Laboratory 9: HOG Detection

Ramón Emiliani
Universidad de los Andes
201125694
rd.emiliani689@uniandes.edu.co

Alejandro Posada
Universidad de los Andes
201227104
a.posada10@uniandes.edu.co

Abstract

Face detection is a widely studied problem in computer vision. Many approaches have been proposed: the Viola-Jones detector, keypoints-based detectors, local invariant descriptors, bag-of-visual-words models, deformable parts models, exemplar models, etc. In this laboratory, we used a multi-scale HOG with a SVM classifier to detect faces on the WIDER FACE dataset. Our results underperformed the face detection baselines: we obtained 0.002, 0.001 and 0.000 at the easy, medium and hard levels of difficulty on the WIDER FACE dataset.

1. Introduction

A popular computer vision task is object category detection, that consists in localizing and identifying objects of a given type in images. This task includes a wide variety of examples, such as detecting pedestrians, cars, animals or shapes of interest in medical images. In order to detect these shapes, it is useful to use a feature set that is robust even under difficult ilumination and cluttered backgrounds. Many feature descriptors have been proposed ([10][9][11][2][8]) to generalize objects in such a way that the object produces a similar feature descriptor when viewed under different conditions.

One ot the must succesfull person detectors are Histograms of Oriented Gradients (HOG). The basic idea of this method is to evaluate local normalized histograms of image gradient orientations in a dense grid. This is done because local object appearance and shape can be described with the distribution of local intensity gradients [1]. The algorithm works by dividing the image into cells and by accumulating a local 1-D histogram of gradient orientations (usually, a 9-bin histogram) over the pixels of the cell. Then, the image representation is formed by combining the histogram entries. Additionally, in order to obtain better invariance to illumination, it is useful to contrast-normalize the local responses. This can be done by accumulating a measure of histogram energy over larger blocks (2×2 blocks with 50% overlap) and by normalizing the cells in the block. The common detection chain consists in tiling a detection window with a dense grid of HOG descriptors and then using a SVM to classify the combined feature vector. Taking into account that objects exist in images at different sizes from the one of the learned template, it is useful to consider different scales. In order to find objects of all sizes, the image can be scaled up and down before searching for the object.

In this laboratory, we approached the problem of detecting faces at different scales. In order to do this, we used a multi-scale HOG detector to detect faces in a small subset of the WIDER FACE dataset [13].

2. Materials and methods

The dataset we used is WIDER FACE, which is composed by 32,203 images with 393,703 labeled faces with a high degree of variability in scale, pose and occlusion. The dataset is organized based on 61 event classes such as ”soccer”, ”parade” and ”picnic”. Figure 1 shows an example of the database’s images. Additionally, in order to improve the enhance the training stage, we added training images from the Labeled Faces in the Wild (LFW) dataset [6]. In particular, we added the subset of deep funneled images, which are alligned images that produce better results for most face verification algorithms [5].

Figure 1: Examples of the images found in the dataset.

Our approach is based on the object category detection code by the Oxford’s Visual Geomtry Group. In order to train the model, we used as positive examples image crops of face occurrences. In total, we used 25,475 positive training images (12,242 crops from the WIDER FACE dataset and 13,233 images from the LFW dataset). In order to collect negative examples, we used 90 images from Caltech101.

HOG was computed with the VLFeat function *vl_hog*. One HOG array was extracted for each extracted image and then these were concatenated in a 4-D array along the fourth dimension. In order for the detector to be multi-scale, we used 15 scales for each image as the HOG-multi-scale hyperparameter. The scales ranged from 0.125 to 2 times the original size of the image. We trained a SVM classifier with $C = 10$. To detect multiple face instances in a same image, we considered the top 1000 detections per image, used non-maximum supression to eliminate redundant detections and for simplicity preserved only the top 10 detections. Finally, we performed hard-negative mining by finding a key set negative examples and adding them to the original negative examples.

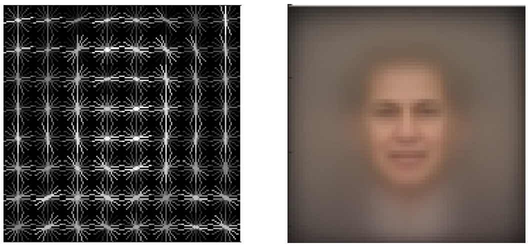
The results were evaluated using the PASCAL VOC criterion by computing Average Precision (AP). In order to evaluate a detection problem, we consider a test image that contains ground truth object instances $(gt_1, \ldots, gt_m)$ and a list $(d_1, s_1), \ldots, (d_n, s_n)$ of candidate detections d_i with score s_i. To convert this data into a list of labels and scores (s_i, y_i) the following algorithm is useful:

1. Assign to each candidate (d_i, s_i) a true (1) or false (-1) label yi by considering a detection as correct if the overlap with the ground truth is greater than 50%.

2. Each ground truth object gt_i that is still unassigned to the list of candidates is added as (gt_j, ∞) with a positive label.

3. Results

Figure 2a shows the HOG model we obtained and figure 2b shows the average of the positive training images.



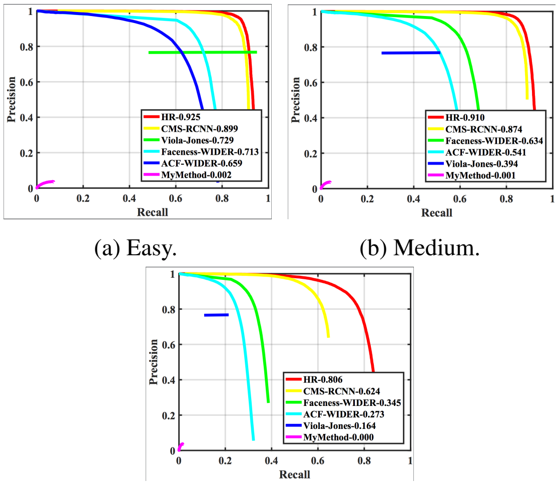
(a)

(b)

Figure 2: a) HOG model obtained after training. b) Average of the positive training examples.

Even though the shape of a face can be perceived in figure 2a, it is important to note that the model presents noise: many of the gradients close to the lateral borders display strong vertical components.

Figure 3 shows the Precision-Recall curves obtained for three levels of difficulty: easy, medium and hard based on the detection rate of EdgeBox [14]. The average recall rates for these three levels are 92%, 76%, and 34%, respectively, with 8,000 proposal per image [13].



(a) Easy.

(b) Medium.

(c) Hard.

Figure 3: Results from the multi-scale HOG detector at three levels of difficulty.

In all cases, our results clearly underperform HR, CMS-RCNN, Faceness, ACF and Viola-Jones. Additionally, our results' recalls are smaller than the average recall rates for the three levels of difficulty [13]. The results that are closest to our method's belong to Viola-Jones, followed by Aggregated Channel Feature (ACF). Indeed, this algorithm is related to HOG since it uses features such as gradient histogram, integral histogram, and color channels that are combined and used to learn boosting classifier with a cascade structure [12]. Viola-Jones method has the advantage of being fast and relatively accurate. The rest of the methods, which are based on deep learning, have significantly superior results since they solve many of the difficulties that our method encounters. For example, HR [4] is very effective at finding small faces while taking into account scale invariance, image resolution, and contextual reasoning.

Figure 4 shows examples of the detections obtained with our algorithm. The negative values are due to the normalization we implemented $\frac{Score-min}{max-min}$ and the fact that we did not iterate over all the set so we did not find the global minimum. Many of the false positives are characterized vi the presence of strong gradients/edges.

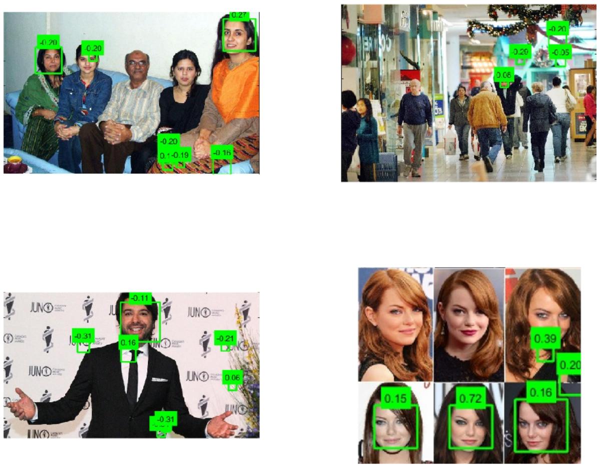
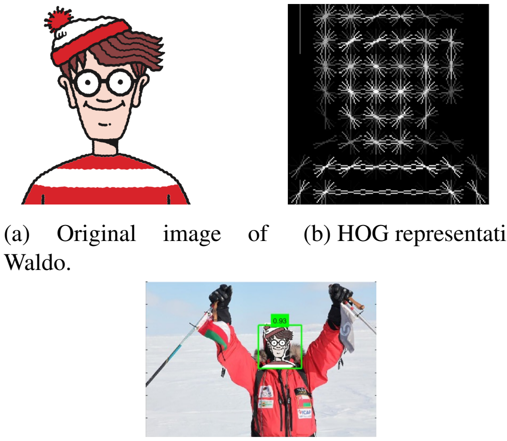


Figure 4: Examples of detections.

The HOG descriptor operates on local cells and is therefore invariant to photometric and geometric transformations. However, it is not invariant to object orientation. In the case of pedestrian detection, the object of interest displays most of the times a similar position: standing upright. However, faces are prone to be tilted and thus HOG's non-invariance to object orientation is a disadvantage. An important disadvantage of HOG in the particular case of face detection is that HOG is color-invariant. Indeed, skin color has been used to detect faces [3][7], so ignoring color-related information is an important downside of using HOG.

As figure 5 shows, HOG can be viewed and used as a variation of temple matching. For example, we used the image of Waldo, obtained its HOG representation and found where he was hiding in the dataset (*13_Interview_Interview_On_Location_13_559.jpg*) by looking at the top detections.



(a) Original image of Waldo.

(b) HOG representation.

(c) Image where Waldo was hidden.

Figure 5

4. Conclusions

HOG is a simple yet effective way to describe image regions. However, the sliding-window detector we implemented showed poor results on the WIDER FACE dataset. The detector could be improved by including more negative examples since we used very few negative training data. Also, incorporating motion information by using optical flow fields or block matching could improve the results. Additionally, faces are variable and can be seen as deformable objects. Thus, using a parts based model could improve the detection results. Finally, instead of using a SVM or an exemplar SVM, the HOG features could be fed into a neural network to perform classification. Following this line of thought, HOG could be computed using neural networks since the HOG computation can be seen as an instance of convolutional neural networks (with some connectivity restrictions and particular kernel functions).

References

[1] 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.

[2] W. T. Freeman and M. Roth. Orientation histograms for hand gesture recognition. In *International workshop on automatic face and gesture recognition*, volume 12, pages 296–301, 1995.

[3] C. Garcia and G. Tziritas. Face detection using quantized skin color regions merging and wavelet packet analysis. *IEEE Transactions on multimedia*, 1(3):264–277, 1999.

[4] P. Hu and D. Ramanan. Finding tiny faces. *arXiv preprint arXiv:1612.04402*, 2016.

[5] G. Huang, M. Mattar, H. Lee, and E. G. Learned-Miller. Learning to align from scratch. In *Advances in Neural Information Processing Systems*, pages 764–772, 2012.

[6] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[7] J. Kovac, P. Peer, and F. Solina. *Human skin color clustering for face detection*, volume 2. IEEE, 2003.

[8] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[9] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. *Computer Vision-ECCV 2004*, pages 69–82, 2004.

[10] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE transactions on pattern analysis and machine intelligence*, 23(4):349–361, 2001.

[11] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2):153–161, 2005.

[12] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Aggregate channel features for multi-view face detection. In *Biometrics (IJCB), 2014 IEEE International Joint Conference on*, pages 1–8. IEEE, 2014.

[13] S. Yang, P. Luo, C.-C. Loy, and X. Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5525–5533, 2016.

[14] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, pages 391–405. Springer, 2014.