Final project proposal by Hela Di

What you are going to implement:

I'm going to implement an image retrieval system that merges a quantization algorithm and a bloom filter. The quantization algorithm is based on a multiple assignment k-means hashing schema. Bloom filter is used as gatekeepers to avoid performing a query if the query features are not stored in the database and speeding up the query process, without affecting retrieval performance. This project is based on the paper "Bloom Filters and Compact Hash Codes for Efficient and Distributed Image Retrieval".

The paper proposed the compact hashing for CNN features. However, I couldn't find the CNN features dataset for now. I looked at the paper it referred to ("Compact hash codes and data structures for efficient mobile visual search") and found they were using the compact hashing code for SIFT features, and I can download the corresponding SIFT descriptors. So, I decide that if I could find the CNN feature image dataset, I would use it, if not, I would use the SIFT to finish the project. This dataset is Oxford 5K images dataset and compressed binary file of SIFT descriptors for the 5K dataset.

How it relates to your work:

I'm very interested in Information Retrieval. Last year I took the information retrieval class. The project of the IR class is text retrieval. As image retrieval is also an important part of information retrieval, I want to take the opportunity to implement the methods discussed in this class on image retrieval filed.

How will you know that you succeeded:

I may not have enough time to do all the comparison (compare MINx with MEAN, UTH and some other methods) by the due time, so I would think the project as succeed if the system I implement could get the right results (return a satisfying ranking of results or return not in the dataset) in satisfying time. As I may not be as experienced as the writers of this paper, I would think 1.5 X time compared to the writers would be OK.

How will you know that the technique is good:

Quantization algorithm: By the time all the features are implemented and the optimizations are done, I could run same query on systems with different hashing methods and compare the precision and speed. If both fields are better, the technique is good. If not, I need to study the tradeoff and determine whether it's a good technique.

Bloom filter: I will try to retrieve the images using the system with (System A) or without (System B) the bloom filter and compare speed difference. I plan to use three different sets of queries. One set is that the query images that are in the set are far more than those are not in the set (Set 1). One set is that the images that are not in the set is far more than those are in the set (Set 2). One set is the almost half to half (Set 3). I will run the 3 sets in both systems and compare the result, if time ((System A – Set 1) + (System A – Set 2) + (System A – Set 3)) is significantly less than time ((System B – Set 1) + (System B – Set 2) + (System B – Set 3)), the bloom filter is proved to be good.