Visual Recognition

Assignment 2

Ishaan Sachdeva

IMT2018508

Task1

In this task we have to click some pictures of a landmark from different viewpoints and stitch them together to create panoramic photograph. This is done using open cv functions like sift_create(), findHomography() , warpPerspective(). The steps followed to stitch the images are:

1) I have used the sift operator to extract the key points and descriptors.
2) I have used BFMatcher() function to find best matches for each descriptor in the other image.
3) Have filter out the outliers from the above obtained matches.
4) Next step is to align the images which is done using findHomography function.
5) The last step is to stitch the the images which is done using warpPerspective()

LEFT IMAGE

RIGHT IMAGE

Final image

SIFT VS SURF

SIFT(Scale Invariant Feature Transform) and SURF(Speed Up Robust Feature) are feature detector algorithms and are very efficient in object recognition applications. Both of them use Difference of Gaussian but SURF performs faster than SIFT without reducing the quality of detected points.

RANSAC VS FLANN MATCHING

Ransac algorithm is used along with findHomography function to reduce the possible chances of error/outliers while matching features in two images.

Flann matching is another interface to perform quick and efficient matching. It contains a collection of algorithms for fast nearest neighbours search in a large datasets.

Task2

In this task we have to implement the BoW(Bag Of Words) and VLAD(Vector of Locally Aggregated Descriptors) approach to build a classification model on the **CIFAR10** dataset.

BoW

This technique is similar to bag of words in NLP. In the NLP's bag of words we count the number of times each word appears in a document and use the frequency of each word to identify the importance. In case of computer vision it is BOVW(Bag Of Visual Words).

In BOVW, image is represented as a set of features. This features consist of keypoints and descriptors. With the help of keypoints and descriptors we construct our vocabulary to represent each image as a frequency histogram of features. From the frequency histogram, we can identify set of features which belong to a specific class and these features can be used to classify different objects belonging to the class. For eg while identifying a face/person in an image: eyes, nose and ears are set of features/words which will help us identify if there is a person in the image. From the frequency histogram we can train our model and predict images.

Steps for BOVW in detail.

1) Using the sift operator the key points and descriptors are identified. Key points are points in the image which are not affected if the image is rotated, shrinked or expanded. Descriptors give the description about the key points. Usually they are of 128 dimension in size.
2) Using the sift operation on every image we have created a visual dictionary. By using KMeans clustering algorithm we can find the visual words which are centre points.
3) Frequency histogram of train and test images are created using the centre points identified in the previous part.
4) The last step is training the model. I have used KNN algorithm.

The accuracy of the model is 16.6%.

The model was trained with 50000 images and was tested on roughly 10000 images out of which it was able identify nearly 1640 images. The model accuracy can be increased to 30% if the more keypoints are identified.

VLAD

This approach is similar to BOW approach. Keypoints/regions are identified in an image using SIFT operator. Then using KMeans algorithm the cluster centres

are identified and each descriptor is assigned to closest cluster. In case of **BOW** we find the number of descriptors assigned to a specific cluster but in case of **VLAD** we record the difference of the descriptor from the cluster centre. These vectors are L2 normalised which are then used to train the model.

The model was trained with 15000 images and was tested on 3000 images. The accuracy of the model was roughly 11 %. The accuracy of the model was low because it was trained with very less samples but its accuracy will be more than BOW nearly 35% when trained using the entire dataset of 50000 images.