CMSC 491/691
Project Report

Pedestrian Detection on Static Images

By,
Arjun Veeramachaneni
CH95135

Introduction:

The goal of this project is to detect and localize humans in static images. This is a challenging problem as people can have a wide variety of poses, distinct appearance and clothing. The background of the scene can be complex as well, and present different kind of illuminations. Human detection in scenes can be useful for applications such as pedestrian detection for smart cars, smart home, visual surveillance. To address the detection problem, we first review different descriptor methods to define a good abstraction of the human model. Such feature descriptors can be scale invariant feature key points descriptors (SIFT), shape context descriptors, Haar based descriptors, histogram of gradient descriptors (HOG). We investigate in more details on how to compute the features and apply to this detection problem. A precomputed approximate model was taken, where a set of positive and negative training images were trained using a intersection kernel SVM classifier. We verify the model by running it over a set of normalized positive testing images, and verify the detection rate with the help of the score in the bounding box. We also apply the model exhaustively on a set of negative images to make sure there are no detections of a human. We also perform a step of non max suppression to remove the noise and get a more accurate detection.

Previous Work:

Different kinds of feature descriptors have been developed for detection purpose. Lowe’s Scale Invariant Feature Transformation (SIFT) approach allows to extract distinctive invariant features from the object we’re trying to detect. A database of key-points features can be generated using a set of training images. Belongie’s Shape Context approach consists of sampling the edges of the object into points of interest, and capturing the distribution of the sampled points on the shape with respect to a given point on the shape. The relationship between a point of the shape and the other points on that shape can be described by a distance and angle measurements. Distances and angles can be binned into different buckets, to generate a histogram. This histogram will capture the relationship between a points p and the other points on the shape. Matching two shapes is equivalent to finding points on each shape that have similar shape context. Another approach proposed by Dalal and Triggs [1] is to use histograms of oriented gradients (HOG) as a template descriptor for human detection. Local object appearance and shape information can be efficiently expressed by the distribution of local intensity gradients (edges) over the object and immediate surroundings. The approach used in this project is different and will be explained in the implementation.

Implementation:

1) Classification Model:

The first stage of the method is to generate a classification model. For this, we use an intersection kernel SVM classifier [2], and train it with a set of well normalized positive training images, as well as a set of negative images. In my case, this is already done where the parameters are computed and an approximate model is generated.
The normalized positive images will present examples of people, all at the same scale, with similar amount of background around the human bodies in these images. For negative training set, we use images that do not contain people, and randomly select areas of these images to generate negative features.

2) Detection:

Initially I started off by creating folders consisting of single and multiple pedestrian’s images in order to check the amount of time it takes to detect single or multiple humans.

The next step is initializing the non max suppression parameters, then we run detector with the stride of 8X8 and varied block sizes. Consider a detection window size 64X128. Now compute the features over the scalespace and create a sampling grid which basically computes the locations of the grid points. We compute the gradients of the image and then normalize it using block size of 16X16. Concatenate the features of the image and then using these features we apply it to our classifier. We kept a threshold of 0 and considered all the indexes with value greater than 0. We then carry out the non max suppression in order to minimize cluster effect and to get a single bounding box around the image. Finally draw a rectangle around the image where the human is present and put the score that is calculated above each image. The score is calculated in the non max suppression [3] function. Below are the figures for single [figure 1] and multiple [figure 2] detections of humans:

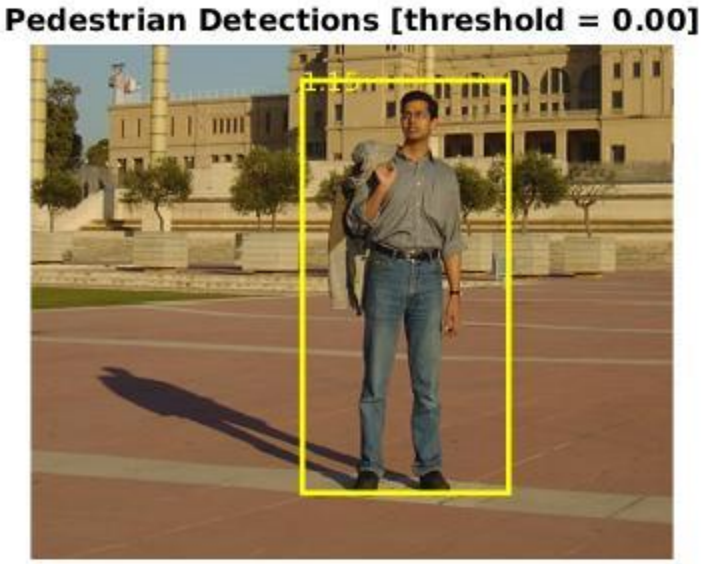

Figure 1: Single Pedestrian Detection

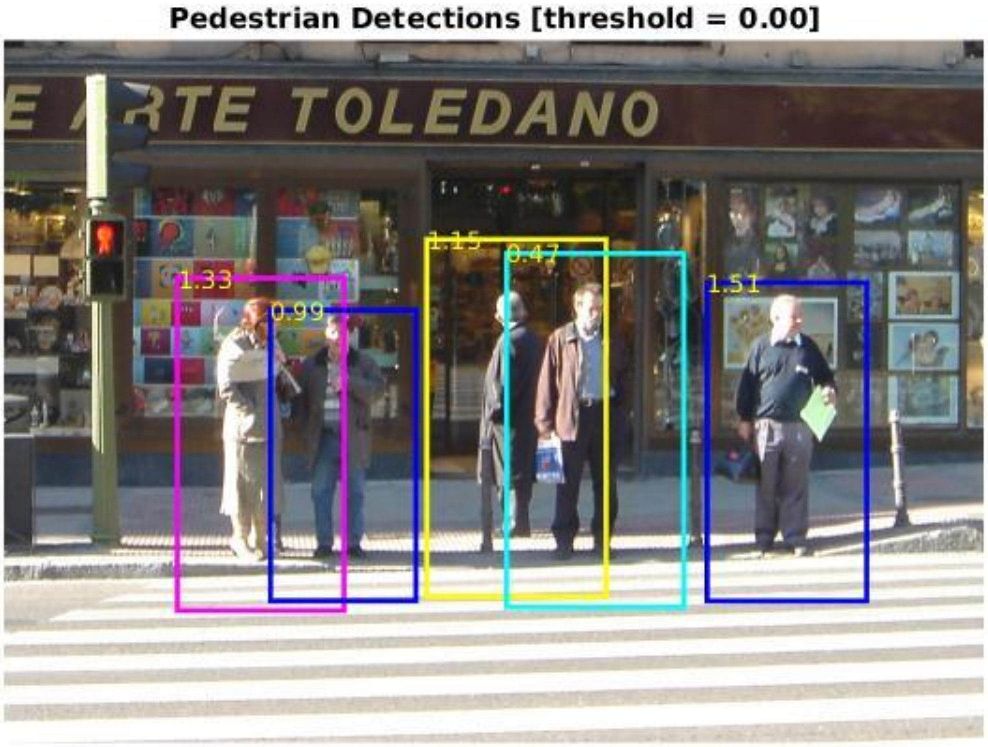

Figure 2: Multiple pedestrian Detection

3) Dataset (INRIA)

The dataset used for this project is INRIA person dataset where the images that should be used for training and testing is provided separately. We could observe that the detector failed to bound the boxes for a few images. A graph [figure 3] is plotted showing the amount of time it took to extract the features for a set of 20 images in single and multiple pedestrian's folder which was taken from the INRIA dataset. The graph is shown below:

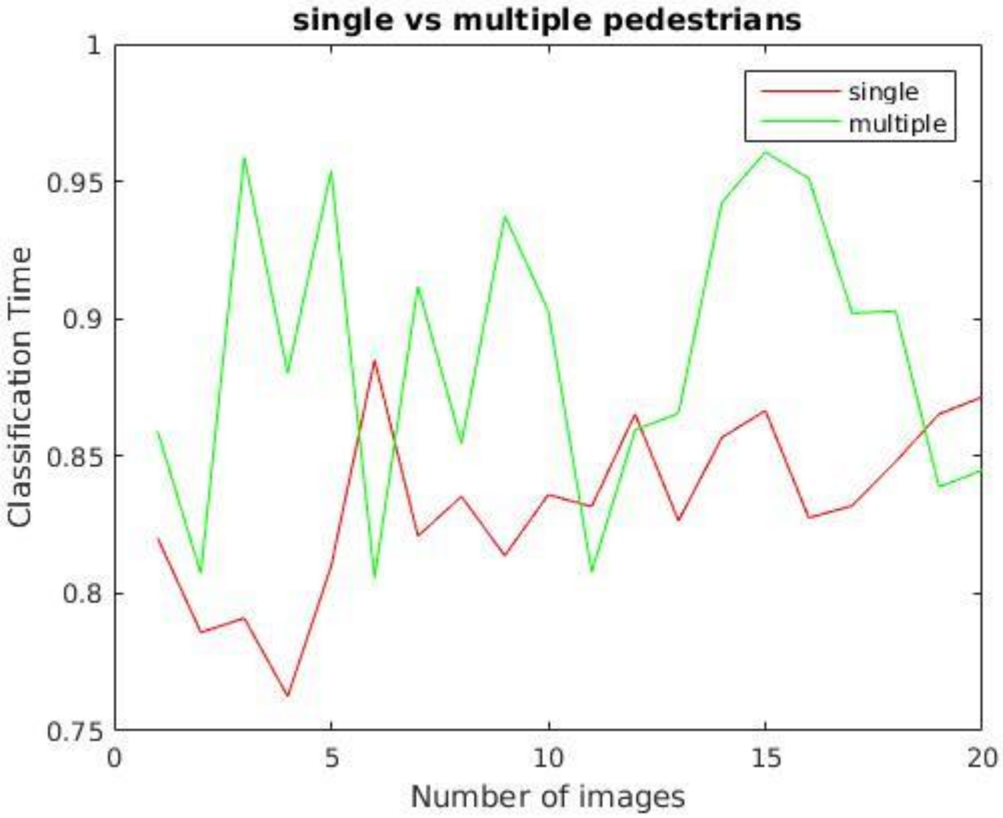

Figure 3: Single vs Multiple pedestrian detector

Conclusion:

We have shown that by computing the features human detection works well, when used in conjunction with an intersection kernel support vector machine classifier, trained with positive and negative examples. Finding the correct scale for the detection is a little tricky. The detection also tends to generate false positive on elements with strong vertical components, such as architectural elements, trees, poles.

References:

[1] Navneet Dalal, Bill Triggs. Histograms of Oriented Gradients for Human Detection. CVPR 2005

[2] Subhransu Maji, Alexander C. Berg, Jitendra Malik. Classification using Intersection Kernel Support Vector Machines is Efficient.

[3] Alexander Neubeck, Luc Van Gool. Efficient Non-Maximum Suppression.