HOG Face Detection

Juan Camilo Martinez
Universidad de los Andes
Departamento de Ingeniera Industrial
jc.martinez10@uniandes.edu.co

Abstract

Histograms of oriented gradients may be used as features for machine learning algorithms, permitting the detection of objects in a scene. Extending a single scale detector to a multi scale detector results in significantly improved performance, as shown in this work.

Keywords: object detection, hog, multi-scale, SVM

1. Introduction

The histogram of oriented gradients (HOG) is a local descriptor similar to SIFT, only it is direction invariant instead of scale invariant. The descriptors may be used as input for machine learning algorithms such as Support Vector Machines (SVMs).

By using a SVM to classify ”patches” of pixels that describe an object, and extracting multiple patches from the analyzed image, the object can be detected regardless of its orientation in the image. This however only works for a limited range of scales, if it works for more than one scale at all. To address this limitation, the descriptor is varied in scale (or more algorithmically accurate, the image is varied in scale) and new patches are extracted fom the image, with the patches exhibiting the highest confidences

Several hyper parameters can be identified in this strategy, such as:

- HOG cell size: The amount of pixels used to generate the descriptor.
- HOG template size: The size of the patches used to describe the object.
- Pixel step: How fine is the movement of one patch to the next for object detection.
- Scale step: How fine is the change in scale when detecting at smaller sizes.

2. Methodology

2.1. Multiscale HOG

Two values where used for the cell size hyper parameter, 3 and 6. The template size remained at 36. Patches of 36x36 pixels were evaluated, as this was the size of the training pictures containing cropped faces.

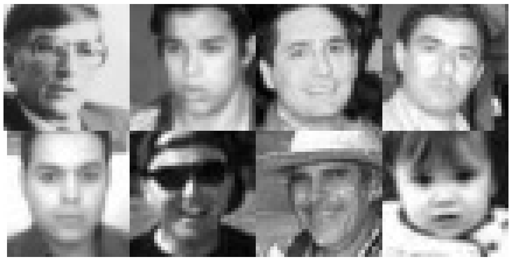

Figure 1. Sample cropped faces from training data

20.000 36x36 patches were extracted from the negative samples obtained using random scales and coordinates on the existing negative training images, all converted to gray scale.

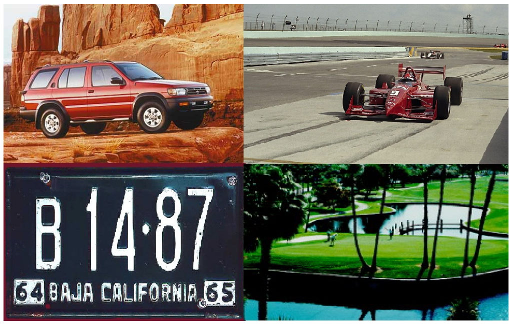

Figure 2. Sample negative patches extracted from training data

Once HOG descriptors were computed for to all cropped faces and negative samples, they were fed as inputs into an

SVM classifier with lambda=0.0001.

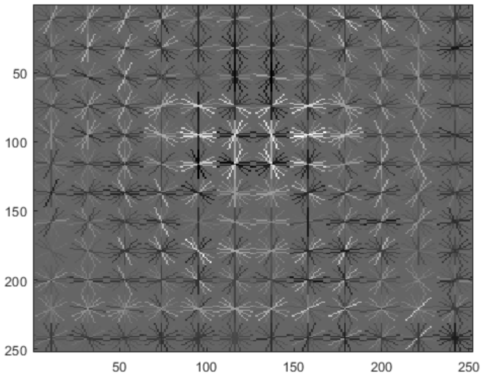

Figure 3. Aggregated HOG descriptor of a face

The HOG descriptor for the cropped images vaguely resembles a human face, as shown in figure 3.

After the training phase was completed, patches were extracted from the test images in an orderly fashioned, from left to right and top to bottom of the picture. This procedure was repeated after scaling down the image progressively to a fifth of the original size, or until the image was too small to extract a 36x36 patch.

The trained classifier was used to assign a confidence value to each patch extracted from the test images. In case this confidence value was larger than 0.75, a bounding box was assigned to that patch. In case several bounding boxes had overlapping centers, all but one of them were removed.

2.2. Evaluation

Test images include bounding box annotations (yellow in figure 4) for all faces, including overlapping ones. Calculated bounding boxes are sorted by confidence values, the ones overlapping with ground truth assigned as true positives (green and dotted in figure 4) and the ones not overlapping but with a comparable or greater confidence values considered as false positives (red in figure 4). Non overlapped ground truths are false negatives. The resulting data can be used to obtain the precision recall curve and the average precision.

2.3. Results

Initially a single scale detector was used, resulting in an average accuracy of barely 0.3.

With ”fast” hyper parameters (cell size = 6, 5000 negative samples, 18 pixels step size and 4 scales considered) average precision increased to 0.573.

Using finer step sizes, a cell cize of 3 and 15.000 negative samples, average precision was increased to 0.759.

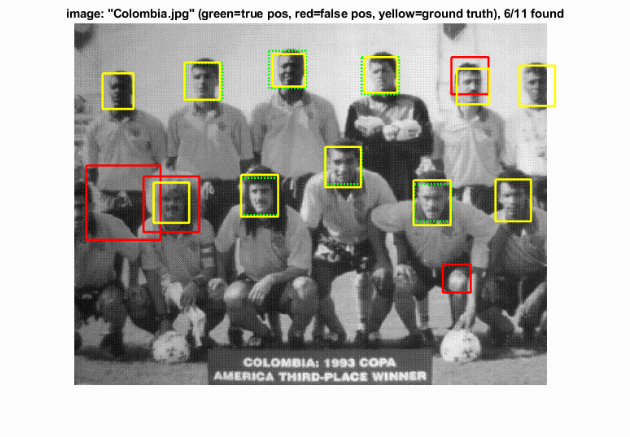
image: "Colombia.jpg" (green=true pos, red=false pos, yellow=ground truth), 6/11 found
COLOMBIA: 1993 COPA
AMERICA THIRD-PLACE WINNER

Figure 4. Detection results example

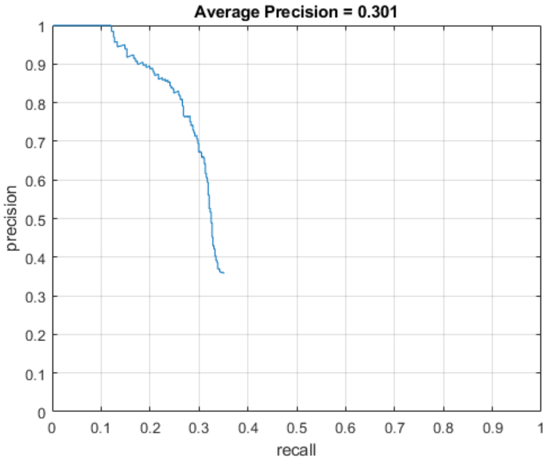

Figure 5. Results for a single scale detector

2.4. Viola-Jones

Paul Viola and Michael Jones implemented a face detection algorithm suited for real time content (videos) [1] that made use of a comparable strategy: a descriptor instead of mere pixel groupings, and a machine learning classifier to separate positive descriptors from negative descriptors.

They used integral images as descriptors of rectangle features and a perceptron for classification. The AdaBoost algorithm was used to select the most relevant features. Where this work uses SVM confidence values to root out false positives, Viola and Jones used what is now known as cascade classification, where features are grouped into different stages of classification, with false detections being

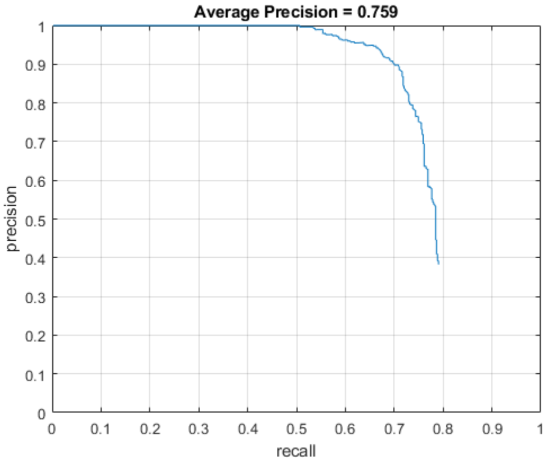

Figure 6. Results for HOG cell size = 3

progressively discarded.

3. Conclusions

The strategy presented produces a significant amount of false positives. Although the evaluation method does not take most of them into account due to the (generally) lower confidence values they present, there is still room for improvement in this regard. One method that promises good results would involve hard negative mining. However, this is not possible with the provided data, as there aren't any annotated training images exhibiting both faces and negative regions. New data with annotated ground truths that produced frequent false positives with the original parameters would need to be introduced.

References

[1] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, May 2004.