Mobile Image Retrieval using Multi-Photos as Query

Xue, Xueming Qian†, Member IEEE, Baiqi Zhang
Xi'an Jiaotong University, Xi'an China, 710049

ABSTRACT

In this paper, we propose a novel image retrieval scheme, where multi relevant images are input as queries to improve the retrieval performance. We exploit sufficient information provided by multi query images to reduce distractor features, quantization loss and learn visual synonyms. During learning synonyms, consisting of visual synonyms detection and visual synonyms expansion, some identical and unique details semantically important to the query are captured. We represent images using a set of visual synonyms, each of which comprises several visual word paths, quantizing a descriptor from the root to a leaf of a hierarchical vocabulary tree. Spatial layout is also introduced for geometry constraint as an information source independent from descriptor space. Hierarchical visual word path and synonyms learning provide multiple choices for feature matching. Finally we evaluate our approach on two image datasets, where images from 5K Oxford building dataset are used as query; a 227K image dataset act as distractor.

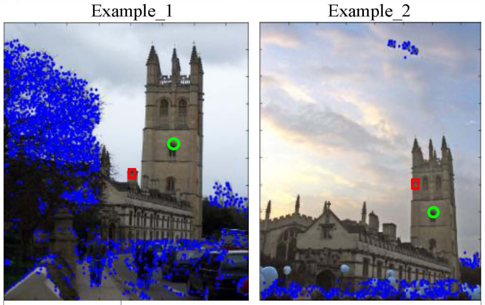
Index Terms—multi query, image retrieval, visual synonyms learning, hierarchical vocabulary tree

1. INTRODUCTION

With the prevalence of image-capture mobile phone and the development in communication techniques, there is an interesting tendency that people are more likely to use mobile phone to take photos or to surf the Internet. Thus a large number of the Internet services have transferred from the PC to the mobile, for example, tourists like to upload photos to the Internet and share with friends, people search on the Internet using their smart phone whenever they have something unknown or curious. Among the massive images available, how to find images satisfying users' interest becomes more and more necessary.

Query By Example (QBE) based image retrieval becomes a hot issue, where users provide an example query and search engines feedback an image sequence sharing common content with the query. The performance of QBE based image retrieval is largely influenced by its query example images. Distinct images sharing the same object can still result in large difference in query representation, due to quantization loss and distractor features introduced by query example image itself. Additionally, discrimination power of image representation seems insufficient, since it only focuses on the presence or absence of visual words, the synonym relation and spatial layout between visual words which contains much semantic for content understanding, is unfortunately ignored.

Fig.1 gives an illustration of different query images. Firstly, many distractor features, which are unreliable and represent irrelevant objects, are extracted in example_1 and example_2. There are another two challenges: quantization loss and visual synonyms. For quantization loss, different image patches are mapped to the same visual word. For visual synonyms, two visual words different in descriptor-space may correspond to the same object in real world.



Distractor features	● 2~3 K in example_1; 1~2K in example_2
Quantization loss	□ Different image pathes mapped to the same word #5062
Visual synonyms	○ Same object mapped to different words #3165 and #5427

Fig. 1. Query by different example images.

For query by singe image, the retrieval performance using the different images of the same object as query can be variable [17]. However, the rich information provided by multi query example images can be probably used to improve retrieval by filtering out distractor features, reducing quantization loss and learning the visual synonyms. Different with query expansion, which refines its query by learning information from returned sequence, multi images as query doesn't rely on a good initial feedback to improve retrieval. For mobile retrieval, where users take and upload photos by their phones, it is convenient to capture several images of an object, users can easily take several photos of a famous landmark, based on which the search engine will return similar images.

The rest of the paper is organized as follows. Firstly we reviews related works; secondly we provides the system overview; finally we gives a description on our approach in section 4 and 5; finally section 6 and 7 present our experimental setup and performance evaluation.

2. RELATED WORK

In recent years, content based image retrieval experiences a rapid development due to the BoW representation [6] and local features, like SIFT [1], and its variants PCA-SIFT [18]

†Corresponding author (qianxm@mail.xjtu.edu.cn). This work is supported by NSFC No.60903121, No.61173109 and Microsoft Research Asia.

and SURF [5]. Researchers have proposed many works e.g. hierarchical vocabulary trees [2,15], visual synonyms [3,4,13], soft quantization [7], query expansion [11,12], embed geometry constraint [8,10,14,16,17,19], etc.

In [9], Yang et al. propose a video based image retrieval system, since the retrieval performance is not reliable enough due to variations in singe query image. Chum et al. propose query expansion [11, 12], which refines the query based on the initial retrieval results. Different with their query process, we use multi images as query to improve retrieval performance.

[3,4,13] explore the application of synonyms relations between visual words in the image retrieval task. Gavves et al. focus on the incoherence of the visual words in bag of words vocabulary and extracted visual synonyms as pairs of independent visual words [3,4]. In our work, we define visual synonym as a set of hierarchical visual word paths, which correspond to the same real world object.

We perform hierarchical clustering to build a vocabulary tree and quantize each local descriptor to a visual word path from root to leaf of vocabulary tree [2, 15]. Thus descriptor information can be largely preserved level by level. Philbin et al. [7] propose soft quantization, which maps each descriptor to a set of words, to reduce the quantization loss.

Many works [8,10,14,16,17,19] introduce spatial layout of visual words for geometry constraint e.g. [8] performs spatial verification by RANSAC; [10,16] construct visual phrases to embed to spatial layout constraints in image retrieval; [14] encodes the spatial relationship among local features. In our approach, we employ geometry constraint in visual synonyms learning to verify the visual word path describing the same object as an information source independent from the similarities in descriptor space.

3. SYSTEM OVERVIEW

Our approach consists of mobile-server architecture. We extract SIFT feature and perform hierarchical clustering to construct a vocabulary tree. Then quantization maps each SIFT descriptor to a visual word path from the root to a leaf of the vocabulary tree.

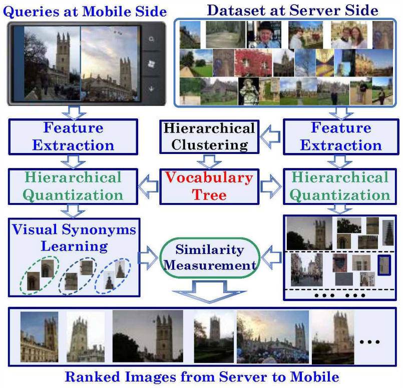


Fig. 2. The framework of our system.

At the mobile side, we perform visual synonyms learning, which finds out "visual synonyms": different visual word paths representing the same real world object (by visual synonyms detection) and estimate the potential visual synonyms (by visual synonyms expansion). At the server side, dataset image is represented with a set of visual word paths, each of which captures an image patch. Finally, the similarity between the two sets of visual word paths are measured as the matching score of query and dataset images. For easy reference, we summarize the notation in Table 1.

Table 1. Notation

Notation	Meaning
Q	query images
M	number of query images
F	branch factor of a hierarchical vocabulary tree
L	depth of a hierarchical vocabulary tree
D	a local descriptor
T	the vocabulary tree
N	number of expanded visual word paths
$O(D)$	orientation of descriptor D
$S(D)$	scale of descriptor D
$Path(D)$	visual word path of descriptor D
$Node^{l}_{Path(D)}$	l-th level node of visual word Path$(D)\{l=1,\dots L\}$
$Sim(\bullet,\bullet)$	Similarity of two elements

4. FEATURE EXTRACTION

To identify the content overlaps between images and capture some unique and representative details in images, we extract local features scale-invariant feature transform (SIFT) [1]. For query images and dataset images, we use the Difference of Gaussian (DoG) detector to find interest points <x, y, *scale*, *orientation*>, and then describe theses interest points using the 128-dimension SIFT descriptor.

We perform hierarchical K-means clustering on the SIFT descriptors to build a hierarchical vocabulary tree, with a branch factor F and depth L. After that, the vocabulary tree consists of about 10^l nodes (visual words) in the l-level. We use hierarchical vocabulary tree to quantize local descriptors, so that each local feature is assigned to an L dimensional vector, which corresponds to node path from the root to a leaf in the hierarchical vocabulary tree. Images are represented with a set of visual word paths.

Hierarchical quantization assigns each descriptor to a visual word path. The hierarchical quantization can capture the difference between descriptors level by level, to reduce the quantization loss. Similar with the soft quantization [7] which assigns a single descriptor to a set of weighted visual words, hierarchical quantization use a node path with L visual words to represent a local descriptor. Differently, the father-son relationship among visual words of a node path is also useful for preserving local descriptor information.

5. VISUAL SYNONYMS LEARNING

Given a query image, its visual words have different weights of importance for the query scene. Some of them may be semantically closer to query scene, others may be noise. We have used hierarchical quantization to reduce the quantization loss. To further filter out distractor features and mine synonyms relationship, we perform visual synonyms learning: visual synonyms detection and expansion

5.1. Visual synonyms detection

Visual synonyms detection is responsible for finding visual word paths representing the same real world object in query images. For a descriptor *D,* we have its quantized visual word path(*D*) and its scale *S*(*D*), its orientation *O*(*D*). To detect visual synonyms, we analyze the common path depth of paths from every query images. And also we take deviation of scale and orientation into account to construct spatial constraint in visual synonyms detection. Firstly we define similarity of two descriptors { D_p^i , D_q^j } as

$$Sim(D_p^i, D_q^j) = cod[path(D_p^i), path(D_q^j)] - ged[D_p^i, D_q^j] \quad (1)$$

where D_p^i is the *p*-th local descriptor in *i*-th image; D_q^j is the *q*-th local descriptor in *j*-th image. *cod* and *ged* are common depth and geometry difference of two visual word paths, which are expressed as follows:

$$cod[path(D_p^i), path(D_q^j)] = \sum_{l=1}^{L} 1\{Node^l_{Path(D_p^i)} = Node^l_{Path(D_q^j)}\} \quad (2)$$

$$\text{where} \quad 1\{*\} = \begin{cases} 1, & \text{if } * \text{ is true} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

$$ged[D_p^i, D_q^j] = O(D_p^i) - O(D_q^j) + \frac{[S(D_p^i) - S(D_q^j)]}{\max[S(D_p^i), S(D_q^j)]} \quad (4)$$

The common depth *cod* of visual word paths describes the similarity of two visual word paths in descriptor space. The bigger *cod* it is, the two visual word paths capture the same real world object in larger probability. Conversely, the geometry difference *ged* of two visual word paths indicate how different the two visual word paths are in spatial level.

Secondly from all the descriptors of *M* query images, we select one descriptor from each image. The selected *M* descriptors, whose total similarity is the largest, are defined as a visual synonym, *vs*={D^i|*i*=1,2,…*M*}. We provide each descriptor of a visual synonym with an importance weight based on its similarity with other *M*-1 descriptors as follows:

$$w(D^i) = w[path(D^i)] = \frac{1}{dev(vs)} \sum_{c=1, c \neq i}^{M} sim(D^i, D^c) \quad (5)$$

dev(*vs*) is the geometry deviation of the visual synonym *vs*={D^i|*i*=1,2,…*M*}. It equals to the deviation of geometry parameter scale and orientation of {D^i|*i*=1,2,…*M*}.

Thus *M* visual word paths are detected as a visual synonym, an importance weight is also available for each visual word path. We represent an image with a set of visual synonyms and image similarity will be measured based on the set of visual synonyms.

In Fig.3, we give the process of visual synonyms detection. Scale and orientation of visual synonyms are also shown by blue circles with different radius and red lines from the centers of the circles. A visual synonym can capture the identical real world object e.g. a window in different images. And the scale or orientation of each visual word path in a visual synonym is similar. Finally we illustrate the importance weight of each visual synonym, where the size of red dot is plotted in proportion to the value of importance weight.

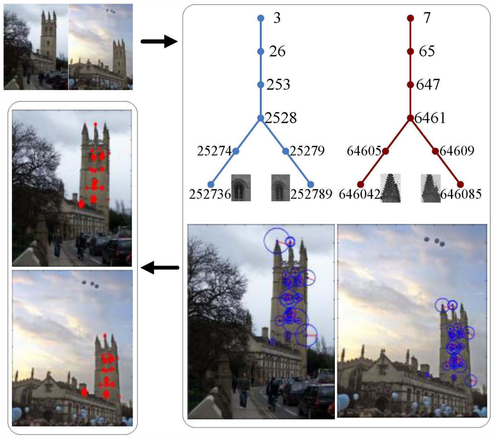


Fig. 3. The process of visual synonyms detection. Numbers in the top of middle part of this figure are BoW index of the SIFT points.

5.2. Visual synonyms expansion

Visual synonyms expansion is to estimate the probability weight of underlying visual synonyms in descriptor space. For a visual synonym containing *M* visual word paths, we expand it by finding another *N* nearest neighboring visual word paths in descriptor space. The *N* expanded visual word paths are assigned with a weight (*W*) based on their distance (*dis*) with the *M* detected visual word paths

$$w(path_E^i) = \frac{1}{M} \sum_{c=1}^{M} [w(path_D^c) - dis(path_E^i, path_D^c)] \quad (6)$$

where $path_E^i$ is the *i*-th expanded visual word path, $path_D^c$ is the *c*-th detected visual word paths. The weight of expanded visual word paths should be lower than that of detected visual word paths. Because the expanded visual word path are estimated in descriptor space; while the detected visual word paths are existing in current images.

6. EXPERIMENTAL SETUP

We conduct our experiments on two image datasets. The first is the 5K Oxford building image dataset (Oxbuild) [8], which comprises 5062 images of buildings in the University of Oxford. We crawl 227K images from Flickr covering 80 attractions with buildings, people, animals, etc. We combine the two image datasets as our retrieval dataset. Images from Oxbuild are query and crawled as distractor.

We perform hierarchical K-means clustering on SIFT descriptors to build a hierarchical vocabulary tree, with a branch factor *F*=10 and depth *L*=6. Our vocabulary tree consists of about 10^l nodes (visual words) in the *l*-th level (*l*=1,2,3,4); 99120 and 97563 visual words in the fifth and sixth level.

7. EXPERIMENTAL EVALUATION

To evaluate the performance of our approach, we carry out a comparison experiment between our Multi Query images retrieval (MQ) with Soft Quantization (SQ) [7], Query Expansion (QE) [11], and Tree Based image retrieval (TB) [15]. We implemented the three relevant approaches: SQ, QE, TB and conduct the comparison experiment on the combined image dataset. During the experiment, three relevant images uploaded by users are used as queries (i.e.

M=3), which is a good tradeoff between computation cost and retrieval precision. Our approach uses multi images as query; while other three methods use only one image as query. We conduct the experiment under the condition of query input. Three images as query for our approach; the same three images for each method (SQ, QE or TB) and the best one among the three retrieval results is taken for comparison with ours. Due to page size limited, no further discussion of M is made in this paper.

As the evaluation metric, we use the precision of the first i returned images which is given as follows

$$Precision(i)=C(i)/i \tag{7}$$

where $C(i)$ is the number of relevant images among the first i returned images. We select 100 groups of three images as query to carry out retrieval for 100 times respectively on SQ, QE, TB and MQ. And the take the average value of *Precision* as final performance of each method. We present the average value of *Precision* in Fig.6.

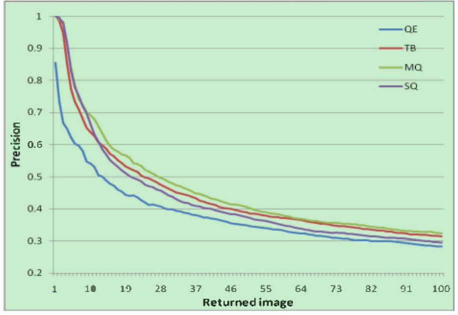


Fig. 6. Performance comparison on the 232K dataset.

From Fig.6, it can be observed that performance of SQ, TB and MQ is very close around 0.85~1 in the top ten images; after that MQ shows better performance. At the beginning of retrieval, SQ, TB and MQ is very close around 0.85~1 in the top ten images; while the accuracy of query expansion is lower. In top 11~64 images, MQ keep about 1.81%, 3.80% and 7.85% higher than TB, SQ and QE in average. In 65~100 images, performance of the four methods are closer. MQ keep about 0.83%, 2.89% and 4.28% higher than TB, SQ and QE in average. In general, the performance of MQ and TB, which are all based on hierarchical vocabulary tree for image representation, are most similar. Probably, due to the visual synonyms learning, MQ performs better especially in 15~45 images.

8. CONCLUSION

Different images sharing the same query object result in variations in retrieval performance, our approach use multi images as query to improve the retrieval performance. We combine the visual synonyms learning and hierarchical vocabulary tree together. Using hierarchical vocabulary tree, we recognize the difference of descriptors level by level to reduce quantization loss efficiently. In the visual synonyms learning, which consists of visual synonyms detection and visual synonyms expansion, we detect a number of visual word paths representing the same real world as a visual synonym, then estimate potential underlying visual word paths to expand the visual synonym. Thus queries are represented with a set of weighted visual synonyms, which largely filter out distractor features and provide multiple matching choices for image similarity measurement in a more efficiently and softly way.

9. REFERENCES

[1] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints”, IJCV, 2004.

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

[3] E. Gavves and Cees G.M. Snoek, “Landmark Image Retrieval Using Visual Synonyms”, ACM, MM, 2010.

[4] E. Gavves, C. Snoek, and A. Smeulders, “Visual synonyms for landmark image retrieval”, CVIU, 2011.

[5] H. Bay, T. Tuytelaars and L. V. Gool, “Surf: Speeded up robust features”, ECCV, 2006.

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

[7] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman “Lost in quantization: Improving particular object retrieval in large scale image datasets” CVPR, 2008.

[8] J. Philbin, O. Chum, M. Isard, J. Sivic and A. Zisserman “Object retrieval with large vocabularies and fast spatial matching”, CVPR, 2007.

[9] L. Yang, Y. Cai, A. Hanjalic, X. Hua and S. Li, "Video based Image Retrieval", ACM, MM, 2011.

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

[11] 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.

[12] O. Chum, A. Mikulík, M. Perdoch and J. Matas, "Total recall II: Query expansion revisited", CVPR, 2011.

[13] W. Tang, R. Cai, Z. Li, and L. Zhang, "Contextual synonym dictionary for visual object retrieval", ACM, MM, 2011.

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

[15] X. Wang, M.Yang, T. Cour, S. Zhu, K. Yu and T. X. Han, “Contextual Weighting for Vocabulary Tree based Image Retrieval”, ICCV, 2011.

[16] Y. Zhang, Z. Jia and T. Chen, “Image retrieval with Geometry-Preserving visual phrases”, CVPR, 2011.

[17] Y. Cao, C. Wang, Z. Li, L. Zhang and L. Zhang “Spatial Bag of Features”, CVPR, 2010.

[18] Y. Ke and R. Sukthankar, “Pea-sift: A more distinctive representation for local image descriptors”, CVPR,2004.

[19] Z. Wu, Q. Ke, Michael Isard, and J. Sun “Bundling Features for Large Scale Partial-Duplicate Web Image Search”, CVPR, 2009.