CS 6670: Final Paper

Detecting Marine Animals

Stephanie Ginting (sag292), Neha Malepati (nm458), Noah Pikielny (np299)

December 12, 2023

1 Introduction

With the onset of climate change, it is becoming increasingly vital to develop tools to assess the health of ecosystems on a large scale to inform conservationists and policymakers better. These tools must be fast and cost-effective to survey large regions. Remote Sensing has been applied to myriad terrestrial ecosystems to quantify the effects of droughts, wildfires, etc on large scales; however, due to refraction and reflections from the surface of the water, such techniques can be difficult and sometimes impossible to apply to marine ecosystems. Marine ecosystems are particularly vulnerable to temperature fluctuations and thus have been severely affected in recent years–over half the corals of the Great Barrier Reef died in 2016 and 2017 and 90% of corals in Florida have died. These ecosystems are vital to the health of the ocean and to many people who are reliant on the habitats for food and safety from extreme weather events.

Through Fluid Lensing (a set of integration techniques) and structure from motion, one can estimate the sea floor geometry free of aberrations given a static monocular video of a region of the ocean in reasonable conditions. These data products encapsulate the health of sessile organisms but eliminate the information about moving animals. Moving animals can be helpful indicators of habitat health as they encapsulate multiple and higher trophic levels of an ecosystem and tend to contain keystone species. We propose to detect and quantify such animals by repurposing the Fluid Lensing data sets to recapture the information intentionally removed by the integration process. Given the nature of the data, we can reframe the problem to be an object detection in static videos–akin to video surveillance–but with significant noise from the ocean's surface. Much of the literature in this area, as discussed in the next section, is founded upon the idea that the static scene elements are unknown and "fuzzy" at best. By utilizing the Fluid Lensing products as our training dataset, we aim to improve upon such methods.

2 Related Work

To learn what previous approaches have been done, we examined several different papers. We decided to base our approach on [1] for automated video analysis and [2]. In [1], Shahare and Shende aim to present and discuss techniques for detecting moving objects in video. We believe this paper will be useful to apply to our problem of detecting marine animals in ocean surveillance videos. The paper discusses a new approach, which integrates object detection and background subtraction into a single process that could be efficient for moving object detection. There are two scenarios, the first of which is detecting moving objects from video sequences of a fixed camera. The proposed method for this scenario computes the estimated foreground (moving objects) and background (static scene) model of the frame specified by rank. To formulate the background model, Shahare and Shende use the SOFTIMPUTE method from [3], and Markov random field (MRF) methods are used for the foreground model, which is effective in detecting a moving object despite a complex background. The second scenario is detecting moving objects from video sequences of a fixed camera. The methods described in this paper could be used in addition to the future avenues described in the second paper to develop a model to detect moving marine animals.

In [2], the authors focus on techniques used to obtain accurate measurements and identify species of whales from drone images. They utilize CNN architectures to automate species identification and length estimation of whales. Specifically, the authors trained a Mask R-CNN model to identify humpback whales, minke whales, and blue whales in aerial images. Of note, the species identification functionality of the model in the paper was reported to have an incredibly high accuracy (98%). This is, however, on a dataset of 58

images to distinguish species within the same type of marine animal (whales). Scrutinizing the data even further, it appears that much of the training data were simply different images of the same individual, which does not truly capture the extent of variation in marine mammals. There were also limited examples of non-whale objects in the training data, which potentially caused the model's trouble in resolving the tips of flukes and fins. What this indicates is that, despite the seemingly high accuracy, the success of such a model in species identification is still rather unclear. In the paper, the model was limited to identifying three species of whales, most of which were the same individual, and had relatively high contrast to the ocean, which would decrease the robustness of the model when applied to other marine species. This calls into question the applicability of this model (i.e. Mask R-CNNs) to our use case, that is, identifying a broader range of marine species from videos of the ocean, which would have more noise and variation in species and habitat compared to the application in [2].

With [1] and [2], our plan for the final project is to evaluate the best method for constructing an annotator for marine animals based on Fluid Lensing data. As stated in the introduction, the detection and classification of marine animals are challenging due to the noise that ocean waves introduce. As such, we hope to utilize the Fluid Lensing data to create an annotator to detect and identify marine animals. There are a few avenues of exploration related to the specific parts of the annotation, detection, and classification pipeline that we can evaluate. Given the scope of the final project and the limited time, we will focus our efforts on evaluating these different methods on the Fluid Lensing dataset, and not necessarily develop a novel architecture for marine animal detection.

3 Data

Most of our data sets are 20-frame videos taken in Guam. These images are roughly 40 megapixels. Given the nature of our data, we are focusing on large animals like turtles and sharks that are roughly 16 pixels long. Fish are present but can be as thin as 2-3 pixels. For each stack of images, there is an accompanying Fluid Lensing result, which can be though of as an estimate of the background. It contains some noise in high-probability glint areas and has decreased resolution in deep areas.

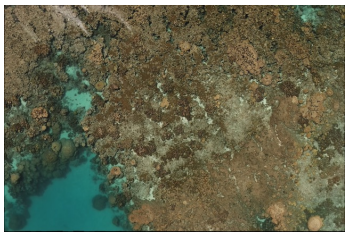


Figure 1: Sample Fluid Lensing product from data set

4 Our Approach

Our main question of interest, then, is which method of object detection would be best for this specified task. Drawing from the first paper, classical methods of object detection from videos utilize the difference between frames and foreground/background estimations; however, this, as discussed in the paper, will pick up noise, meaning that it will require additional filtering for marine animals in low signal regions (deep animals). We can utilize these classical video surveillance techniques to extract objects and annotate our data. The second paper makes use of Mask R-CNN, though this method was applied to photos instead of videos, and outputs a mask of the object in question, in this case, marine animals. The flexibility of this learned model instead of the classical methods above may make it more accurate when generalized to other

types of marine animals (though the paper only uses the model to classify whales). This provides a second method to detect objects.

Methodologically, we will first separate our data into training and validation sets. We will then use the two aforementioned methods to obtain object masks for marine animals and finally compare the results. The output of this project aims to provide preliminary findings in our evaluation of the approaches, and aims to answer the question: what is the best pipeline to develop a detector for marine animals from Fluid Lensing data?

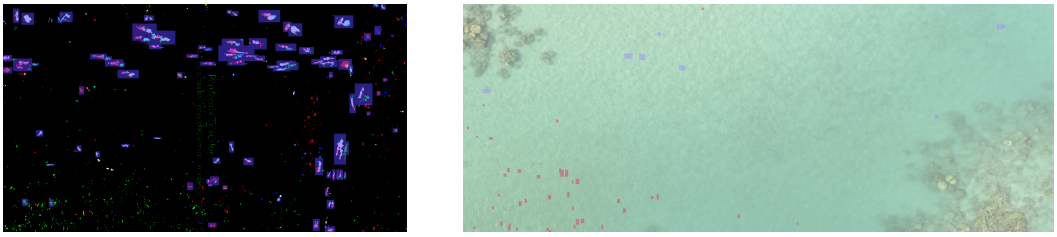
5 Evaluation

5.1 Part 1: Classical Methods

We implemented two methods that leveraged morphological operators as a baseline for what we deem classical methods. The first method is entirely comprised of morphological operators and thus focuses on filtering objects by size. There is significant noise at the local level, so the use erosion and dilation operations emphasizes signals that persist over greater spatial regions. For the second method, we assume that the deviation from the background estimate, when no animals are present, will approximate a normal distribution. We can thus calculate the probability that a deviation violates a null hypothesis given some threshold α. For both methods, we have a final speed filtering pass; we assume that all animals move slower than the surrounding wave events. Moreover, waves move at such a speed that in adjacent frames, their bounds do not overlap. Thus, using a logical and operator between binary masks of adjacent frames removes some wave events. We expect it may be an issue for small fish, but given our focus on larger animals like turtles, sharks, and predatory fish, it is not an issue. Lastly, the results are pipelined into a flood fill algorithm to find bounding boxes for detected objects. These boxes are filtered by a minimum number of corresponding pixels, which is left as a hyper-parameter. Any bounding boxes that do not intersect with other bounding boxes on adjacent frames are removed.

We also considered using k-means clustering on images with a high prevalence of glint to remove the glint and increase our model's accuracy, however, due to the scope of our project and the limited time we believed this could be an avenue for future work instead.

Given time constraints and difficulties with applying the classical methods to the formatted learned data set, we were unable to perform IoU analysis with the classical methods. The Triathlon data set 2a, is a drone video of a triathalon where swimmers are watched by people on kayak and there are many sharks in the bottom left portion. This data set lacks high-frequency background information, so it is an easier data set than would normally be expected. On the Triathlon data, the gaussian method had a precision of 68% and a recall of 93%. On a more complex data set with high-frequency background and glint 2b, the guassian method achieved a precision of 0.4% and a recall of 100%; however, the model was able to pick out animals that were not originally detected by the naked eye.



(a) Triathlon (b) Cocos Turtles

Figure 2: a: Swimmers at a triathlon near sharks with proposed bounding boxes and masks. b: Turtles in Cocos Lagoon with proposed bounding boxes projected onto the original image with correct annotations in blue and incorrect in red.

5.2 Part 2: Learned Methods

For our second approach, we focus on learned methods of object detection. Like in [2], we started with a Mask-R-CNN model as proposed in [4] with a ResNet-50-FPN backbone, which we then fine tune on on our dataset. The model was trained on 365 images of the ocean floor from the Fluid Lensing dataset, split into 300 for training and 65 for validation. The multi-task loss was optimized using AdamW with a learning rate of 1×10^{-5}. We trained the model for 100 epochs, which took around 2 days to complete.

We evaluate our model using intersection over union (IoU) to measure accuracy. On our validation set, we found that the model had an accuracy of 0.04, which is quite low. Generally, the model only did well when an object was very close to or on the surface of the water, and where the ocean floor did not have any rocks as added noise. We observe this in Figure 3, where swimmers that are close to the surface are identified correctly by the Mask-R-CNN. However, a majority of our data had animals that were farther from the surface of the water and a lot of rocks along the ocean floor, leading to a very low accuracy. We can see in Figure 4 that the model has trouble distinguishing non-rock items from rocks. Though a little difficult to see, there is a fish in the bottom right corner of the image that the model fails to capture. Interestingly, when we pass in a cropped image of the fish as seen in Figure 5, the model did manage to identify the fish, but still included the noise from rocks. Of note, the fish was swimming over sand with no rocks, which made it easier for the model to identify.

(a) Swimmers close to the surface (b) Mask output from Mask-R-CNN

Figure 3: Mask output from the Mask-R-CNN on a surface with a lot of swimmers close to the surface. Here, we see the model correctly detects the objects.

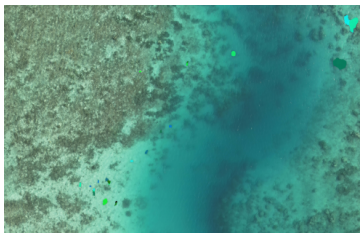
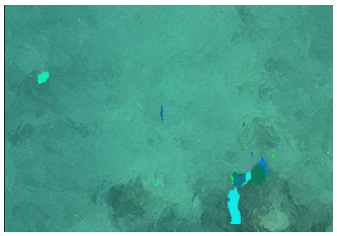


Figure 4: Mask output from Mask-R-CNN on an ocean surface with lots of rocks. We see that the output incorrectly predicts rocks as animals. Though difficult to see, there is a fish in the bottom right corner that the model misses.

(a) Cropped image of fish



(b) Mask output from Mask-R-CNN

Figure 5: The fish from Figure 4, croppped so we can better see it. Now, the model captures the fish, though it also incorrectly identifies the rocks as animals.

Because of the low accuracy, the Mask-R-CNN is unlikely to be useful in isolation for detecting animals in our data set, though it might be more useful in conjunction with the classical methods outlined above. Though the model did well when detecting objects that are close to the surface, we are more interested in animals that are deeper in the ocean that may be obfuscated by waves and rocks. Additionally, a Mask-R-CNN model was not as successful because it considers images in isolation from the videos they were found in. Even as a human annotator, looking for animals based on a still image was difficult, such as trying to identify the fish from Figure 5 in Figure 4. The key indicator was the movement of animals by contrasting it with the motion of the waves and the stillness of the rocks. As such, since this Mask-R-CNN only evaluates still images, it will likely continue outputting inaccurate masks, even with more data and training.

6 Discussion

At its current state, the classical method does not have adequate precision to do annotation directly, as there are too many false positives; however, it has proven a useful tool in annotating as it is able to highlight anomalies in deep regions that are difficult to detect without assistance. Moreover, these deep anomalies tend to be easy to filter out, for a human, since they have more temporal coherence than false positives. Thus, for large-scale processing, while it could not produce a great result by itself, it could greatly augment the human annotation process. This also suggests, that stronger constraints on coherence (matching bounding boxes across frames) could greatly increase the precision of the method.

Overall, we find that the classical method has more success through a higher rate of precision and recall compared to the learned methods. As stated previously, this is likely because the classical method receives as input the entirety of a video from the Fluid Lensing data, while the Mask-R-CNN only looks at still images. With some assumptions and estimations, such as the calculations used to isolate animals based on their movement, the classical method was able to correctly identify larger animals. In comparison, the Mask-R-CNN was unable to identify objects that contrasted poorly with the ocean floor (e.g. the presence of rocks) or were deeper underwater which resulted in larger noise from waves. Because of this, the classical method is the better choice to develop a pipeline to aid in identifying animals in the Fluid Lensing dataset. However, we did see some promise in the Mask-R-CNN, especially when given a localized area that may have objects of interest. Though the Mask-R-CNN in isolation cannot be used to detect marine animals well, we may be able to use it in conjunction with the classical methods. For example, we can first use the classical method to determine areas that may have smaller objects, which, as mentioned in Section 5.1, it may have issues with. Then we can crop said areas and pass it into the Mask-R-CNN to output a final mask.

6.1 Challenges

One major challenge we faced was a lack of computational power while running the Mask-RCNN model. Training the model took much longer than expected due to the RAM limitations of our devices and Google Colaboratory. This slowed down our work on the learned methods we wanted to use for building our annotator

and caused us to re-evaluate our scope to better work within the confines of our machines. Furthermore, creating the manual annotations required to train the Mask-RCNN model took hours of manual labor. Many pictures in our dataset did not even have an animal that we could annotate to train our model. The limited availability of data combined with the large amount of sheer manpower to dissect the data we could use proved to be a challenge as we explored learned methods.

6.2 Future Work

There are several avenues of future work to build off of our research. First, the scope of this paper is focused on evaluating the best approach to develop a detector for marine animals from Fluid Lensing data. We conclude that the classical methods performed the best, but we also believe some form of combination of classical and learned methods may also perform as well, or even better, than just the classical methods.

Additionally, the methods described in this paper could be expanded to a variety of species to help conservation and ecology communities better understand and detect species in natural habitats. Given that terrestrial ecosystems lack water artifacts, these methods should have significantly increased performance for similar tasks. The learned method would need new training data, but the classical methods should generalize perfectly, assuming properly tuned hyper-parameters. We could also include additional object parameters, such as detecting a blowhole, or other key features that could help in the identification of marine animals (even though these may be even more difficult to spot than the animals themselves from an aerial view). As mentioned in Section 5.1, we believe another avenue for future work is to use k-means clustering on images with a high prevalence of glint to remove the glint and increase our model's accuracy. Moreover, implementing more complicated methods for filtering bounding boxes based on the continuity of movement vectors should yield better filtering of false positives in the classical methods, as most of the false positives exist for very few frames and have very unstable positions. The best way to accomplish the goals that we defined is to gather more aerial images of marine animals to better train our model.

References

[1] D. Shahare and G. H. Raisoni, "A survey on moving object detection in static and dynamic background for automated video analysis," *International Journal for Scientific Research and Development*, vol. 1, pp. 2050–2054, 2013.

[2] P. Gray, K. Bierlich, S. Mantell, A. Friedlaender, J. Goldbogen, and D. Johnston, "Drones and convolutional neural networks facilitate automated and accurate cetacean species identification and photogrammetry," *Methods in Ecology and Evolution*, vol. 10, 06 2019.

[3] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *Journal of Machine Learning Research*, vol. 11, no. 80, pp. 2287–2322, 2010.

[4] P. D. Kaiming He, Georgia Gkioxari and R. Girshick, "Mask r-cnn," *Proceedings of the IEEE international conference on computer vision*, 2017.