Visual Recognition

Kotha Tejas (IMT2016112)

Q1.1. Play with Image Stitching

Ans. Please refer to Panorama folder in the submission

Steps:

- Find all the descriptors for the two given images using xfeatures2d of SIFT.
- To find all the possible stitches use BruteForce matching
- Then based on a threshold, we discard the unnecessary matches.
- Compute the homography of the images.
- Finally, use wrap_perspective to merge the images.

Q1.2. Explain how SURF is different from SIFT (10 sentences)

Ans. SURF uses square-shaped filters as an approximation of Gaussian smoothing. (The SIFT approach uses cascaded filters to detect scale-invariant characteristic points, where the difference of Gaussians (DoG) is calculated on rescaled images progressively.)

Scale: In SIFT Difference of Gaussian (DoG) is convolved with different size of images with same size of filter. While in SURF, Different size of box filter(Laplacian of Gaussian (LoG)) is convoluted with integral image.

Key Point Detection: SIFT uses local extrema detection, applies Non maxima suppression and eliminates edge response with Hessian matrix. Meanwhile, SURF determines the key points with Hessian matrix and Non Maxima suppression.

Orientation: In SIFT, image gradient magnitude and orientations are sampled around the key point location, using the scale of the key point to select the level of Gaussian blur for the image. Orientation of histogram is used for this while in SURF, a sliding Orientation window of size $\pi/3$ detects the dominant orientation of the Gaussian weighted Haar Wavelet responses at every sample point with in a circular neighbourhood around the interest points.

Descriptor: In SIFT, the key point descriptor allows for significant shift in gradient positions by creating orientation histograms over 4 x 4 sample regions. In SURF an orientation quadratic grid with 4x4 square sub regions is laid over the interest point. For each square, the wavelet responses are computed from 5x5 samples.
Descriptor of SURF is $V= (\sum d x , \sum d y , \sum |d x| , \sum |d x|)$

Descriptor Size: Descriptor Size in SIFT is 128 bits while in SURF it's just 64 bits.

Reference: https://ijarcce.com/wp-content/uploads/2016/04/nCORETech-7.pdf

Q1.3. Briefly explain the main principles of RANSAC/FLANN matching (5 sentences)

Ans. **Random sample consensus** (**RANSAC**) is an iterative method to estimate parameters of a mathematical model from a set of observed data. It is a non-deterministic algorithm in the sense that it produces a reasonable result only with a certain probability, with this probability increasing as more iterations are allowed.

RANSAC uses the voting scheme to find the optimal fitting result. Data elements in the dataset are used to vote for one or multiple models. The implementation of this voting scheme is based on two assumptions: that the noisy features will not vote consistently for any single model and there are enough features to agree on a good model. The algorithm is essentially composed of two steps that are iteratively repeated:

1. In the first step, a sample subset containing minimal data items is randomly selected from the input dataset. A fitting model and the corresponding model parameters are computed using only the elements of this sample subset.
2. In the second step, the algorithm checks which elements of the entire dataset are consistent with the model instantiated by the estimated model parameters obtained from the first step. A data element will be considered as an outlier if it does not fit the fitting model instantiated by the set of estimated model parameters within some error threshold that defines the maximum deviation attributable to the effect of noise.

FLANN: Keypoint matching is the process of finding correspondences between two images of the same scene or object. FLANN performs fast nearest neighbour search in high dimensional features. It uses randomized kd tree algorithm and does a priority search using k-means tree algorithm. The search can quickly eliminate the part of the search space by using the tree properties. It splits the data into M multiple regions and recursively partitioning each zone until each of the leaf node has no more than M items. Then picks up the initial centers in random manner.

Reference:
https://books.google.co.in/books?id=zguCDwAAQBAJ&pg=PA600&dq=flann+and+ransac&hl=en&sa=X&ved=0ahUKEwiluZCA0ODgAhUMEbwKHV3RBpYQ6AEIKjAA#v=onepage&q=flann%20and%20ransac&f=false

Q2. Implement Bike – Horse classification using bag of visual words (SIFT/SURF)

Steps

- Extract SIFT/SURF features and append them into an array. We get an array of size (m, 128) for each image where ‘m’ is the number of descriptors of the image and 128 is the descriptor vector. Make another array and append all descriptors into it.
- Convert the array into a dataframe and apply Minibatch KMeans with cluster size 20. This will group all the descriptors from all the images into 20 clusters/bags. This is necessary because SIFT descriptors are very unique to the image and can’t be generalized for a particular class, hence can’t be used as features directly. This makes clusters of similar descriptors.
- We iterate through each image and predict which descriptor of the image lies in which cluster.We get an array of cluster numbers corresponding to each descriptor. By taking a histogram of these clusters, we get the number of descriptors of an image fall into a specific cluster.
- These histograms are our feature vectors for training.
- Then we use SVM or Logistic Regression or KNN to classify the images.

Cross validation scores (roc_auc)
SVM => 0.90875
Logistic Regression => 0.92014
KNN => 0.911875

Observation
The scores are really promising, considering very less feature engineering and the use of traditional image processing and machine learning tools.

Q2b. Implement CIFAR10 classifier with Bag of visual words (SIFT/SURF)
Steps

- The CIFAR10 data is pickled, so we unpickle it.
- The data is divided into 5 parts, each of size (10000*3072). We parse through this to get images of dimensions (32*32*3) .
- Follow the same procedure as in horse-bike classification. There are around 125 images in the CIFAR10 dataset which have no descriptors. We ignore those images for training purposes.
- I used train_test_split instead of cross validation to reduce the amount of computation needed for the results so as the code could execute smoothly.

Test results with train_tes_split(0.9:0.1) (accuracy)
SVM => 0.264676
Logistic Regression => 0.27068
KNN => 0.15808

Observation

The results are bad primarily because the images are very small. These images have very minute amount of details, which our descriptors fail to realize.