Scooby-Doo Face Detection and Classification

Contents

1 Problem Description

Given a set of 4000 training images containing characters from *Scooby-Doo* (1000 images per character), the goal was to detect and classify faces belonging to **Daphne, Fred, Shaggy, and Velma**. The dataset also contained a small number of unknown faces, which were treated as background.

The task consisted of:

- detecting all faces in a test image;
- classifying each detected face into one of the known characters.

2 Face Detection

2.1 Initial Attempts

Face detection was initially implemented using a **sliding window** approach combined with **Histogram of Oriented Gradients (HOG)** features. In addition to positive examples, negative samples were generated from random patches that did not overlap significantly with faces.

One early mistake was attempting to keep the number of positive and negative samples similar. In practice, there should be significantly more negative samples, as non-face patches vastly outnumber face patches in an image.

Initial experiments achieved an average precision (AP) between **40% and 50%**. A window size of **69×69** was chosen, as it represented the mean face size in the dataset. The window dimension for an image of size $h \times w$ was defined as:

$$\sqrt{h \cdot w}$$

Other heuristics such as $\frac{(\min + \max)}{2}$ were also tested, with similar results. The motivation behind using the mean window size was to minimize resizing artifacts that appear when windows are too small or too large.

All experiments used image pyramid scaling to detect faces at multiple resolutions. Non-maximal suppression (NMS) was applied to reduce overlapping detections.

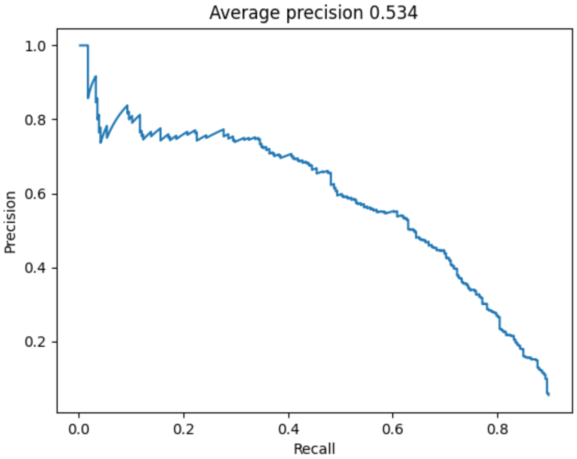

Figure 1: Detection results for 69 × 69 windows, HOG cell size 6

Experiments with HOG cell sizes of 4 and 8 produced similar results. Increasing the number of pyramid scales significantly improved recall but reduced precision.

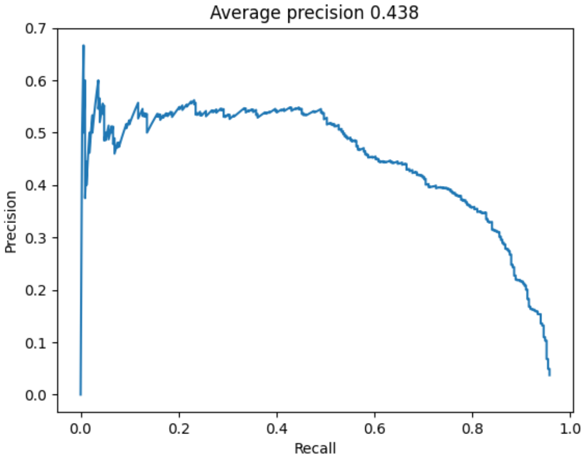

Figure 2: Detection results using additional pyramid scales

2.2 Hard Mining

Hard negative mining was introduced by re-running the detector on training images and selecting false positives that overlapped less than 0.3 IoU with any ground-truth face. Only the top-k detections (highest confidence scores) were kept.

Better results were obtained by lowering the overlap threshold. Initially, overlaps between 0.15 and 0.3 were retained to include near-face examples, but this appeared to confuse the model. The final threshold was set to **0–0.25 IoU**.

Unfortunately, early hard-mining experiments were implemented incorrectly and realized too late, so intermediate results are not shown.

In addition to hard negatives, **hard positive mining** was also introduced by selecting false negatives with IoU greater than 0.6. As before, keeping a balanced number of hard positives and negatives proved suboptimal.

2.3 Multiple Window Sizes, Multiple Models

Since mining initially appeared ineffective, a multi-model approach was adopted. Two separate detectors were trained:

- a small-face detector (windows between 18 × 18 and 64 × 64);
- a large-face detector (windows between 72 × 72 and 170 × 170).

The corresponding window sizes were 36 × 36 and 96 × 96. At this stage, pyramid scaling was corrected by progressively resizing the image using fixed scale factors:

- downscaling: 0.9;
- upscaling: 1.1 and 1.2.

The positive-to-negative ratios were adjusted to 1:2 for the small detector and 1:3 for the large one. Mining was also randomized across training images to avoid repeatedly selecting the same samples.

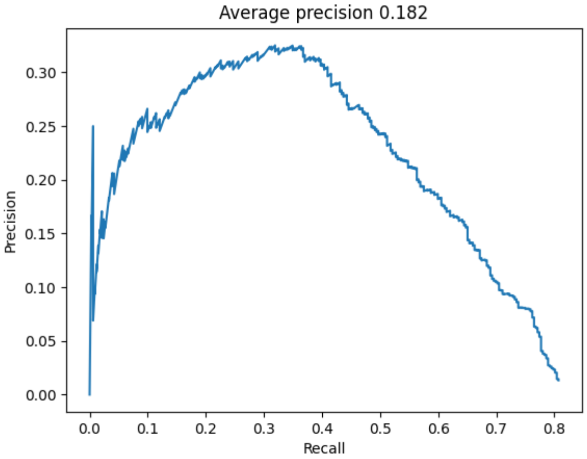

Figure 3: Initial results for the small-face detector

After four mining iterations, the small-face detector achieved its best performance:

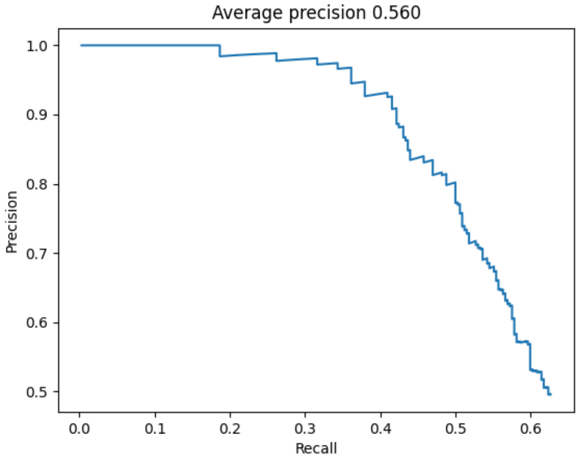

Figure 4: Best results for the small-face detector

The large-face detector used a HOG cell size of 8:

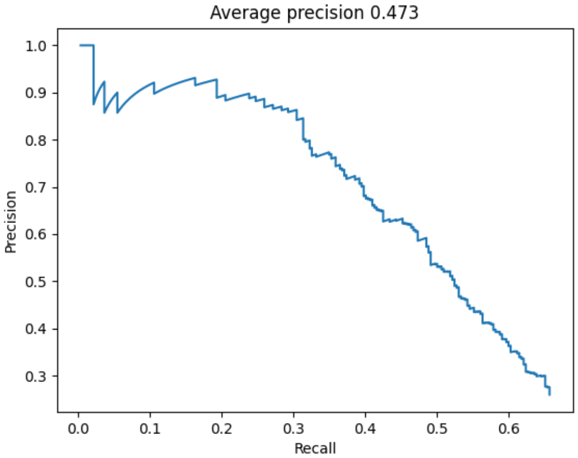

Figure 5: Initial results for the large-face detector

After seven mining iterations:

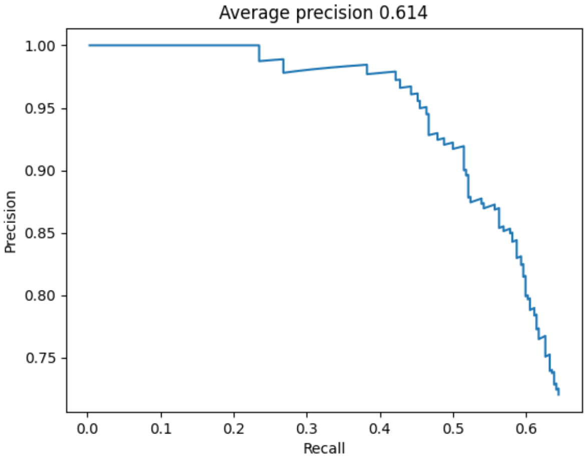

Figure 6: Best results for the large-face detector

After merging detections from both models and applying NMS (with score normalization), the final detection AP reached **86%**:

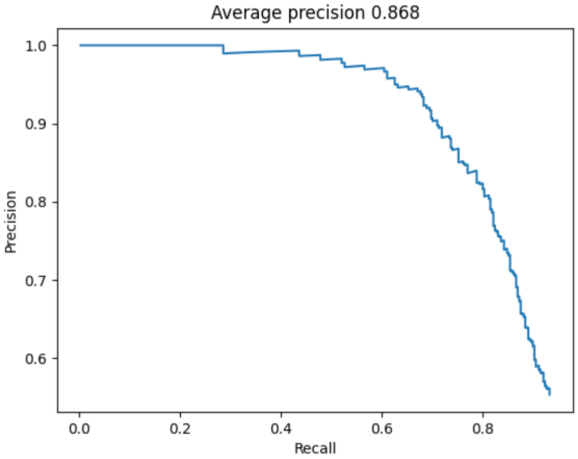

Figure 7: Merged detections from both models

Without NMS, the AP dropped to approximately **80%**.

2.4 Other Failed Attempts

Local Binary Patterns (LBP) were tested as additional descriptors alongside HOG. This resulted in very poor performance (approximately 2% AP), likely due to insufficient data or implementation issues.

CLAHE (Contrast Limited Adaptive Histogram Equalization) was also tested to enhance facial details, but this reduced AP to 33% on the baseline model and was abandoned.

Other experiments included tuning positive-to-negative ratios and HOG cell sizes. Larger window sizes generally benefited from larger cell dimensions.

2.5 Efficiency Considerations

Efficiency improvements included:

- computing HOG descriptors once per image and sliding over the feature map;
- caching HOG descriptors to speed up mining iterations.

Caching reduced processing time per image from approximately 5 seconds to 1 second. Hard-mined samples were also saved for potential reuse.

3 Face Classification

3.1 Multi-Model Approach

Following the success of multi-model detection, a similar strategy was tested for classification. Initial results (without NMS) are shown below:

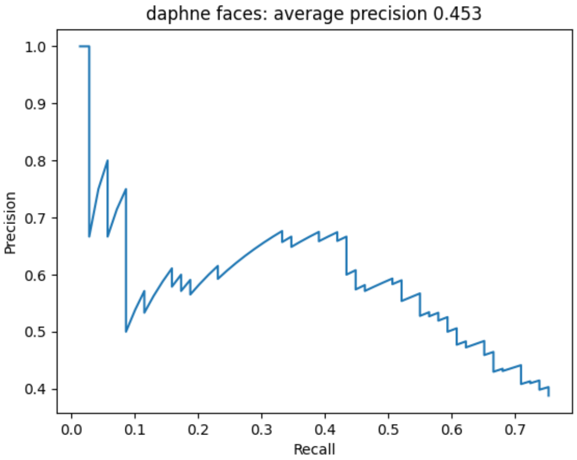

Figure 8: Two-model classification results: Daphne

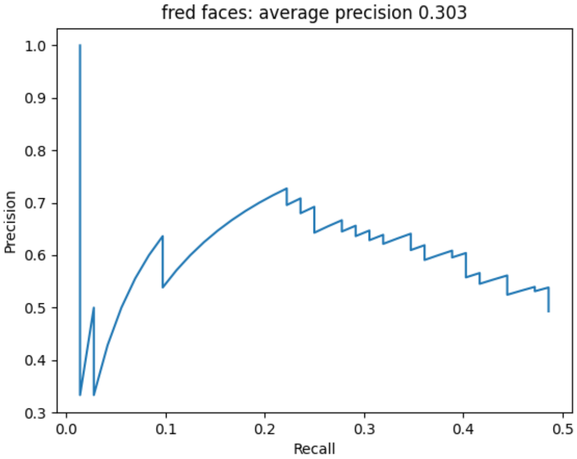

Figure 9: Two-model classification results: Fred

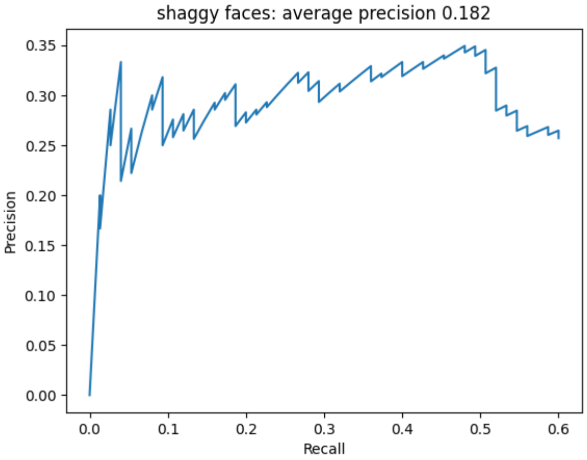

Figure 10: Two-model classification results: Shaggy

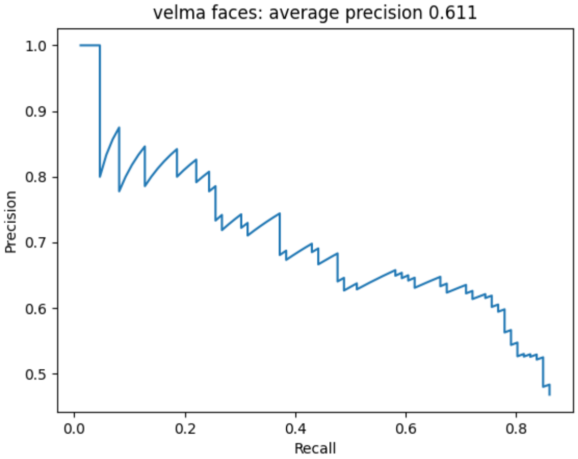

Figure 11: Two-model classification results: Velma

Female characters consistently proved easier to classify. Increasing training data slightly improved results, suggesting that a single model might be more effective under data constraints.

3.2 Single-Model Approach

A single classifier was trained on window sizes between 18 × 18 and 170 × 170, using a 64 × 64 window and HOG cell size 6.

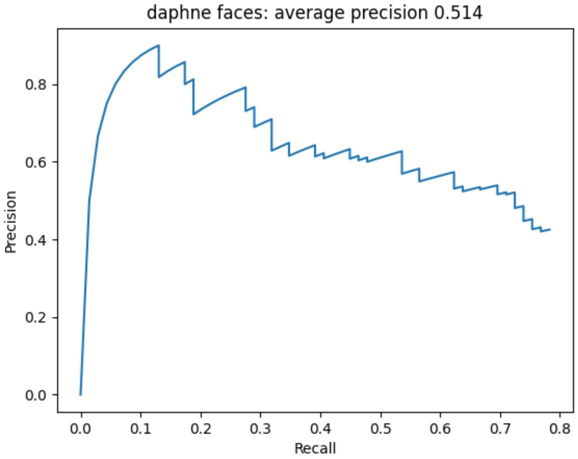

Figure 12: Single-model classification results: Daphne

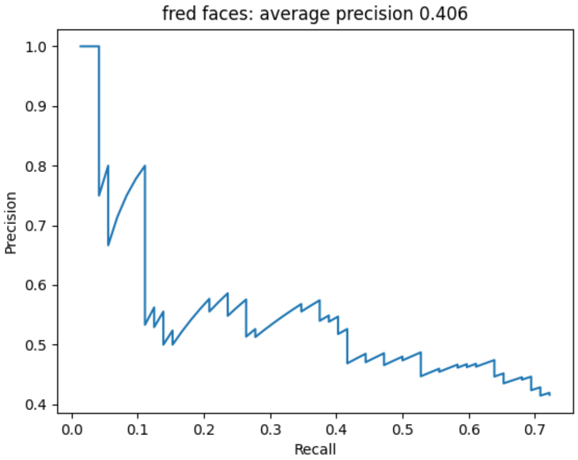

Figure 13: Single-model classification results: Fred

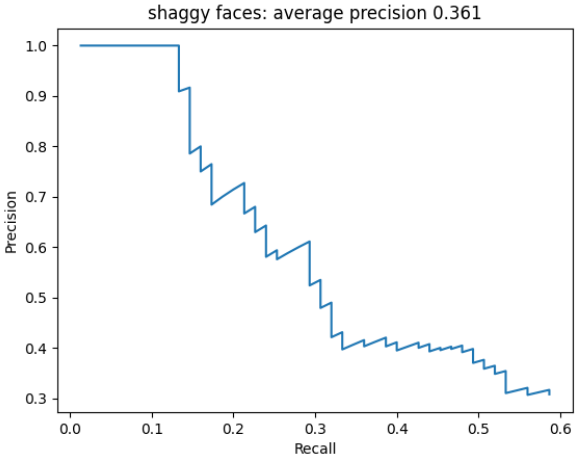

Figure 14: Single-model classification results: Shaggy

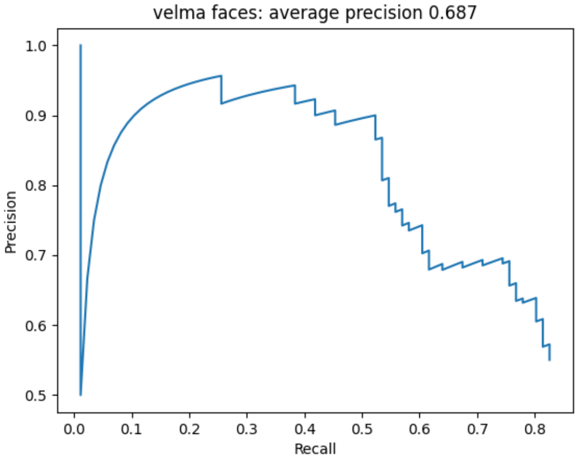

Figure 15: Single-model classification results: Velma

Other parameters such as larger windows or different HOG cell sizes were tested but did not yield significant improvements.

4 Pre-trained Networks (YOLO)

A YOLOv8-nano model was fine-tuned on the dataset. After 10 epochs, the following results were obtained:

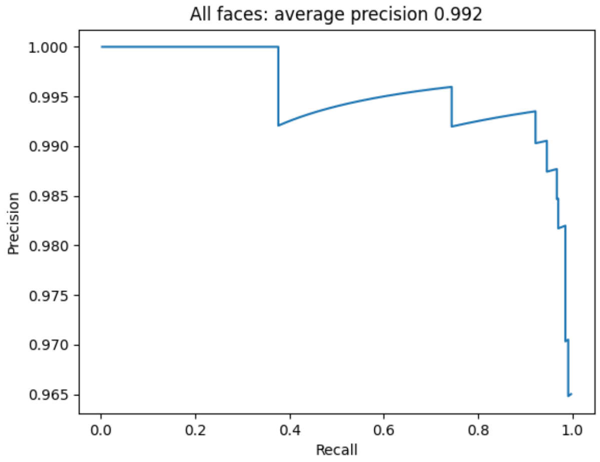

Figure 16: YOLO face detection results

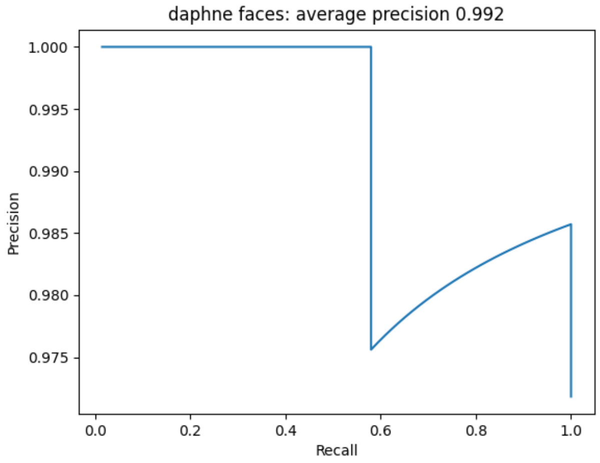

Figure 17: YOLO classification results: Daphne

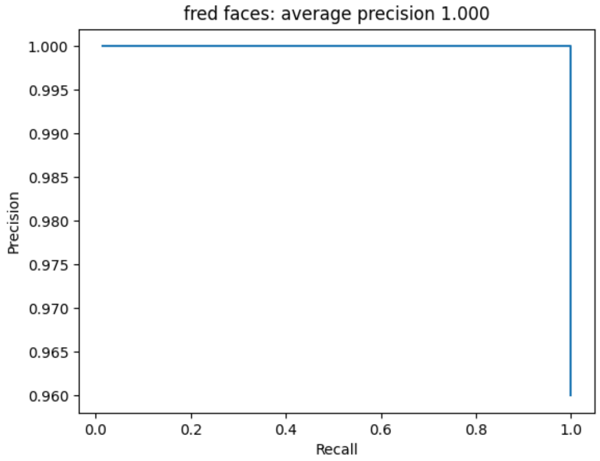

Figure 18: YOLO classification results: Fred

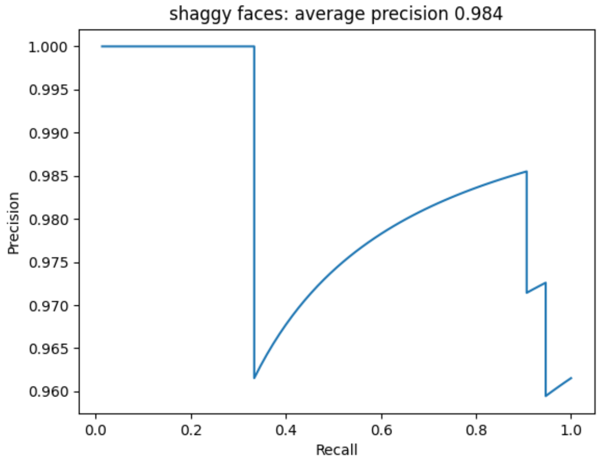

Figure 19: YOLO classification results: Shaggy

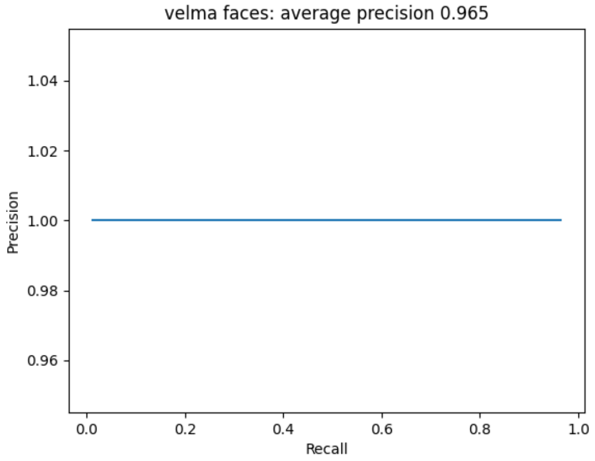

Figure 20: YOLO classification results: Velma

Interestingly, Velma achieved the lowest AP among the characters, although overall performance remained excellent.

5 Conclusion

- **Best classical face detection**: 86% AP using two specialized models.
- **Best classical face classification**: 49% mean AP using a single global model.
- **YOLO face detection**: 99.2% AP.
- **YOLO face classification**: 98.5% mean AP.