Statistical Machine Learning

Bonus Assignment Report

By - Kaustav Vats (2016048)

Data Collection

Training Data contains 10K images, Test data contains 1K images. Data contains 20 classes.

Data Visualization and Preprocessing

In first image we can see that see that frog boundary is very pixelated and lot of pixel intensity change is visible from

 Original Image	 Gray scale image with gaussian blur

I also tried histogram equalization, which basically stretch the histogram of the image.

Pre-Processing

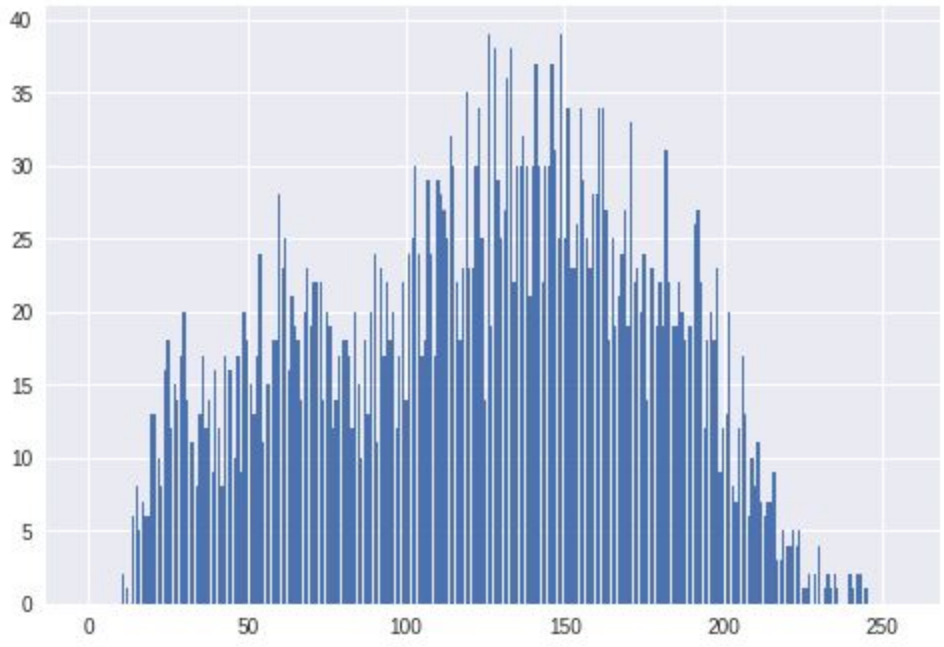
- Gaussian Blur to remove noise

- Contrast Enhancement
- Image Normalization (MINMAX)
- Z-score Normalization

I tried combination of above methods. I observed that Gaussian Blur reduces accuracy, by removing high frequency from the images.

Feature Extraction

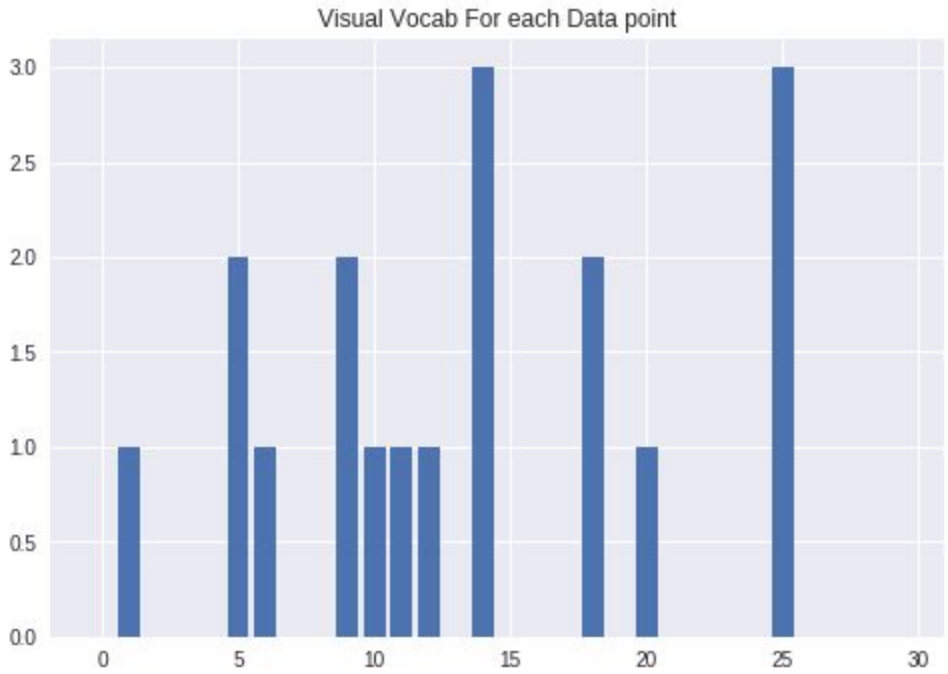
Below is the histogram of the image i used for classification



I tried various methods like bag of visual words which uses SIFT Descriptor .

Sift identifies the keypoints from the image and a corresponding descriptor of size (1x128). Then vertically stacked all file descriptors of the images. Then passed this feature set in k means, which clusters similar file descriptor in same cluster node. Based on these cluster, we predict cluster number of each file descriptor of the image and created

a histogram features vector of size number equal to number of



clusters.

Then i used Random forest classifier with around 1K estimators. I also tried Support vector machine, but was not able to perform best for given data.

Methodology

- Histogram Equalization (Pre-Processing)- This basically stretch the histogram of the image.
- Then i extracted features from the image. I used Histogram of gradients.
- Converted image to hsv space. Then created histogram for each channel.
- Flatten the hsv image and use as a features vector.
- Finally i vertically stacked all features in a single feature vector.

- Which was passed in the Random forest classifier with around 1K estimators. I also tried with different number of estimators like 1096(“optimal”), 1200. I observed that sometimes it works but sometimes it doesn’t and overfit the model.

Observation

I observed that classifier was sometimes overfitting with large number of estimators. And was performing bad with less number of estimators.
I observed that there was some noise in some of the images on the dataset.
For which i tried gaussian filter. But after gaussian filter number of features reduced and some high frequency details was also not available. Tried LBP features, but it performed worse than the simple histogram of the image.
I tried to reduce number of features by doing PCA, with eigen energy 99. But it was reducing the overall accuracy so i removed it.

Results

Approach	PCA	Classifier	Accuracy
Histogram of the image	-	Random Forest	36.66%
Hog image	-	Random forest and Logistic regression	39.00%
Extracted Hog features, hsv image, histogram of hsv image	-	Random forest	42.00%

I observed that accuracy was better with Hog features, Hsv space and histogram of the hsv image.