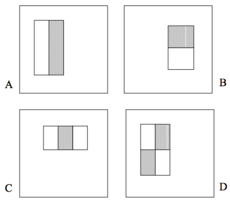
Joseph Bai: jb16450
Oliver Goldstein: og14775

Image Processing Report

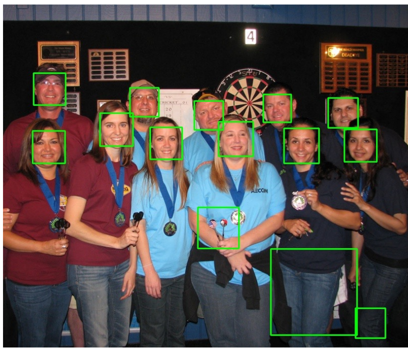
Introduction

The Viola-Jones method is a landmark paper, which demonstrates a rapid object detection, using the concept of an integral image, the adaBoost algorithm and a degenerate tree of classifiers, called a cascade. Iterations of the algorithm over weak classifiers, which have the smallest error rates when tested on a training set, collectively contribute to a strong classifier. The Viola Jones method works with Haar features (see right), where the differences between light and dark regions act as a feature. This can be calculated for each Haar region in 4 operations, regardless of the region size. There are a few different region patterns used as Fig 1.0 displays.

Fig 1.0 Haar features.



Dart 5.jpg



Dart 4.jpg

Dart 14.jpg

Dart 13.jpg

Testing the classifier on faces

For regions with bright and dark patterns similar to those of faces, (e.g. bridge of nose vs eyes) the classifier is more likely to detect these as faces. For instance, in Dart 5.jpg there exists a shape similar to a dark set of eyes by the woman arm in dark blue. This was correspondingly picked up as a face.

Dart 15.jpg

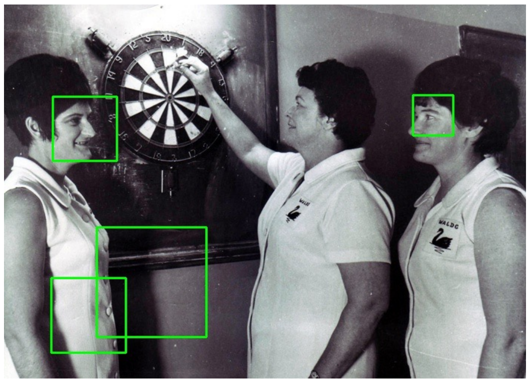


Table 1.0 details the positive and negative counts of the given classifier on images Dart 5.jpg and Dart 15.jpg. The true negative counts have been ignored due to the complexity of calculating them. However, the true negative rate is extremely good as out of the possible number of regions, numbering well over 100,000, most regions have been disregarded.

Technical analysis

The true positive rate for dart 5 appears to be 100% and for dart 15 appears to be 33%. The true positive rate is not necessarily a good reflection of the utility of the classifier, as the trivial classifier that always predicts true has a true positive rate of 100%. In addition, the decision to assess the classifier as correct when it predicts a subsection of a face, or a rotated face, as positive, can be difficult to assess. In the case of dart 15, we classified this as a false positive, as the whole face was not detected. A final difficulty in classifying faces is to provide a definition of a true negative. There is no objective method to describe the lack of a face to a computer, as it is an ambiguous empirically determined concept and so must be hand labelled by humans, which is also difficult to get an objective answer to.

Table 1.0 Dart 5

	Predicted +
Actual +	11
Actual -	3
Total	14

Table 1.0 Dart 15

	Predicted +
Actual +	1
Actual -	3
Total	4

The true positive rate is the proportion of predicted faces, which are, actually faces divided by the actual number of faces. In order to make sure that it is not the trivial classifier mentioned above, precision and more importantly the F1 score can be used. The F1 score includes precision and recall.

Formula 1.0

$$2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

Given ground truth (i.e. the location of the faces), the precision can be calculated by dividing the actual faces that were predicted by the total number of predictions we made. The recall can be calculated by looking at the cardinality of the set of correctly predicted faces and dividing that by the total number of faces that existed in the given image. Coming up with a meaningful set of rules is a problem in computer vision. Formula 1.0 depicts the calculation for the F1 score we used. The F1 Score must be calculated objectively and consistently. Therefore, if the intersection of the area of the boxes and the ground-truth exceed 65% and proceed 30%, numbers we intuitively felt were appropriate, the box is classified as a true positive.

Explaining the dartboard classifier

Stage 1 terminates with a TPR and FPR of 1.0. The classifier starts by classifying everything as a face, implying non faces are classified as faces leading to a false positive rate of 1.0.

After stage 1, the classifier has produced a meaningful set of features that help distinguish faces from non faces. The false positive rate now drops by a factor of over 40 (it gets better at removing non faces). The second training iteration drops the false positive rate by a further factor of 23. This seems to be approaching a (potentially local) optimal solution - the classifier may be overfitting on the training data. An overfitted classifier, can trivially achieve a 100% TPR and a 0% FPR.

Fig 1.2 shows the number of false positives are far greater than the number of true positives. The largest number of false positives occurs in the image Dart 14, which has the highest resolution and size, with a high dynamic range in terms of intensity. This variation in intensity is analogous to the pattern of the alternating bright and dark regions that occur within a dartboard. This may explain why many false positives exist here. This image has one of the lowest F1 scores, but not the lowest, as it has two dartboards rather than one, so it is relatively more precise. Dart 12 has a low resolution and is a relatively small image, with the smallest difference between the predicted positives and the actual positives. It has the highest F1 score. A disadvantage of the F1 score is that it does not explain image complexity or size.

The images on the right convey the robustness of the classifier with respect to occlusion (Dart 11), and perspective (Dart 10) of the dartboards, even if there are also many false positives. The F1 score takes into account the number of false positives versus the number of true positives, recall and precision making it useful for measuring performance of a classifier.

Fig 1.1 - Log graph: training the classifier.

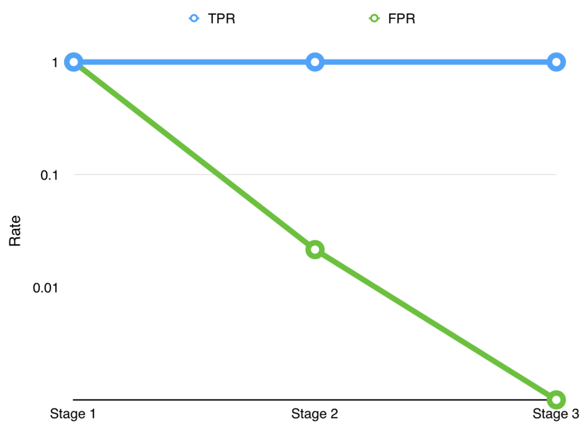
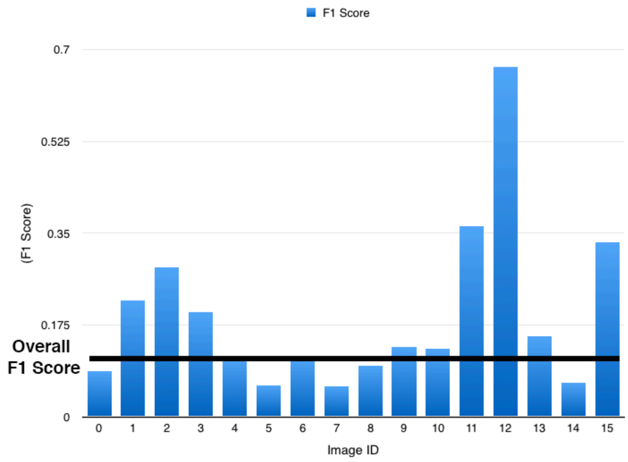
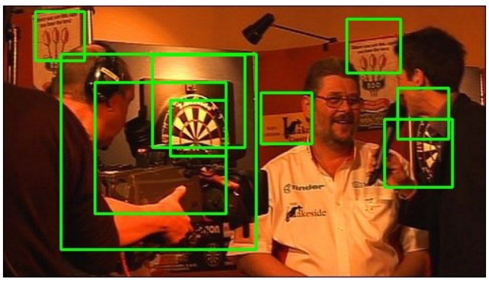


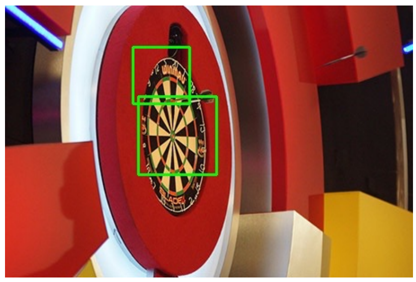
Fig 1.2 - Results of the classifier on test data.



Dart 11.jpg

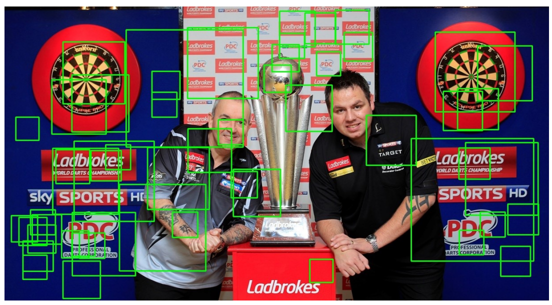


Dart 12.jpg



Dart 10.jpg

Dart 14.jpg



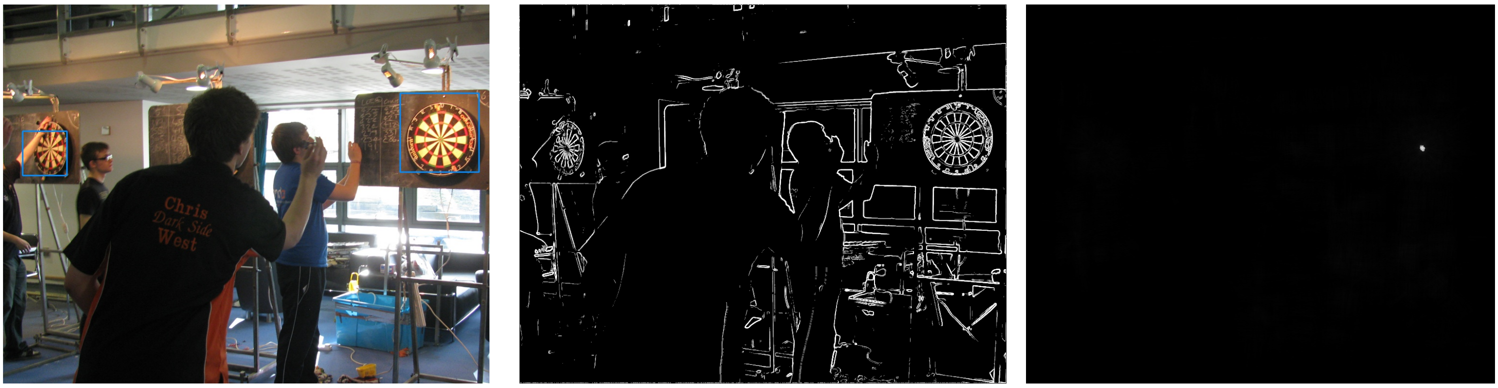
Comparing graphs

The false positive rate in Fig 1.1 is defined differently than the false positive rate in Fig 1.2. In Fig 1.1, the FPR uses only one detection window, with at most one dartboard per image, the true negatives are also on an image level. The FPR is then calculated over all images. There are 1000 true negatives used, to achieve such a low FPR, the number of false positives must be around 1. In 1.2, if we apply a similar thinking to the images we have and look at a false positive on the image level, we achieve orders of magnitude more false positives. In addition in 1.1, the classifier must only give a binary response in relation to the existence of an image. In order to codify a set of rules, we had to come up with a threshold to determine whether a classifier was correct or not. In the method that we devised we used 16-fold stratified cross validation in order to find a threshold value, for which, if the percentage of the area of the predicted dartboard exceeded the area of the true bounding box, which we hard coded for each dartboard, then it was a true positive. After doing this, it happened to be the case that all of the true positives, that we could visually identify, were correctly classified as true positives by the codified rule based method we implemented. With this in mind; The false positive rate for our images is far greater than the training data. This points to the potential of the classifier to be overfitting on the training data or that the image used to train our classifier is not representative of the test data that we receive i.e. it does not generalise. Given time, we seek to potentially add training data that is more reflective of the true data we test on.

Dart 8 shows the merits of the system as the dartboard on the left is not picked up by the hough circle detector, yet it is picked up by the hough lines detection process highlighting the benefits of a parallel approach. On the right, both the intersection and the circle were very clearly picked up by the hough spaces. In particular for the circle, the peak in the image is very clear. Dart 11 shows shortcomings, not only is the occluded left circle very noisy in the hough space, resulting in an inaccurate detection, the rightmost occluded dartboard is not detected. In addition, minor distortion occurs at the top of the hough space.

Bounding Boxes　Thresholded Magnitude　Dart 11　Hough Space

Dart 8



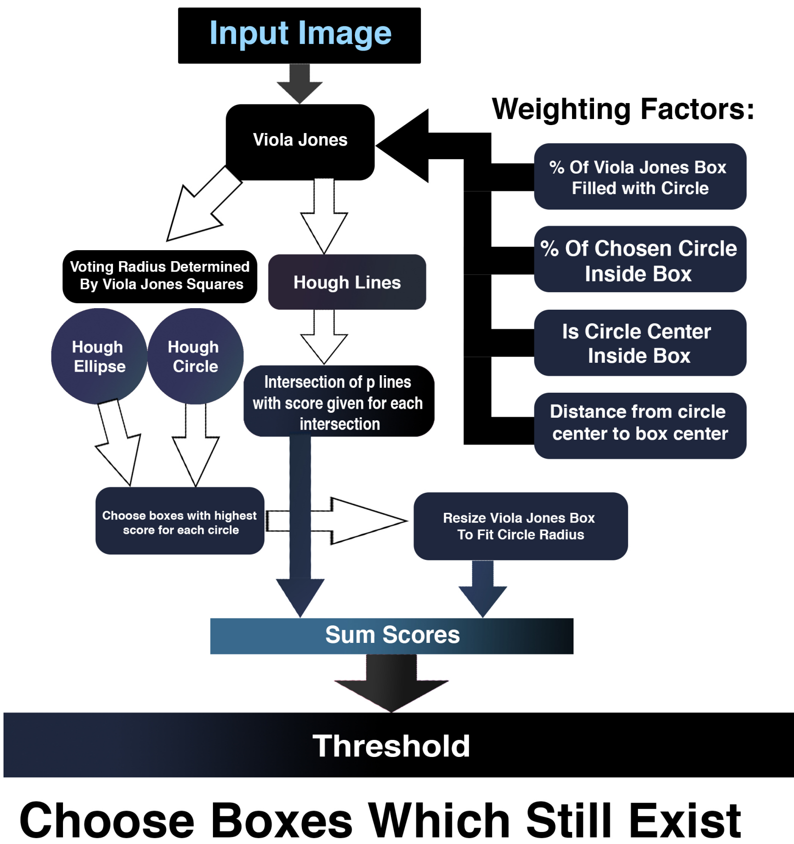
The overall F1 Score is 0.97561 and the overall recall and precision values are 1 and 0.9524, respectively, implying a correspondingly high accuracy. The ratios were the same as Subtask 2 to create a consistent and fair way of judging the overall performance of the classifier. This shows a vast improvement on the above. Testing on additional data revealed that, surprisingly it overfitted less than we thought initially, with similar F1 scores on data similar to those given to us.

Key merits of the implementation include:
1) Can detect vertical ellipses by allowing more votes on gradient with a direction that is towards image top/bottom.
2) Parameterising the voting radius of the hough space for circles to be between at most slightly over half of the largest viola jones box in the image and just under half of the smallest.
3) Relative thresholding for the circle hough space e.g. values less than 50% of the highest value are reduced massively, removing the effect of noise.

Key shortcomings include:
1) A fixed distance to prevent circles being too near.
2) Two or more dartboards that share an intersection do not cancel each other out. (highest score chosen).
3) The hough space for circles can distort at image edges picking up non existent circles at edges.
4) Many finely tuned parameters exist in the model which may lead to overfitting.

Rationale behind the combination:
1) Pruning boxes is fruitful, as Viola Jones always captures the dartboards (and many false positives). Circles and intersections are not necessarily always captured. Conditioning on either of them incurs lost true positives, so a parallel linear combination of different weights is most appropriate.
2) For each circle, we choose the relevant box with the highest score (assuming it passes the threshold), which results in the most appropriate viola jones detection.
3) Resizing improves precision as circle radius is often more accurate measure of dartboard size.

Fig 1.3 Flow Diagram



A note on subtask 4:

Performed failed extensions: We attempted applying a median blur to the hough space in order to remove the effect of salt and pepper noise in the hough space itself. This distorted the correct radius and circle centre rendering this approach ineffective. In addition we combined viola jones classifiers together, especially in light of the idea that the initial dart.vec would not be representative of all real world data we test on. We generated many different classifiers with bigger dartboard images, and some smaller; for instance one contained just the centre of the dartboard and some just half of a dartboard, in order to account for noise in the outermost regions, but each of them yielded worse results overall than the dart.vec file that we were initially provided with. In addition they often agreed on false positive regions, making the process difficult. I suspect given a much larger amount of test data, more classifiers would be better in the long run than just one.

Potential extensions:
Given time, we could have combined many Viola Jones classifiers into one, potentially improving the overall accuracy of the classifier on different image sets. This could have been done in such a way that reduces the potential for noise by allowing more than one classifier to agree with each other in order to classify a dartboard as a true positive. This would remove the number of false positives. In addition, HOG or SVM classifiers could have been used.

References: 1) Wikipedia F1 score 2) Original viola jones paper.