09: HOG Face Detection

Sergio Steven Leal Cuellar* and Mateo Rueda Molano†,
Department of Biomedical Engineering, University of the Andes, Bogota, Colombia
Email: *ss.leal10@uniandes.edu.co, †ms.rueda10@uniandes.edu.co,

***Abstract*—The Caltech Web Faces dataset was used to train a face detection algorithm using a SVM model classifier and HOG descriptors at several scales. In addition, a Viola-Jones face detection algorithm was developed in order to compare with the results of our algorithm. The results obtained show that the algorithm that was developed presents similar performance to the Viola-Jones algorithm with a slightly improvement in the precision obtained in contrast to a worse recall. The unsatisfactory result obtained in the recall might be caused by limitations of the algorithm itself. In conclusion, procedural classification process can be done in order to improve the precision by diminish the amount of false positive detections and the usage of other features to detect faces might increase the number of them that are being recovered thus improving the recall of the algorithm.**

I. Introduction

Face detection is a very important task to recognize a person by using a computer algorithms. Actually, many approaches have been developed to make this detection task more easy but in real world scenario it is a complex task due to background, variations in scale, pose, color, illumination and among others. Because of its popularity many applications use it such as surveillance systems, digital camera, access control, human-computer interaction, social media and so on [1]. Face detection is the step stone to all facial analysis algorithms, including face alignment, face modeling, face relighting, face recognition, face verification/authentication, head pose tracking, facial expression tracking/recognition, gender/age recognition, and many many more [2].

Face detection consists of: given an arbitrary image, the goal is to determine whether or not there are any faces in the image and, if present, return the image location and extent of each face [3]. There have been hundreds of reported approaches to face detection. Early Works (before year 2000) had been nicely surveyed in [3] [4]. For instance, Yang et al. [3] grouped the various methods into four categories: knowledge-based methods, feature invariant approaches, template matching methods, and appearance-based methods.

This article will emphasize on feature invariant approaches which are very popular and used [2]. Dalal and Triggs [5] proposed a called histogram of oriented gradients (HoG), which became a very popular feature for human/pedestrian detection. In [6], the authors proposed spectral histogram features, which adopts a broader set of filters before collecting the histogram features, including gradient filters, Laplacian of Gaussian filters and Gabor filters. The histogram features in [6] were based on the whole testing window rather than local regions, and support vector machines (SVMs) were used for classification.

In the spectral histogram representation (HOG), local features of an image are captured through filtering as the responses of individual filters depend on local structures and the global structures are implicitly captured by the constraints imposed by the histograms of different filtered images. The representation is non-parametric in nature and is effective to characterize different kinds of patterns. One distinctive advantage of the spectral histogram representation is that two images do not need to be aligned in order to be compared due to that the spectral histogram representation is not sensitive to perturbations of local image features. To specify a spectral histogram representation, one needs to choose a set of filters and in this lab we use filters from vl_hog library [6]. Because there are a large number of face sizes, it is necessary to train with different scales in the patterns, so a multi-scale HOG is usually proposed, in which a sliding window and scaling of the test image are obtained to get the patterns different sizes of faces. Finally, the patterns are classified through Support Vector Machine with a chi-square kernel.

With respect to the previous methods, in this article we evaluated face detection based on HOG multi-scaled feature. Our method was compared with the Viola-Jones algorithm to observe the power and performance of the attributes of HOG for this problem. The database used was Caltech Web Faces and the performance of both algorithms was evaluated by the ROC curve and the area under the curve (AP).

II. Materials and Methods

A. Dataset description

The Caltech Web Faces dataset is based on several images of human faces taken from google image search. The original dataset was made up from 10,524 images, the images that are going to be used for the training of the algorithm are just 6713 images taken from the original dataset. Each of those training images are in gray scale and have a size of 36x36 since these images are the positive features from faces. On the other hand, the negative feature images are 275 RGB images with multiple sizes none of them are objects related with human faces. The train set is conformed by 130 gray scale images with multiple sizes. In addition, 7 images taken from the web were used to evaluate the performance of the algorithm trying to avoid the bias introduced by the dataset. Some of those images are photos that are not upright and forward facing and others are cartoon representation of human faces.

B. Methods description

In order to the develop a efficient face detection algorithm a HOG multi-scale strategy was adopted. Our algorithm started

by extracting the HOG features (using vl_hog from vlfeat library) for positive face instances and negative face instances (any other thing that does not have a face) in the train set. In the images without faces (negative instances) to obtain the HOG attributes, a random number of windows of the same size of the positive cropped faces positive was selected. This with the aim of training the SVM in train with attributes of the same size. In the test set, the image was scaled top-down by varying the sliding window and evaluating the training model for each window, obtaining the bounding box for each face in positive windows. In order to avoid multiple detections over a same face, a non-maximum suppression was done. This strategy helps to conserve the bounding box with the higher confidence in a set of overlapping detections.

Some relevant parameters on the implemented multi-scale HOG detector are the scales of the sliding window, threshold and lambda of the SVM classification, the number of negative samples and the cell size. The first of them, scales for the sliding window, determines how bigger or smaller are the new windows that are going to be run on the detection step, this parameters determines the size of the faces that are going to be find (e.g. faces on portrait images). The second parameter is the threshold which establishes the minimum value of confidence of in the SVM samples in order to determine whether a window can be classified or not as a face one, this parameter modifies how restrictive the classification will be. Third, a higher number of negative samples provides the SVM model with more samples of windows that are not similar to faces (in a HOG feature). Lambda is defined as lambda = 1 / (C * (numPos + numNeg)), C corresponds to the confidence of the classifier and numPos / Neg corresponds to the number of positive and negative attributes. This value was left fixed at 0.001 on the recommendation of the main author of this method James Hays in the Pascal VOC toolkit. Finally, the cell size parameter was set at low values found in the literature of [7].

The algorithm was evaluated using the precision-recall curve which relates the number of true positive successful detections, noise detections and non-detected faces. In general, it provides relevant information to determine how well the algorithm works depending on how many faces are being detected, how many are not being detected and how many misdetections are being done. The metric used to measure the precision-recall curve was the average precision (AP) or the area below the curve.

III. Results

A. Caltech Web Faces

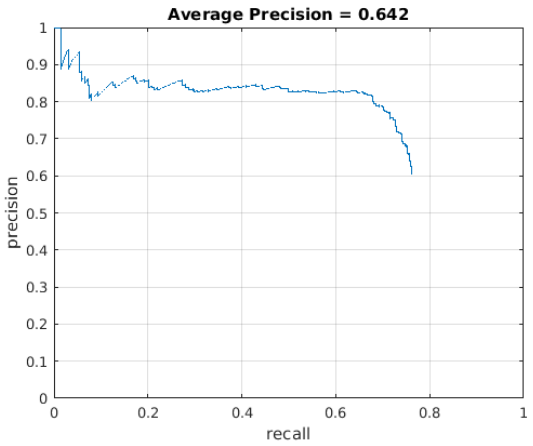


Figure 1. Average precision obtained on the Caltech Web Faces dataset using our algorithm with a maximum scale of 1 and a threshold of 0,9.

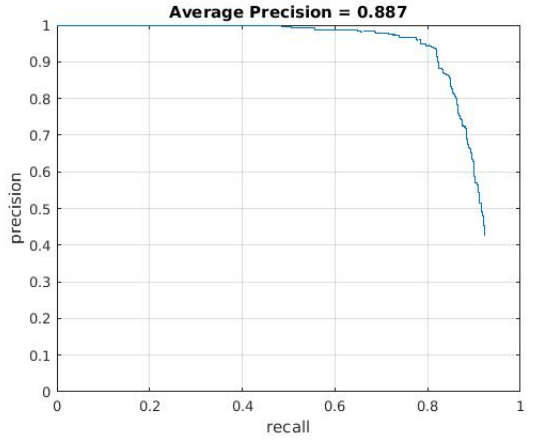


Figure 2. Average precision obtained on the Caltech Web Faces dataset using our algorithm with a maximum scale of 2 and a threshold of 0,9.

Figure 3. Detection obtained on the Caltech Web Faces dataset using our algorithm with a maximum scale of 2 and a threshold of 0,9.

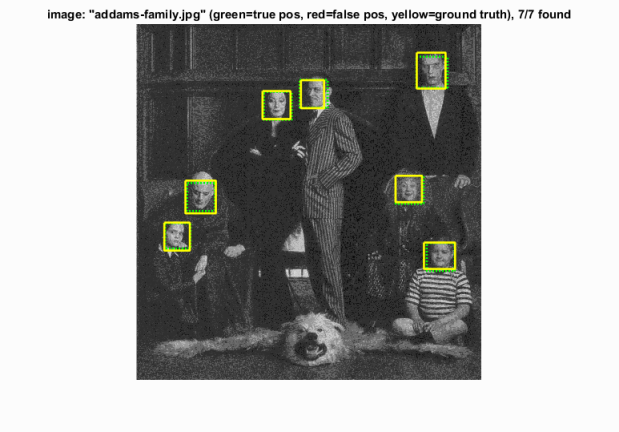


Figure 4. Detection obtained on the Caltech Web Faces dataset using our algorithm with a maximum scale of 2 and a threshold of 0,9.

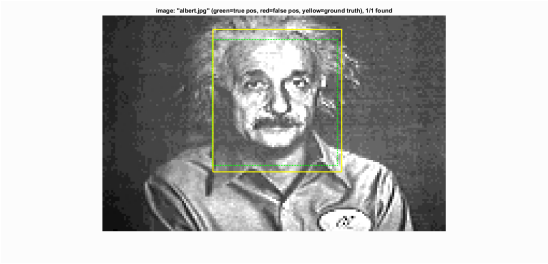


Figure 5. Detection obtained on the Caltech Web Faces dataset using our algorithm with a maximum scale of 2 and a threshold of 0,9.

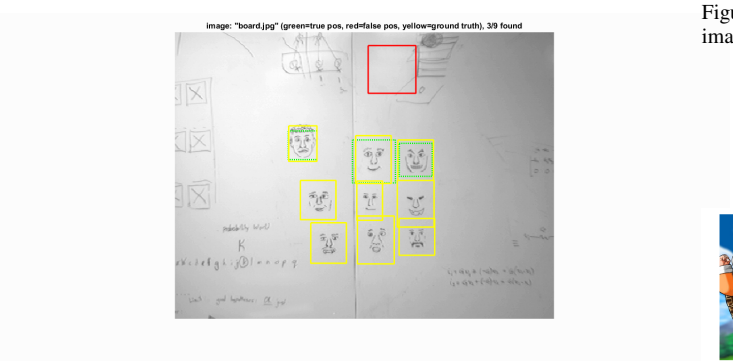


Figure 6. Detection obtained on the Caltech Web Faces dataset using our algorithm with a maximum scale of 2 and a threshold of 0,9.

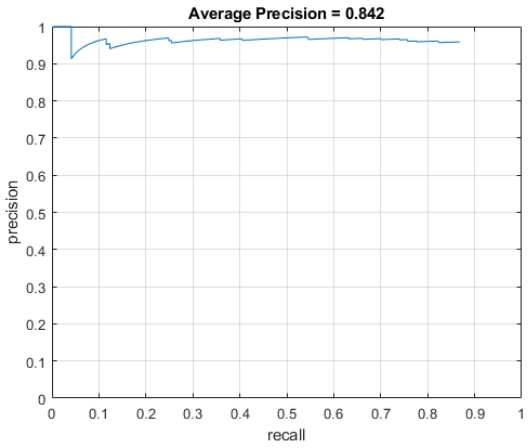
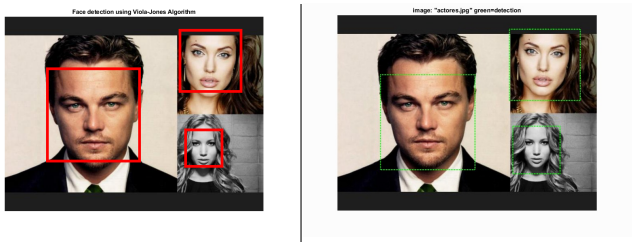


Figure 7. Average precision obtained on the Caltech Web Faces dataset using the Viola-Jones Algorithm.

B. *Extra scene images*

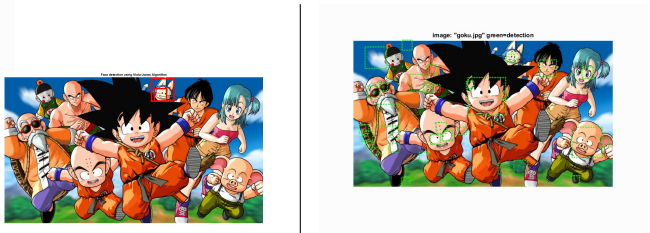


(a) Viola-Jones method (b) Our method

Figure 8. Comparison between our method and Viola-Jones on a single-face image.

(a) Viola-Jones method (b) Our method

Figure 9. Comparison between our method and Viola-Jones on a multiple-face image.



(a) Viola-Jones method (b) Our method

Figure 10. Comparison between our method and Viola-Jones on a cartoon image.

(a) Viola-Jones method (b) Our method

Figure 11. Comparison between our method and Viola-Jones on a big-face image.

IV. Discussion

The overall results of our multiscale HOG method were quite good in the Caltech database even better than the Viola-Jones method in this database. As we see the detections in several types of images are accurate and have few false positives with respect to the groundtruth. The algorithm detects faces at different scales (fig 3 vs fig 5), detects images with noise 5 and even animated faces and some drawings with certain characteristics similar to a face 6. However, there are also artifact problems, objects that have face form and affect the result of the detections. In the extra scene images the task is more challenging and presents more variability than the images of Caltech. Extra scene images have different positions and sizes of the faces so it is not detected as accurately as in the Caltech database.

As can be seen in figure in 1 the parameters set gave an average precision of 0,642 which is for a maximum scale of 1 (same size of the window) and a threshold of 0.9. In comparison, in figure 2 can be seen that considering a bigger scale for the sliding window results in a considerable improvement on the recall. This improvement can be explained because initially the algorithm was not considering enough scales (sizes) of the sliding window in order to detect as many faces as possible and as consequence bigger faces were not being detected as can be seen in figure 11. In addition, due to the scale and to the relative shape similarity of small objects, like the little lamp at the top left corner on figure 9 is being detected as a face (same happens with the shoulder of the man due to its rounded shape).

Additionally, as expected, many cartoon faces of extra scenes are presenting many false positives, as seen on figure 10, due to the patterns on their clothes. This phenomenon may suggest that the threshold that is being used is not restrictive enough and many noise objects are being mis-classified by the SVM. However, it is necessary to verify that in the train set there are images of cartoons because without them it is not possible to classify them in an optimal way Additionally, some other faces are not being detected, this can be caused because of the face orientation, i.e. upright and forward facing.

Comparing the general performance of both methods average size faces, upright and forward facing are being detected satisfactory by our method and it's performance is slightly better (in terms of precision) to the method of Viola-Jones, as seen in figure 8 and figure 7. Viola-Jones method tends to fail with small sized faces, a failure that is compensated with it's computation time and higher recall. Our algorithm lacks in it's general recall as seen in figures 1 and 2, in which the precision value tends to be higher than the recall obtained. This means that our algorithm detects on average accurately but fails to detect some objects. That can happen because there are faces in different orientation, very blurred faces and faces that are actually drawings as in fig 6. However, accuracy is not perfect and artifacts are generated on objects with a similar shape to a face. In the extra scene images the accuracy tends to decrease due to the greater variability in the images and false positives increase. In that order of ideas, increasing the threshold will diminish the recall and increase the precision, which is not satisfactory at all because the algorithm will become specialized in detecting pretty simple faces (rarely will fail detecting upright and forward facing medium sized faces) and really bad detecting the strange faces.

The Viola-Jones algorithm highlights its fast processing and considerable good results in face detection. This algorithm is based on extracting "integral image" features and implementing a "cascade" of classifiers which allows noise zones of the image to be easily omitted and focusing in regions similar to faces. There are several differences between Viola-Jones and our algorithm, one of them is the way the generate features from every pixel "integral pixel" in which every pixel is equal to the sum of the values of the other pixels to the left and above. With that sense the sum of all the pixels inside a rectangle is given by the sum of its corners, this makes the computation a lot less expensive which makes this algorithm that fast. In contrast, our algorithm get features for every positive and negative window and as a result the detection process is slower than the Viola-Jones method. Then, in Viola-Jones method, from small pictures several features can be extracted due to the Haar features (based on convolution kernels) extraction that is done and AdaBoost process allows to get just the relevant features (the best features over the other). On the other hand, in our method single features are being calculated for every sample which restricts the amount of information in relation to Viola-Jones method. Finally, the classifier with several stages "cascade" processively discards the windows that are not faces and let the ones that probably are faces pass and let the subsequent classifier to determine if it's a face or not. In comparison, our method does not have such a step-by-step classification refining process, in our method the model classification is just being discriminated using a threshold in order to diminish false-positives detections but directly affecting the recall of the method.

V. Conclusions

The results presented determine that our algorithm is having an overall good performance when more scales are considered. Nevertheless, even if the threshold is increased the recall will not increase significantly due to the limitations of the algorithm and the lack of variance in the train set. The lack of variance causes that the extra scene images were not detected precisely because they have more orientations, shapes, artifacts like lenses or hats, etc.. In order to overcome those limitations a region proposal algorithm might be helpful

in order to establish face-like probabilities to regions on the image. Having such a region proposal method some detection refinement processes, like the cascade classifier done by Viola and Jones, may help to diminish the amount of false positive detections that are being done by our algorithm. In addition, considering more features to identify face regions, like SITF descriptors or texture, may help to increase the overall recall due to the amount of information that is being used to detect faces. Finally, it is needed a bigger and updated dataset to detect extra scene images. With Caltech Faces the detection of these images are not optimal and they are of less quality than the detections of the Caltech test set.

REFERENCES

[1] P. Viola and M. Jones, "Robust real-time face detection," *International journal of computer vision*, p. arXiv:137.154, 2004.

[2] L. Cerna and D. Cámara-Chávez, G.and Menotti, "Detection: Histogram of Oriented Gradients and Bag of Feature Method," *Computer Science Department, Federal University of Ouro Preto*, p. arXiv:36.44, 2013.

[3] M. Yang, D. Kriegman, and N. Ahuja, "Detecting Faces in Images: A Survey," *IEEE Transactions on pattern analysis and machine intelligence*, p. arXiv:0162.8828, 2002.

[4] E. Hjelmas and B. Kee, "Face Detection: A Survey," *Computer Vision and Image Understanding*, p. arXiv:236.274, 2001.

[5] N. Dalal and B. Triggs, "Histogram of oriented gradients for human detection. ," *Proc. of CVPR*.

[6] C. Waring and X. Liu, "Face detection using spectral histograms and SVMs," *IEEE Trans. on Systems, Man, and Cybernetics – Part B: Cybernetics*, p. arXiv:467.476, 2005.

[7] L. Yongmin, G. Shaogang, J. Sherrah, and H. Liddell, "Support vector machine based multi-view face detection and recognition," *Image and Vision Computing*, p. arXiv:413.427, 2004.