Lab 09 - HOG

Jhony A. Mejia
Universidad de Los Andes
Biomedical Engineering Department
ja.mejia12@uniandes.edu.co

Abstract

HOG is a feature descriptor that was used for pedestrians detection. This method showed decent results for this application. This is the motivation for using HOG for other applications, for example, for faces recognition. One of the most used databases for face detection is WIDER FACE, and this database is the one studied in this paper. The process of using HOG for detection is divided in two main phases, HOG extraction and construction of the model. For this two phases scales, cell size and number of iterations for hard negatives are of high importance for producing a good multi-scale HOG model. Fifteen scales, three cell sizes and three different number of iterations. The performance obtained for face detection was extremely poor. For using HOG as a detection algorithm, a lot of pre-processing steps must be taken into account for producing good results. The presence of additional annotations are also useful for using HOG as a detection algorithm.

1. Introduction

HOG (Histogram of Oriented Gradients) is a feature extraction process that can be used for detection problems. Roughly speaking, HOG can be understood as a detailed silhouette/shape descriptor. The first step of HOG is to calculate the gradients of the image (usually done with basic sobel filters). After that, the image is divided in cells of MxN (usually 8x8) and then histograms of gradients are calculated in 9 bins. Each bin represents a gradient orientation (0, 20 and so on until 180), and the number of counts represents the magnitude of a given orientation in the MxN cell. After that a normalization process is done. This process' objective is to make HOG 'invariant' to changes in intensity. The normalization process is done in cells of 16x16 (4 cells of 8x8, each containing its corresponding histogram) which gives as result a representation of 36x1 per cell (of 8x8).

After performing all the previous steps, a overall description of the shapes in the image is observed in HOG. However, this description is highly dependent of the image's resolution. For high resolutions little shapes can be easily identified, while in low resolution images only big images will be seen.

HOG's first application was done in identifying pedestrians. HOG usually works for detection problems because it easily identifies pattern's of shape in a given category. For example, a normal face should fulfill basic patterns as a big ball (head), two little balls (eyes), one triangle (nose) and an ellipse (mouth). HOG work's well with changes of intensity (for example it can identify a black and a white person's face as a face). However, HOG doesn't work well for changes in orientation. For example, a person looking to a side will only have 'one eye', which breaks the patterns of shape previously mentioned. Anyway, HOG has a good performance with little orientation variations in detection problems.

2. Materials and Methods

2.1. Database: WIDER FACE subset

The database used for studying HOG performance as a detector was a subset of *WIDER FACE: A Face Detection Benchmark*. The subset contains all of the original categories, or as they call it 'events'. The total of 'events' are 61, and each event contains a different number of images, with different number of faces per image. The subset of WIDER FACE chosen contains 'relatively easy to detect faces' as their size is larger than 80x80.

The subset contains the original images (both for train and validation sets), and some cropped faces for the train images. The size of the cropped images was similar and had its mean in 135x100. Additionally, an evaluation folder contained codes used for evaluating the proposed method's performance. That evaluation folder also contained ground-truths that could be used for the training process. These ground-truths had the names of the original images, the corresponding bounding boxes for the faces in each image (xmin, xmax, width and height), and some sub-division of the images depending on the difficulty of the face that had

to be detected. Some of the difficulties were classified by occlusion, pose, illumination, blurriness and expression.

2.2. HOG extraction

The HOG extraction process was performed using the algorithms proposed by the Oxford Visual Geometry Group [1]. One of the most important hyper-parameters of HOG is the size of the cell. Three sizes were considered (8 pixels, 4 pixels and 16 pixels). The size of the cell is important for detecting very local patterns or more global information. Another important hyper-parameter is the scaling that will be considered for the multi-scale HOG. 15 sizes were taken into account (beginning in 0.5 and ending in 8 times the original image's size). The scale is important for detecting little faces (in high scales) or big faces (in low scales).

2.3. Model training

The model was trained based on a hard negative mining algorithm, also proposed by the Oxford Visual Geometry Group [1]. Three different number of iterations were considered for obtaining the hard negatives (5, 10 and 20). Each of the iterations contained 1, 2, 5, 10, 20, 50 or 100 train images for obtaining the hard negatives. After obtaining the negatives, an SVM was trained with both positives (cropped faces) and the hard negatives.

2.4. Model evaluation

The model's performance was evaluated with Precision - Recall curves. The algorithm used for evaluating was the one provided by WIDER FACE. The Precision - Recall curve was created taking into account the overlap between the detections predicted and the ground-truth (real bounding boxes). The overlap between boxes had to be more than 50% for being considered a true positive. Also, if the detection with highest confidence corresponded with a ground-truth, the AP remained in (1,1). The AP started to diminish as the highly confident predictions didn't match with a ground-truth.

The model was evaluated in the validation subset. Results shown are only the ones of the 'easy' faces.

3. Results

Re-scaling the cropped faces

The cropped faces had similar dimensions and were roughly squared. However, small variations were detected and those variations could make more difficult the process of finding patterns using HOG. That is why the images were re-scaled considering the statistical indicators of the raw cropped faces.

Dimension	Mean	Mode	Std Dev.
Rows	135	134	28
Cols	100	100	0

Table 1. Original statistical indicators.

As it can be seen in Table 1, the faces were cropped taken into account the number of columns (fixed to 100 pixels). The standard deviation of the number of rows was low. The average number of rows and cols were 135x100. However, these dimensions do not perfectly fit in the HOG cell sizes that were used (8, 4 and 16 pixels per cell). That is why the images were resized to the closest multiple of those cells. All of the cropped faces were resized to 128x96.

3.1. HOG-based Model test in Wider Face validation subset

The HOG model was built taking into account three different sizes of the cells (8, 4 and 16 pixels per cell). The idea of changing the size of the cells is obtaining more local or more global data. Additionally, the number of iterations for finding hard negatives were 5, 10 and 20 iterations. 15 different scales were considered for detecting little or big faces. The Precision-Recall curves for each hyper-parameter can be seen in Fig. 1.

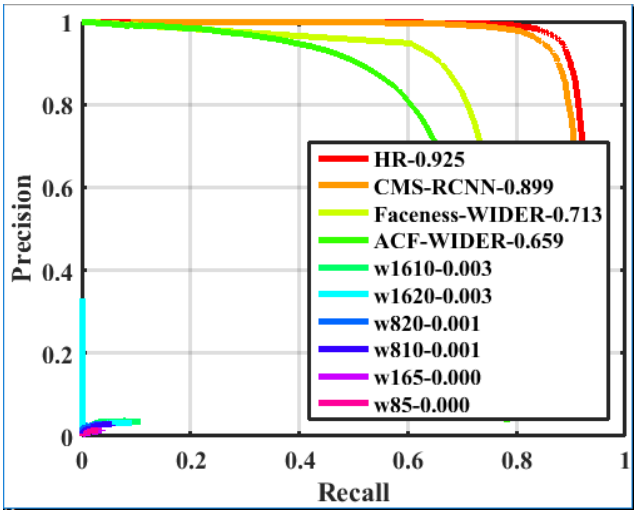

Figure 1. Precision-Recall curves for six of the proposed hyper-parameters values. w1610 (16 cell size and 10 hard negative mining iterations), w1620 (16 cell size and 20 hard negative mining iterations), w820 (8 cell size and 20 hard negative mining iterations), w810 (8 cell size and 10 hard negative mining iterations), w165 (16 cell size and 5 hard negative mining iterations) and w85 (8 cell size and 5 hard negative mining iterations).

The previously presented results were tested on the 'easy' subset of WIDER FACE. The method ran out of

memory for HOG cell size of 4 pixels. A tendency for obtaining higher precision is observed with higher number of iterations for hard negatives. This makes sense, the idea is that the SVM learns what looks like a face but is not a face. This takes as result that the model has less false positives, as the common mistakes are actually identified as mistakes.

The AP obtained was extremely poor (only 0.3%). A possible explanation for this is that the database had badly-aligned faces. This produced that the model developed couldn't find any pattern in the training data, and neither in the validation data. Actually, this is the case, as it can be seen in Fig. 2

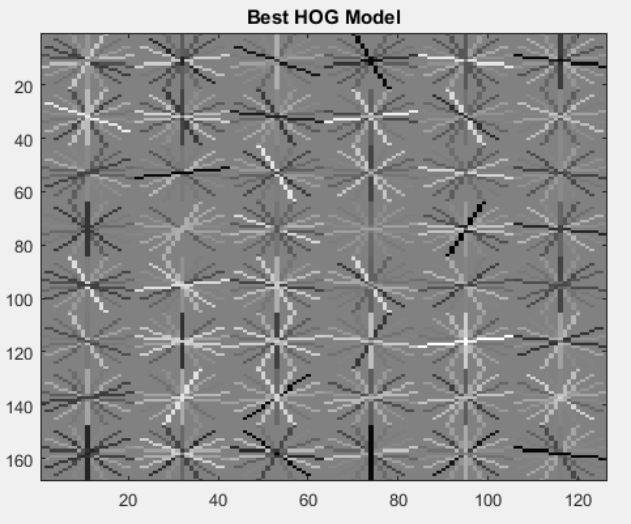

Figure 2. Best HOG Model, obtained with cells of 16 pixels per image and 10 iterations for hard negative mining.

In the image above it is impossible to find any pattern that looks like a face. If an extreme effort is made, a round geometry can be seen in the center of the image, which could represent the head. However, there are no patterns for eyes, nose, ears or mouth. This makes it impossible for the model to be descriptive enough to recognize faces. For comparison, the HOG of a single image is shown in 3.

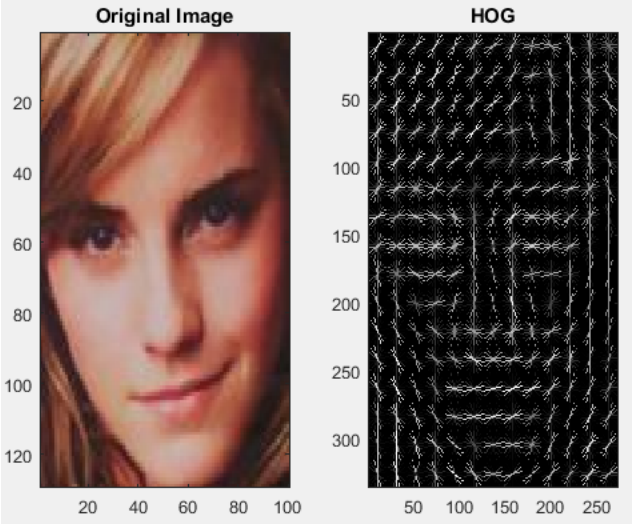

Figure 3. HOG for a single face.

It is way easier to recognize features in that single image. The head is easily recognizable, as well as the eys, nose, mouth and even hair. However, not all the images look like the image shown before. One of the most variated in pose 'events' was 'hand-shaking' (Fig. 4).

Figure 4. Faces from the 'hand-shaking' category.

As it can be seen, the faces change hugely in orientation. Some of them are straight to the cam, others are facing one side, others are facing the other side, some have a tilt on their face. A simple HOG descriptor (the mean of HOGs) was computed for the images shown of the hand-shaking category and its result is shown in Fig. 5.

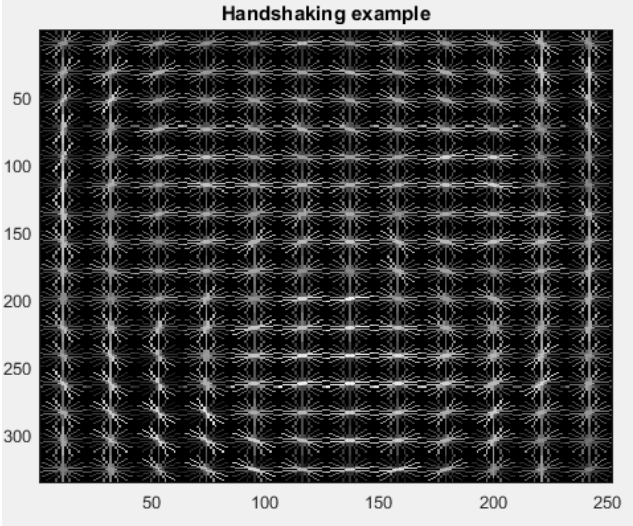

Figure 5. HOG from faces from the 'hand-shaking' category (Fig. 4).

The result obtained is similar to the one obtained in the best model. There are no patterns in the image. If an effort is made, a circle in the mouth and two circles for the eyes can be seen. However, with a larger number of images those details will be lost, and only the shape of the head will be kept as in the model.

3.1.1 Common mistakes

Inferring from the model, the easy to recognize faces will be straight faces with a very round geometry. However, this also leads to the creation of false positives to things that are roughly round.

The faces that are more difficult to find are the ones who are looking aside, as their shape looks more like a rectangle than like a circle. However, the model representation is really poor and this approximations could not be real.

3.1.2 Limitations of the method proposed

HOG is extremely sensible to changes in orientation. The simple fact of a face looking to a side or another changes everything. HOG would work good if all the faces are like the ones taken for identification cards (ID, Visa, etc). In this type or photos the person is looking straight to the cam and is looking serious. HOG could be useful for finding identifying a person as a person on an ID. For other cases, HOG will not work well.

In real life faces are not always looking straight to the cam and with a serious expression. Changes in the orientation of the face and in the internal expression of the face will affect the results obtained by HOG.

3.2. Further improvements

I didn't perform any regrouping process and that was a huge mistake. The process of leaving all the images in the same resolution could have been useful for diminishing the variations of size, but a technique for dealing with face orientation was not considered.

It would be useful to manually select a group of images (30 per orientation) to know if the person is looking to the cam, to the loft or to the right. After that, a model must be built for each orientation. Later, each model should be tested in the rest of train cropped images to see which category does the face belongs to. Finally, all three models should be run in the test images and the higher scores should be kept. This should hugely lower the orientation variety between faces, produce more robust models and detect better in the test images. This would be a less complex approach to voting models of SVMs (Excentered SVMs).

Another possibility is to compute the HOG representation on each image, and then run a clustering algorithm for separating things that look alike. Something like k-means would be useful. After using k-means, each cluster should have its SVM model. this alternative is similar to the previously mentioned.

Both proposals are thought to lower the intra-class variability. However, there will always be variability whether the subclasses are found manually or automatically. That is why these possibilities might not solve the problem completely.

More robust solutions are related to Deformable Parts Models or Poselets. Again, both options' objective is to somehow explain or lower the intra-class variation.

4. Conclusions

HOG is not robust enough to be used as a detector of everyday things. It could be useful for finding IDs, magazines car or things that have low intra class variation.

HOG is extremely sensitive to changes in orientations or poses. With a high variety in the orientations HOG will not be able to find patterns, as in this paper.

For using HOG as a detector it is suggested to perform multiple pre-processing steps for trying to lower the variation of a class. Changes in size of a global image can be decently solved by considering a multi-scale HOG. Changes in orientation are way more difficult to describe.

References

[1] Vedaldi, A., Zisserman, A. *Object category detection practical*. Oxford Visual Geometry Group. Oxford University. 2018.