Multi-scale HOG recognition in a subset of the WIDER FACE dataset

Stephannie Jimenez Gacha
Universidad de los Andes
Cra 1 Este No 19A - 40 Bogotá, Colombia
s.jimenez16@uniandes.edu.co

Sergio Galindo León
Universidad de los Andes
Cra 1 Este No 19A - 40 Bogotá, Colombia
sa.galindo10@uniandes.edu.co

Abstract

Face detection is a well developed technology used in photography, security, military engineering, animation and security with considerable importance in these areas. Also, in computer vision, detection corresponds to one of the main problems. For this, a face detection model using a multi-scale histogram of oriented gradients (HOG) descriptor using a cell size of 8 and 4 scales and a linear SVM with hard negative mining was implemented in a subset of the WIDER FACE dataset. The results obtained show a very poor performance of the model with a maximal F measure of 0.0% and a minimum quantity of detections. It seems that the HOG descriptor with the actual settings is not appropriate for face description and detection, also, the classifier used might not be adequate for the very unbalance class that face detection represents.

1. Introduction

Detection corresponds to one of the three main problems of the computer vision whose objective is to localize all possible instances of objects belonging to a semantic category[2]. This task can be seen as a specification of the classification problems where the idea is to classify all windows of the image in one of two categories: instance container or not. Detection algorithms have been used widely in biology, engineering, military industry, surveillance and medical diagnosis as an important tool to reduce the amount of time required to identify objects as well as to improve the actual performance when human criteria is not consistent or the image characteristics are limited[1].

One of the most commonly used descriptor for detection is the histogram of oriented gradients (HOG). This feature descriptor describes the objects shape based on the number of gradient orientations in determined regions of the image utilizing local contrast information and generally, a grid to generate the image partitions[3]. This descriptor has been used for many applications including pedestrian detection, digit recognition and handwritten letters detection with acceptable and slightly poor results.

The pure original HOG strategy had some disadvantages as the scale and rotation dependence of the descriptor. For that, a pyramidal HOG (PHOG), that analyzes the gradient orientations at different scales was proposed, leading to the improvement of the detection performance considerably. This new descriptor solved the scale issues,without fixing the rotation dependence which is still a concern. PHOG was implemented for the same type of applications, specially, for pedestrian detection, one of the most important categories in the detection datasets[2].

One important recognition problem regarding PHOG corresponds to faces detection, commonly used in cameras and optical systems for photography and design. This task was particularly challenging as faces, despite being homogeneous and non deformable objects have variable shapes given by pose, scale and orientation. Face detection is particularly useful in animation, photography, security and military engineering, for that, in order to implement and evaluate the PHOG descriptor for face recognition, a detection algorithm by means of this descriptor in a WIDER FACE dataset subset is proposed.

2. Materials and Methods

In order to implement the PHOG descriptor a very small subset of the WIDER FACE dataset, from the Chinese University of Hong Kong, was used for train, validation and test purposes. Also, the multi-scale HOG descriptor implementation of the Visual geometry group at Oxford was used. Finally, for detection purposes the SVM approach and implementation from the same group was used, and for evaluation, we used a modified version of the Oxford benchmark, for small results.

2.1. Dataset

The subset of the WIDER FACE dataset was composed of 7245 images from the 61 original classes with unbalance distribution, 12242 faces crops from the train images, divided into the same 61 classes, and 3226 validation images (from the same classes too) which were used for evaluation

purposes only. All train and test images dimensions corresponds to 1024xM while all train face crops dimensions are 100xN. The images are found in .jpg format in both landscape and portrait orientation.

The categories present in the database are unrelated and correspond to substantives enclosing common activities of daily life. Some of the images were digitally modified changing color contrasts or including test. The groundtruth of the dataset consist of 3 different. mat files containing the image name, number of faces, the coordinates of the left superior corner of the detection and the width and height of the detection.

2.2. PHOG implementation

The PHOG implementation was performed using the multi scale HOG implementation found in the exercise 4 of the visual geometry group. First, the face crops found in the train folder are loaded into the workspace as they corresponds to the selected and preferred train images rather than the whole images. This is because the HOG descriptor of the class *faces* should be correctly computed from the crops directly offering an homogeneous model to represent faces.

After that, the HOG features of the patches were computed using a cell size of 8, 8 orientations equally separated in the polar plane and 4 different scales, doubling its resolution at each level. The positives corresponds to face crops while negatives to the remaining image windows of the same size as crops. Then a hard negative mining was performed using a linear SVM for 7 iterations to finally evaluate the model using the evaluateModel function provided, which calculates the F measure over the train images.

To obtain detections in the images, the image is resized n times (n=number of scales) using a factor of 1/scale and the HOG features using a cell size of 8 and 8 orientations (the same parameters of training) are calculated using the vlhog function from the vlfeat library. Then, a convolution of the image with the SVM trained positives is performed using the vlnnconv followed by a normalization.

After that, a non maxima suppression in the response map was performed and an elimination of the boxes with an overlap of more than 50% was done. Finally, a .txt is constructed for each image containing the image name, number of faces detections, confidence and the description of the detection bounding boxes with the left superior coordinates, width and height.

2.3. Evaluation

The evaluation of the methodology was performed using the widerEval function of Shuo Yang which uses the .txt constructed previously to compare the detected bounding boxes with the groundtruth; a detection is considered a true positive if the intersection of the boxes is more than 50%, false negative if there is not bounding box intersecting a face and false positive if the detection does not intersects with a face groundtruth box.

Finally, the function constructs and returns a PR curve by thresholding the detections respecting the confidence level and calculating the respective precision.

3. Results

The PR curves for the face detection in the easy, medium and hard difficulty are presented below. It can be seen the number of detections are very small which represents a precision of nearly zero respecting the nearly 12000 faces present in the dataset for any threshold. For none of the difficulties the multiscale HOG strategy worked well respecting the state of art models which are plotted in the same figures when using the widerEval function provided. The state of art methods posses a maximum F measure of 0.802.

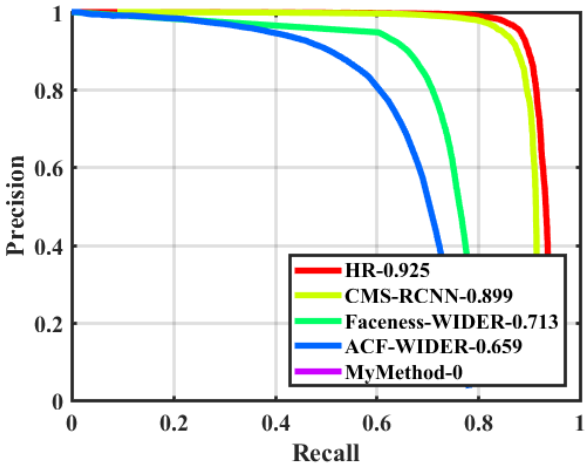

Figure 1. Easy difficult faces PR curve

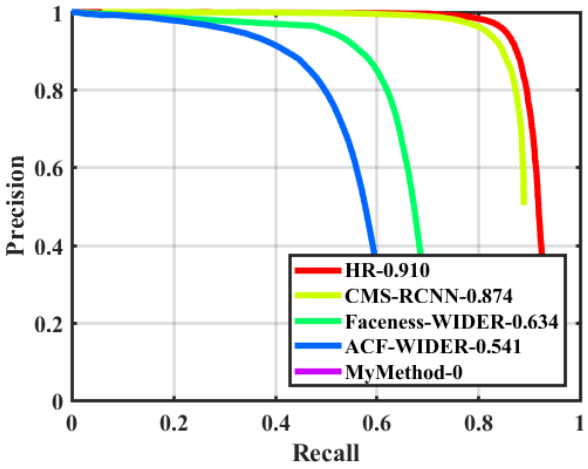

Figure 2. Medium difficult faces PR curve

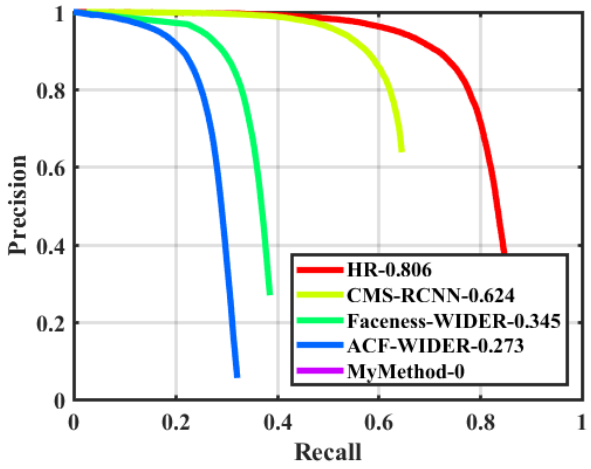

Figure 3. Hard difficult faces PR curve

4. Discussion

The Histogram of oriented gradients (HOG) for detection corresponds to a methodology that aims to construct a shape descriptor model of a semantic category based on its' gradient orientations when appearing in a image. The main idea is that the shape of the object in different instances might be preserved, such that, the contrast with the different objects and background allows to represent the object with a particular gradient orientation histogram distribution. This distribution, ideally, should be different enough from other objects in order to train a classifier capable of distinguish between HOG belonging to different classes.

Due to the scale dependence of the original HOG descriptor, which limits the detection performance, the multiscale HOG methodology was proposed. This method consist on applying the original HOG strategy to the images at different resolutions such that, objects belonging to the class of interest can be detected despite their scale in the images using the same HOG trained model. However, the number of scales considered and the size of the HOG cells are still some important characteristics of the methodology that can affect the performance.

Some of the most important hyper-parameters of the multiscale HOG corresponds to the number of scales, their relative differences and the cell size. At first, it should be noticed that the multiscale HOG methodology works by adjusting the image resolution to the trained model, which remains of a predetermined size. This implies, the number and magnitude of scales must be sufficient enough to cover from low to high resolution instances and also, they must be correctly distributed. If the number of scales remain low, the number of false negatives will increase, as the objects detectable in other scales will not be considered. Also, if the scales are not well distributed it is possible that, in the selected resolutions, the descriptor does not describe properly the objects (i.e setting a very high resolution an a very low resolution might cause a model to small to detect in high resolution but to big to detect in very small resolution).

On the other hand, the cell size determines the degree of detail considered to describe the objects in the images. A large cell size will produce very rough models giving many detections with reduced confidence (underfitting) while a small size will produce very detailed but rigid model that might offer little high confidence detections (overfitting). During training, it is necessary to posses an adequate cell size to acquire some flexibility that allows to construct an aligned, general and not so noisy model. Also, during test, it is necessary that the cell size remains coherent with the cell size used for training. Despite the fact that multiple scales can partially deal with this differences, if the train and test cell sizes remain very different, the model an descriptor will become ineffective to detect the desired objects as they will be capturing different representations of the image.

A general detection problem can be evaluated using PR curves and the maximal F measure. The PR curve is constructed by computing the precision while changing the recall by means of varying a confidence level respecting the detections. Both measures belong to the 0-1 interval. Also, it is necessary to establish some criteria for considering true positives, true negatives, false positives and false negatives like an overlapping percentage threshold. Two of the mostly used quantitative measures are the maximal F measure, calculated as $max(2 * P * R/(P + R))$ and the area under the PR curve, known as the AP.

The main limitations of the implemented methodology correspond to the pose related shape of the human face as there are considerable differences when the face is captured frontally or laterally and when the subjects is looking in different angle. This difficults the construction of a general HOG based model capable of detecting any face instance while adding noise to the model. Also, a limited number of scales considered in HOG respecting the infinite number of face scales due to little changes in the distance to the camera corresponds to a great limitation. Also, the descriptor is non invariant to the frequent transforms as rotation which is frequent in human poses.

Respecting the results, the false positives follow a characteristic pattern described by a rounded edge with linear elements inside like wheels, balloons or sewer that can be properly described with the HOG descriptor and hardly classified with linear SVM. On the other hand, the false negatives follow a very characteristic pattern: they do not present a pattern and are found everywhere. There is not a characteristic face that can be easily detect using the multiscale HOG implemented. This partially explains the very bad results obtained of maximal F measure of 0.0% in the dataset.

The algorithm results can be improved by using much

more scales than the 4 used in the current implementation. Also the usage of a different classifier like a nonlinear SVM or a random forest might offer better results. Also, utilizing more than 8 orientations and a bigger cell size can be useful as the current HOG descriptor could be capturing mainly contrast noise that does not offer a discernible face description from the rest of the image.

5. Conclusions

the histogram of oriented gradients corresponds to shape descriptor based on the gradient orientations, which can be translated into an edge and contrast descriptor. This characteristics of an object, should produce, ideally, a characteristic and unique histogram for an object class. However, in the current implementation, it does not represent a robust descriptor.

The multiscale HOG methodology for detection, in which the HOG descriptor is calculated at different resolutions and the resulting windows are given to train a classifier offers very bad results with a 0.0% as maximal F measure. This can be explained due to the face variability due to pose, the infinite number of scales at with faces appear in the images and a noisy HOG model trained from non homogeneous and non aligned faces of the images. The classifier has an enormous number of negative window examples while the positive windows are minimum, resulting in a very difficult tasks for any classifier.

Finally, the methodology can be improved by means of using an increased number of scales, a different classifier, and bigger cell size with more orientations and by increasing the number of positives including an equal number of positive windows (face crops) to train the classifier. Despite these modifications will increase the computational time needed to train the model any result above 0.0% can be considered an improvement. For the face detection task, any other descriptor is recommended.

References

[1] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan. Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction. In *2003 Conference on Computer Vision and Pattern Recognition Workshop*, pages 53–53. IEEE, 6 2003.

[2] R. Radke, S. Andra, O. Al-Kofahi, and B. Roysam. Image change detection algorithms: a systematic survey. *IEEE Transactions on Image Processing*, 14(3):294–307, 3 2005.

[3] P. Viola and M. J. Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

6. Code and Images

The code is available at the team's repository at *https://github.com/steff456/IBIO4680/tree/master/09-HOG*