Face detection with PHOG strategy on Wider Face dataset

Catalina Gómez
Universidad de los Andes
Cra 1 #18a-12, Bogotá, Colombia
c.gomez10@uniandes.edu.co

Diana Herrera
Universidad de los Andes
Cra 1 #18a-12, Bogotá, Colombia
ds.herrera10@uniandes.edu.co

Abstract

Detection is a relevant task in the Computer Vision field, the goal is to find all the instances of an object category in an image using a bounding box usually. A correct detection is considered by a minimum accuracy in the overlap between the annotation and obtained bounding boxes. Afterwards, precision recall curves are built using various thresholds to evaluate overall performance. To tackle this problem, Dalal and Triggs developed the PHOG algorithm, evaluated in this opportunity on the WIDER FACE dataset, in which faces must be detected. We added a few variations such as increasing the epochs for hard-negative mining and reducing the number of negative instances in order to balance the classes to train the SVM. Performance was measured with the area under the precision-recall curve, and its magnitude was around 10^{-6} *and with the best method* $10^{-}4$*. This bad results arose from the difficulties inherent to detection problems, specially the intra-class variability due to the different scenarios where faces could be detected.*

1. Introduction

Detection is a relevant task in the Computer Vision field. The goal is to find all the instances of an object category in an image using usually a bounding box. However, there could be within-class appearance variations and different scales that make this a challenging problem to study, in addition with other aspects such as background clutter, occlusion, changes of illumination, and the projection of a 3D world into a 2D image. In this task, we have previous knowledge of the class to be detected and training is done with examples of punctual positives, so as to learn features of the particular object to be detected instead of the whole image, with a large area where there are no objects of interest [1].

Evaluation must be done in two steps. First of all, each detection must be compared with ground truth to check the overlap between them through calculation of intersection over union; only if the result exceeds a given threshold the detection is considered a true positive. After calculating the amount of true positives, true negatives, false positives and false negatives, a precision recall curve is built by varying the threshold to determine a true positive. Precision is calculated as true positives over true positives plus false positives, whereas recall is found through the division of true positives over true positives plus false negatives.

To tackle this problem, Dalal and Triggs developed the Pyramid of Histograms of Oriented Gradients (PHOG), initially devoted to detect pedestrians [2]. In the algorithm, each image is represented through one descriptor based on histograms of oriented gradients of cells in the image. First of all, the image is convoluted with filters to find borders, and using them, gradient magnitude and direction are calculated, leaving for each pixel the color channel with greatest magnitude as final gradient. Then, the resulting image is divided into cells (usually size 8x8), and histogram of oriented gradients for each cell is built assigning each pixel to the closest gradient orientation within 18 bins ranging from 0° to 180°. The 18 orientations are used to identify the contrast sensitive features but similarly, the procedure is repeated with only 9 orientations to bring information about contrast insensitive features. Both features are normalized 4 times for each neighbors blocks and then average is calculated. Lastly, texture features are added through the sum of the magnitudes among all orientations and the normalization of the data obtained 4 times. The three vectors are concatenated into a 31 feature vector to represent each cell in the image. Based on these vectors, an initial Intersection Kernel Support Vector Machine (IKSVM) is trained to classify candidates from bounding boxes between positives and negatives (windows without the object of interest obtained randomly). Afterwards, to improve results, the IKSVM is trained iteratively, but this time negatives used will be "hard negatives", those detections that are usually classified as positives but are not, which can be identified manually or based on the score the SVM assigns to each detection. This process is called Bootstrapping and helps the algorithm to discriminate better among true positives and those shapes that are commonly misclassified but are not true positives.

Finally, the algorithm is used in the test set in which a sliding window scans all positions from the image and evaluates whether they are detections or not through the previously trained IKSVM. During this process, a pyramid of scales is scanned because objects can be in different scales in a single image. To do so, the image is subsampled and resized through a pyramid with various levels, but bounding box is always the same size used to train the model, otherwise a new SVM would be needed to be trained for each size of the box [2].

Nevertheless, a common problem is that many bounding boxes can be created to identify the same instance of a single object. To solve the problem, non maximum suppression is used. Detections (bounding boxes) are sorted according to their classifier score and afterwards, bounding boxes overlapping over a certain threshold with another detection with higher score are suppressed. This way, a single detection (bounding box) remains for each instance of the object in the image [2].

Although the PHOG algorithm was initially used on pedestrian detection, it is a generalized algorithm and can be used to detect other categories, due to the fact that the representation is done based on the histogram of gradients of the positives of the class to be detected, thus giving a broad example of the gradients of different instances belonging to the same class, hence considering intra-class appearance variation and differences in point of view. Based on these representations, the IKSVM trained is capable of classifying detections considering a broad range of possibilities in the same class. Additionally, the permutation between the dimensions of the feature vector can be done and represents also changes in the perspective of the objects, giving additional representations for the training stage. Furthermore, in detection tasks it is paramount to search objects in different scales, and the pyramid of each image through which sliding window is analyzed, allows to use the same features obtained in training at a single scale to evaluate detections made at smaller scales. As can be seen, the PHOG strategy is not particularly thought for a particular class to be detected but for any, because the representation is based on the examples of the class of interest.

Another algorithm developed for face detection is the Viola Jones algorithm. Made public in 2003, it was created to be a Robust Real Time-Face Detection method, processing images extremely rapid with high detection rates. It is based on three important points developed by other authors; the first is the "Integral Image" representation that computes the features used by the detector very quickly, the second is an efficient and simple classifier built using the AdaBoost learning algorithm and the third, is a method that combines classifiers in a cascade to discard quickly background regions and focus computation only in promising face-like regions [4].

In this opportunity, face detection was made in a small part of a dataset called WIDER FACE. We used 2379 train images each one with at least one face, 10500 positives which are faces individually cropped from train images, 2379 test images with at least one face, and obtained over 80000 negative random crops from train images.

2. Materials and Methods

The VLFEAT open source library was used to run PHOG algorithm within the subset of the WIDER FACE dataset used. PHOG strategy implemented was the one authored by Andrea Vedaldi and Andrew Zisserman on 2014, available at Oxfor Visual Geometry Group [3]. Based on PHOG representation of 10500 positives (individually cropped faces) and over 80000 negatives (obtained randomly from 2379 train images and manually revised to delete faces), an IKSVM was trained to classify detection candidates (bounding boxes). Test was done considering scales from -1 to 3 times the original size. Afterwards, bootstrapping was made using the negatives randomly found with the highest scores. Three epochs were used for bootstrapping procedure. Lastly, non maximum suppression of detections was performed to eliminate all bounding boxes that overlapped over 25% with a higher scored detection. Evaluation was done through precision recall curve.

The most relevant parameters for the PHOG strategy in WIDER FACE dataset included the number of negatives to do bootstrapping, the number of epochs for bootstrapping process, the number of scales analyzed in test, C parameter for IKSVM (10 initially), threshold to suppress overlapping detections and the number of detections to be kept with highest scores.

After testing the algorithm with the default values to improve the results, we tried to change some parameters according to the hypotheses explained here. We diminished the number of negatives used, in order to balance positives and negatives, and increased the SVM C parameter because precision results were extremely low, and we presumed that doing so, false positives would decrease by punishing them more. We also increased the number of iterations for Hard negative mining to improve the results of this process.

We also tested Viola Jones algorithm on the dataset, based on the method implemented at Matlab vision library to compare results with PHOG. For each detection made, a random score was assigned.

Lastly, using the algorithm available in Matlab for *Object Detection in a Cluttered Scene Using Point Feature Matching* we found Waldo :) To succeed, we rescaled the template of Waldo to 30% the original size and used a *Match Threshold* of 0.2 to identify only the most perfect candidate in the dataset.

Table 1. Scenarios evaluated in WIDER FACE.

Default Scenario	
Train Images	2379
Train Positives	10500
Train Negatives	80200
Test Images	2379
Bootstrapping epochs	3
Scale Range	-1 to 3
SVM C	1
Threshold suppress overlapping detections	25%
Our Scenario	
Train Images	2379
Train Positives	10500
Train Negatives	13400
Test Images	2379
Bootstrapping epochs	5
Scale Range	-1 to 3
SVM C	10
Threshold suppress overlapping detections	25%

3. Results

Results for Precision Recall with PHOG are shown for the default scenario evaluated in the subset of WIDER FACE. The results are shown in magenta and the AP of the method in the legend has an exponent of $x10^{-6}$.

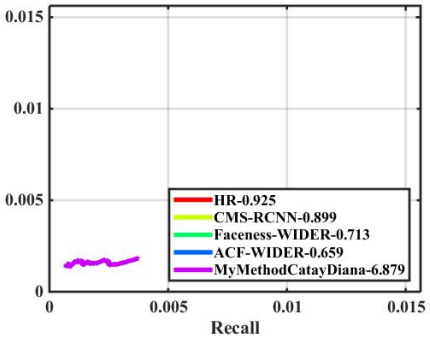

Figure 1. Precision-Recall for default scenario in easy.

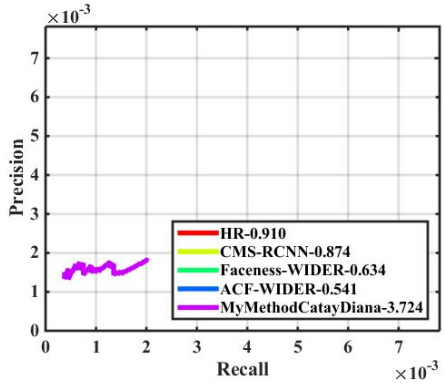

Figure 2. Precision-Recall for default scenario in medium.

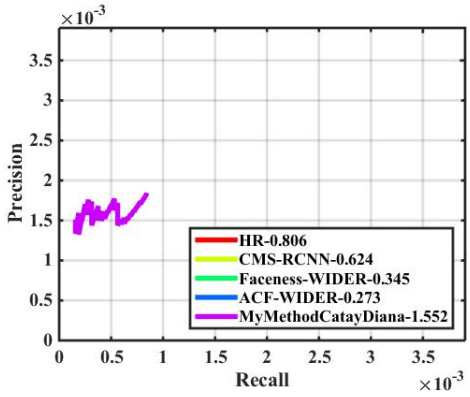

Figure 3. Precision-Recall for default scenario in hard.

Below, we show the results for Precision Recall with PHOG for the modified scenario with balanced negatives and positives and increased C SVM parameter. Our results are shown in purple.

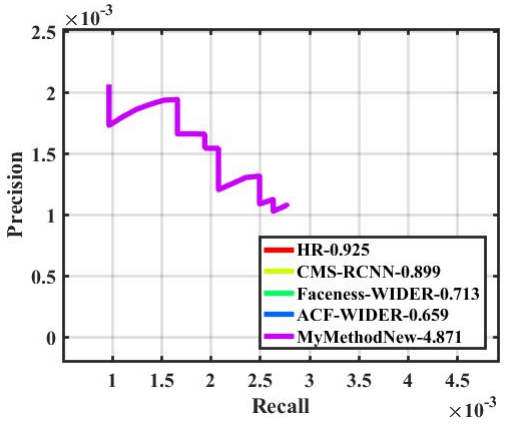

Figure 4. Precision-Recall for modified scenario in easy.

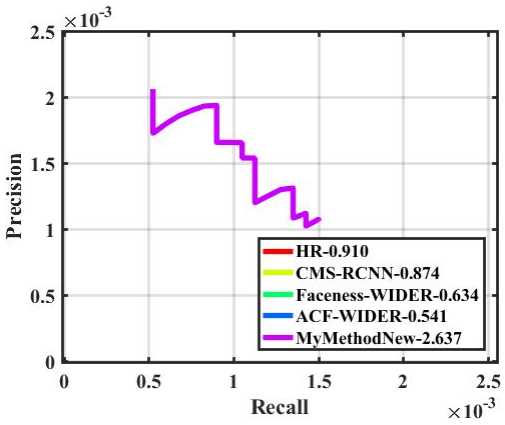

Figure 5. Precision-Recall for modified scenario in medium.

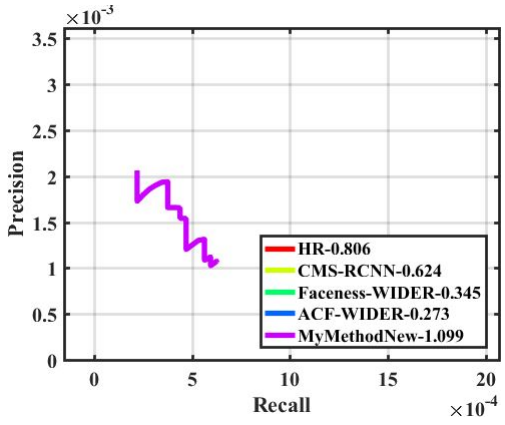

Figure 6. Precision-Recall for modified scenario in hard.

Lastly, the results of Viola Jones algorithm are shown in magenta.

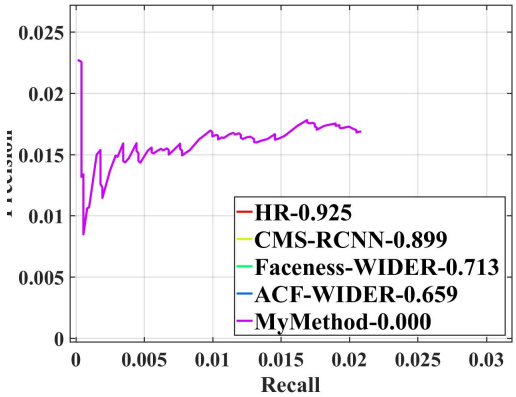

Figure 7. Precision-Recall for Viola Jones Method in easy mode.

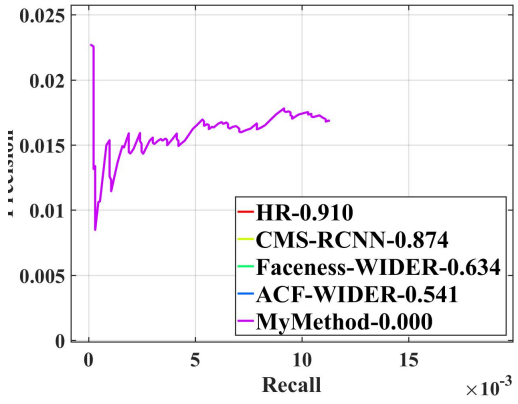

Figure 8. Precision-Recall for Viola Jones Method in medium mode.

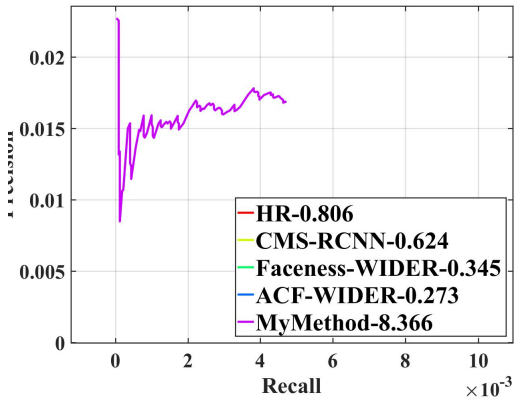

Figure 9. Precision-Recall for Viola Jones Method in hard mode.

Average Precision for all algorithms are shown explicitly in the table below.

Table 2. Average Precision for scenarios evaluated in WIDER FACE.

Level	PHOG: Default Scenario	PHOG: Our Scenario	Viola Jones
Easy	$6.879x10^{-6}$	$4.871x10^{-6}$	$3.6x10^{-4}$
Medium	$3.724x10^{-6}$	$2.637x10^{-6}$	$1.6x10^{-4}$
Hard	$1.1552x10^{-6}$	$1.099x10^{-6}$	$7.2x10^{-5}$

And lastly, the real probe that we found Waldo is shown in the following figure.

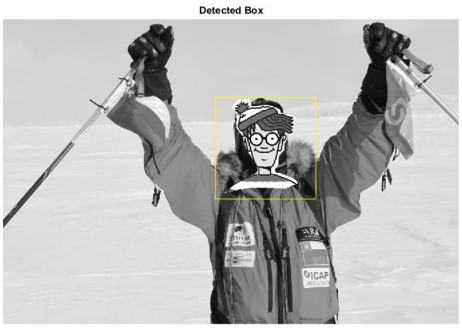

Figure 10. WALDO WAS FOUND.

4. Discussion

The results obtained with both algorithms tested and all scenarios were not good, especially compared to the ones of the other algorithms. This might be due to the fact that the dataset is hard and has a wide range of scales in which faces must be detected and a cluttered background with loads of face-like objects that are detected as faces when using PHOG features. Nevertheless, it can be clearly seen that Viola Jones algorithm overperformed PHOG in both default and modified scenarios in at least one order of magnitude for easy, medium and hard levels and on both precision and recall. This might be due to the fact that Viola Jones is a more advanced algorithm that uses three procedures to optimize calculation and improve results, explained in detail previously. These contributions have an impact in both precision and recall because discarding the background and focusing on potential candidates detection has great impact on precision, and so does using a powerful cascade classifier. The combination of these together with the fast Integral Image representation algorithm increases the probability of not missing so many detections. In fact, a better result was expected for Viola Jones implementation.

Furthermore, when comparing the results between the default PHOG and our modified scenarios, the formers are slightly better than the laters. Although we expected the opposite, because the negative and positive classes were almost balanced, this can be explained because with more negatives, the SVM learned better what it should not detect as a face, and because although using a higher SVM C parameter led to a change in precision results, increasing the range of the values, it did not improve the overall precision. This happens because with a higher value, points inside the margins are highly penalized and precision tends to have a slight increase, but recall may decrease because some of those points are actually faces. This is also why recall is bigger in the default scenario, because SVM C is lower and more detections are considered positive.

Despite the fact that HOG representation has many advantages such as capturing specific edge or gradient structure of local shape and it has a controllable degree of invariance to local geometry and photometric transformations [2], the results obtained with this strategy were not as good as expected. The bad results could be due to the fact that the images in WIDER FACE dataset are extremely different between classes and even inside classes. There are occlusions, point of view variations, background cluttering, different scales, illumination changes and intra class variations that make this a challenging problem and dataset.

The hyperparameters of PHOG must be carefully chosen because not only precision and recall are affected but also processing time. Increasing the number of negatives for bootstrapping increases the time to train the IKSVM and could lead to a better trained classifier if the negatives are well chosen, otherwise they could worsen the classifier performance if faces are in some of the negatives or simply do not improve the performance if negatives used are not truly hard negatives. Since HOG method takes into account orientations of local shape as descriptors of the positive instances, objects with similar appearance (negative instances) could lead to confusion and misclassification. Out negatives patches were chosen randomly, but we were careful removing the patches that contained faces, but not the objects that could be confused with faces. However, hard-negative mining was used to deal with this problem. In the hard-negative mining, the number of epochs of the process increase considerably the time taken to train because it is equivalent to the number of times that the SVM must be trained. Nevertheless, with each iteration the training is done with less hard negative but with the best ones, because only those with highest scores are kept, therefore an optimum level must be found empirically.

Additionally, the C parameter for the IKSVM can not be too high because then precision might increase but recall will decrease, as points inside the margins would be severally punished. Likewise, if the C value is too low, precision will decrease although recall might increase. The threshold to suppress overlapping detections may have great impact on performance depending on the image tested; if there are a lot of instances to be detected and a low threshold is used, it is probable that some instances located close by wont be detected, because their bounding boxes will have an overlap. If the level is high, then precision will decrease because there will be some detections with more than one bounding box that do not match with annotations, hence leading to increase false positives.

Lastly, the number of detections to be kept with highest scores is a problematic hyperparameter because the number of instances on each image is not known and thus, the chosen value will limit the amount of detections made on a image, either leaving some instances without any chance to be recognized or keeping errors as detections.It must be taken

into account that for crowded images the problem difficulty increases since this task is also hard for humans to annotate the images, and therefore match the annotations.

In order to improve the results, it would be extremely useful to know how many faces are there in each image, so as to keep only the same amount of highest scores, and also to know the range of scales through which faces appear, that way the scale analysis can be done only in that range and avoid making detections in outer scales. in addition, more positives instances could be used to trained the model in order to take into account a more generalized case of faces in any scenario.

5. Conclusions

- Evaluation of the hyperparameters used in a detection algorithm must be done considering not only PR result but also time cost required to train the model. In addition, in this type of problems where a particular object wants to be detected, it is important to chose well the negative instances to train the model to discriminate better between positive and negative instances.
- Neither PHOG nor Viola Jones performed well on the selected subset of WIDER FACE dataset. This might be due to the fact that the images are extremely different between classes and even inside classes and there are occlusions, point of view variations, background cluttering, different scales, illumination changes and intra class variations that make this a challenging problem and dataset.
- Viola Jones overperformed PHOG significantly. We presume it is due to the three powerful contributions of the former.
- We found Waldo! The method available in Matlab for Object Detection in a Cluttered Scene Using Point Feature Matching demonstrated to be a powerful algorithm when Match Threshold parameter was properly tuned as well as the scale of the template.

References

[1] S. Ilic, "Tracking and detection in Computer Vision", 2009-2009 TUM.

[2] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection", 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05).

[3] A. Vedaldi and A. Zisserman, "Object category detection practical", 2014 [online] Available at `http://www.robots.ox.ac.uk/%7Evgg/practicals/category-detection/`

[4] P. Viola and M. Jones, "Robust Real-Time Face Detection", International Journal of Computer Vision, vol. 57, no. 2, pp. 137-154, 2004.