance for different *thr* value.

6.4.2 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 server terminal. We use 20, 50,100, and 200 SVWs for retrieval respectively. Figure 5 shows that more SVWs result in better performance. However, when the data volume reaches 100 SVWs, the improvement in precision decelerates. And we find that 20 SVWs are enough for retrieval, for SVWs are pertinent to the crucial content of the query image.

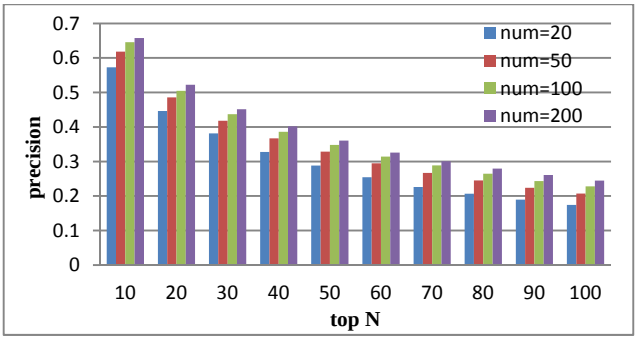

Figure 5. The comparison for different data volume.

7. CONCLUSION

In this paper, we propose a novel mobile image retrieval scheme based on mining salient visual words from multiple relevant photos. Our approach achieves better performance with less data. 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.

8. ACKNOWLEDGMENTS

Our thanks to ACM SIGGHI for allowing us to modify the templates. This work is supported by NSFC No. 60903121, No. 61332018, No. 61173109, Microsoft Research Asia.

9. REFERENCES

[1] J. Sivic, A. Zisserman. Video google: a text retrieval approach to object matching in videos. ICCV, 2003.

[2] D. G. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 60(2): 91-110, Nov. 2004

[3] S. Zhang, Q. Tian, G. Hua, Q. Huang and S. Li. Descriptive visual words and visual phrases for image. ACM MM, 2009.

[4] M. A. Fischler and R. C. Bolles. Random sample consensus. Comm. ACM, 24(6):381–395, 1981.

[5] W. Zhou, Y. Lu, H. Li, Y. Song and Q. Tian. Spatial coding for large scale partial-duplicate web image search. ACM MM, 2010.

[6] A. Qamra and E. Chang. Scalable landmark recognition using EXTENT. Multimedia tools and Applications, 2008.

[7] D. Nistér and H. Stewénius. Scalable Recognition with a Vocabulary Tree. CVPR, 2006.

[8] Y. Xue and X. Qian. Visual summarization of landmarks via viewpoint modeling. ICIP, 2012.

[9] S. Zhang, Q. Huang, G. Hua, S. Jiang, W. Gao and Q. Tian. Building contextual Visual Vocabulary for large-scale image application. ACM MM, Oct. 2010

[10] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: automatic query expansion with a generative feature model for object retrieval. ICCV, 2007.