

Image Classification

Detailed Description:

Develop predictive models that can determine, given an image, which one of 11 classes it is.

Traffic congestion seems to be at an all-time high. Machine Learning methods must be developed to help solve traffic problems. We will analyze features extracted from traffic images depicting different objects to determine their type as one of 11 classes, noted by integers 1-11: car, suv, small_truck, medium_truck, large_truck, pedestrian, bus, van, people, bicycle, and motorcycle. The object classes are heavily imbalanced. For example, the training data contains 10375 cars but only 3 bicycles and 0 people. Classes in the test data are similarly distributed.

The input for the analysis will not be the images themselves, but rather features extracted from the images. An image can be described by many different types of features. In the training and test datasets, images are described as 887-dimensional vectors, composed by concatenating the following features:

- 512 Histogram of Oriented Gradients (HOG) features
- 256 Normalized Color Histogram (Hist) features
- 64 Local Binary Pattern (LBP) features
- 48 Color gradient (RGB) features
- 7 Depth of Field (DF) features

Since the dataset is imbalanced the scoring function will be the F1-score instead of Accuracy.

Data Description:

The training dataset consists of 21186 records and the test dataset consists of 5296 records. The attributes are floating point values and are presented in a dense matrix format within train.dat and test.dat. The numpy loadtxt function can be used to read the data in Python. The data are included in the **data.zip** file. While the file is fairly small (17MB), note that it expands to over 550 MB. Ensure you have enough space on your drive before expanding the file. Moreover, you will need to think carefully about how you will organize computations so you do not run out of RAM during training.

- **train.dat**: Training set (dense matrix, samples/images in lines, features in columns).
- **train.labels**: Training class labels (integers, one per line).
- **test.dat**: Test set (dense matrix, samples/images in lines, features in columns).
- **format.dat**: A sample submission with 5296 entries randomly chosen to be 1-11.