Target Detection Trade Study

Tiger Rescue

Table of Contents

[**Background**](#_93b8mtep3mca) **3**

[**Models**](#_n0g46mt1r30g) **3**

[**Assessment Criteria**](#_gko5q434merd) **3**

[**Procedure**](#_oukwaqjcd3lf) **4**

[**Results**](#_38vuv9xv7i6l) **4**

[Quantitative Criteria](#_w9itajq6z0fd) 4

[Qualitative Criteria](#_rj1w24u42cc4) 5

[HOG Default People Detector](#_a1104lk70zpk) 5

[Cascade HAAR Full Body](#_lp33p0sxgmin) 5

[Caffe Okutama-Action Pedestrian](#_tqf1rmf7jx2d) 5

[**Results Analysis**](#_z7hxpqhztm5s) **5**

[**Initial Conclusion**](#_hkafn3i289ue) **6**

[**Final Conclusion**](#_qn4xhay87yc) **6**

## 

## Background

The system must be able to detect human targets in real-time from the drone’s streamed video feed. To accomplish this, we needed to investigate candidate Computer Vision (CV) models and assess their performance. As part of our area coverage approach, the selected target detection model must perform well at high altitudes and steep gimbal angles, and have a reasonable processing speed and footprint.

## Models

The following models were selected for assessment based on ease of use and established effectiveness:

* [HOG](https://en.wikipedia.org/wiki/Histogram_of_oriented_gradients) Default People Detector
  + Model: Embedded in OpenCV
* [Cascade](https://en.wikipedia.org/wiki/Cascading_classifiers) [HAAR](https://en.wikipedia.org/wiki/Haar-like_feature) Full Body
  + Model: <https://raw.githubusercontent.com/opencv/opencv/master/data/haarcascades/haarcascade_fullbody.xml>
* [Caffe](https://caffe.berkeleyvision.org/) [Okutama-Action](http://okutama-action.org/) Pedestrian
  + Model: <https://drive.google.com/drive/folders/0BydaU2Imk1zjVXZBWE9yTWlQdDA>

Java OpenCV was leveraged to run each of these models. Only CPU-based implementations were explored due to hardware constraints.

## Assessment Criteria

Quantitative criteria:

* Processing frames per second
* CPU usage (%)

Qualitative criteria:

* Detection accuracy
  + Rate
    - Missed detections
    - False positives
  + Factors
    - Angle / altitude / distance
    - Contrast

## Procedure

Three ~7 minute videos were recorded at Mendon Ponds Park from a Parrot ANAFI drone. These videos included a variety of altitudes - ranging from 25ft to 400ft - and gimbal angles - ranging from 0° to 90°. Subjects were recorded in both a snowy landscape and an asphalt parking lot to provide a range of contrast. The subjects wore a variety of clothing colors, again to provide a range of contrast.

The implementations of the candidate models were run against the entirety of each assessment video. For each model/video combination, the quantitative metrics were first collected as averages with detection results being discarded at each frame. Next, the video was run through the model again, but this time with detection rectangles being drawn to the frame and output to a new video file. For the models that provided confidence levels, the rectangles were colored based on the provided confidence level - red for low (15-30%), yellow for medium (31-70%), and green for high (71-100%). In some cases the model parameters needed to be adjusted based on the detection results. When this occurred, the assessment criteria were re-collected for each video and the previous results were discarded.

The detection result output videos were then watched to evaluate the qualitative criteria.

****Figure 1: Example of detection output

## Results

### Quantitative Criteria

These metrics were collected on an Intel Core i7-9750H CPU @ 2.60GHz.

|  |  |  |
| --- | --- | --- |
| **Model** | **Processing FPS (avg)** | **CPU % (avg)** |
| HOG Default People Detector | 8.01 | ~40 |
| Cascade HAAR Full Body | 2.83 | ~60 |
| Caffe Okutama-Action Pedestrian | 0.68 | ~46 |

### Qualitative Criteria

#### HOG Default People Detector

This model was assessed to perform well at low altitudes (<100 feet), low-medium gimbal angles (0-60°), and small distances to targets. In these conditions for the varying contrast scenarios, targets generally maintained high confidence levels. False positives were overall uncommon and usually at a low confidence level. At high altitudes (100+ feet) and steep gimbal angles (60-90°), the model gave zero detections with the exception of occasional false positives.

#### Cascade HAAR Full Body

This model was assessed to generally perform well at any altitude up to ~300ft, low to steep gimbal angles (0-70°), and most target distances. However, this was observed to come with the cost of (comparably) many false positives. Some missed detections were observed in low-contrast scenarios (e.g. yellow shirt with snow background). Unfortunately, the model does not provide direct access to percentage confidence levels as the others did, so confidence of detections was not assessed.

#### Caffe Okutama-Action Pedestrian

This model was assessed to perform best at low altitudes (<50ft), low angles (0-30°), and small target distances. However, even in these conditions, valid detections never exceeded a medium confidence level. Outside of these conditions, valid detections were rare. Additionally, sustained false positives of all confidence levels were fairly common across all scenarios.

## Results Analysis

Our assessment results show that each model comes with tradeoffs, and that no single model can be applied to our desired high-altitude, high-angle viewing scenario.

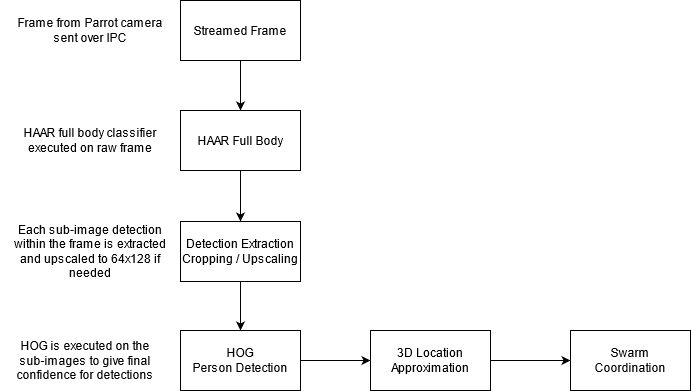
The HOG Default People Detector and Caffe Okutama-Action Pedestrian models were found to perform poorly at high altitudes. We explored upscaling the full frames to try to compensate for this. In the case of HOG Default People Detector, the images needed to be upscaled so that the targets filled most of a 64x128 box (the unconfigurable detection window size). However, this approach was found to introduce an unreasonable performance impact for both models.

The Cascade HAAR Full Body model was found to perform reasonably well in our desired viewing scenario, with the main drawback being frequent false positives.

## Initial Conclusion

We concluded that our CV pipeline should consist of two of the assessed models: Cascade HAAR Full Body and HOG Default People Detector. These models will operate in stages - not concurrently. First, the HAAR model will be run on the full, unmodified frame. Then, each detection sub-image will be extracted from the frame. These sub-images will be scaled to fit the HOG detection window size (64x128) when needed. The HOG model will then be run independently on each of these sub-images to provide a final confidence level for the detections.

By combining these models in this way, we are able to confidently perform detections in most viewing scenarios at a reasonable processing speed.

  
Figure 2: Proposed CV Detection Flow

## Final Conclusion

Once the Rochester snow melted, we found that our initial proposed solution did not perform as expected in a grass field. We determined that the models involved in that solution must have greatly benefited from the contrast of a snow background.

Additional analysis footage in a grass field was collected to briefly re-assess each of the models. It was found that the previously poor-performing Okutama-Action model performed *very* well with the contrast of a grass field, with a false positive rate of about ~30% when using a 50% confidence threshold. While the processing FPS of this model is significantly lower than the others, in our analysis footage we found that the target was always visible long enough to be detected multiple times - alleviating the risk that the model won’t have a chance to see them.

Given these observations, we changed our target selection solution to consist only of the Okutama-Action model to support the more probable use case of a non-snowy environment. If support for snow environments is desired, the system would need to take the environment type as an input and select a CV model accordingly.