CS231A Project report

Cecile Foret

March 19, 2014.

Introduction

The goal of this project is to detect and localize humans in static images. This is a challenging problem as people can have a wide variety of poses, distinct appearance and clothing. The background of the scene can be complex as well, and present different kind of illuminations. Human detection in scenes can be useful for applications such as pedestrian detection for smart cars, smart home, visual surveillance.

To address the detection problem, we first review different descriptor methods to define a good abstraction of the human model. Such feature descriptors can be scale invariant feature key points descriptors (SIFT), shape context descriptors, Haar based descriptors, histogram of gradient descriptors (HOG). We investigate in more details the HOG descriptor, and apply this descriptor to this detection problem.

Once the feature descriptor is defined, we use a set of positive and negative training images to train a detection model, using a linear SVM classifier. We verify the model by running it over a set of normalized positive testing images, and verify the detection rate. We also apply the model exhaustively on a set of negative images to find a series of hard cases for false detection and retrain the model with these hard cases.

We then apply the detection method to a set of non normalized images. Since the HOG descriptor is not scale invariant we generate a pyramid of scaled images, and search each scaled image exhaustively, using a sliding window. We also perform a step of non max suppression to remove the noise and get a more accurate detection.

Previous work

Different kinds of feature descriptors have been developed for detection purpose. Lowe's Scale Invariant Feature Transformation (SIFT) approach allows to extract distinctive invariant features from the object we're trying to detect. A database of key-points features can be generated using a set of training images. Detection is done by matching feature key-points in a testing image to a dictionary of key-points extracted from training images for the object that we are trying to detect.

Belongie's Shape Context approach consists of sampling the edges of the object into points of interest, and capturing the distribution of the sampled points on the shape with respect to a given point on the shape. The relationship between a point of the shape and the other points on that shape can be described by a distance and angle measurements. Distances and angles can be binned into different buckets, to generate a histogram. This histogram will capture the relationship between a points p and the other points on the shape.

Matching two shapes is equivalent to finding points on each shape that have similar shape context. This method is working well for text recognition, finger print recognition, but is sensitive to issues such as occlusions.

Another approach proposed by Dalal and Triggs [1] is to use histograms of oriented gradients (HOG) as a template descriptor for human detection. Local object appearance and shape information can be efficiently expressed by the

distribution of local intensity gradients (edges) over the object and immediate surroundings. We will describe this approach in more details.

Overview of the method:

The first phase of the method is to generate an efficient feature descriptor for human body, and generate a series of templates to amount for the differences of human poses, appearances, clothing. The descriptor should present good robustness towards illumination, scale, background clutter that can be seen when detecting human in normal scenes. We will use the HOG descriptor, as presented by Dalal and Triggs [1].

The next step is to train a linear classifier (SVM) and generate a 2 class model (people vs non people) using positive and negative training images. We then construct an classification model, which we use to predict whether or not a specific portion of a testing image contains a human body.

Once the model is trained, we generate a series of scaled images, and run a sliding detection window over each of the scaled images. The descriptor corresponding to the areas of the image covered by the detecting window is generated, and then classified using the model previously generated. This tends to generate a lot of positive match when the window partially overlaps the object of interest. For better accuracy of the detection, we apply a non-max suppression stage after the initial detection phase.

Implementation details:

1) HOG descriptor

The first step of the implementation is the calculation of the HOG descriptor. We use a detection window for the human template of size 64x128 pixels. The detection window is tiled into smaller size cells, typically 8x8 pixels. For each cell, the pixel gradient is calculated, and the orientations of the gradient for each pixels in that cell are then partitioned into a certain amount of bins (number of orientations). A one dimensional histogram of gradient orientations is calculated using the gradient orientations of pixels in that cells. The vote in the histogram for a specific orientation is weighted by the corresponding gradient magnitude. This puts more emphasis in the descriptor on stronger edge information. For humans, the wide range of clothing and background colors make the signs of gradient rather uninformative, and the gradient orientations are partitioned into [0-180] degrees into N orientations. The gradient orientations are bi-linearly interpolated before being binned into the N orientations. We also apply a histogram normalization stage using a larger block size that overlaps multiple cells, for better invariance to contrast and illumination changes.

The final HOG descriptor is the concatenation of all normalized cell histograms inside the detection window. For a detection window of 64x128, with cells of size 8x8, normalizing blocks of size 16x16, number of orientation bins set to 9, we get a final descriptor vector of (8 x 16 x 4 x 9 = 4608) elements. To calculate the HOG descriptor, we use the vl_hog() API of the vl_feat library.

The size of the detection window is directly related to the resolution of the normalized training images in the data set. For our experiment, we used the INRIA people data set. Different parameters of the HOG descriptors were tested, to find the

optimum settings for this detection problem. We find that coarse spatial sampling, fine orientation sampling and block normalization are contributing factors to the efficiency of the detection algorithm.

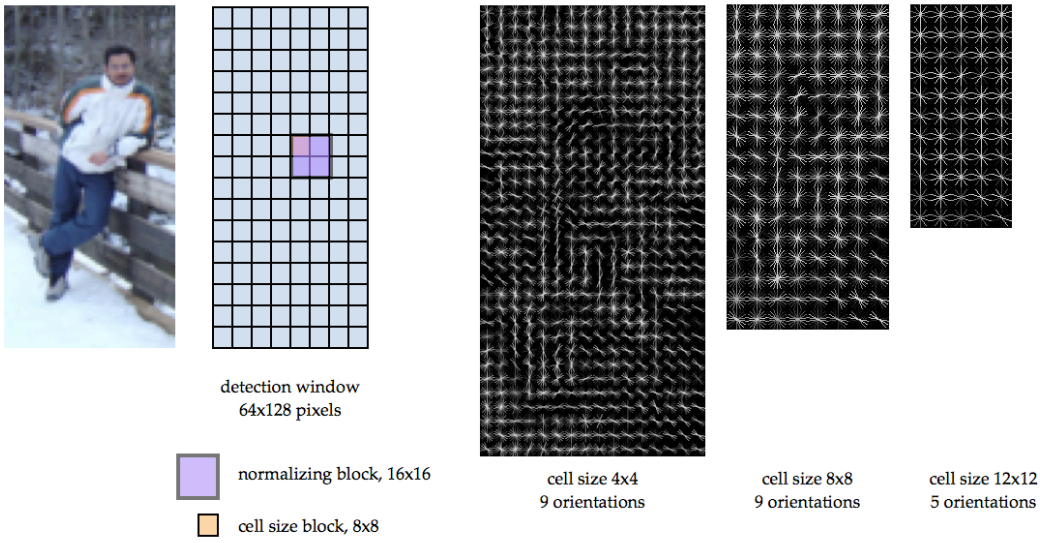

Figure 1: HOG descriptor representation, for different cell sizes, and number of orientation bins.

2) Classification model

The second stage of the method is to generate a classification model. For this, we use a linear SVM classifier, and train it with a set of well normalized positive training images, as well as a set of negative images (Figure 2).

The normalized positive images will present examples of people, all at the same scale, with similar amount of background around the human bodies in these images. For negative training set, we use images that do not contain people, and randomly select areas of these images to generate negative descriptors. The areas selected in the negative images have the same size as the normalized positive training images and generate HOG descriptors of similar size as the positive images.

We use a linear kernel for the SVM classifier, with soft margin C = 0.1, and use a set of 2416 positive descriptors and 12180 negative descriptors. We then test the classifier by running it on a set of normalized testing positive images (also from the INRIA dataset) and by running it exhaustively on the set of negative images. We calculate the miss-rate by counting the number of normalized positive testing images for which the detection was missed, and divide that count by the total number of normalized positive training images. We also calculate the false positive rate by sliding a window exhaustively across all negative testing images, counting the windows that triggered a false detection, and dividing that count by the number of window searches over all negative testing images.

We do this exercise with different variants of HOG descriptors (different cell size, different number of orientations) to look at the impact of the descriptor parameters on the detection (Table 1a, 1b).

We observe that the detection performance is best with cell size of 8x8, which provides the smallest rate for false positive detection. Fine sampling of the gradient orientation is also important, and we see that using 11 bins to characterize the gradient orientation provides the smallest miss rate and false positive rate. Strong block normalization is also an important factor of the efficiency of the detection, although we did not explicitly verify that.

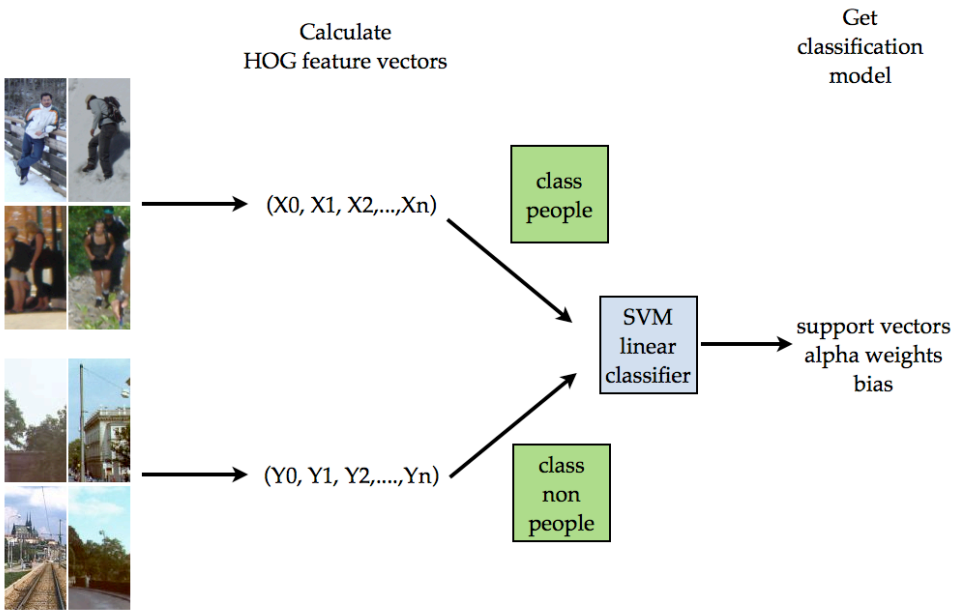

Figure 2: training the detection model using HOG descriptors, using two classes: people vs non people

3) Detection on non-normalized images

We use the non-normalized positive testing images from the INRIA data set to test the detection algorithm. In this non-normalized images, humans can be of various scales, different from that of the normalized training images we used to generate the classification model. The HOG descriptor is not scale invariant, so we need to generate a pyramid of scaled images.

We scale the testing images down from their original resolution by a factor 1.2, and repeat that process until the size of the scaled image reaches the size of the detection window. We then exhaustively search each scaled image in the pyramid with a sliding window of same size as the detection window, over a grid of 8x8 pixels, calculate the HOG

cell size	miss rate	false positive rate
8x8	0.0530	0.0054
12x12	0.0521	0.0063
16x16	0.0530	0.0084

Table 1a: number of orientation bins for the HOG descriptor is set to 9

number of orientation bins	miss rate	false positive rate
5	0.0636	0.0064
7	0.0592	0.0056
9	0.0530	0.0054
11	0.0512	0.0048

Table 1b: cell size for the HOG descriptor is set to 8x8

descriptor on each position of that search grid and classify it into human or non human class using the previously trained SVM classifier.

A couple of observations are made from the preliminary testing over various images: there are some instances of false positive detections over each image. The detector will tend to report a positive match on elements of the image that have strong vertical components, such as columns, poles, trees. We also usually see a cluster of positive matches around human bodies (Figure 3). Human instances that are partially occluded are not consistently detected.

To minimize the amount of false detections, we can refine the initial model by running it over the negative training set exhaustively, retrieve the areas of these images that are generating false positive and use the corresponding HOG descriptors as additional negative descriptors (hard negative test cases). Adding these hard negative descriptors sometime complicates the training of the SVM classifier, and this re-training stage can require additional iterations and a lower soft margin to allow the SVM classifier to converge.

To minimize the cluster effect around positive matches, we calculate the actual prediction weight for each HOG descriptor (instead of getting the binary people/no people classification result) and apply a 2D non max suppression stage on these detection weights, as described in [6]. We partition the 2D image of weights results for the HOG descriptors at to the sampling positions of the sliding window in cells of n * n points. We find the candidate for local maximum by finding the position in each cell with maximum weight value. For each candidate, we then search a [-n, n] x

Figure 3: exhaustive search using sliding window over a pyramid of scaled images, using SVM classifier

[-n, n] window around that candidate, and mark it as a local maximum if it remains a maximum value within that region. The choice of the n parameter has some impact on the non max suppression results, especially in the case where we have multiple distinct humans rather close from each others in the scene (Figure 4). In the case of a single detection, or for multiple very distinct humans, the non max suppression method works relatively well (Figure 5).

Conclusion

We have shown that using HOG descriptors as a template for human detection works well, when used in conjunction with a linear classifier, trained with positive and negative examples. The nature of the descriptor allows detection of people under various illumination conditions, over different kinds of backgrounds, with some degree of occlusion and variation in pose. However, the descriptor is not scale invariant, and detection requires analysis of the image at multiple scales. Finding the correct scale for the detection is a little tricky. The detection also tends to generate false positive on elements with strong vertical components, such as architectural elements, trees, poles. On the other hand, subjects that are occluded for a large part are not very well detected. The approach of using a root template for people detection could be enhanced with a part based approach, as described in [5]. This approach seems promising for the detection of humans with larger range of poses, or that are partially occluded. It could also help ruling out some of the false positive detection cases, by doing a finer analysis of the parts relationships within the root template.

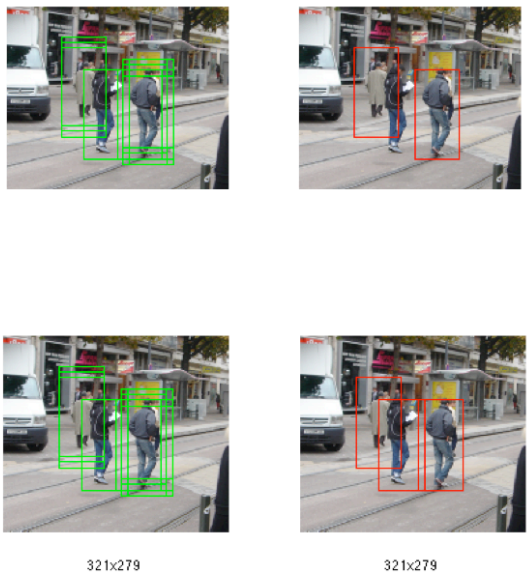

Figure 4: non max suppression with different window sizes, when looking for local maximum of weight cost.

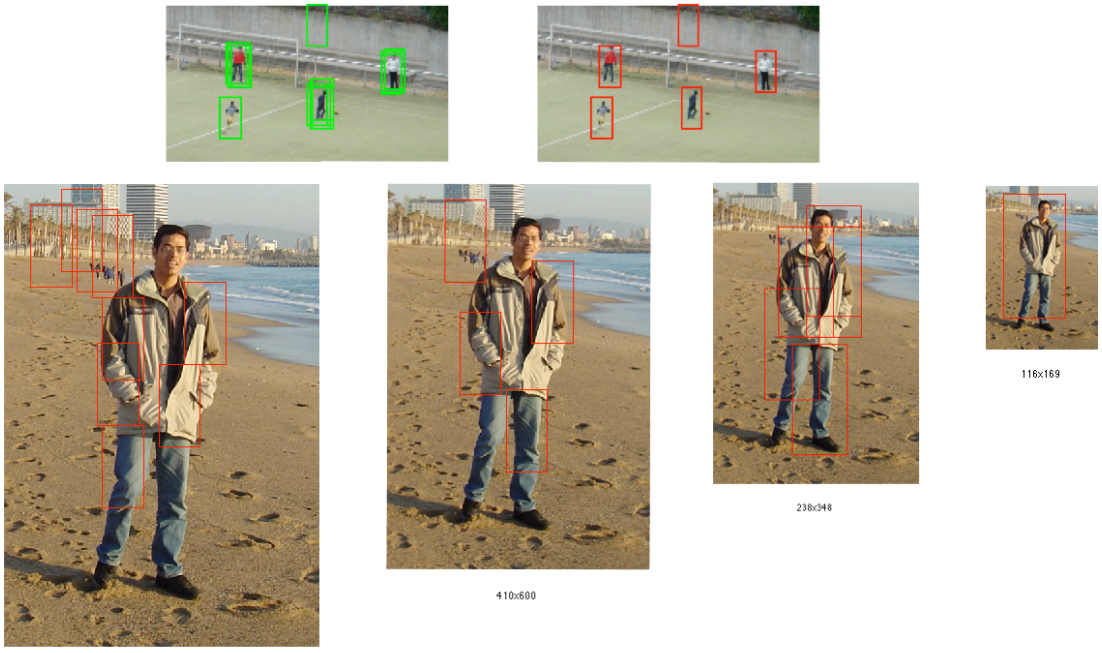

Figure 5: non max suppression with single or multiple distinct matches. Primary detection is done using a model retrained with hard negative cases

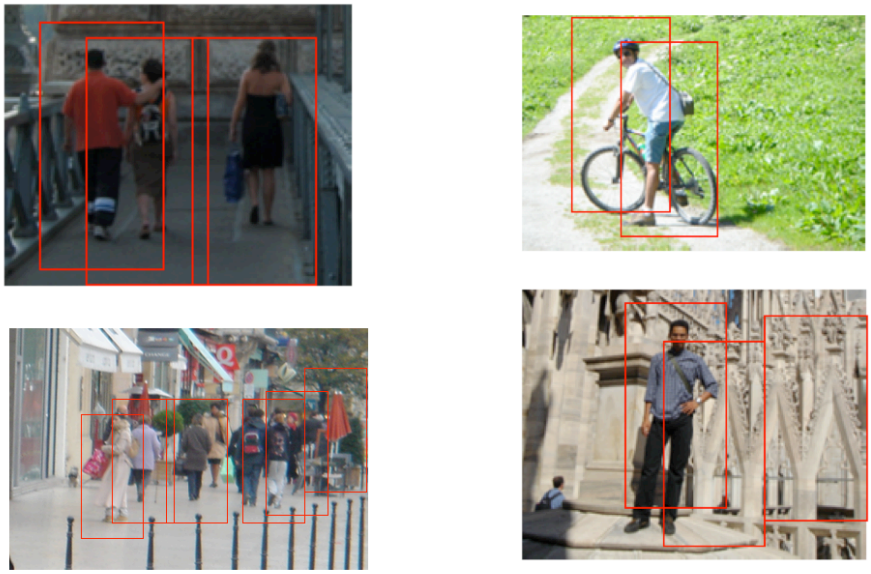

Figure 6: detection with different illumination, pose, scale, occlusion scenarios.

Datasets

http://pascal.inrialpes.fr/data/human

References

- [1] Navneet Dalal, Bill Triggs. Histograms of Oriented Gradients for Human Detection. CVPR 2005
- [2] Paul Viola, Michael J. Jones. Robust Real-time Object Detection. IJCV 2004
- [3] D.G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. IJCV 2004
- [4] Serge Belongie, Jitendra Malik. Shape Matching and Object Recognition using Shape Context. IEEE transactions on Pattern Analysis and Machine Intelligence. April 2002.
- [5] Pedro Felzenszwalb, Ross B. Girshick, David Mc Allester, Deva Ramanan. Object Detection with Discriminatively Trained Part-Based Models. IEEE transactions on Pattern Analysis and Machine Intelligence. September 2010.
- [6] Alexander Neubeck, Luc Van Gool. Efficient Non-Maximum Suppression.