Histograms of Oriented Gradients (HOG) for face detection

Mariajosé Serna Ayala
Universidad de los Andes
m.serna10@uniandes.edu.co

Henry Daniel Torres Acuña
Universidad de los Andes
hd.torres11@uniandes.edu.co

Abstract

Face detection in computer vision is a well-defined and interensting problem since is real-world application of years of research in the field. This challenging situation is addressed by following a two basis phases: construction of a descriptor and classification. In the series of experiments presented in this report we tested the multi-scale Histogram of Oriented Gradients (HOG) strategy for constructing the descriptor and an support vector machine with intersection kernel (IKSVM) for the classification task. The experiments were performed on a subset the WIDER FACE dataser. For this purpose we used the VL_Feat library created by Andrea Vedaldi [4] and followed the Object category detection practical of the Oxford Visual Geometry Group[2].

1.. Introduction

Pedestrian detection detection on natural scenes has been one of the major achievements of computer vision in the early 2000s since it became a real-world application of the theory. Navneet Dalal and Bill Triggs addressed this problem using a method called Histogram of Oriented Gradients (HOG)[5]. The pedestrian problem was studied in the INRIA dataset, which contains 1800 natural scene images with annotated bounding boxes containing upright persons. The composition of this dataset allowed to study the problem considering variation in view-point, illumination, occlusion, scale, background clutter and intra-class appearance.

The first step in the HOG object detection technique is to map the given image into the proposed feature space. This is done by convoluting the image with several edge filters which will output the magnitude and direction of the gradient, expected to be maximum on those regions that resemble the silhouette of the object. Then, the image is divided into 8x8 pixels regions and then place each pixel's direction in one among 18 predefined oriented gradients and then build per-cell histograms using magnitudes. It is worth to be noted that the histogram is constructed using a softbining strategy. These histograms are then normalized and the final descriptor is the concatenation of all the cells in the image. Since there are different object sizes, representing different depths, the proposed descriptor should address the scale problem. For this, it is desirable to do the processing with the image at different sizes. In the resulting image pyramid when the image is smaller the recognizable object is bigger and the opposite happens when the size of the image is larger. Finally, since the goal is to is to detect the region containing the object it is necessary to use a classifier. Among various classifiers the chosen one is the SVM, because for its simplicity and robustness on a binary problem (object or no object). The final part of the algorithm is to train and IKSVm which take advantage of the fact that the descriptor is basically an histogram.

It is possible to improve the results of the classifier by applying bootstrapping and non-maximun suppresion methods. The bootstrapping technique allows to tune the parameters of the classifier to obtain good results even on the hardest conditions. The chosen classifier (IKSVM in our case) is first trained with random negatives from the database. Then, the this trained model is tested on the train set and retrained with the negatives that were most difficult to classify. On the other hand, due to the use of the sliding window technique the classifiers can output several boxes that contains the same object but with a displacement in either direction. To correct this situation only the boxes with the maximum response make it to the final result.

In the detection problem the goal is to localize in a given images all the instance of an specific object. In our case, the detection problem can be seen as a binary classification problem because the task is to find the 100x100 pixels boxes that contains a face. This kind of problem is evaluated using precision-recall and specifically the average precision measure. Precision means the rate between the real detections and the predictions. Recall refers to the percentage of positives that the detection algorithm is capable of predict. The area under formed curve is the average precision and is often used as a summary statistic for the performance of the classifier.

For the problem of face detection the aforementioned multi-scale HOG algorithm can be tested on the WIDER FACE dataset. This dataset provides faces with high degree of variability in scale, pose and occlusion. This is a large dataset with 32.203 images and 392.703 labeled faces; however, for the implementation of the multi-scale HOG algorithm we will use a small subset (aprox 9.000 images) of WIDER FACE composed with faces whose area is larger than 80x80 pixels[3].

2.. Materials and Methods

For the development of this task we used VL_Feat library, a series of functions re implemented in Matlab by Andrea Vedaldi. Also we follow the instructions from "*Object category detection practical*" practical from the Oxford Computer Vision Oxford Visual Geometry Group computer vision.

As mentioned before we used a HOG representation to train an SVM with positive and negative samples of this Face Detection task (Face or no Face). These training include a Hard -negative-mining, which means that after running, iteratively, the detection in our train images we only choose as negatives the ones that generate a greater confusion, which means, the ones that seems ro be closer to the category face than to the no-face one.
Initially we used image crops with faces to train the positive case (face), these ones where provided. For the negative cases we, based on the annotations of the complete images (whole scenario), we remove the faces and after that we took crops from each modified image. These are square image with an aleatory size between 100 and 160.

The next part was developed based on the Oxford Vision Group Object recognition Laboratory. First we represented all crops (positive and negative) with HOG with the function vl_hog. With these and their respective labels (face/no-face) we trained our model to predict, or detect a face inside an sliding window if that window corresponds or not to a face. The evaluation was performed over a train set which annotations where the upper left coordinate of each bounding box containing a face and the size, in both directions, of the bounding box. The detection was intended to be in a multiscale way, this means that it is able to recognize faces at multiple scales (small, medium and large faces).

3.. Results

We where able to represent such as positive and negative face examples and to train our model with these representations. Even though we could not be able to detect in larger images all the instances of the category face. This is because we couldn't make the Hard negative mining implementation work.

4.. Discusion

For the multi-scale HOG strategy we identified various important parameters that directly affect the result of the algorithm. For the IKSVM model we used the multi-scale HOG representation of 100x100 pixels regions containing faces (positive) or any other object (negative). the first important parameter is the size of these regions since ti directly defines the minimum scale at which the classifier is able to detect a face. Another important parameter is the size of the cell in which every histogram of oriented gradients is calculated. we worked with a HOGcell size of 8x8 pixels; however, when the size of the cell increases more individual pixels will be represented by the same HOG meaning the descriptor will tend to generalize the response on these pixels and if the size of the cell is too small the descriptor will tend to over-represent the response of the pixels. The multi-scale limit is also an important parameter to the algorithm since it represents the number of scales in which the descriptor can encode a face and then the classifier recognize it. Finally, worth mentioning that the C parameter of the IKSVM models the confidence on the classification given by the distance to the formed hyperplane. The C is the coefficient of the summatory of the errors allowed on the margin between the support vectors for each category. Thus, the manipulation of the C parameter is a way to control the propagated error in the classification process. A high value of C leads toward a much restrictive classifier and a low value to a more relaxed one, influencing directly the accuracy of the classifier.

The convolutional neural networks (CNN) often use a simple classifier (softmax) on the final layer to predict the label of an input image. Currently the best image classifications methods in computer vision are based on CNN and the excelent results they present can be understood by looking into what is learning a single neuron. On the first layers the learned features are basically edge filter, but as the depth of the networks increases the learned features become more specific for a certain type of object. This evidence that the key on classification or detection (binary classification) is to build a better feature descriptor. Considering this, we propose to improve the multi-scale HOG descriptor by taking in count other information like colour and texture.

These two new features can provide useful additional since faces (and skin in general) have a very distinctive texture and well-defined colour patrons. We propose to use the textons descriptors to create a texton map of the input image and the feed this texton map to the multi-scale HOG algorithm. Also we suggets to consider the 3 dimensions of the L*a*b* colour space when constructing the HOG descriptor by concatenating it with a colour histogram.

5.. Conclusions

- HOG is an easy approach in the object detection task, by taking into a count the form. Even though, it could be way more interesting if this representation is combined with color or color distribution information to have a more precise detection, or to divide a category into other categories.
- There are a lot of parameters to establish during the whole process which makes it confusing when having to give it values. Because we don't have another logic other that that it was the author(s) used.

References

[1] R. Szeliski. Computer Vision: Algorithms and Applications. Chapter 14:Recognition. 2011

[2] A. Vedaldi and A. Zisserman, Object category detection practical, 2014a, http://www.robots.ox.ac.uk/~vgg/practicals/category-detection/

[3] Yang, Shuo and Luo, Ping and Loy, Chen Change and Tang, Xiaoou, IEEE Conference on Computer Vision and Pattern Recognition, WIDER FACE: A Face Detection Benchmark, 2016, http://mmlab.ie.cuhk.edu.hk/projects/WIDERFace/

[4] A. Vedaldi and B. Fulkerson, ”VLFeat : An Open and Portable Library of Computer Vision Algorithms”,2008, http://www.vlfeat.org/

[5] N. Dalal and B. Triggs, Histograms of Oriented Gradients for Human Detection, CVPR 2005.