Machine Learning Options for Magnificent Frigate Bird Identification and Counting using CCTV

Ellen Xiao

Mark Reynolds, Du Huynh
Computer Science and Software Engineering
The University of Western Australia

Sandip Deshpande, Andy Watt
CEED Client: Woodside Energy Technologies Pty Ltd

Abstract

Magnificent Frigate Birds have been nesting on unmanned offshore platforms. The birds resting and roosting on the platforms pose significant operational risks, especially for helicopters. A general increase in bird population has also worsened the situation. To identify mitigation strategies, the first step is to understand the behavioural pattern of the birds. In this project, the aim is to utilise the available CCTV footage from the concerned platforms to identify machine learning techniques for Magnificent Frigate Bird detection and counting. The methodology of this project includes 1) identification and comparison of the state-of-the-art approaches to detection and counting problems; 2) frame extraction from videos and manual data labelling, and 3) fine-tuning the model to optimize the performance based on the accuracy of the bird counts. The deliverables of this project include a report, labelled data, source code, a trained model, and preliminary insights in the collected bird activity data. With this knowledge, bird dissuasion and relocation mechanism can be designed and implemented. This will reduce the likelihood of bird-related incidents, thereby improving safety for operations and personnel on the platforms.

1. Introduction

1.1 Problem Statement

Magnificent Frigate Birds (hereafter referred to as birds) shown in Figure 1 are large seabirds known for their impressive wingspan, reaching up to 2.3 metres. The birds are primarily black, with females and males exhibiting distinct characteristics. Females have a white chest and dark head, while breeding males are entirely black except for a bright red throat pouch, which may not always be visible (Cornell Lab of Ornithology, 2017).

The primary site affected by bird activity includes five Wellhead Protector Platforms (WPPs), which are structures specifically designed to protect wellheads from environmental hazards and operational damage. Several of these platforms are of particular concern, as birds frequently rest and roost on them. Figure 1 shows a snapshot of the CCTV footage recorded at one of the helidecks. Bird activity significantly increases the risk to helicopter operations, particularly during critical and time-sensitive tasks such as medical evacuations (Medivac). The presence of birds raises the risks of bird-strike incidents, potentially leading to catastrophic outcomes, including loss of human lives.

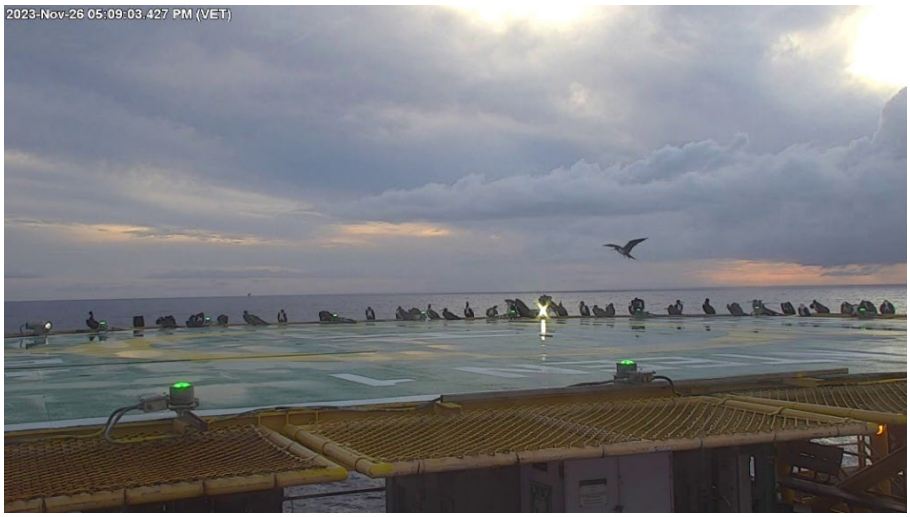

Figure 1 A frame extracted from the CCTV footage showing bird activity on the helideck of an offshore platform.

As a result, knowledge of the distribution and behaviour of birds is vital for the design and evaluation of mitigation strategies. Moreover, this understanding can help raise helicopter operator's awareness of bird activities, leading to more effective and safer flight planning.

1.2 Background Information

To improve the safety of operations, the health of personnel, and environmental outcomes for birds, mitigation plans are in development. However, designing and implementing strategies presents risks and constraints, and data collection has not been systematic. Currently, there is no standardised bird detection and counting strategy; the identification of bird numbers and behaviour relies solely on human observation and interpretation. As a result, analysing the effectiveness of mitigation strategies and making targeted improvements and enhancements remains challenging.

As the images for detection are sourced from CCTV footage covering the helideck, the birds appear quite small within the images. On average, each bird can be enclosed within a 45×45 pixel bounding box. Therefore, the specific problem falls under the category of small object detection, a challenge in computer vision and machine learning. Recent advancements in these fields have shown promise in addressing such challenges. By leveraging these technologies, it is possible to develop automated systems for bird detection and counting, significantly enhancing the accuracy and efficiency of data collection with minimum human intervention. The problem of bird detection and counting can be addressed using two primary approaches: density map estimation and object detection.

1.2.1 Density Map Estimation

Density map estimation algorithms are primarily utilized for estimating the number of humans in surveillance video frames. According to the research by Lempitsky and Zisserman (2010), instead of focusing on detecting and localizing individual subject instances, a density map estimation algorithm reframes the problem by estimating image density over a given region and predicting the total count from that density map. This approach is particularly advantageous in highly congested scenes where the subjects are severely occluded.

1.2.2 Two-stage Object Detectors

In two-stage object detectors, the first stage generates region proposals, extracting object regions independent of their classes. The second stage computes Convolutional Neural Network (CNN) features on these regions and classifies them. This approach was first proposed by Girshick et al. in the original R-CNN in 2014, using a selective search algorithm to extract around 2000 region proposals. However, R-CNN was computationally inefficient, as each proposal had to be processed separately through a CNN.

Fast R-CNN, introduced by Girshick in 2015, integrated the region proposal and classification stages into a single network, using a Region of Interest (RoI) pooling layer to extract fixed-length feature vectors from each proposal, thus improving efficiency. Faster R-CNN, proposed by Ren et al. in 2017, further enhanced this by introducing the Region Proposal Network (RPN), which shares convolutional layers with the detection network. The RPN generates region proposals that are refined and classified by the Fast R-CNN network, significantly improving both speed and accuracy. These advancements have made two-stage object detectors more practical for real-time applications.

1.2.3 One-stage Object Detectors

One-stage object detectors, like YOLO (You Only Look Once), introduced by Redmon et al. (2016), reframe object detection as a single regression problem. YOLO predicts bounding boxes and class probabilities directly from the image in one evaluation, significantly improving detection speed. YOLO runs at 45 FPS with a mean Average Precision (mAP) of 52.7%, while Fast YOLO achieves 155 FPS with a mAP of 63.4%.

The evolution from YOLO to YOLOv10 has brought significant advancements. YOLOv2 (Redmon & Farhadi, 2017) improved the original model by incorporating batch normalisation, anchor boxes and dimension clusters. YOLOv5 (Jocher, 2020) transitioned to PyTorch, optimizing model size and performance. YOLOv8 (Jocher et al., 2023) builds on the success of previous versions, offering even greater accuracy, efficiency and flexibility. The latest YOLOv10 (Wang et al., 2024) provides real-time object detection advancements by introducing an end-to-end head that eliminates Non-Maximum Suppression requirements.

RetinaNet (Lin et al., 2018), another one-stage detector, introduced Focal Loss to address class imbalance, focusing on hard, misclassified examples. This is effective for detecting birds in CCTV footage, where birds are a minority class against the background.

2. Methodology

2.1 Data Collection

The data used in this project consists of CCTV footage provided by the client. The footage is sourced from the concerned WPPs at different times, predominantly in the afternoons. There are a total of seven usable videos, and the total duration of these videos is 33 minutes and 3 seconds. In this project, frames are extracted from the CCTV footage at a rate of 1 frame per 2 seconds. This extraction rate is chosen for several reasons: redundancy reduction, as the CCTV footage has a frame rate of 600 FPS, where consecutive frames often contain redundant information. By extracting 1 frame every 2 seconds, the extracted frame is likely to contain new information about bird activities. It also makes the annotation process more efficient, as

annotating every frame from high-frame-rate footage would be extremely labour-intensive and time-consuming. Reducing the number of frames to be annotated makes the manual labelling process more efficient. Additionally, this rate achieves a balance between data volume and processing efficiency. From a total video length of 33 minutes and 3 seconds, extracting frames at this rate will produce 969 images, generating a sufficient amount of data for training while maintaining manageable data volumes.

2.2 Data Labelling and Pre-processing

Manual data labelling is crucial for creating a high-quality dataset that is essential for effective model training and accurate predictions. Each bird in the extracted frames is annotated with a bounding box using CVAT. While time-consuming, this step is critical for model accuracy.

After labelling, image tiling is applied as a pre-processing step. High-resolution images are divided into smaller, overlapping tiles to increase the relative size of birds, preserving detail and enhancing the detection of small objects that might be missed in full images.

To further enhance model robustness in real-world scenarios, data augmentation is performed. Techniques such as rotation, flipping, scaling, random distortion, and colour adjustments generate diverse variations of each image, reducing model overfitting and improving generalization to unseen data. The full data preprocessing pipeline is shown in Figure 2.

2.3 Model Training and Refinement

After data augmentation, the selected models are trained on this enhanced dataset. The model considered include Faster R-CNN (Ren et al., 2017), RetinaNet (Lin et al., 2018) and YOLOv8 (Jocher et al., 2023), which are state-of-the art object detection models. The training process includes several key steps. First, transfer learning is utilised, which involves using pre-trained models on larger datasets like COCO to expedite training and enhance accuracy. Next, hyperparameter tuning is employed, using grid search or random search techniques to find the optimal settings for learning rate, batch size and training epochs. Regularization and checkpointing are also implemented to prevent overfitting and ensure the best model is saved during training.

2.4 Model Comparison and Evaluation

Model comparison of the selected models will be conducted to evaluate their performance. The best model will be chosen based on its accuracy in counting birds and its consistency across different environments. The evaluation will focus on metrics such as mean Average Precision (mAP) and mean Average Recall (mAR), which measure the average precision and recall across all object classes and various Intersection over Union (IoU) thresholds.

This comparison will provide insights into how different models perform on our dataset, guiding the choice of a best model. Efficiency and extensibility will also be considered, which allows for potential upgrades to a real-time detection and counting system.

Finally, the best fine-tuned model will be used to analyze the CCTV footage and collect bird counts over time. The results will be visualized, including a line chart showing bird counts over 24 hours, to clearly demonstrate the method's effectiveness and the types of insights it can provide.

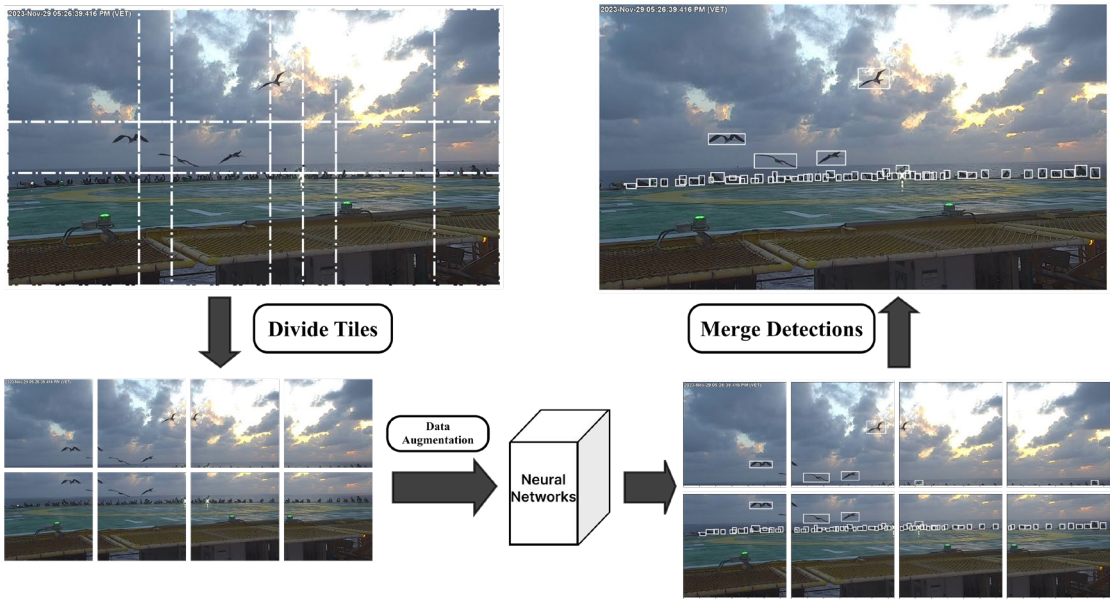

Figure 2 The image is divided into overlapping tiles (20% overlap) for model training, followed by data augmentation. Detections from each tile are then merged to improve small object detection.

3. Results and Discussion

The Faster R-CNN model with MobileNetV3-Large FPN backbone was evaluated for bird identification and counting in CCTV footage. Its performance was measured by accuracy, recall, and speed. The model achieved a mean Average Precision (mAP) of 31.4% across Intersection over Union (IoU) thresholds from 0.5 to 0.95. These thresholds represent increasing levels of required overlap between predicted and actual object locations, with 0.5 being lenient and 0.95 demanding near-perfect alignment. The mean Average Recall (mAR) was 39.1%, reflecting its ability to detect birds. Additionally, at an IoU threshold of 0.5, the model reached 81.3% accuracy in counting birds. The model processed each image in 366 milliseconds on average, analysing nearly three images per second. These results are summarised in Table 1.

Metric	Faster R-CNN
mAP@[.5: .95] (%)	31.4
mAR@[.5: .95] (%)	39.1
Count accuracy (%)	81.3
Avg. Inference Time (ms)	366

Table 1 Performance metrics of the Faster R-CNN model, presenting accuracy (mAP, mAR) and speed.

These preliminary results demonstrate a balance between accuracy and speed. Potential improvements include hyperparameter tuning and testing different backbones. As training continues, these metrics may improve, and further comparisons with other models like YOLOv8 and RetinaNet will help identify the most suitable model for accurately detecting and counting birds in CCTV footage.

The original Faster R-CNN with VGG-16 backbone achieved 42.1% accuracy at lower IoU thresholds and 21.5% at stricter thresholds, processing images in 200 ms each. This comparison highlights potential for optimizing the current model's accuracy and efficiency.

4. Conclusions and Future Work

This project aims to develop an accurate and efficient bird detection model using CCTV footage. Leveraging state-of-the-art object detection models such as Faster R-CNN, RetinaNet and YOLOv8, the goal is to accurately identify and count birds for further analysis. Current results are preliminary, with ongoing training focused on improving accuracy. The next phase involves completing model training and obtaining final results. It is anticipated that continued training and optimisation will enhance the model's performance. Future work includes implementing semi-automatic labelling techniques to streamline the annotation process, potentially reducing manual effort and improve labelling accuracy. This approach could significantly accelerate dataset preparation for future iterations.

5. Acknowledgements

I would like to express my sincere gratitude to my academic supervisors, A/Prof. Du Huynh and Prof. Mark Reynolds, for their invaluable guidance and support throughout this project. I thank my client mentors, Sandip Deshpande and Andy Watt, for the research opportunity and ongoing support. Special thanks to Dr. Emily Barker for HPC resources, Mohamed Ibrahim and the team for technical assistance, and the CEED team – A/Prof. Jeremy Leggoe and Kimberley Hancock – for their support during this project. Lastly, I am deeply grateful to my family and friends for their unwavering encouragement.

6. References

Girshick, R. (2015). Fast R-CNN. *IEEE International Conference on Computer Vision (ICCV)*, pp. 1440-1448. https://doi.org/10.1109/iccv.2015.169

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587. https://doi.org/10.1109/cvpr.2014.81

Jocher, G. (2020). Ultralytics YOLOv5. https://doi.org/10.5281/zenodo.3908559

Jocher, G., Chaurasia, A., & Qiu, J. (2023). Ultralytics YOLOv8. https://github.com/ultralytics/ultralytics

Lempitsky, V., & Zisserman, A. (2010). Learning To Count Objects in Images. *Advances in Neural Information Processing Systems*, **23**.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2018). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**(2), pp. 318-327. https://doi.org/10.1109/TPAMI.2018.2858826

Magnificent Frigatebird Identification, *All about Birds, Cornell Lab of Ornithology*. (2017). Allaboutbirds.org. https://www.allaboutbirds.org/guide/Magnificent_Frigatebird/id

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779-788. https://doi.org/10.1109/cvpr.2016.91

Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, Faster, Stronger. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6517-6525. https://doi.org/10.1109/cvpr.2017.690

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**(6), pp. 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., & Ding, G. (2024, May 23). YOLOv10: Real-Time End-to-End Object Detection. ArXiv.org. https://doi.org/10.48550/arXiv.2405.14458