Fusing Region Proposal Network with AVED Tracking: Counting Machine Learning System

Raymond Esteybar, California State University Monterey Bay

Mayra Ochoa, California State University Monterey Bay

Danelle Cline and Duane Edgington, Monterey Bay Aquarium Research Institute

Summer 2018

ABSTRACT

From oceanographic research, the Monterey Bay Aquarium Research Institute has recorded a collection of over 23,000 hours of video footage. Professional video annotators have obtained observations for the Video Annotation and Reference System (VARS). Over the years, MBARI scientists and researchers have utilized this data for analyzing behaviors in the area that commonly demands a lot of energy and time from professional video annotators. In fact, the increase of imagery has conducted engineers to intermediate in this. Due to this large amount, an opportunity to explore other possibilities of utilizing this data would provide an understanding of what species are found within a certain area. Over the course of 28 years, annotators have accumulated a vast amount of classifying animals from video footage. Due to the large amount of time spent doing so, an implementation of a tracking algorithm would improve the amount of efficiency to classify sea life along with these annotators. The project includes using deep learning algorithms and annotation methods that approach detection of creatures including classification and a counting system.

INTRODUCTION

Over the years, professional video annotators have studied and annotated video from across the Monterey Bay area. They have provided this useful data to Video Annotation and Reference System (VARS) at the Monterey Bay Aquarium Research Institute. MBARI also collects data using Remotely Operated Vehicles (ROVs) that go underwater approximately 300 times per year, producing vast amounts of underwater imagery. These operated vehicles include high quality cameras, and MBARI makes use of them in areas that provide the minimal environmental impact.

This vast collection of high quality imagery is ever increasing, so software engineers at MBARI started collaborating with professional video annotators to use automation to augment their annotation tasks.

Convolutional neural network object detection methods such as Faster R-CNN has shown promise for accurate localization and classification of animals (Nathaniel Yee, 2017). Faster R-CNN, however, is designed for operating on single images. We explore an algorithm that fuses Faster R-CNN and a tracking system to provide an efficient way to both count and classify animals in the footage for professional video annotators.

Project code is located here at: https://github.com/mbari-org/avedac-mbarivision

MATERIALS AND METHODS:

We explored the past code that our mentors, Danelle and Duane, provided us. The first thing we explored was an algorithm with Faster R-CNN and multiples tracking algorithms. From there, we were able to start with our approach utilizing tools from the written code itself.

Modifying Saliency Maps

As part of the exploration, saliency maps, a matrix based on the series of other maps, contain three main levels in order to display the output category values changing with respect to a modification in the input image pixels. These three main levels are color, orientation and intensity which provide the deep learning model the intuition of attention of what is visually interesting from each frame. The tracking system contains a saliency map that scans the frame to obtain the highest value, then to the next highest value, and so forth. The class detector obtains each value as a point and then locates the boundaries around it. In the other tracking system that incorporates saliency maps, it helps to find a winner and a mask around that point. However, in this new implementation of the tracking system, it will be utilizing a Rectangle rather than the mask.

Introducing Rectangles

Indeed, saliency maps had provided points to the mask for the tracking systems. Although, as mentioned before, a Rectangle is utilized instead of a mask for the points. Since saliency maps do not have a sense or indication of what the interesting objects are, they only provide their location. Rectangles store the x-ymin and x-ymax points that surround the object with a box to help predict what that interesting object is. Also, between supervised and unsupervised learning, creating Rectangle objects, it practices supervised machine learning where the coordinates are mapped to the points to create a rectangle. Even though Rectangle class has boundaries to create an object, it provides a certain localization of an interesting thing that may benefit the Voting Classifier to keep track of each Rectangle object created and avoid repetition.

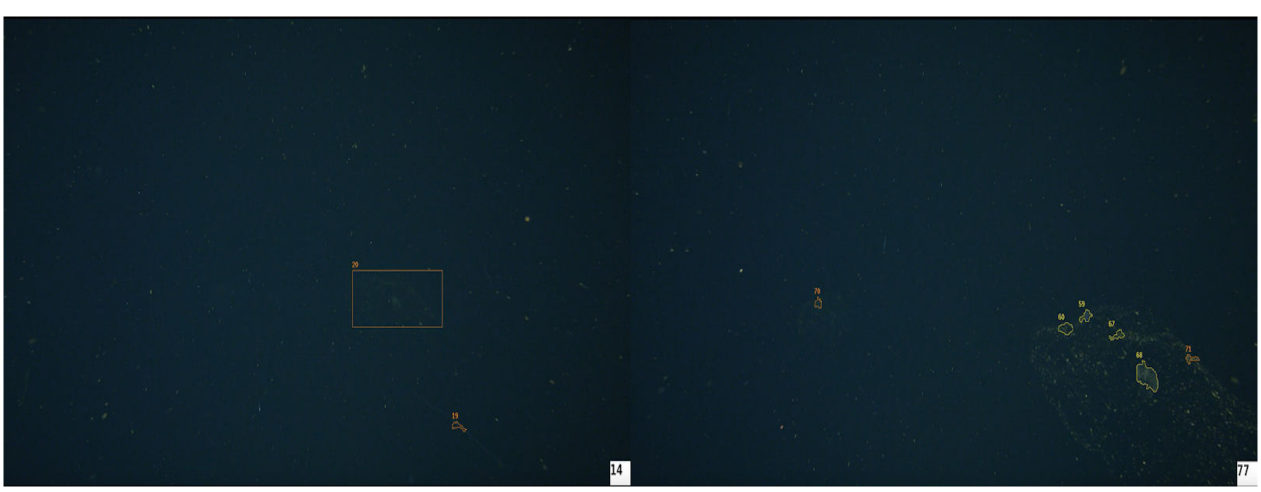
i2MAP Midwater Test-Data

To test different types of re-scalings and deep learning methods, we utilize data from more than 6 minute long video clip from the i2MAP dataset. i2MAP is a high resolution video camera in the cone of the Autonomous Underwater Vehicle (AUV) named Dorado. Data is collected from Monterey Bay ocean areas, known as a mesopelagic zone, with the purpose of conducting transects of deep water in the area to provide imagery that can be identified and understood regarding marine life. These transects run at a depth between 50 meters to 1000 meters deep. Data is then processed through the Faster R-CNN model providing us a total of 22 018 images from VARS.

RESULTS

As part of the process, we created a new detection class considering the boundaries that counting animals will have. Two of the results are shown down below where on the left, is frame number one, and on the right, is frame number of four of the i2MAP Midwater Test-data where the program is identifying a few species. They include, the Mitrocoma and the Krill.

Smaller objects in the video footage would display an accurate detection of the sea animal, but would face difficulties when viewing larger objects such as the Bathochordaeus. For example, on the left, frame number 14, we have the program detect the Bathochordaeus house with an orange square from a far. As it comes closer, when reaching frame number 77, it detects pieces of the Bathochordaeus house and the Bathochordaeus filter on the bottom right.



DISCUSSION

With the process set in motion, we have yet to discover if re-scaling the images to a smaller size than the original would provide better tracking results. By re-scaling the image, we are able to run the program much faster with the low amount of pixels needing to be overlooked when locating objects. In addition to resizing the image, we must also take note of the program's performance in detecting the objects per frame. If the program does not offer the same amount of objects found with a 25% rescale reduction compared to a 50% rescale reduction, we would know that we could compare the 50% reduction to around 75%.

CONCLUSION

We present new arguments for re-scaling i2MAP data. This implementation shows positive results, with faster runtime and more accurate tracking.

SUGGESTIONS FOR FUTURE WORK

Counting machine learning system

Re-scaling the original images provide a consistent result in the detection of every species and creating the bounding box. For this reason, adding more specific work for the tracker may provide promising results to this project.

1. **Test different arguments to rescale images into more trackers -** After rescaling the original images by approximately 50%, more animals were detected from the Faster-RCNN bounding box. Finding the rescale point in which the image intersects provides a faster runtime with an accurate detection. The rescale argument can be used in the tracking system to improve the detection accuracy.
2. **Voting classifier -** According to the observation that re-scaling images provide a better detection in every object, implementing a voting classifier would track the amount of instances a sea creature has been found throughout the whole video footage. With information such as, the class name, class probability, and bounding box, we could utilize this information to understand if that sea creature is the same one as the instance in the previous frame. If so, it would not add to the total number of instances of that creature, but be accounted for already. To classify if the object was indeed what the program thinks it is, we would use the concept of probability weighting. This concept will help determine if the unique instance of a sea creature was what it is identified as. The process to determine its identity is by utilizing the amount of frames it has been found in and the probabilities it has gathered to produce a percentage of classifying the animal. This method could be utilized online (throughout the video), updating the class prediction at the next frame, or at the end of the video to classify the species' identity.

ACKNOWLEDGEMENTS

We would like to thank Duane Edgington and Danelle Cline, our mentors, for providing us the opportunity to work with them by providing guidance, motivation to learn new tools, and advice for our future in the tech field. We also want to thank Rob Sherlock for providing us all the data we needed for our project, and George Matsumoto for creating an incredible intern experience and encouraging us to do a great work. Last but not least, we want to thank the MBARI interns for bringing in a sense of community as young students.

References:

1. Walther, D., Edgington, D., Koch, C., Detection and Tracking of Object in Underwater. Video. In Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, vol.1, I-I(IEEE, 2004).
2. Ren, S., He K., Girshick R., & Sun, J., Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. ArXiv e-prints(2015).
3. Yee, N., Workflows for Automated Detection and Classification of Unlabeled Deep Sea Imagery. Monterey Bay Aquarium Research Institute (2017).
4. Gough, A., & Dalit, M., (2017, August 11). Investigations of imaging for imaging for midwater autonomous platforms.
5. MBARI. i2MAP. https://bitbucket.org/mbari/avedac-i2map/src