Multi-scale Histogram of Oriented Gradients for human face detection

Johana M. Ramirez Borda
Los Andes University
jm.ramirez11@uniandes.edu.co

Francisco A. Rozo Forero
Los Andes University
fa.rozo1843@uniandes.edu.co

Abstract

HOG descriptors are based on the concatenation of a number of histograms of gradients calculated in a determined number of cells that divide one image. This descriptors are then used to train a classification algorithm, for example an SVM, to detect the presence of certain object (faces) within an image. In this work, the WIDER FACE dataset was used in order to test a model based on HOG and SVM methodology to detect faces. The results obtained showed to be almost useless in the sense that little or none correct detections could be performed.

1. Introduction

In 2005, Dalal and Triggs [**?**] proposed a method called Histograms of Oriented Gradients (HOG) for the problem of Human Detection. In order to extract the HOG features of an image it is necessary to divide the image into cells and then calculate the gradients within each cell. Later, a histogram is created in which an angle is one bin and the value added to that bin corresponds to the magnitude of the gradient in that direction. These histograms are then grouped into blocks of 2x2 cells and are normalized within these blocks so they wont be susceptible to changes in brightness. Finally, the histograms are concatenated into a single HOG descriptor [3], [6], [7].

This methodology can be applied to a detection problem because each extracted image feature (HOG) is then used to train a support vector machine (SVM) algorithm. This method was firstly proposed for classification problems of two classes. With this in mind, it can be used to detect faces since it will classify the images between 2 categories, "face" or "not face".

Moreover, since real objects can be found at different sizes from the one learned, this detection can be performed at multiple scales. This is done by scaling the images up and down and using the descriptors to find the object in all possible scales. Also, it can be used to find not one but many objects within one image. To do so, there are selected x number detector responses. This number must take into account many responses (if there are more than one object) but also, since one object may generate multiple responses at nearby locations, it's important to reduce the number by deleting redundant detections. This can be done using the non-maximum suppression algorithm.

Within the main parameters used in HOG are the following: cell size, which defines the number of pixels that will be first grouped to calculate the gradient in each one and then to calculate the histogram in that cell; block size, which defines the cells that are going to be used to normalize the histograms; size of overlap or threshold to delete a possible repeated detection; finally, the bins that are used to organize the direction and magnitude data could be varied to study its influence.

2. Materials and methods

2.1. Dataset

The Chinese University of Hong Kong Multimedia Laboratory is one of the most recognized institutes worldwide on deep learning. Their department of Information Engineering developed a WIDER FACE dataset which consists on benchmark dataset that contains 32203 images with 393703 labeled faces organized in 61 event classes. For each class, there are 40% belonging to train, 10% to val and 50% to test sets. Figure 1 was provided by this laboratory and helps to illustrate some examples of the images and its categories [8].

Figure 1. Example of the images and categories found within the database and their annotations [8].

As it can be seen, there are several categories that provide a high grade of variability in terms of scale, pose, oc-

clusion, illumination, etc. Also, Figure 1. illustrates the annotations which are a bounding box surrounding the faces. Finally, the evaluation methodology used in this dataset is the same employed in PASCAL VOC dataset. This consists in the submission of a bounding box, with a level of confidence, for each detection. Then, the precision/recall curve is computed and the AP (average precision) is calculated.

2.2. HOG

The methodology of this laboratory is based on the Object category detection practical [1] provided by the Oxford Visual Geometry Group and authored by Andrea Vedaldi and Andrew Zisserman. In this sense, the procedure implemented is described below:

1. Selection of positive training data: this step was already provided in the form of a directory and consists on the extraction of the patches that correspond to the object of interest within the train images. Then, those patches were resized to the same dimensions.

2. Selection of negative training data: non-object patches were extracted uniformly and saved into a directory.

3. Selection of cell size: it was set to 8 as the default value, and more important, the one recommended in the Dalal and Triggs work [3]. Once it was selected, the horizontal and vertical gradients were calculated in each pixel within the cell.

4. Calculation of the direction and magnitude of each gradient. In color images, the magnitude is selected as the maximum of the magnitude of the three channels, and the angle is selected as the one of the maximum gradient.

5. Creation of an histogram per cell: this histogram is built using the angles of the direction as bins and the magnitude as the values assigned to each bin. When an angle of a gradient is not the same than one of the bins, its contribution was divided between the two closest bins proportionally.

6. Normalization: One block is formed when 4 cells are grouped. Then, the block normalization is performed in order to avoid susceptibility for light changes, since this process makes them invariant to multiplications of the pixel values.

7. Obtaining of one HOG descriptor per image: all the histograms are concatenated into a single feature that will represent each image.

8. Training: once the images features have been extracted, a SVM was trained in order to obtain a model of classifier to detect the faces.

9. Evaluation in training data: given the model obtained before, it was tested and the hard negatives were found and used in the next negative training data. Later, the repeated negatives were eliminated.

10. Detection: using the model created, it was used to detect the objects (faces) in the test images. To do so, a for loop was created to read all the test images.

11. Non-maximum suppression: this algorithm was used to keep just those detections that, starting with the highest response, its overlap didn't pass a default threshold-0.25.

2.3. Evaluation

Similar to the PASCAL VOC dataset, the evaluation methodology used in Wider Face dataset (and the one used commonly in a detection problem) is based on the precision-recall curve and the average precision. In this sense, *recall* is the proportion of all positive examples ranked above a given rank and *precision* is the proportion of all examples above that rank which are from the positive class. The AP is defined as the mean precision at a set of eleven equally spaced recall levels [5]. Also, there was a metric used in order to compare the bounding box of a candidate and the groundtruth: this was the overlap metric and is defined as ratio of the area of the intersection over the area of the union of the two bounding boxes [1].

3. Results and discussion

While running the code, the evaluation of each train and test image was showed. In Figure 1. it can be seen some examples of the detection when evaluating in training data before collecting the hard negatives and retraining the model with them.

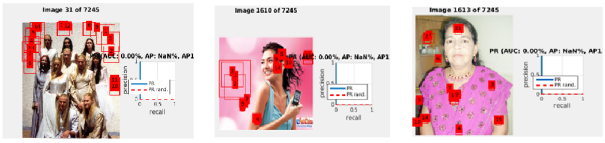

Figure 2. Example of the detection when evaluating the training images.

As it can be seen, there are results when detections appear everywhere except in faces and others where one or a little more parts of the face are detected. Although these are not the expected results, they may be reasonable since we hadn't found yet the hard negatives and the model still detects a lot of false positives. This process was repeated iteratively with almost no significant change. For example, even though in some cases there were more than one detection in the area of the face, many more were detected outside this area.

This first train of the algorithm spent a lot of processing time, as it process one image every 3.5 seconds, so a wide number of training images would require several hours or even days to complete the process. In general terms, observing the detections, we can not conclude that there is a pattern within the false positives nor the false negatives. For example, there were found several false positives in different areas such as, background, clothes, walls, shadows, etc (see Figure 2, left image.). A similar situation happened with the false positives because it was never found a complete positive object, it means, it was only found one or little parts of the face but not completely as it is found in the annotations (see Figure 2, middle and right images).

This results may be caused by the selection of the parameters to use in both HOG and the SVM classifier. For example, it was noted that the cell size parameter influence a lot the obtained results: when this parameter is too low, a face won't be detected because it is not identified as one by the algorithm, on the contrary, if the value of this parameter is too big,the train features will be very specific for the train images and will not give good results in the test set. In this sense, wrong values of parameters like the block size may be influencing in the false positives caused by changes in illumination that the model detects as one face when there is not. Finally, other parameters such as the number of bins may be considered since increasing this value will provide more information about the image. However, in this case one must be careful because even though one will get more information, the time processing will increase considerably.

3.1. Limitations

Although HOG has many advantages, some of its problems are related with the fact that it uses a rigid template that usually is not enough to represent a whole category. For example, many objects may be rotated, deformed or just look very different from different viewpoints. Moreover, it may not differentiate very well between parts of edges or contours and textures within objects. For example, Doersch and Efros [4] in their work called "Improving the HoG descriptor" affirmed that "intuitively these two patterns should be separated, since they tend to have different semantic meanings, but HOG is only sensitive to the direction of gradients". Finally, it uses many parameters like cell size, block size, size of overlap and number of bins that plays important roles in the process and, at least initially, must be chosen subjectively.

3.2. Further work

Some proposed methodologies to overcome the problems of HOG using linear SVM are the "exemplar SVM" and the "part based models". For example, exemplar SVM train a separate SVM for each positive instance. This makes that it can handle intra-category variance naturally without using a complicated model and can make use of negative data and train a discriminative object detector. Furthermore, as there is a explicit correspondence between the detection and the training example, it can be obtained much more information, like 3D and orientation information [2]. Moreover, a way to deal with 3D orientations is to train using a neural network instead of an SVM. Therefore, a series of neurons/nodes should correspond each to a different orientation for a single object and should be weighted differently. Hence, for multiclass problems where several categories contain images that can be found with different orientations, would require a major selection of nods and layers. Though the problem would require a big network, the results may improve the HOG's performance.

4. Conclusions

Even though it has been found that using locally normalized histograms of gradients as descriptors in a dense overlapping grid gives very good results in detection [3], our results didn't show the same information. All of the recommended tips by the Oxford Visual Geometry Group were performed in order to improve the detection: implementing the non-maximum suppression to reduce the number of false postives (overlaping and repeating detections); evaluating firstly with the train images in order to find hard negatives and then use them; re-scaling the image, etc. However, the results didn't improve. For that reason, and noticing the big impact of the parameters, we propose a exhaustive study of the values that must be given to them in order to obtain better results. Moreover, we suggest the implementation of HOG using others classifiers or the implementation of other algorithms like exemplar svm or neural network.

References

[1] A. Z. Andrea Vedaldi. Object category detection practical, may 2018.

[2] P. Arbelaez. Recognition 03. In *Computer Vision*. https://sicuaplus.uniandes.edu.co/webapps/blackboard/execute/content/file?cmd= 252927_1&$course_id =_1 54774_1$.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[4] C. Doersch and A. Efros. Improving the hog descriptor.

[5] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[6] C. McCormik. Hog person detector tutorial, may 2013.

[7] L. OpenCV. Histogram of oriented gradients, dec 2016.

[8] S. Yang, P. Luo, C. C. Loy, and X. Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

All the codes created and used are attached.