Face detection with multiscale HOG

Julio Nicolás Reyes Torres
Universidad de los Andes
jn.reyes10@uniandes.edu.co

Juan David Triana
Universidad de los Andes
jd.triana@uniandes.edu.co

Abstract

This article presents the implementation of a multiscale sliding model for face detection. The algorithm is implemented according to the method presented by "Machine Learning and Vision group" from Texas University. The main idea is to use a HoG (Histogram of Gradients) representation to extract the features of each image, both faces and non-faces, and a SVM (support vector machine) to classify them. The way to detect the faces is running a sliding window that compares the characteristics in different scales of the image. Finally, face detection algorithm is evaluated using a ROC curve, precision-recall curve, and average precision score, some examples of face detections are showed to subjectively analyze the accuracy of the method.

1. Introduction

Face detection is a challenging problem but also a well studied one. The aim is to detect a face independently of the position, illumination, and the features of each people as glasses, hats, skin color. Different groups in computer vision have worked on this approach, there are some relevant researches that have improved significantly detection results and the processing time. For instance, Viola-Jones (2001) which developed a frontal face detection system and was distinguished because of its ability to detect faces extremely rapidly (approximately 15 faster than any previous related work) [4].

This article presents the implementation and evaluation of "Sliding window detector" of Dalal and Triggs (2005) [2]. The goal of this detector was to made a robust feature set of human detection even in difficult case with cluttered backgrounds and bad illumination. They demonstrated that normalized "Histogram of Oriented Gradient (HOG)" descriptors provide an excellent performance to extract human features. The way to classify both a face and a non-face is using a linear SVM. Finally, the method implement a sliding window which slices across the image which is changing between different scales [3].

2. Methods

2.1. Data description

In this practice we implemented the famous "Caltech Web Faces data set" [1], which is divided into three main folders and one extra set. It contains "positive" images that represent cropped faces, "negative" images that contain non face images and a test set for evaluation. Data set implement images with a very wide range of conditions including: illumination, scale, pose, and cam- era variation

2.2. Multi-Scale HOG Strategy

Histogram of oriented gradients is a very well-known descriptor algorithm to extract features of images. The work was thought for study a better way to describe human features, they found that descriptors called "Histogram of Oriented Gradient (HOG)" provide an excellent performance. This proposed method is a brief of edge orientation histograms, SIFT descriptors and shape contexts, besides it is computed on a dense grid of uniformly spaced cells, also to improve performance is used overlapping local contrast normalizations [2]. The features are then stored in a histogram form using different weights depending on each dimension's value, this procedure is implemented on every training image, which are cropped faces and "negative" images, to identify the features of the object and the background (non-object). Afterwards, a linear Support Vector Machine (SVM) was implemented to train the algorithm based on these descriptors. This can be very useful for detection in bigger images, since there can be subsamples per test image to be compared against the SVM model. But this has a complication, because faces may vary in size on the test images. This is where a scale variation must be implemented, by using a Gaussian pyramid on each of the test images. For our detection algorithm we resized every image onto different scales so the sliding window didn't have to be resized.

2.3. Main Strategy Parameters

- HoG cell size: size of each cell on the different images when applying HoG algorithm
- Sliding window step: number of pixels from which the sliding window jumps for detection
- Number of "negative" faces
- Image scales: images resize in the multiscale detection.

3. Results

Learned HoG detector is shown in figure (1), it represents the mask that is compared among all the images to detect the different potential faces.

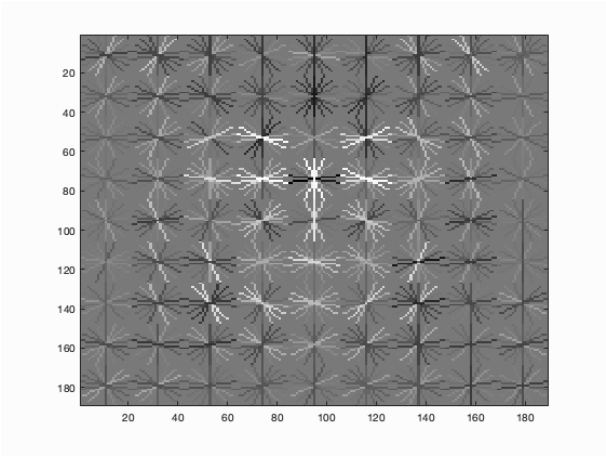


Figure 1: **HoG descriptor**, the mask to be compare among test dataset to find the human faces.

3.1. Strategy data

The most sensitive parameters that represents notable changes in detecting faces were varied to find the best combination of them. Those parameters are presented below, they were chosen changing one by one each parameter without vary the others, and so one each best was determined.

- Lambda: It corresponds to the number of iterations in SVM classifier.
- Step: The jump among pixels that HoG window slides.
- Threshold: The score limit to accept possible faces.

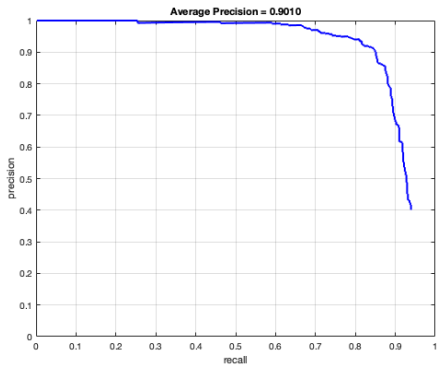
In the Table (1) are presented the more representative proved cases to analyze and in highly the one which present the best result.

Case	Parameters			Average Precision
	Lambda	**Threshold**	**Step**	
1	0.001	0.9	1	0.8968
2	0.001	0.9	3	0.6202
3	0.001	0.9	6	0.3093
4	0.001	0.85	1	0.8985
5	**0.001**	**0.8**	**1**	**0.9010**
6	0.001	0.7	1	0.8981
7	0.0001	0.9	1	0.8901
8	0.0001	0.85	1	0.8928
9	0.0001	0.93	1	0.8901

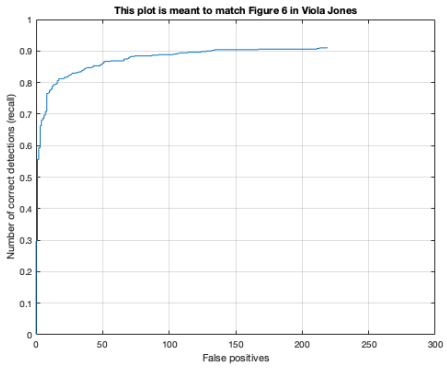
Table 1: **Representative parameters:** Lambda, Threshold and Step are the most representative parameters. Average precision (orange) is the score that represent the best accuracy combination of parameters for human face detection.

3.2. Precision Results

Following are presented the best "precision-recall curve" (2.(a)), and "ROC curve" (2.(b)) that represent the behavior of the implemented algorithm in face detection.



(a) **Precision-Recall curve**, the best result for case 5 presented in Table (1). Average precision: 0,9010.



(b) **ROC curve,** Recall vs False positives.

Figure 2: Precision-Recall curve (a) and ROC curve (b). Results for the best condition (case 5).

(a) (b) (c) (d) (e) (f) (g) (h)

Figure 3: **Faces detected by HoG.** Image (a): Colombian team 1993 (9/11 face detected). Image (b): Mickey mouse painting a human face (2/2). Image (c): A man sitting, behind him a face painted (2/2). Image (d): Adam's Family (7/7). Image (e): German soccer team (11/11). Image (f): Woman face (1/1). Image (g): Monalisa (1/1). Image (h): Friends (6/7).

The method was also proved with a set of extra images, which are familiar images without ground truth. In figure (4) is showed the accuracy of face detection in those images.

(a) (b)

Figure 4: Algorithm proved in the extra test set images. Image (a), 4/5 faces detected. Image (b), 7/7 faces detected.

3.3. Viola-Jones algorithm comparison

As it was explained before, Viola-Jones algorithm was developed to detect frontal faces extremely rapidly in real-time. Our implementation is compared against this algorithm to analyze who works best. Figure (5) shows the ROC curve result of Viola-Jones method.

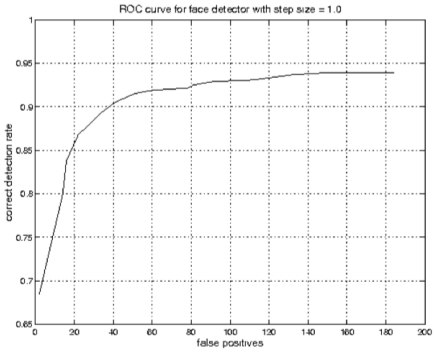


Figure 5: ROC curve for Viola-Jones method.

4. Discussion

Table (1) consolidate the most representative variations done over parameters. The information extracted is very usefull, for example, the parameter "step" is the one which present the higher change in the Average precision score, the more the step was increased the more bad was it. That is, with step = 1 better results are obtained, and it's logical because the sliding window is going to detect more pixels to determine the possible number of faces.

Lambda parameter didn't show a significant variation between the two values analyzed (0,001 and 0,0001), instead Threshold was determinant to identify finely the best Average precision score. A small value of threshold leads to accept a huge number of possible faces but a lot of non-faces are marked too. Otherwise, a high number of threshold is going to limit and be restrictive with the number of possible faces. That is, the suitable range is around 0,8 - 0,93 . Different combinations among parameters were analyzed, the best one was the case 5 (Lamda=0,001, Threshold=0,8 and Step=1) that presented an Average precision score equal to **0,9010**.

The results that directly show the face detector are presented in Figure (3). As it can appreciate, there are 3 colors of squares, yellow ones represent the ground Truth, the greens are the well face detected by the algorithm (true positives) and the reds are wrong faces detected (false positives). The chosen images have different cases of illumination, occlusion, scale, in general, those are appropriate to analyze the behavior of detector in difficult conditions.

Figures (3.f) and (3.g) are "easy" to detect, in both cases the face is well defined with good conditions of illumination, scale, shades, position, in general, the detector works very well. Figures as (3.a), .d, .e, .h, shows faces well defined but with some problems in illumination. Specifically, in figure (3.a) the algorithm never could identify the two players at the edges of the image, because of the shades. Finally, figures (3.b) and (3.c) present two curious situations, the first one present a Mickey Mouse drawing a man face, the algorithm didn't recognize the mouse face to identify because it shape doesn't correspond to a human. In this image in fact there are two copies of the man face in different scale, the algorithm was able to identify both. In the other image (3.c) there are two representations of human face, a real man sit and a draw in the board corresponding to a human face. Again, the algorithm worked appropriately to find the human faces.

The algorithm proved with extra-images which are different without ground truth (figure 4) was able to identify all the faces, however, it show other zones that don't correspond to faces. Anyways, it had the ability to determine the faces in all the set of extra-images proved. This also shows one of the weaknesses of the algorithm. Figures like shoes, and pant silhouettes were identified as faces. This means that there is a very high chance that a vast variety of objects can be similar to that of a face. This means that our algorithm should include a skin detector regarding color and not only shape. This way, the algorithm can exclude other objects that have a similar HOG representation. Obviously this needs to be applied on regular images (no color variation).

Moreover, the non-face set needs to implement a more broad amount of images that are face-like, in a way that the algorithm can identify more objects as non-faces. This is the case of figure 4 b) where the algorithm detected a pair of sandals as a face. In this manner one can expect that there are certain patterns placed upon some false face detections (false positives). More specifically the shape of the object is way more representative than its colour but with no details. This means that the detector is not recognizing specific face attributes, such as eyes, noses, mouths and so on. On the other hand, false negatives (faces not detected) also have some similar patterns. As it was said before, some specific illumination, occlusion and more, aspects affect the recognition of the images, since there is low similarity with the descriptors found and its interpretation on the linear SVM results.

5. Conclusions

- A correct adjustment of parameters are fundamental to choose the best condition for the detector. The best value to Step is setting it in 1, because despite that it waste more time, the results are really affected with values higher that 1. According with our observation with different proves, Lambda does not represent determinant as long as it be small $\approx (0.001 - 0.0001)$. The most important parameter to adjust finely the robust of the detector is the threshold, which is the one that limit the quantity of "possible faces" detected by the algorithm.

- It could be demonstrated that "Sliding window detector" of Dalal and Triggs (2005) is a robust detector that implements an excellent performance descriptor to human features (Histogram of Oriented Gradient (HOG)). The method is able to detect the specific features even in difficult cases with cluttered backgrounds and bad illumination. In figure (3) different images were proved and in general Independent of each one, the method was able to detect almost all the objectives.

- Images (2) and (5) show the result of ROC curve respec-

tively, Viola-Jones method show a higher "correct detector rate" than "Sliding window detector", however it is important to recognized that multiscaling was defined in a set of ten different scales. It could be improved with a higher number and also changing the HoG window.

6. Credits and Extra Credits

Matlab computation was based on the work of gary30404 on his public github repository.

Waldo is found on image 13InterviewInterviewOnLocation13558 of the Waldo dataset.

References

[1] Caltech Vision Lab. Caltech 10,000 Web Faces.

[2] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. Technical report, 2005.

[3] A. Rosebrock. Sliding Windows for Object Detection with Python and OpenCV, 2015.

[4] P. Viola and M. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. Technical report, 2001.