Fast Retrieval and Matching of Objects

Lina Zhu

23 May 2018

In this article, we propose a large object retrieval system. The user provides the query object by selecting the region of the query image, and the system returns an ordered list of images containing the same object retrieved from the large corpus. We use Oxford landmarks as queries, and data from more than one million images captured from photo-sharing sites shows our system's scalability and performance. Due to the size of the data set, constructing image feature vocabularies is a major time and performance bottleneck. To solve this problem, we compared the different scalable methods used to construct the vocabulary, and introduced a new quantization method based on the random tree. We have in fact demonstrated a wide range of fundamentals that surpass current state-of-the-art technologies. Our experiments show that quantification has a significant impact on the quality of the search. To further improve query performance, we have added a valid space validation phase to re-order the results returned by our model and show that this consistently improves search quality, but when the visual vocabulary is

large, fewer The margin. We believe this work is a promising step towards a larger "web-scale" image corpora.

1 Retrieve objects from large corpora

Our motivation is to retrieve a subset of images containing query objects from a large number of image sets[1]. In practice, no algorithm can determine whether an image is located in a query subset. In fact, even human judgment may be due to occlusion, distortion, and other reasons. Do not agree. Therefore, we have solved the slightly different problem of ranking each image in the corpus to determine its likelihood of containing the query object and aiming to return to the user some of the prefixes in this ranking list, in descending order. A naive and inefficient solution to this task is to develop a ranking function and apply it to each image in the data set before returning to the ranking list. This is computationally expensive for large corpora, and the standard method in text retrieval is to use a bag of vocabulary models that are effectively implemented as a reverse file data structure. This serves as an initial "filter", greatly reducing the number of documents that need to be considered.

[1]from"The query object is specied by a user selecting part of a query image, so it is really a query region however we will refer to it as an object to avoid overloading the term region."

2 Datasets to evaluate and implement Oxford 5K datasets

In order to evaluate performance when comparing different visual vocabularies and spatial rankings, we collected a set of images of 11 different Oxford "landmarks" - here we refer to specific parts of the building - together with distractors. Use the "Oxford Christian Church" and "Oxford Radcliff Camera" queries to retrieve the image of each landmark. We also search for more disturbing images by searching "Oxford" separately. The entire data set consists of 5,062 high resolution (1024 x 768) images. The sample image in the data set is shown in Figure 1. For each landmark, we chose five different query areas, as shown in Figure 2. Five queries are used so that retrieval performance can be averaged based on any single query characteristic.We manually acquired ground truth by searching the entire data set of 11 landmarks. The image is assigned one of four possible labels:
(1) Good - Clear picture of object/building.
(2) OK - More than 25% of the objects are clearly visible.
(3) Rubbish - Less than 25% of the object is visible, or there is very high occlusion or distortion.
(4) Absence - The object does not exist. The number of different landmarks appears between 7 and 220 times.

Image data sets and ground truth tags can be found. [1] In addition to this set of tags, we also use two other data sets to stress test the retrieval performance. These include the images that are crawled in the most popular tag list. Then we assume that these datasets do not contain the objects that

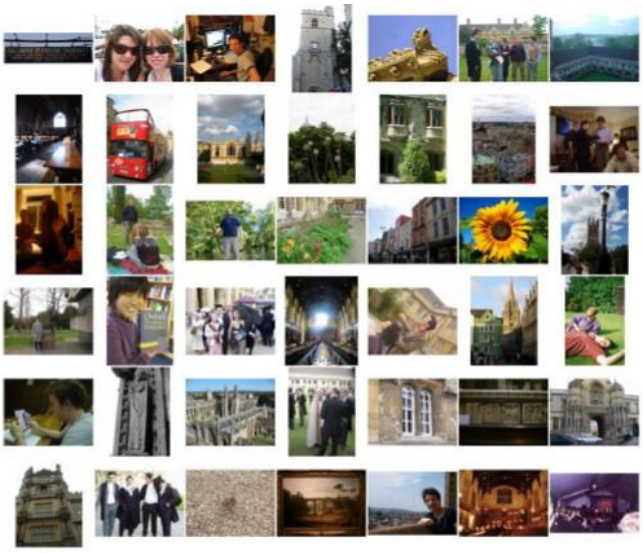

Figure 1: **42 randomly sampled images from the 5K dataset. Note that the dataset contains difcult distractors which may easily be confused with those used in the query set**

are being searched, so they act as interfering objects and test the performance and scalability of our system. 100K data set. This data is captured from Flickr's 145 most popular tags and contains 99,792 high resolution (1024 x 768) images. The data was captured from the 450 most popular tags and contained 1,040,801 medium resolution (500 x 333) images. Table 1 summarizes the relative sizes of the data sets.

3 Conclusions and further work

We show a scalable visual object retrieval system that uses photo corpora obtained from public websites. [2] The system returns the photos contained in the corpus of the query object. Despite significant differences

Figure 2: **All 55 query images used in the ground truth evaluation. Each row shows different queries for the same scene landmark. Note the large variation in scale of the query regions and the variation in position, lighting, etc. of the images themselves.**

between the lighting, perspective, image quality, and occlusion between the query and retrieval images, we need to continue exploring.

Table 1: **The number of images, features and descriptor sizes for each dataset.**

Dataset	images	features	Size of descriptors
5K	5,062	16,334,970	1.9 GB
100K	99,782	277,770,833	33.1 GB
1M	1,040,801	1,186,469,709	141.4 GB
Total	1,145,645	1,480,575,512	176.4 GB

References

[1] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, 2007.

[2] Luiz Andr Barroso, Jeffrey Dean, and Urs Lzle. *Web Search for a Planet: The Google Cluster Architecture*. IEEE Computer Society Press, 2003.