Assignment 2

Shourabh Payal (MT2020054)

1. ***Panorama stitching***

   (a) **Explain how SURF is different from SIFT**

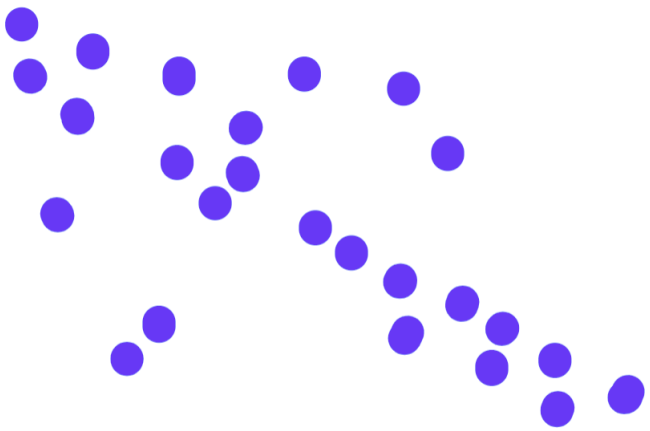
   In SURF we approximate the LoG using box filter. The advantages of using this is that convolution with box filter can be easily calculated with the help of integral images. And it can be done in parallel for different scales. Also the SURF rely on determinant of Hessian matrix for both scale and location.

   SURF feature descriptor usually has total 64 dimensions. Lower the dimension, higher the speed of computation and matching.

   Another improvement is the use of sign of Laplacian for underlying interest point. It adds no computation cost since it is already computed during detection.

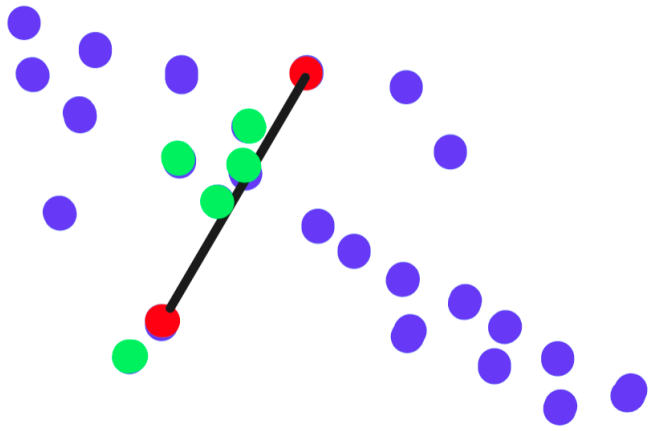
   (b) **Briefly explain the main principles of FLANN matching and RANSAC**

   Ransac stands for random sample concensus. It is a simple and effective algorithm to deal with outliers in our dataset. For example when we work with sensor data, the sensor is in generally effected by noise which causes outliers. The algorithm tries to group our dataset into an inlier and outlier set, so that we can forget about outliers and work with inliers.

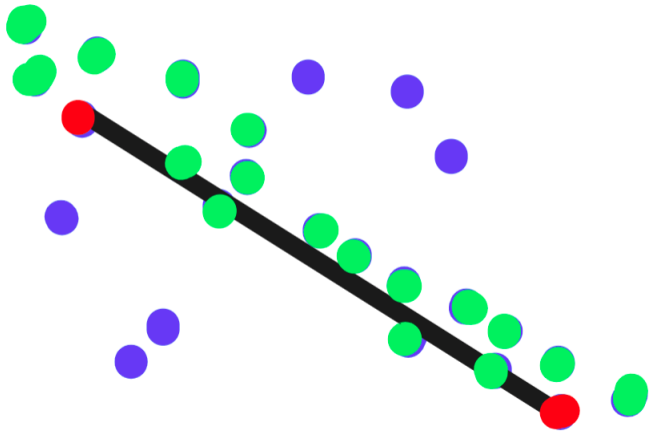


Let's consider the data points given in the above image. We want to fit a line through this data set. Some data points lie on the line and some of the points are outliers.

Let's say we randomly pick 2 points and draw a line through those points. As shown in the below image we picked the red points to fit the line. We observe that some of the points which lie closer to the line can be marked as inliers (marked in green). So we can assign a score of 5 to this fit.



We again sample some data points and draw a line through them to calculate the score.



We again calculate how many data points agree with this line and calculate the score. We repeat this process of sampling, fitting and scoring multiple times. We then select the model which has the highest score. As a result we get our inlier and outlier set.

To put it simply RANSAC is a 3 step process.

i. Sample a small subset of data points. Treat them as inliers.
ii. Compute the model.
iii. Score the model.

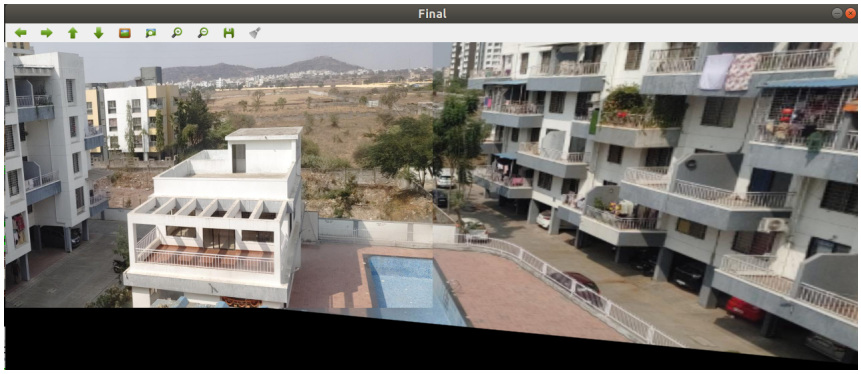
How often should we repeat the above process?
If we want to succeed with a probability p, outlier ratio in our data points is e and we sample s points in the first step then the below formula gives us the numer of trials T we need to perform.

$$T = \frac{log(1-p)}{log(1-(1-e)^s)}$$

FLANN stands for Fast Library for Approximate Nearest Neighbors. It contains a collection of algorithms optimized for fast nearest neighbor search in large datasets and for high dimensional features. It works more faster than BFMatcher for large datasets.

FLANN builds an efficient data structure (KD-Tree) that will be used to search for an approximate neighbour. It only finds an approximate nearest neighbor, which is a good matching but not necessarily the best.

Below are samples from image stitching exercise.



2. ***Bag of Visual words v/s VLAD***

(a) **Bag of Visual Words**

Bag of visual words (BOVW) is commonly used in image classification. We count the number of each word appears in a document, use the frequency of each word to know the keywords of the document, and make a frequency histogram from it. We treat a document as a bag of words (BOW). In short we use image features as the words. Image features are unique pattern that we can find in an image.

Approach

i. For each image in our data set we perform appropriate pre processing on it such as converting to gray scale, resizing (to detect interest points easily in later stages).

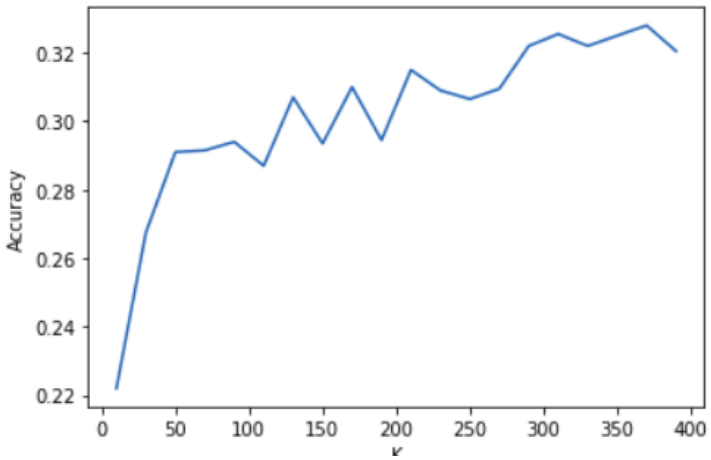
ii. For each processed image we will detect a set of interest points using SIFT.

iii. We collect all the interest points from all the images and perform k-means on them. These k clusters act as our bag of words.

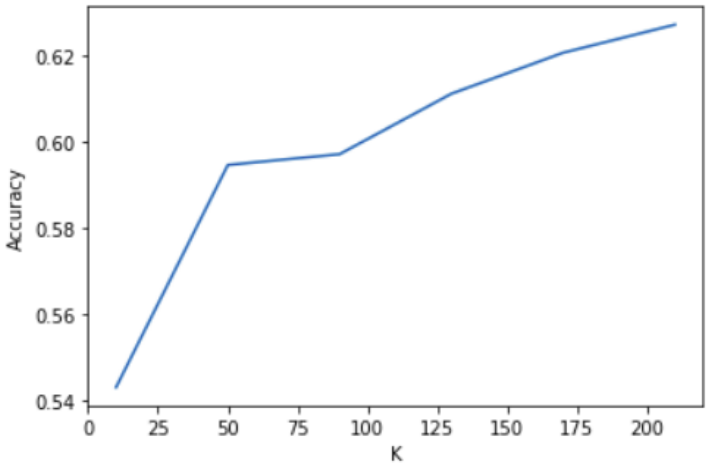
iv. For each image we will get the cluster assignments corresponding to each interest point descriptors obtained earlier. We make a bag of words representation using these.

v. We train standard ML algorithms like Logistic regression and predict image labels against it.

The image below shows the plot of choice of K with accuracy obtained. The data set used was 10K random samples of cifar-10.



Below image represents accuracy among top 3 results against k clusters.



(b) **VLAD**

Vector of Locally Aggregated Descriptor also known as VLAD is an extension of Bag of Visual Words. In BoVW we store a scalar value for the occurrence of descriptors. In VLAD we actually store the aggregated sum of residual vectors which maps to the cluster centre.

Approach

i. We perform first three steps same as Bag of Words.

ii. For each image and each descriptor present in the image we calculate a residual vector which equals to the difference between the nearest cluster centre and the descriptor.

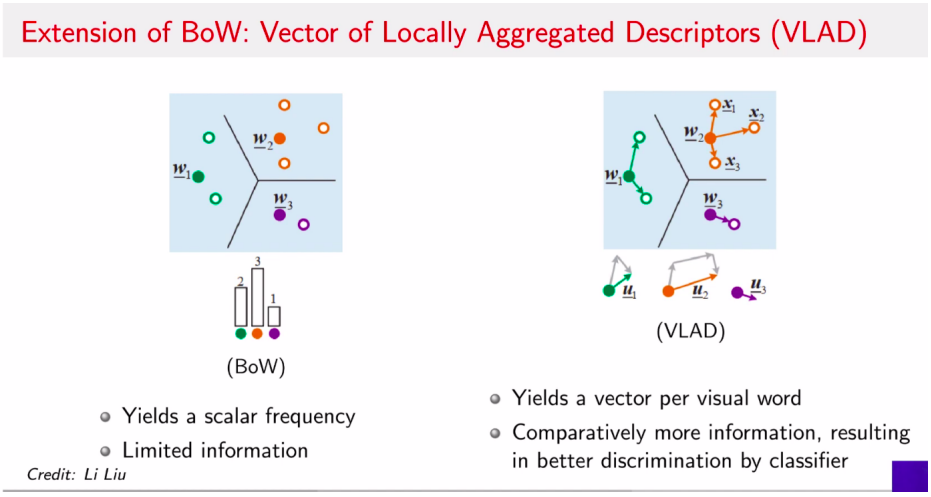
iii. This residual vector is aggregated over all words in the bag for every image.

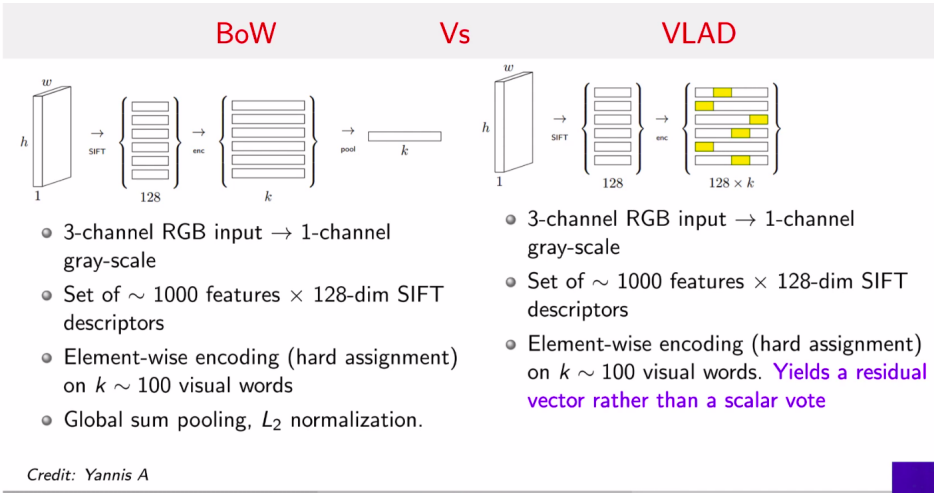
iv. We now have a 64d (SURF usage) residual aggregated vector for each descriptor

v. For each image we have a kx64 (SURF usage) vector. In bag of words we only had a k dimension vector corresponding to each image.

vi. For training we need to flatten/reshape this kx64 vector into 64kx1 vector.

vii. We now proceed with conventional ML models for training and prediction.





From points to images NPTEL

Results

Using of VLAD resulted in an accuracy boost of 0.1 for large set of images.
The dimension of the training vector increased to a large extent due to residual aggregation vector. This resulted in longer running time for training algorithm. As a result working on large set of images became difficult. Plus this meant that it took more time to select the right k using the line plot.

Below image shows the plot for accuracy vs choice of k on a small data set = 1000 images.