
FACE AND OBJECT DETECTION BY CASCADE CLASSIFIERS

MY PERSONAL NOTES ON

By

0xLeo (github.com/0xleo)

OCTOBER 29, 2019

DRAFT X.YY

MISSING: ...

Contents

1	Viola Jones detection	2
1.1	Problem statement and idea	2
1.2	Integral image	3
1.3	Haar features	3
1.4	Adaboost training	5
1.5	Classifier cascade	7
1.5.1	Training the cascaded classifier	8
1.6	Face and object detection in OpenCV	9
1.7	How to train your own cascade classifier in OpenCV	10
1.7.1	Prepare the directory structure	10
1.7.2	Collect positive and negative training images	10
1.7.3	Augment the positive dataset	11
1.7.4	Generate the metadata and the vector file	12
1.7.5	Do the training	12
1.8	A trained hand classifier	14
A	Appendices	15
A.1	Viola Jones in OpenCV	16
A.2	Real-time cascade detector in OpenCV	17

1 Viola Jones detection

1.1 Problem statement and idea

Detecting faces is complicated since they have so many features, such as forehead, eyes, lips etc. Faces can also be in any size and position through the image, they can be non-uniformly illuminated, have multiple expressions or be occluded. Viola Jones takes advantage of the fact that every face has certain characteristics, that such the nose usually brighter is darker than the surrounding cheek area, the eye pupil is darker than the sclera, the nose, lip and eye position relative to each other etc. Viola Jones works in real time by combining the following methods:

1. Integral image. This speeds up the feature computation time.
2. Haar features. These are simple rectangular masks that resemble with Haar wavelets.
3. Boosting for feature selection (“AdaBoosting”). Iteratively classifies good and bad features.
4. Cascade of weak feature classifiers to quickly reject windows without faces (Attentional cascade).

Viola Jones scans a window that searches features over the image, incrementing pixel by pixel. The window size remains constant – the authors of VJ suggest 24×24 window the image is considered at various scales – the authors suggest a scaling factor of 1.25. All the scalings along with the original image are called *image pyramid*. Each level in the pyramid is obtained by downsampling the previous. For example, downsampling of 1.25 implies that every 5 pixels we save 4 and reject the 5th.

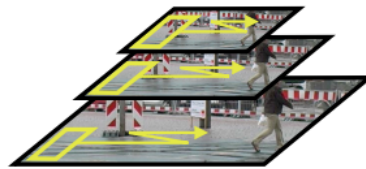


Fig. 1. Same window scans different image scales.

The window contains a different feature detector each time, e.g. one that might detect an eye or one that might detect a nose. *Haar features* determine what we want to compute each time, e.g. an eye, a forehead etc. One might think that the larger the window, the larger the number of operations done to compute the feature inside it. We will see that the number of computations is always constant thanks to the idea of *integral image*, whose computation is the first step of the algorithm. Why integral image is useful will be shown later, next section explains how it's computed. To summarise, the whole algorithm is explained by the following high-level pseudocode.

```
for number of scales in image pyramid do
  downsample image by one scale
  compute integral image for current scale
  for each shift step of the sliding detection window do
    for each stage in the cascade classifier do
      for each filter in the stage do
        filter the detection window
      end
      accumulate filter outputs within this stage
      if accumulation fails to pass per-stage threshold do
        break the for loop and reject this window as a face
      end
    end
    if this detection window passes all per-stage thresholds do
      accept this window as a face
    else
      reject this window as a face
    end
  end
end
end
```

1.2 Integral image

Given the source image $ii(x,y)$ where we want to detect faces, a new, integral image $i(x,y)$ is created such that each new pixel is the sum of the pixels above and to the left of it, including the original pixel itself:

$$ii(x,y) = \sum_{x' \leq x, y' \leq y} i(x',y') \quad (1.1)$$

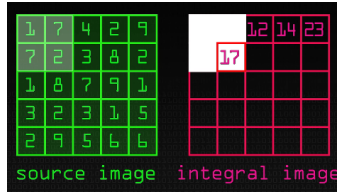


Fig. 2. Computing a single pixel of the integral image.

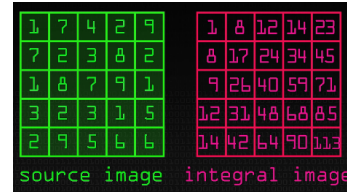


Fig. 3. The whole integral image computed.

This technique allows constant time ($\mathcal{O}(1)$) calculation of the sum of all pixels within a rectangle.

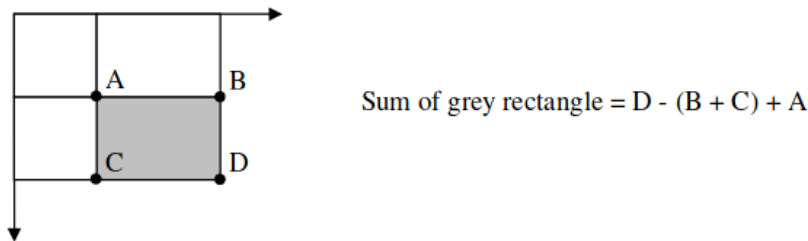


Fig. 4. Rectangle sum calculation. The rectangles from the origin to A, B, C, D are pre-computed by the integral image. A is counted in both B and C so in the end we subtract it.

Sums of pixels are computed in order to find the Haar features described in the next sections. Since we have tons of Haar features, pre-computing the integral image is necessary.

1.3 Haar features

But what exactly are the Haar features and how do they look like? They are simple rectangular masks and each one is supposed to roughly resemble with a certain face feature, e.g. the left end of the lip. They contain certain combinations of zeros and ones. Ones should match bright areas of the feature and zeros dark. Haar features are always located in the sliding window and they are used similarly to convolution kernels.



Fig. 5. Some Haar features for line and edge detection.

For example, the edge feature in the above figure could roughly match an eyebrow and the vertical line feature could match the nose bridge, as shown below. Of course, line features can also be diagonal to detect rotated faces.

Haar features are supposed to be similar to simple face features found in the nose, eyes, etc.

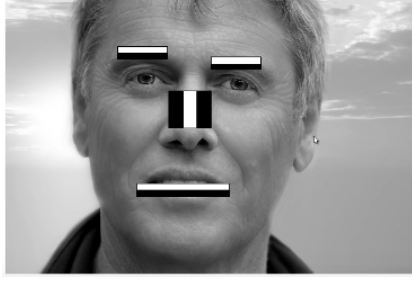


Fig. 6. Some Haar features that match face features.

To measure the quality of a feature, i.e. its similarity to the overlapping Haar cascade, we simply subtract the pixels under the black ones from those under the whites and divide the result by the size of the Haar feature. The higher the result Δ , the more likely a feature to belong to a face. This method is also called “difference of adjacent boxes”. Below, n of the number of pixels of the Haar feature.

$$\Delta = \text{dark} - \text{white} = \frac{1}{n} \sum_{\substack{(x,y) \\ \text{under white}}} i(x,y) - \frac{1}{n} \sum_{\substack{(x,y) \\ \text{under black}}} i(x,y) \quad (1.2)$$

The two figures below explain the Δ calculations. For the feature in Fig. 8, we have ¹

$$\Delta = \frac{1}{16} (209 + 214 + 207 + 199 + 183 + 196 + 200 + 212 - 181 - 77 - 105 - 67 - 68 - 101 - 183 - 196) = 40.125$$

The closer the value of Δ to 255, the better the feature. A region of random pixels would have $\Delta \approx 0$. This calculation seems quite fast but it needs to be repeated for a huge number of times. When the Haar rectangle is large, it adds up a lot of computation time.

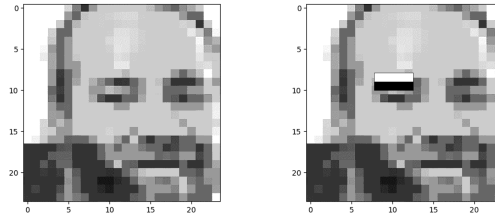


Fig. 7. The sliding 24×24 window captures a downsampled face. A line feature is overlaid the eyebrow.

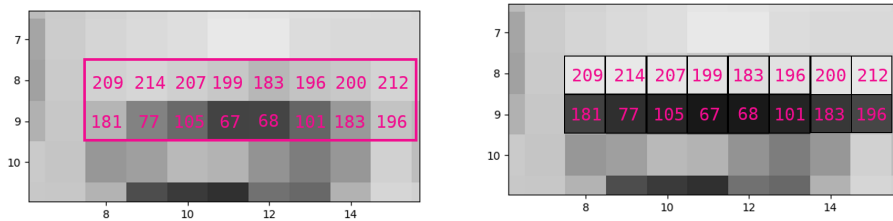


Fig. 8. A closer look at the intensity values of the eyebrow. Pixels under the black region will be subtracted and those under the white region will be added.

This is where integral image discussed in the previous section comes in handy. Instead of calculating the sum naively, for Fig. 8 we would find the sum inside the white and black rectangles using the integral image. Therefore each black or white sub-rectangle only requires 4 additions.

¹Images are represented as 8-bit integers so a negative result would be clipped to 0 and one that cannot be represented by 8 bits to 255.

Integral image speeds up the Haar feature computation.

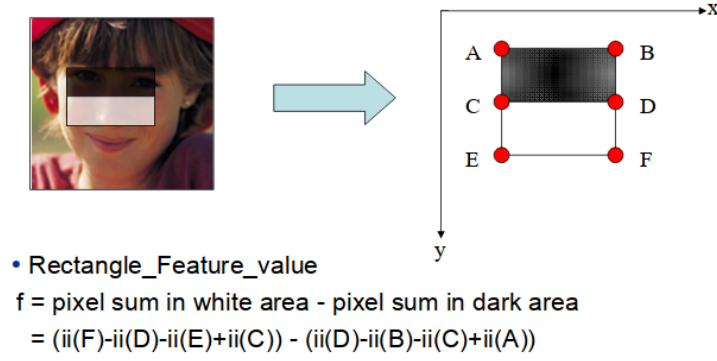


Fig. 9. Computing an example Haar feature using the integral image ii .

By comparing each Δ value to a threshold, e.g. 30 we can tell if the region under the window is likely to be a face, although this procedure will give lots of *false positives*. If the result is less than 30, then that feature in that window position will not be considered. If it's greater, more Haar classifiers will be applied. That's called a *weak classifier* and the algorithm combines multiple of them. "Weak essentially means better than random classifier, so in this case as long as the classifier outputs a value roughly greater than zero, then the window is likely to contain a face:

$$y(\Delta_\theta) = \begin{cases} 1, & \Delta_\theta > T \\ 0, & \text{else} \end{cases} \quad (1.3)$$

, where θ refers to a set of parameters such as the number of rectangles, their positions, etc. The problem with using Haar features inside the window is that for a 24×24 window there are about 160,000 of them, so computing them all is inefficient. The algorithm needs to select and combine in advance some good Haar features. This is done by a machine learning algorithm called "AdaBoost".

1.4 Adaboost training

The solution to the fact that the sliding window contains over 160,000 features is to select a few of them – the most reliable ones that can tell whether the sub-image contains a face. This is done by AdaBoost (adaptive boost), which is a topic on its own, so it won't be discussed in too much detail but this section includes the overview.

For a certain position of the sliding window and given its features f_1, f_2, \dots such as position of black boxes, coordinates, threshold etc, the idea is that we have a set of data and outputs $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \{-1, 1\}$ (-1 : is not face, 1 : is face) and we want to train perceptrons to correctly classify the output points, i.e. we want to linearly classify them. The goal is to find the features that best separate the positive and negative classes.

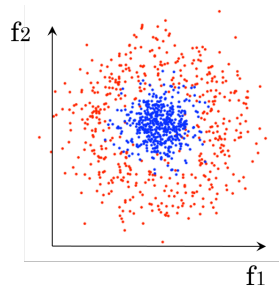


Fig. 10. Various data points $((x_{i1}, x_{i2}), y_i)$ in the 2D feature space f_1, f_2 . Each class is represented by a colour.

Adaboost aims to separate the data in two classes, minimising some error function. It constructs a "strong" linear classifier as a combination of T simple "weak" simple linear classifiers $h^{(1)}(x), \dots, h^{(T)}(x)$:

$$f(x) = \sum_{t=1}^T \alpha^{(t)} h^{(t)}(x) \quad (1.4)$$

, where:

- $h_t(x)$ is a “weak” or base classifier – can be a perceptron, or decision tree etc, as long as it can perform linear binary classification.
- $H(x) = \text{sgn}(f(x))$ is the final, strong, classifier.
- $\alpha^{(t)}$ is a coefficient we set at each iteration depending on how good the base classifier $h^{(t)}$ is.

The algorithm maintains a weight distribution $\mathbf{w}^{(t)} = (w_1^{(t)}, \dots, w_m^{(t)})$ over the data points². The weights $w^{(t)}$ are initialised uniformly, and are updated in each iteration, keeping their sum constant, e.g. normalised. The goal of the base learner is to minimise the *weighted (base) error*

$$\epsilon^{(t)} = \sum_{i=0}^m w_i^{(t)} I \{h^{(t)}(\mathbf{x}_i) \neq y_i\} \quad (1.5)$$

, where $I(\text{true}) = 1$, $I(\text{false}) = 0$ so $\epsilon^{(t)}$ can be considered as the sum of the weights of the wrong examples³, i.e. $\sum_{h^{(t)}(\mathbf{x}_i) \neq y_i} w_i^{(t)}$.

Having found the total error, the weak classifier weight $\alpha^{(t)}$ is set as

$$\alpha^{(t)} = \frac{1}{2} \ln \frac{1 - \epsilon^{(t)}}{\epsilon^{(t)}} \quad (1.6)$$

So the larger the error $\epsilon^{(t)}$, the smaller the weight $\alpha^{(t)}$. The final pseudocode is shown below. We denote the training set as $\mathcal{D}_m = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$

Algorithm 1 AdaBoost algorithm outline.

```

1: procedure ADABOOST( $\mathcal{D}_m$ , Base(.),  $T$ )
2:    $w^{(1)} \leftarrow (1/m, \dots, 1/m)$  ▷ initial weights
3:   for  $t \leftarrow 1, \dots, T$  do
4:      $h^{(t)} \leftarrow \text{Base}(\mathcal{D}_m, \mathbf{w}^{(t)})$  ▷ apply base classifier; e.g. a simple perceptron
5:      $\epsilon^{(t)} \leftarrow \sum_{i=0}^m w_i^{(t)} I \{h^{(t)}(\mathbf{x}_i) \neq y_i\}$  ▷ weighted (total) error of base classifier  $t$ 
6:      $\alpha^{(t)} \leftarrow \frac{1}{2} \ln \frac{1 - \epsilon^{(t)}}{\epsilon^{(t)}}$ 
7:     for  $i \leftarrow 1, \dots, m$  do ▷ re-weight each training point
8:       if  $h^{(t)}(\mathbf{x}_i) \neq y_i$  then ▷ error
9:          $w_i^{(t+1)} \leftarrow \frac{w_i^{(t)}}{2\epsilon^{(t)}}$  ▷ increase the weight
10:      else ▷ correct classification
11:         $w_i^{(t+1)} \leftarrow \frac{w_i^{(t)}}{2(1 - \epsilon^{(t)})}$  ▷ decrease the weight
12:   return  $\text{sgn}(f(\mathbf{x})) = \text{sgn}\left(\sum_{t=1}^T \alpha^{(t)} h^{(t)}(\mathbf{x})\right)$  ▷ weighted “vote” of best (a.k.a. base) classifiers

```

It’s important to remember the weights are normalised for each iteration of the algorithm, i.e.

$$\sum_{i=1}^m w_i^t = 1 \quad (1.7)$$

²Weights $\mathbf{w}^{(t)}$ and $\alpha_1, \dots, \alpha_m$ are different.

³There are better error functions than this sum but we don’t discuss them here for simplicity.

AdaBoost essentially applies a series of linear classifiers, increasing the weights of the errors and decreasing the correct ones each iteration.

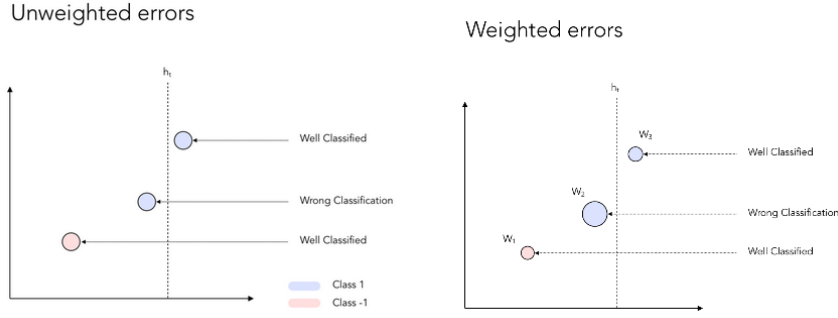


Fig. 11. Weights of wrong examples are increased and weights of correct ones are decreased.

The figure below shows how the weights are updated at every iteration.

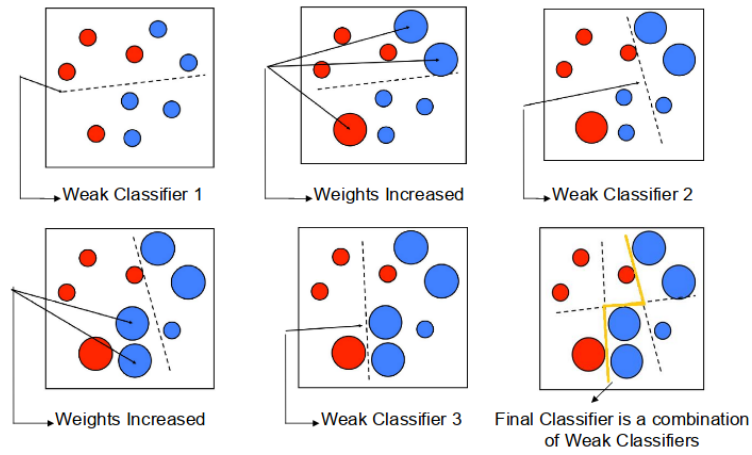


Fig. 12. At every iteration t the weights of false classifications are increased and the correct ones are decreased. Then the base classifier $h^{(t)}$ is applied again.

1.5 Classifier cascade

So far, we applied various Haar masks of different sizes, formations, and locations, etc (there are the features $\mathbf{x}^{(i)} \in \mathbb{R}^d$, where d the total number of sizes, formations, and locations etc), compared the summation result of masking calculated using the integral image to a threshold, and depending on the comparison result we classified the output $y^{(i)} \in \{-1, 1\}$ as “is in face” (1) or “is not in face” (-1). Then we separate all positive and negative results in two groups using a strong classifier calculated by AdaBoost. We had been dealing with only a fixed image size (no scales).

For a large image, and considering the image will have to be scaled and rescaled to account for differently sized faces, this amounts to using Viola-Jones a large number of times. If the system needs to detect faces in real-time, the computation needs to be fast. It is for this purpose that Viola and Jones introduced the attentional cascade.

The attentional cascade uses a series of strong classifiers (see previous section), each progressively having more features, to classify an image. An image is only put through by the n th classifier (larger image size) if the $n - 1$ th classifier (smaller size) classifies it as a positive example. If at any point a classifier does not think the image is a positive example, the cascade stops and the region is rejected as a face. The reason behind it is that in an image we will have only a few faces and thousands of non-faces. Being able to quickly discard non-faces saves time.

Each part of the cascade operates on a different input scale.

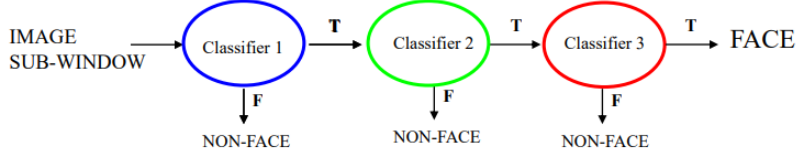


Fig. 13. How multiple strong classifiers are chained to form a cascade one. The first cascale belongs to large scale (smallest image), the second to smaller scale, etc. Classifier 2 is activated by 1 and 3 by 2.

We study how the length of the cascade improves its quality. Let f_i be the False Positive Rate (FPR) of each classifier and d_i is detection rate (True Positive Rate – TPR⁴). Assuming the stages are independent, then the *overall* false positive rate F and the overall detection rate D is

$$F = \prod_{i=1}^K f_i \quad (1.8)$$

$$D = \prod_{i=1}^K d_i \quad (1.9)$$

Assume $K = 10$. If every stage (assume they all preform roughly the same) has TPR $d_i = 99\%$ and FPR $f_i = 40\%$, then the cascade has $D = 0.99^{10} = 90\%$ and $F = 0.4^{10} = 0.001\%$.

High FPR does not really affect the cascade's quality. However, TPR needs to be high.

1.5.1 Training the cascaded classifier

Therefore having high FPR at each stage is not a bad thing but having low TPR is undesirable. Hence, training the whole cascade is summarised as follows.

1. Set target detection and false positive rates for each stage.
2. Keep adding features to the current stage until its target rates have been met.
 - (a) Need to lower AdaBoost threshold to maximise detection (as opposed to minimising total classification error).
 - (b) Test on a validation set.
3. If the overall false positive rate is not low enough, then add another stage.
4. Use false positives from current stage as the negative training examples for the next stage.

Training could, for instance, include $5 \cdot 10^3$ positives, $350 \cdot 10^6$ negatives, and could result in a real-time cascaded detector with 38 stages and 6061 features along all layers. The figure below how individual classifiers, each with its own Haar features, could look like.

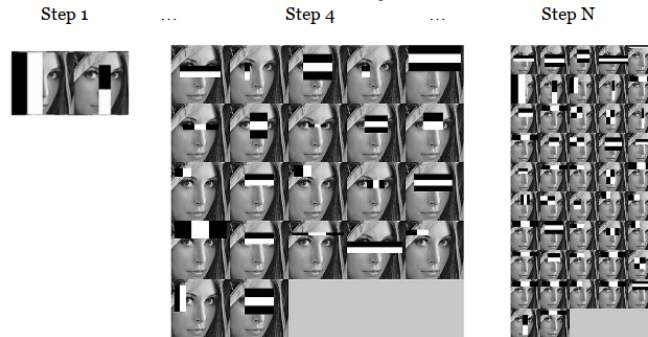


Fig. 14. The first weak classifier has in this case 2 simple features , outputting lots of FP. The rest have progressively more features.

⁴FPR and TPR are defined as follows. P are the total positive and N the total negative examples. $TPR = \frac{TP}{P} = \frac{TP}{TP+FN}$, $FPR = \frac{FP}{N} = \frac{FP}{FP+TN}$.

1.6 Face and object detection in OpenCV

As mentioned before, the cascade Viola Jones classifier can work in real time thanks to the cascade of strong classifiers, which are applied sequentially. Although we could have thousands of features, they are not applied at the same time. The first classifier contains the fewest features, the second more etc. If at any stage the input image does not pass one of the classifiers, it's rejected. This saves huge amounts of computation time.

Object, not only face, detection in OpenCV is very straightforward. First, we initialise a `cv2.CascadeClassifier` object:

```
cv2.CascadeClassifier(xml_file) -> cascade_detector
```

- `xml_file` is the path to an `.xml` file that contains training data, i.e. the Haar features and threshold of each classifier stage. It contains data in a special format. Some samples can be found in [OpenCV's repository](#). Note that the `.xml` files contain also features trained to detect face profiles, eyes, bodies, etc.
- `cascade_detector` is the object that will actually run the Viola-Jones algorithm. More on it next.

```
cv2.CascadeClassifier.detectMultiScale(image[, scaleFactor[, minNeighbors[, flags[, minSize[, maxSize]]]]) -> objects
```

- `image`: Greyscale image that contains objects to be detected.
- `scaleFactor`: Specifies how much the image size is reduced at each image scale. Default value is 1.2.
- `minNeighbors`: Specifies how many neighbors each candidate rectangle should have to retain it. Default value is 3.
- `flags`: Parameter with the same meaning for an old cascade as in the function `cv2.HaarDetectObjects`. It is not used for a new cascade.
- `minSize`: A tuple. Minimum possible object size. Objects smaller than that are ignored.
- `maxSize`: A tuple. Maximum possible object size. Objects larger than that are ignored. If `maxSize == minSize` model is evaluated on single scale.
- `objects`: An array of rectangles. Each rectangle is, in turn, an array `x0, y0, width, height`.

If the input image has low contrast, OpenCV suggests histogram pre-processing before applying the face detector.

`scaleFactor` may need to be calibrated depending on how large the face to detect is relative to the image size. Recall that scale factor defines how much the image is downsampled from one pyramid level to another. For example, if the input image has resolution 1280×800 and `scaleFactor = 1.2`, then the pyramid base level has size 1280×800 and the next has 20% of it, so 1168×730 and so forth. Reducing the scale factor increases the chance of finding a face but also increases computation time and also false positives.

Decreasing `minNeighbors` increases the number of detection rectangles that can overlap each other.

The code is listed in A.1 and a result is shown below.

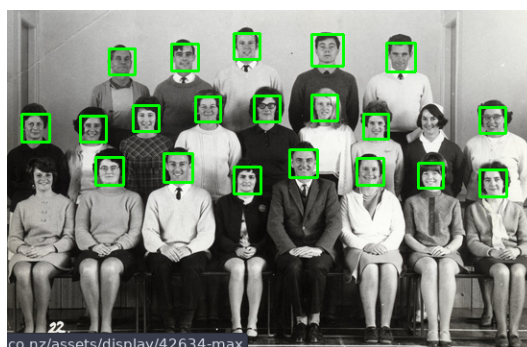


Fig. 15. Applying Viola-Jones for `scalingFactor = 1.1` and `minNeighbors = 5`.

1.7 How to train your own cascade classifier in OpenCV

OpenCV 3.x allows you to train your own cascade classifier. Cascade training has been dropped in version 4.x. The way it works is

1. Prepare the directory structure.
2. Grab some positive and negative images and store them in the directory.
3. (Optional) Augment the positive datasets, e.g. randomly rotate, flip, or add noise to each positive image.
4. Generate the metadata, such information files regarding where the images are and the object location in the positive images.
5. Do the training. This will eventually generate an `.xml` file that is usable for the target object detection.

Before deciding on an object to train the classifier for, it is worth asking yourself whether it's susceptible to Haar feature detection. The object to detect must consistently have certain features at certain positions relative to each other, so that when Haar rectangles cover them, they are able to produce high differences. For example, for a clenched fist detector these features could be the dark ridges between the fingers, the thumb, the nails at the end etc. For an orange detector it would be hard to come up with such features as the shape is generic and other than the pores, not many indicators exist for the classifier to tell it apart e.g. from a ball. There are many similar ways to do the training but below is the one that works for me.

1.7.1 Prepare the directory structure

On github there is already an empty repository that's fit for cascade training. Clone it

```
git clone https://github.com/OxLeo/opencv-haar-classifier-training.git
cd opencv-haar-classifier-training/
```

Remove all the `.gitkeep` files so that they're not accidentally included as training images.

```
find . -type f -name .gitkeep -exec rm -f {} \;
```

After cloning, the directory structure should look as follows:

```
.
+-- bin
|   +-- createsamples.pl
+-- classifier
+-- LICENSE
+-- negative_images
+-- positive_images
+-- README.md
+-- samples
+-- tools
|   +-- <files>
+-- trained_classifiers
    +-- <files>
```

1.7.2 Collect positive and negative training images

Remember two things when downloading training images.

1. The positive images should contain only the object of interest and as little margin around it as possible.
 - (a) The reason we want the background margin around the OOI to be as narrow as possible is to facilitate object annotation in the next sections.
 - (b) Cropping may be required to get a good representation of the OOI. After cropping, remember that all positive images (OOI) must have roughly the same width : height ratio. The reason for this will be clear in the augmentation section.

2. The negative images can contain anything but the object of interest (OOI). Ideally, they contain objects or scenes that can be found in the background.

. To quickly crop the object, a tool such as [batchcrop](#) is recommended. Unfortunately, this only works on Windows.

A way of collecting positive images is by downloading pre-compiled datasets and extracting their RGB images and cropping the OOI. Alternatively, videos can be converted to images. A few thousand positive and a few thousand negatives should be fine. Once they have been collected and cropped and follow guidelines (1) and (2), dataset augmentation can begin. When done collecting the images and cropping the positives, copy them to `positive_images` and `negative_images`.

After copying them, it is recommended to convert the negative images to grayscale. This is done by configuring and running the augmentation tool found in `tools` directory. Then edit the tool's config, found in `augment_config.yaml` so that only grayscale conversion is enabled, run the tool from `main.py`. If there are any non-grayscale images (they do not contain `gray` in their name) left in the `negative_images`, delete them manually.

```
edit augment_config.yaml
config:
  directory: ../negative_images/
  grayscale: yes
  equalise: no
  rotate:
    flag: no
    how_many: 6
    min_angle: -25
    max_angle: 25
  flip:
    flag: no
  noise:
    flag: no
    var: 0.6
  resize:
    flag: no
    width: 128
    height: 128
./main.py
```

Tip: christopher5106⁵ on github recommends that “each (negative) image should be (but not necessarily) larger than a training window size, because these images are used to subsample negative image to the training size”. The reason he recommends it is that so that only the exact negative image can be used during training, and not random parts of it. I tried it, but it *didn't improve my results* compared to no resizing – it only introduced more FPs.

Tip: I found that the more positives used for training, the more FP the classifier is likely to detect and the more negatives, the more FN the it's likely to detect. So a workflow is to test the trained classifier, check whether it detects too many FP and FN and balance the dataset accordingly. You can begin with a ratio `pos:neg = 3:2` and check the trained classifier's results.

Negative images need to be converted to grayscale.

1.7.3 Augment the positive dataset

Image data augmentation is a technique that can be used to artificially expand the size of a training dataset by creating modified versions of the positive images in the dataset. It therefore allows the model to generalise its results. Augmentation can be done by the `opencv_createsamples` command but that didn't work for me so I used my own tool.

If you have collected e.g. a few hundred positive images, augmentation first performs pre-processing (conversion to grayscale followed by histogram equalisation) and create a flipped copy of each one, randomly rotate it a number of times and add noise to the result if necessary, therefore generate new positive images based on the existing ones. The class to perform augmentation is found in `tools/augment.py` and its config

Run the augmentation tool on positive images.

⁵ <http://christopher5106.github.io/computer/vision/2015/10/19/create-an-object-detector.html>

in `tools/augment_config.yaml`. You need to modify the latter file to meet your needs. The file contents were also listed in the previous section. When done, augmentation of positive samples can begin by running:

```
chmod +x main.py
./main.py
```

When done, the `positive_images` folder should only contain grayscale images of the same size. They will contain either `resized` or `augm` in their filename. If anything else exists in the folder, remove it manually.

1.7.4 Generate the metadata and the vector file

OpenCV needs a file to list positives, one to list negatives and one to list obj locations in positives images.

```
cd <project_root>
find ./positive_images/ -iname *.png > positives.txt
find ./negative_images/ -iname *.png > negatives.txt
```

Positive images should only contain `.png` images. If you haven't converted the negative images to grayscale by the tool and they still contain `.jpg`, `.tiff` etc, you need to append them like so:

```
find ./negative_images/ -iname *.jpg >> negatives.txt
```

Tip: None of the images, either positive or negative, must contain spaces in its file name. This is guaranteed for positive images as they should have been generated by the augmentation tool, but if negatives contain spaces remove them manually.

Now generate the file that contains the object positions as rectangles (info file). Since cropping is as tight as possible, we assume that the OOI takes the whole image. Info file is based on `positives.txt` so in the project root, copy the latter:

```
cp positives.txt info.dat
```

In the end, we want each line of `info.dat` to look as follows:

```
<path_from_root_to_positive_image> <number_of_objects> <x0> <y0> <x1> <y1>
```

The starting position of the rectangle around the OOI is (0,0) and the end is (*image_width*,*image_height*). For example, if all images have been resized to (128,128), objects have been cropped correctly and each one contains only one object, then if you open the info file, each line of `info.dat` should look like:

```
edit info.dat
./positive_images/resized_00212_rot_03.png 1 0 0 128 128
```

Once annotating is done, generate the `.vec` file, which encodes all positive images together in a vector given the `info.dat` file. However, the vector file should further resize the positives to the training size, otherwise training will be slow. In this case 128×128 is too large for training, so shrinking the positives is necessary. For this, we need the `opencv_createsamples` program, which is part of OpenCV 3.x. The command to create a vector file (`samples.vec`) containing all positive images downsampled to 36×36 is:

```
NPOS=`cat positives.txt | wc -l`
opencv_createsamples -info info.dat -vec samples.vec -h 36 -w 36 -num $NPOS
```

To make sure the vector is created correctly, view the generated images with the following command and press escape to close the window:

```
opencv_createsamples -vec samples.vec
```

1.7.5 Do the training

Before the training begins, make sure the directory where the cascade classifier will be writing (specified by the `-data` options below) the training weights for each stage and the final classifier is empty:

```
mkdir -p classifier
rm classifier/*
```

Use the following command for training. Adjust the parameters to your needs.

```
opencv_traincascade -data classifier -vec samples.vec -bg negatives.txt
-acceptanceRatioBreakValue 0.00001 -minHitRate 0.995 -maxFalseAlarmRate 0.35
-numPos 2850 -numNeg 1678 -w 36 -h 36 -mode ALL -precalcValBufSize 1024
-precalcIdxBufSize 1024 -featureType LBP
```

Tip: If your minHitRate is too high, you may get the following error:

```
imagestorage.cpp:159: error: (-5) Can not get new positive sample.
The most possible reason is insufficient count of samples in given vec-file.
```

This means that the classifier does not have enough information to reach the required minHitRate or maxFalseAlarmRate. To continue, either decrease minHitRate and increase maxFalseAlarmRate and continue the training or get more positive samples, augment them, annotate them, and restart the training. The parameters that need to be explained are the following. For more, see the help menu of the command.

- data: The output file where the training weights and the final .xml, called cascade.xml will be written.
- numStages: Number of stages of cascade classifier.
- minHitRate: Minimum TP rate of each stage of the classifier. It needs to be high as, if it's e.g. 0.99, for 16 stages the final TP rate would be $0.99^{16} \approx 0.851$.
- maxFalseAlarmRate: Maximum FP rate of each stage. If for one stage FPR is 0.35, then for 16 stages the final is $0.35^{16} \approx 0$.
- numPos: Number of positive images used for learning at each stage.
- numNeg: Likewise, for negative images.
- w, h: The width and height of images stored in the .vec file.
- featureType: can be either LPB or HAAR. LPB stands for Local Binary Patterns and it's a similar type of features to Haar as it uses Haar features. LPB uses integer arithmetic and HAAR floating point, therefore LPB is several times faster but slightly less accurate.

Tip: According to the OpenCV documentation, if acceptanceRatio BreakValue has been set for training, A good guideline is to train not further than $1e-4$, to ensure the model does not overtrain on your training data. For me good, non-overfitting results we

A good rule of thumb when to stop the training is when $1e-4 \leq \text{acceptanceRatioBreakValue} \leq 1e-5$

While the training is ongoing, it will display an output similar to the following.

```
===== TRAINING 0-stage =====
<BEGIN
POS count : consumed    2300 : 2300
NEG count : acceptanceRatio    430 : 1
Precalculation time: 13
+---+-----+-----+
| N |      HR |      FA |
+---+-----+-----+
|  1 |         1 |         1 |
+---+-----+-----+
|  2 |         1 |         1 |
+---+-----+-----+
|  3 |         1 |         1 |
+---+-----+-----+
|  4 |         1 | 0.909302 |
+---+-----+-----+
|  5 |         1 |         0.9 |
+---+-----+-----+
|  6 |         1 | 0.695349 |
+---+-----+-----+
|  7 |         1 | 0.723256 |
+---+-----+-----+
|  8 |         1 | 0.634884 |
+---+-----+-----+
```

```
| 9| 1| 0.327907|  
+---+-----+-----+  
END>
```

Training until now has taken 0 days 0 hours 0 minutes 24 seconds.

When it's done, it will generate the trained classifier at `classifier/cascade.xml`. Then, for example the program at A.1 can be used to test it. If you want to train a new classifier, delete the contents of the classifier directory:

```
rm classifier/*
```

In the next section there are some real-world results of my Haar hand detector.

1.8 A trained hand classifier

A classifier that detects the following hand gestures

1. clenched fingers/ fist
2. index finger raised
3. index and middle finger raised

, best at indoor light conditions, has been stored as .xml file at ⁶. To try it, or another trained classifier, on a webcam, use the code at A.2. Below are some of its outputs.

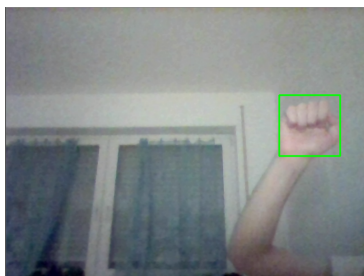


Fig. 16. Fist with light background.

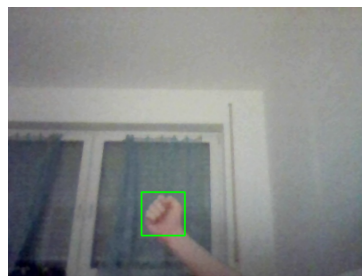


Fig. 17. Fist with dark background.

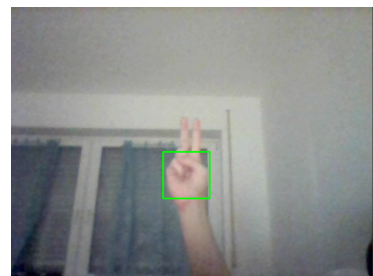


Fig. 18. Victory sign.

⁶ https://github.com/0xLeo/opencv-haar-classifier-training/blob/master/trained_classifiers/hand_44x44_19_stages_lbp.xml

A Appendices

A.1 Viola Jones in OpenCV

Listing 1: Python implementation (src/viola_jones_cv.py).

```
1 import cv2
2 import numpy as np
3
4 # download the xml from repo is search it in your installation
5 faceCascade = cv2.CascadeClassifier('./haarcascade_frontalface_default.xml')
6
7 im = cv2.imread('screenshot.png')
8 grey = cv2.cvtColor(im, cv2.COLOR_BGR2GRAY)
9
10 face_rects = faceCascade.detectMultiScale(
11     grey,
12     scaleFactor = 1.1,
13     minNeighbors = 1)
14 import pdb; pdb.set_trace()
15
16 for (x,y,w,h) in face_rects:
17     cv2.rectangle(im, (x, y), (x+w, y+h), (0, 255, 0), 2)
18 cv2.imshow("detected", im)
19 cv2.waitKey(0)
20 cv2.destroyAllWindows()
```

A.2 Real-time cascade detector in OpenCV

Listing 2: Detecting an object in real time given an .xml file (src/viola_real_time.py).

```
1 import cv2
2 import sys
3 import os
4
5
6 def test_cascade(xml, vid_file = 0):
7     assert os.path.isfile(xml), "Invalid filepath to %s" %xml
8     obCascade = cv2.CascadeClassifier(xml)
9     video_capture = cv2.VideoCapture(vid_file)
10
11     while True:
12         _, bgr = video_capture.read()
13         im = cv2.cvtColor(bgr, cv2.COLOR_BGR2GRAY)
14         # minSize: minimum size for which it was trained
15         objects = obCascade.detectMultiScale(
16             im,
17             scaleFactor = 1.1,
18             minNeighbors = 3,
19             minSize = (44, 44)
20         )
21         # Draw a rectangle around the objects
22         for (x, y, w, h) in objects:
23             cv2.rectangle(bgr, (x, y), (x+w, y+h), (0, 255, 0), 2)
24
25         cv2.namedWindow('det', cv2.WINDOW_NORMAL)
26         cv2.imshow('det', bgr)
27         if cv2.waitKey(10) & 0xFF == ord('q'):
28             break
29
30     video_capture.release()
31     cv2.destroyAllWindows()
32
33
34 def main():
35     xml = 'hand_44x44_19_stages_lbp.xml' if len(sys.argv) == 1\
36         else sys.argv[1]
37     test_cascade(xml)
38
39
40 if __name__ == '__main__':
41     main()
```