Detection Theory and Industrial Applications - Final Project
A Contrario Detection of Faces

Sara El Khbir sara.el_khbir@ens-paris-saclay.fr

April 15, 2024

1 Introduction

We will study the face detection algorithm used in the papers [1] and [2] using a contrario approach.
The a contrario methodology is a mathematical formalization of the non-accidentalness principle proposed for perception (sometimes called Helmholtz principle). In this approach, an observed structure is considered relevant if it rarely occurs by chance. This is implemented assuming a null-hypothesis for the data where no detections should occur, which is called the a contrario model. The rarity or non-accidentalness of a structure is quantified then as the probability of observing that structure under the null-hypothesis hypothesis.
This approach has been widely used in computer vision through Gestalt theory in various applications such as detecting specific structures such as lines, ellipses and circles in images and contours within images, identifying moving objects in videos, and more.

The papers [1] and [2] aim to demonstrate the adaptability of the a contrario formulation to Viola and Jones' face detection method proposed in [3] and that involves quickly calculating Haar-like features from input images, identifying the most discriminative features through AdaBoost training, and implementing a cascade of classifiers to achieve high detection rates while minimizing false alarms. The studied papers present an enhancement to the original Viola-Jones method using an a contrario methodology in the detection phase, aiming to boost the detector's performance by improving detection rates, minimizing false alarms, and reducing computational demands. Furthermore, an adaptive thresholds unique to each input image is employed. These thresholds, rather than being fixed and pre-learned during training, are dynamically determined from the detection values across the entire image.

In this report, we will start by giving the definition of the problem in section 2. Following that, we'll outline the problem in a formal manner in Section 3 by discussing the a contrario model , then we will present algorithms and experimental outcomes in Section 4. Section 5 will delve into comparison with other face detection methods and potential enhancements for further discussion.

2 Problem definition

2.1 The Viola-Jones Face Detector

The Viola and Jones face detection algorithms rely on the following steps:

- Defining of the Haar-like features which are rectangular patterns used to characterize different aspects of a given image. These features are defined by the difference in intensity between adjacent rectangular regions.

Figure 1: Haar-like feature masks (From [4])

- A weak classifier $h_k(x)$ is defined with association with each Haar-like feature k. This classifier operates on a sub-image x and computes a feature value as the difference between the sums of intensity values within 'white' and 'black' feature masks. $h_k(x) = 1$ if the feature value at x is above/below a learned threshold; otherwise = 0.
- The strong classifier is constructed by combining K weak classifiers associated with Haar-like features. It is defined as

$$h(x) = \begin{cases} 1 & \text{if } \mathrm{v}_{det} = \sum_{k=1}^{K} \alpha_k h_k(x) \geq T = \frac{1}{2} \sum_{k=1}^{K} \alpha_k, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

  Where α_k is the weight of h_k in the final strong classifier

2.2 Problem definition and link with gestalt theory

Gestalt theory, originated in the early 20th century, posits that humans perceive objects and scenes as organized wholes rather than the sum of their individual parts. This theory shows the role of perception in creating meaningful patterns. Gestalt theory relies on manys laws such as proximity, similarity, closure, good continuity, and figure-ground segregation (ie distinguishing objects from their background), etc..

In the context of face detection, this can be linked to Gestalt theory through the concept of emergent features. In fact, Gestalt principles suggest that certain configurations of visual elements (such as eyes, nose, and mouths) create a perceptual whole (the face) that is more than the sum of its parts.

In the "a contrario" approach, the emergence of a face-like pattern from statistical noise or random arrangements of pixels indicates a meaningful structure that aligns with Gestalt principles of perceptual organization.
As has been stated in the introduction, the a contrario approach relies on establishing a stochastic model for the data where the desired structure is absent and can only be noticed by chance. In [1], the authors derived the stochastic model by analyzing how the classifier responds to images that don't contain faces. They examined the distribution of detection values across tested subwindows within a specific image for various strong classifiers with differing numbers of features. The results show that for the images wihtout faces, the distribution of detection values tends toward a normal distribution, the more features we consider. In images containing faces, the distribution of values remains Gaussian, but the detection values corresponding to face-containing

sub-images are notably higher than those of the Gaussian distribution. The example in Fig 2 shows the distribution of detection values for a 80-features classifier where the red dots indicate the detection values for the sub-windows actually containing a face.

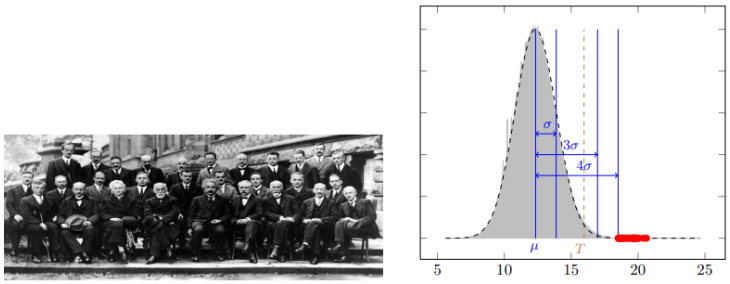


Figure 2: Example of the distribution of detection values for a 80-features classifier. (From [1])

As the parameters of the gaussian changes for each image, the authors propose to adapt the threshold to each particular distribution of detection values associated to each image rather than using the default fixed detection threshold T . This will mainly allow to avoid false positives in the images (corresponding to the values above T). This will be done by applying the a contrario detection principle to test the presence of a face in a subwindow against a noise or a contrario model where the face is not present.

3 Formulation of the Problem

3.1 The a contrario model

The goal of the a contrario framework is to exclude false detections, given a background model H0 (called the noise model). If a detection is likely to happen under the noise model H0 it is said to be accidental and is considered to be a false detection. This permits to set the detection thresholds automatically for each image.
A stochastic background model H0 needs to be defined, where the structure of interest (faces) is not present and can only arise as an accidental arrangement. In this context, we want to test the presence of a face in a subwindow against the a contrario model where the face is not present. This is equivalent to doing the following hypothesis test:

H0 (null hypothesis) : the subimage does not contain a face
H1 (alternative hypothesis) : the subimage contains a face

A Gaussian distribution of the detection values is assumed for H0 (the distribution of detection values for the nonfaces subwindows is Gaussian). The underlying assumption of the a contrario model is that only a small fraction of sub-windows in any image actually correspond to faces, if any at all. This means that the distribution of detection values for the entire image aligns,roughly, with the distribution of values under the null hypothesis.

3.2 Definition of the Events and tests

Generally, in the a contrario approach, for an event of interest e, e is considered meaningful if its expected occurrence is low under a stochastic background model $H0$. The expectation of an event

is quantified by its number of false alarms (NFA), defined as:

$$\mathrm{NFA}(e) = \mathrm{N}_{\text{test}} \times P_{H0}(e)$$

Where N_{test} represents the number of possible occurrences of e, $P_{H0}(e)$ signifies the probability of event e happening under the $H0$ model. A relatively low NFA indicates that event e is rare under the a contrario model, thus making it meaningful[5].

We can then state the following for our problematic:

- **The event of interest e :** Accepting a sub-image as a face.
- **Number of tests** N_{test} **:** the number of tested subwindows in the image.
- **NFA :** The expected number of false positives in an image.

3.3 Number of False Alarms (NFA)

According to the lecture notes [6], the NFA is valid if it guarantees a bound on the expectation of its number of false alarms under H0. In fact the smaller the NFA, the more unlikely the event e is to be observed by chance in the background model H0; thus, the more meaningful. The a contrario approach prescribes accepting as valid detections the candidates with NFA < ϵ for a predefined value ϵ , in which case the event e is termed ϵ-meaningful. As a result, ϵ gives an a priori estimate of the mean number of false detections under H0.

According to the equation (2.4) formula in [6], the NFA formula is:

$$\mathrm{NFA} = \mathrm{N}_{\text{test}} \times P_{H0}(v_{\text{det}} > \Theta)$$

where *v_det* is the detection value associated to the subimage and θ is the rejection threshold.
θ can be written as a function of μ and σ (the parameters of the Gaussian distribution, they can be estimated from the empirical values of the histogram) : $\theta = \theta_s = \mu + s\sigma$, and where s is a parameter.
Therefore:

$$\begin{aligned} P_{H0}(e) &= P(\text{False positive}) \\ &= P(v_{\text{det}} > \Theta_s | H0 \sim \mathcal{N}(\mu, \sigma^2)) \\ &= \frac{1}{2}\mathrm{erfc}\left(\frac{\Theta_s - \mu}{\sqrt{2}\sigma}\right) = \frac{1}{2}\mathrm{erfc}\left(\frac{s}{\sqrt{2}}\right) \end{aligned}$$

Where the erfc is the complementary error function.
Thus:

$$\mathrm{NFA} = \mathrm{N}_{\text{test}} \times \frac{1}{2}\mathrm{erfc}\left(\frac{s}{\sqrt{2}}\right)$$

Example and analysis of the NFA We can see that θ_s is an adaptive threshold, since it depends on the detection statistics (μ and σ) of the input image. Furthermore, we can observe that the NFA is inversely propostional to s. The bigger we take the parameter s, the less NFA we observe on the image.

A toy example would be to consider the image in Fig 3 with no face in it, where the histogram of detection values is computed with 200 features and that yields μ = 26.18 and σ = 2.16.
With N = 51709 sub-images, a value of **s = 5** yields **NFA = 1.48** and a value of **s = 4** yields **NFA = 163.76**.

The detection threshold can be set such as we guarantee a value of NFA below a predefined upper bound ϵ. This yields a value of θ_s:

$$\Theta = \mu + \sqrt{2}\,\mathrm{erfc}^{-1}\left(\frac{2}{N}\cdot\epsilon\right)\sigma \tag{2}$$

The choice of the value ϵ= 1 is reasonable. It allows for less than one false detection on an image, which is quite tolerable. This can lead however to miss to detect some faces by the detector. Conversly, as ϵ increases more false positives appear but also more faces are detected.

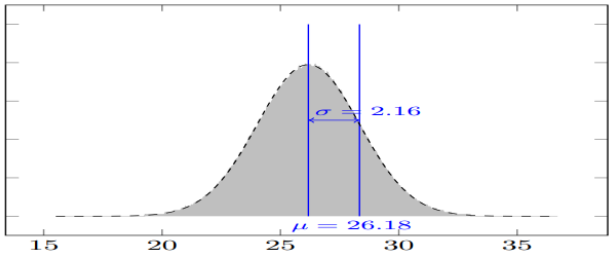


Figure 3: Image with no face and histogram of detected values obtained with 200 features. (From [1])

4 Algorithm for Face Detection and Results

4.1 The algorithm proposed

The algorithm contains mainly 4 steps:

Inputs: Gray-level image I.

1. Get a set S of all the image sub-windows to consider on the image with a total of N sub-window on the image I.
2. Compute the detection values V(S) of all the subwindows and compute mean and standard deviation of these detection values (μ and σ).
3. Compute the detection threshold θ in Equation (2).
4. Loop all over the sub-windows in S and add the sub-window to the detections set D if V(S) $> \theta$

Outputs : The set D of image sub-windows containing a face.

4.2 Non redundant detections

For each true detection, many squares of similar sizes are found by the above algorithm. These detections are centered around the detected face and form a thick frame around it.
To deal with this, in the paper [1] and [2] are proposed two post precessing steps are applied to the sub-windows in the set D:

1. Keep only the stable detections (the stability of a detection is measured in the number of detections in the whole group of detections it represents). This is done by retaining solely the detections that test positive in a mirror version (horizontal flip) of the input image.
2. The stable detections are further simplified. When two square regions overlap significantly (if their intersection area is more than half the size of either detection), we consider them part of the same face. We then keep the detection with the highest score and discard the other.

The exclusion principle The exclusion principle answers the question of how to detect the best rectangle, both explaining and masking the redundant detections where the most meaningful observed structure (the one with smallest NFA) is kept as a valid detection. Then, all the basic elements that were part of that validated group are assigned to it and the remaining candidate structures cannot use them anymore. The NFA of the remaining candidates is re-computed without counting the excluded elements.
We can adapt this to the above algorithm where we don't compute the NFA but rather the adaptive threshold :

- **Initialization:** Mark all sub-windows in the set D containing a face as Available and initialize an empty list L for final detections.
- **Iterative Selection:** Select a candidate sub-window s from D with the highest detection value V(s) that is greater than Theta then add s to the list of detections L and remove it from S.
  Exclude the pixels in s from being used by the remaining sub-windows in D (as each pixel of the image can vote for only one final face detected) and recompute the detection values V(s) for all remaining Available sub-windows in D.
  Reselect the sub-window with the highest detection value.
- **Output:** List of detections L.

4.3 Results

4.3.1 Example following the a contrario model

In this part, we take an example following the a contrario model, an image that doesn't contain a face in it and we apply the algorithm on it (with redundancy suppression).
We evaluate the results for $\epsilon = NFA_{max} = 1$ and $NFA_{max} = 1000$.

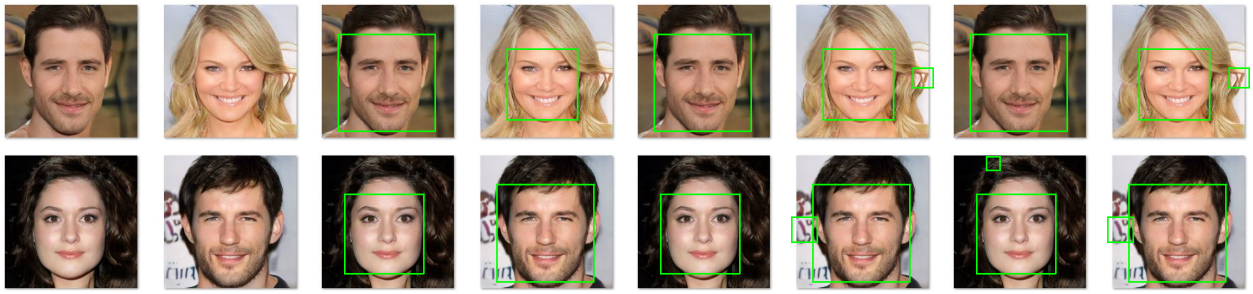
Initial image | Detections with $NFA_{max} = 1$ | Detections with $NFA_{max} = 1000$

Figure 4: Example of an image following the a contrario model H0

We observe that for $NFA_{max} = 1$, no faces are detected on the image, but for $NFA_{max} = 1000$, three false detections appear on the left side of the image.
This is normal as with a higher values of $NFA_{max} = 1$, we have a smaller value of the threshold θ_s allowing more detection values from the histogram, that don't necessarily belong to faces in the image.

4.4 Simple case

Single face detection In this part, we show the results for a simple case, with frontal faces. We run the algorithm on four images from the CelebA HQ Dataset[7] using values of $NFA_{max} = 1$, 100, 1000.



Initial image | $NFA_{max} = 1$ | $NFA_{max} = 100$ | $NFA_{max} = 1000$

Figure 5: Example of frontal face detection

The results are satisfying with $NFA_{max} = 1$. The four faces are detected. In fact, as the classifier was trained with the same set of frontal 24×24 faces used in the original Viola and Jones paper [8], therefore, we get good detection results. We observe as has been said above, that increasing the value of NFA_{max} increases the number of false positives detected in the images.

Multiple face detection We show here a picture from [9] and run the algorithm with different values of NFA_{max}. The results are stable but as the parameter increases, more detections appear, although they are not always correct.

Initial image $NFA_{max} = 0.1$ $NFA_{max} = 1$ $NFA_{max} = 100$

Figure 6: Example of multiple frontal face detection

4.4.1 Complex Examples

Occluded faces In this part, we test the algorithm on a set of occluded faces (by glasses, sunglasses, closed eyes and hidden mouth).
We test with multiple NFA_{max} values. With a value of 1 , already 5 out of the 8 faces are detected. The cases of the faces non detected are harder examples. With a value of $NFA_{max} = 1000$, we have detections on all the faces but of which 2 are considered negative (top one on right and second one on the bottom starting from the left). In fact, a detection is considered positive only if (i) it contains both eyes, (ii) they are located above the center of the detection subwindow, and (iii) the size of the subwindow is less than five times the distance between the eyes.

Initial image $NFA_{max} = 1$ $NFA_{max} = 10$ $NFA_{max} = 1000$

Figure 7: Example of occluded face detection

Multiple face detection Here, we show an example for a more complex multiple face detection case. The picture shows many pedestrains walking.
We can see that with an $NFA_{max} = 1$, only 3 faces are detected in the picture. We can see that as we lower the threshold (by increasing NFA_{max}), more faces are detected in the picture. For $NFA_{max} = 1000$, 10 faces are detected of which 3 don't correspond to faces.
These results can be explained by the fact that the algorithm is trained mainly on frontal faces and that in the picture, mainy of the faces are not frontal and moreover have some parts of them that are hidden.

Initial image $NFA_{max} = 1$ $NFA_{max} = 100$ $NFA_{max} = 1000$

Figure 8: Example of a complex multiple face detection

4.4.2 Case of failure

An example of failure would be a side picture of the face. As we can see, the algorithm is unable of detecting the face for all the values of NFA_{max} and rather generated false detections. This is more a shortcoming of the Viola-Jones Face Detector as no detection value corresponds to a face no matter how low the threshold is set. And also a due to the nature of the training data (mainly frontal faces).

Initial image $NFA_{max} = 1$ $NFA_{max} = 100$ $NFA_{max} = 1000$

Figure 9: Example of occluded face detection

5 Improvements and Extensions

5.1 Extensions

An extention of the algorithm proposed in section 4.1 is to consider a cascade of classifiers (multiple strong classifiers in a sequence) instead of just one. Each classifier takes as input the positive detections from the previous one. The early classifiers in the cascade focus on using a limited set of features to reject obviously non-face objects, while later stages use a more comprehensive set of features to only pass true face detections. And each classifier is trained using negative examples that were mislabeled by the previous classifier, leading to increasingly stringent detection thresholds.
This would yield in improved detection precision and faster computation. This last result is because most sub-windows can be rejected early in the cascade by the simpler classifiers, reducing the number of features that need to be checked in later stages.

Moreover, the same principles might be applied to more recent face detectors by exploring the use of integral channel features trained using faces in various poses/views. This would generate a more robust algorithm mainly to partial occlusion and without restrictions on the pose of the face in the image.

5.2 Comparison with other face detection techniques

The Viola-Jones face detection algorithm[8] is one of the earliest and widely used face detection algorithms.
Another widely used method for face detection is the **Histogram of Oriented Gradients (HOG) combined with the support vector machine (SVM) classifier**. HOG is a feature descriptor used for object detection tasks like face detection. It works by calculating gradients (changes in pixel intensity) in localized areas of the given image and then creating a histogram of the gradient orientations. This histogram is useful for representing the shape and texture of objects.
HOG features provide a detailed representation of object characteristics, then the linear SVM

efficiently classifies regions of the image as containing a face or not based on these features. This face detector is effective for detecting faces with varying orientations and scales in comparison with The Viola-Jones algorithm.

In contrast, these limitations are less present with the use of deep learning models that excel at detecting faces in various poses, scales, and lighting conditions. Deep learning models can handle occlusions and complex backgrounds better than traditional methods.
Of the most famous deep learning algorithms, we have:

- **MTCNN (Multi-task Cascaded Convolutional Networks):** this neural net consists of three stages: a proposal network that generates candidate bounding boxes, then a refinement network that refines these boxes, and finally, a classification network determines if each box contains a face or not. This cascaded approach allows it to efficiently detect faces of different scales in images.
- **YOLO (You Only Look Once):** YOLO uses a single CNN to predict the bounding boxes and class probabilities directly from full images in one evaluation by dividing the image into a grid and predicting the bounding boxes and class probabilities for each grid cell. YOLO is mainly known for its real-time detection applications.
- **RetinaFace:** RetinaFace[10] is a face detection and alignment algorithm that uses a single-stage fully convolutional neural network to detect faces. It also permits to detect the landmarks on a face (like eyes, nose, and mouth) simultaneously. It adopts a multi-task loss function to handle face detection, face alignment, and occlusion prediction in a unified framework. It is highly robust to variations in scale and orientation, and efficient for real-time face detection applications.

In Fig 10, we show the result of face detection using the RetinaFace model. We can see that it effectively detects mostly all faces on the image and detects the landmarks on them, even the faces that are far on the back, side faces and partially occluded ones.

Figure 10: Face detection using RetinaFace

The deep learning models are however more computationally demanding than traditional methods and need large amounts of training data and longer training times compared to traditional methods.

6 Conclusion

Trough this report, we have demonstrated the effective use of the a contrario methodology to enhance the efficacy of the traditional Viola-Jones face detection algorithm. Using the Gaussian distribution as a background model for evaluating the presence of faces within an image and introducing a technique to adjust the detection threshold of a classifier based on a fixed value of NFA to manage false positives more effectively in images A key benefit of this methodology lies in its efficient time complexity compared to the original algorithm while providing similar detection rates. However, the performance of the algorithm can be additionnally improved to be robust to different orientations and scales in the image as well as to occlusions that can occur in order to compete with the advancement in deep learning models.

7 Bibliography

[1] JL Lisani, S Ramis, FJ Perales, "A contrario detection of faces: a case example", SIAM Journal on Imaging Sciences 10 (4), 2091-2118, 2017.

[2] JL Lisani, S Ramis, "A Contrario Detection of Faces with a Short Cascade of Classifiers", Image Processing On Line 9, 269-290, 2019, $https://www.ipol.im/pub/art/2019/272/article_lr.pdf$

[3] P. Viola and M. Jones, Rapid object detection using a boosted cascade of simple features, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001, pp. 511–518, https://doi.org/10.1109/CVPR.2001.990517.

[4]Y.-Q. Wang, An Analysis of the Viola-Jones Face Detection Algorithm, Image Processing On Line, 4 (2014), pp. 128–148, http://dx.doi.org/10.5201/ipol.2014.104.

[5] Boshra Rajaei, and Rafael Grompone von Gioi, Gestaltic Grouping of Line Segments, Image Processing On Line, 8 (2018), pp. 37–50. https://doi.org/10.5201/ipol.2018.194

[6] Morel, J.-M., and Grompone von Gioi, R. Detection Theory and Industrial Applications. Lecture Notes.

[7]https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

[8] P. Viola and M. Jones, Rapid object detection using a boosted cascade of simple features, in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, 2001, pp. 511–518. https://doi.org/10.1109/CVPR.2001.990517.

[9]https://www.publicdomainpictures.net/en/view-image.php?image=280567picture=butch-cassidy-vintage-photo

[10] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, Stefanos Zafeiriou, RetinaFace: Single-stage Dense Face Localisation in the Wild, https://arxiv.org/pdf/1905.00641.pdf.