Image Retrieval Using Visual Words and Spatial Re-ranking

Project of Visual Data Retrieval and Detection

Ke Lin[*]
School of Software
Tsinghua University
Beijing, China
linke0113@gmail.com

Chen Chen[†]
School of Software
Tsinghua University
Beijing, China
chenchen_9207@163.com

Qingfu Wen[‡]
School of Software
Tsinghua University
Beijing, China
thssvince@163.com

ABSTRACT

The "bag of words"(BoW) model is very popular in large-scale text retrieval and image retrieval. SIFT descriptor is also wildly used in large scale image retrieval. In this paper, we implement an image retrieval algorithm based on SIFT and BoW. Traditionally, we can obtain an ordered list of top N images from the dataset which have the highest relevance to each query instance. However, the performance of naive BoW method is not ideal because it lacks spatial information. In this paper, we perform a spatial re-ranking on the ordered list to recalculate the similarity between each image and the query instance. In our experiments based on the **Oxford 5K** dataset, we can utilize spatial verification to achieve better retrieval accuracy at manageable computational cost.

General Terms

Algorithms

Keywords

Image Retrieval, SIFT, Bag of Words, Spatial Re-ranking

1. INTRODUCTION

Along with the fast increasing of image data, large-scale image retrieval has becoming more and more significant in both academia and industry. The problem of searching images according to their semantic content, which is also called content-based image retrieval (CBIR), is very challenging s-ince many factors can affect the performance of CBIR, like resolution, illumination variations and occluded objects.

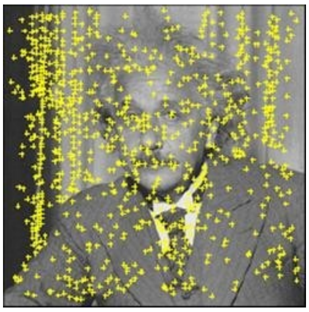
Albeit global features like colors, textures, and shapes have been used to represent images, they still have a plethora of limits. Scale Invariant Feature Transform (SIFT)[2] was proposed for describing several salient patches around key points within the images. Based on this local descriptor vectors, a popular approach, bag-of-words (BoW), was proposed in [4]. It is noted that the BoW approach, although is simple and directly borrowed from text retrieval community, has shown excellent performance not only for CBIR task but also for other vision tasks like object recognition, image classification and annotation.

2. SIFT & BOW

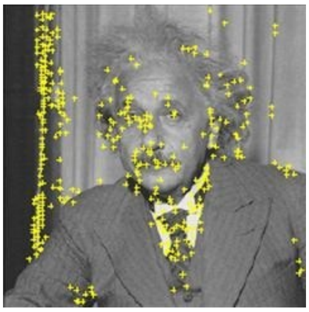
2.1 SIFT

SIFT features demonstrating great discriminative power are invariant to the changes of rotation, scaling, translation and distortions. SIFT descriptor can be extracted in the following 4 steps.

- Construct scale space, test interest points and get scale invariant feature.



- eliminate unstable points.

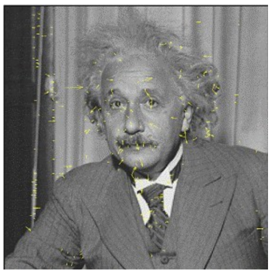


[*] This author do some experiments and revise paper

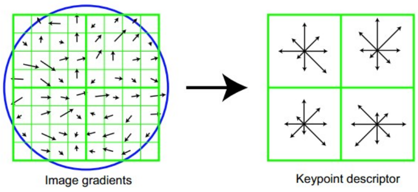
[†] This author do some experiments and revise paper

[‡] This author write code, do some experiments and write paper

- obtain orientation for each key point.



- generate a 128-D descriptor for each point.



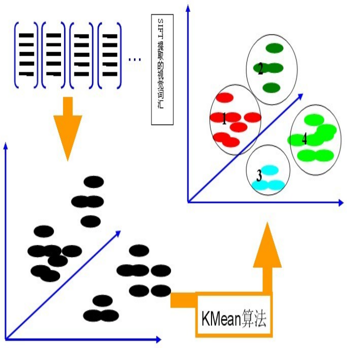
After these steps, we can gain a 128×n descriptor for each image.

2.2 BoW

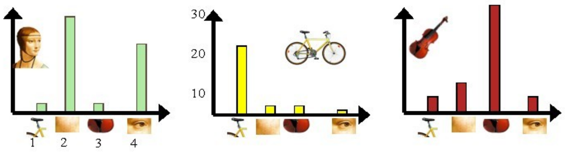
BoW model is a very common method in information retrieval. We will use an example to demonstrate BoW in image retrieval. For example, there are 3 categories of images–face, bike and guitar.

The first step of BoW is to extract SIFT features and extract visual words from all the images.

Second, generate word list using K-Means.



Finally, map each image to the word list. We can use a word list to represent an image.



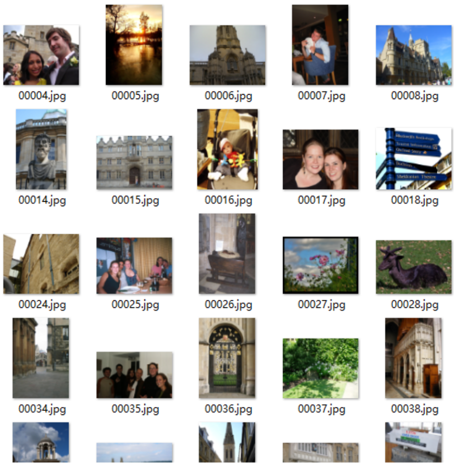


Figure 1: Randomly sampled images from the 5K dataset. Note that the dataset contains difficult distractors which may easily be confused with those used in the query set.

3. DATASET

The entire dataset consists of 5,062 high resolution (1024×768) images, comprising 11 different Oxford "landmarks" - a particular part of a building. Sample images from the dataset are shown in figure 1.

4. IMPLEMENTATIOIN

We implement this experiment using VLFeat[1] in MATLAB on Windows Server 2008 with 32G memory and 16-core CPU.

4.1 Extract SIFT features

For each image in the dataset, we can extract about 3000 128-D SIFT descriptors.

4.2 Generate visual vocabularies

In previous step, we may obtain about 15M SIFT descriptors. We perform a K-Means clustering on these descriptors with K = 100000. Then we can 100000-D visual vocabulary for all the image words.

4.3 Describe image width visual vocabulary

We try to assign each image to the 100000-D vocabulary using KD-Tree. Normally, we can use a 100000-dimension vector to describe each image.

4.4 Query

[1] http://www.vlfeat.org/

Category	top100	top200	top400	top800	top1000
all_souls	0.2594	0.2830	0.3025	0.3137	0.3161
ashmolean	0.1863	0.1888	0.1906	0.1925	0.1929
balliol	0.1911	0.1956	0.1962	0.1970	0.1974
bodleian	0.5620	0.5677	0.5713	0.5722	0.5728
chris_chur	0.4197	0.4449	0.4587	0.4629	0.4641
cornmarket	0.3179	0.3190	0.3206	0.3223	0.3227
mAP	**0.3227**	**0.3397**	**0.3400**	**0.3434**	**0.3443**

Table 1: Average precision without re-ranking

In this stage, we can naturally assign a query image to a 100000-D vector. Now, we compute the distance[2] between a query image and each image in dataset. we sort the list to get top-N images by distance.

4.5 Re-rank

At last, we try to re-rank this top-N list using Spatial information. We try to match the query instance and dataset images with SIFT again. Experimental results show higher accuracy after this step.

5. RESULTS

We run this experiment twice for comparison. First, we run it without re-ranking. You can see the mAP of each category in table 1. Second, we run it again to gain top 100 images with re-ranking. Results show in table 2.

6. CONCLUSIONS AND FURTHER WORK

We try to demonstrate a framework for image retrieval using SIFT and BoW with re-ranking. From the experimental results, we can see pretty good performance but not perfect. Unfortunately, because of the limitation of time, we did not pay much attention on the K-Means step. A more accurate K-Means algorithm may perform a better result in retrieval, like Approximate K-Means[3]. Meanwhile, the re-ranking stage has much work to research[1].

7. ACKNOWLEDGMENTS

We are grateful to Prof. Guiguang Ding who taught us a lot of useful visual data processing and analysis methods. In addition, thanks to our TA, Shijiang Chen, who provided us with guidance materials, we can finish our project successfully. Finally, thank you for reading this paper written in our poor English.

8. REFERENCES

[1] Y. Avrithis and G. Tolias. Hough pyramid matching: Speeded-up geometry re-ranking for large scale image retrieval. *International Journal of Computer Vision*, 107(1):1–19, March 2014.

[2] D. G. Lowe. Object recognition from local scale-invariant features. *ICCV*, 2:1150–1157, Sept. 1999.

[3] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

Category	no re-ranking	re-ranking
all_souls_1	0.0543	0.0883
all_souls_2	0.2514	0.3009
all_souls_3	0.2269	0.2919
all_souls_4	0.3974	0.4913
all_souls_5	0.3672	0.5158
ashmolean_1	0.0780	0.2029
ashmolean_2	0.3009	0.4768
ashmolean_3	0.0421	0.0883
ashmolean_4	0.1938	0.2328
ashmolean_5	0.3167	0.4202
balliol_1	0.3483	0.6784
balliol_2	0.1200	0.1855
balliol_3	0.1362	0.3333
balliol_4	0.1075	0.3333
balliol_5	0.2433	0.5470
bodleian_1	0.6158	0.6429
bodleian_2	0.0491	0.2083
bodleian_3	0.3488	0.5112
bodleian_4	0.8921	0.9479
bodleian_5	0.9044	0.8770
christ_church_1	0.5318	0.6517
christ_church_2	0.5919	0.6745
christ_church_3	0.5014	0.6044
christ_church_4	0.4287	0.5379
christ_church_5	0.0448	0.1026
cornmarket_1	0.4593	0.5556
cornmarket_2	0.1128	0.2222
cornmarket_3	0.4554	0.5556
cornmarket_4	0.2286	0.3333
cornmarket_5	0.3333	0.4444
mAP	**0.3227**	**0.4352**

Table 2: Average precision @top100

[2] the number of words both appears in image A and image B

[4] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. *Proceedings of the IEEE International Conference on Computer Vision*, 2:1470–1477, Oct. 2003.