Large Scale Image Deduplication

Tzay-Yeu Wen
Stanford University
tywen@stanford.edu

Abstract

With the rise of internet and personal digital camera, it becomes easy for researchers to get image data in mass quantity. With these large amount of image data, it is impossible for humans to examine each image and insure the quality of the dataset. Therefore it is crucial to develop algorithms that can process large amount of data.

This paper will focus on a particular problem related to image dataset, image deduplication. We propose an efficient and scalable method to find near-duplicate images in an image collection. Our method includes 3-steps, first, extract compact features from each image. Second, use a fast clustering algorithm to reduce possible image match. The clustering algorithm should be CPU/memory efficient and can scale through multiple machines easily. We will use the method proposed by [3] that uses map-reduce to implement approximate nearest neighbor.

Finally, apply a more accurate method on the clustered images to find duplicate images in each group. We will use method proposal by [6].

1. Introduction

Near-duplicate image detection is a special kind of image retrieval problem, which is relatively easy and well studied compared to other computer vision problems. Several image features, for example [4] and [1], have been proposed to calculate the similarity of two image or image parts. Those features are robust to noise and many image transforms; Therefore are more than enough for duplicate image detection. Using feature aggregated from SIFT, [5] has proposed a method which will find similar images in the image dataset. Their experiments showed that their method maintains high accuracy for up to 1M of images.

The challenge however, is to be able to handle massive amount of images using reasonable computation resources. The amount of data an image retrieval algorithms, like K-mean or KNN, can processed are constrained by the amount of available memory. Therefore many methods that can reduce the memory footprint or scale to multiple machines had been proposed.

[2] proposed a method that reduce the feature representation of each image into less than 100 bytes. This increase the limit on a single machine but the result is less accurate. [6] proposed another features aggregate algorithm that can take advantage from both local and global features. Their algorithm uses visual words to represent an image and inverted index to search though the dataset.

Another approach proposed by [3] is to compute the approximate nearest neighbor on features using Map-Reduce, which can process large amount of data with less accuracy.

2. Experiment

We will evaluate our method using two different image dataset, one for accuracy and one for performance.

For accuracy measurement we will use a dataset produced by [5], which contains 6376 images with groups of four images that are photos of the same object. In addition to those images, for each image we will add 2 images that are produced by random translation, cropping and rotation. Following [6] we will use mAP as our evaluation metric.

For performance measurement we will use a dataset provided by ILSVRC2010, which contains about 1.2M images that are unlabled. The accuracy of the result will be manually verified. The performance will be evaluate by the CPU and memory usage.

References

[1] M. S. Extremal, J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from. In *In British Machine Vision Conference*, pages 384–393, 2002.

[2] H. Jegou, M. Douze, C. Schmid, and P. Prez. Aggregating local descriptors into a compact image representation. In *CVPR'10*, pages 3304–3311, 2010.

[3] T. Liu, C. Rosenberg, and H. Rowley. Clustering billions of images with large scale nearest neighbor search. In *Applications of Computer Vision, 2007. WACV '07. IEEE Workshop on*, page 28, feb. 2007.

[4] D. Lowe. Object recognition from local scale-invariant features. pages 1150–1157, 1999.

[5] D. Nistr and H. Stewnius. Scalable recognition with a vocabulary tree. In *IN CVPR*, pages 2161–2168, 2006.

[6] Z. Wu, Q. Ke, M. Isard, and J. Sun. Bundling features for large scale partial-duplicate web image search. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 25 –32, june 2009.