Whirligig beetles few-shot segmentation and tracking

Project Report

Nadav Misgav - 208521658
nadavmisgav@mail.tau.ac.il

Yanay Danan - 315524157
yanaydanan@mail.tau.ac.il

October 13, 2022

1 Abstract

In this paper, we propose a series of procedures to identify and track swimming whirligig beetles and the waves they create in the water. The data we used contains multiple unlabeled videos of the beetles swimming in an arena. The problem we faced includes several challenges; 1) Identifying the beetles despite their small size in relation to the arena. 2) Variety of patterns the waves create due to different movements and rotations the beetle performs. 3) Find the subtle change the waves make in water. 4) Deal with a noisy background containing a bright light source and spots in the arena itself. Furthermore working with unlabeled data, neither the beetles nor the waves were labeled. Our proposed solution consists of two steps. Segmenting the beetle using a classical connected components method performed in 3D space, secondly segmenting the waves inside a bounding box surrounding the beetle using a few-shot foreground segmentation network. In our work we notice for both the beetles and the waves, the state-of-the-art methods for unsupervised segmentation we tried, such as [1, STEGO] and [2, COS-Net] failed the task. With the few-shot learner [3, FgSegNetV2], based on only 30 manually tagged frames of the waves, and our classic method, the segmentation matched well visually both to the beetles and the waves in the video. The waves segmentation yields an average overall F-Measure of 0.42. An implementation for this paper can be found in Section 9.1.

2 Introduction

Swimming whirligig beetles often make conspicuous wave patterns on the surface of ponds and quiet streams, apparently it is presumed that they use the reflections of the waves to detect objects at a distance to sense their surroundings [4]. As far as we know, there is no automatic tool to identify and track the beetles and the waves they create. To assist and promote the zoological research of these beetles we propose an automatic segmentation approach for both the beetles and the waves they create.

Despite the many challenges in this problem, there are several properties such as time continuity of successive frames, a static background, and the beetle's characteristics that we can take advantage of. Unfortunately, any state-of-the-art we have tried could not solve this problem end to end and some couldn't even solve any part of it at all. In this paper, we propose an approach to solve this problem end to end, by separating it to smaller problems and using different methods for each of the problems, to achieve a complete solution at the end.

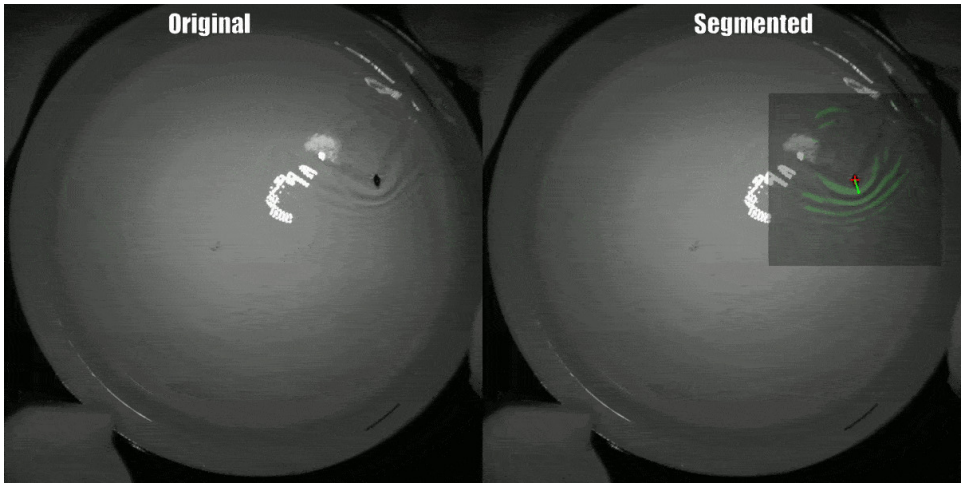

Figure 1: Whirligig Beetle Segmentation Example

For the beetle segmentation, we used a series of classical procedures which take advantage of all the properties mentioned above. Whilst for the waves segmentation we used a few-shot supervised learner for foreground segmentation [3, FgSegNetV2], after some preprocessing of the frame. To utilize a few-shot supervised learner, about 30 ground truth masks of waves have been manually tagged. However due to the manually tagging, the

ground truth was not perfect, as we tagged mainly the most dominant and big waves, while the subtle and smaller waves have not been tagged. Surprisingly the network was able to generalize to finer more subtle and small waves very well.

By using these steps and overlaying our solutions we were able to solve end-to-end this segmentation problem, both for the beetle and its waves.

3 Related Work

As far as we know, no previous work has been made to solve our problem. However, our problem can be generalized and can be solved with generic state-of-the-art tools and methods. One way is to solve it end to end is with an unsupervised image segmentation network [1, STEGO] or an unsupervised video segmentation network [2, COSNet], if some ground truths are given more powerful supervised learning tools for foreground segmentation can be used [3, FgSegNetV2], both for the beetles and the waves segmentation.

Another approach is to divide the problem into smaller ones and use different tools to solve them separately, such as background removal to emphasize the foreground, which can be solved classically or with deep network architecture [5, U2NET], or calculate the deformation field between frames [6, VoxelMorph] to emphasize moving targets. In addition to all of these, there is work that has been made to track coastal waves [7, Wave-Tracking], which is more similar to our scenario.

The performances of all of the unsupervised segmentation methods we tried were insufficient. Therefore we decided to use a background subtraction technique and work with the foreground for better results. Surprisingly, the classical background subtraction technique outperformed the deep network method. Unfortunately, also after the background subtraction both the unsupervised segmentation methods and the calculation of the deformation field failed to segment, and we realized that a classical method to segment the beetles can be more robust and perform very well while also allowing for better preprocessing for the wave segmentation. After all of the unsupervised tools above failed to sufficiently segment the waves, we decided to use a supervised foreground segmentation that theoretically should outperform all the unsupervised ones. Because no tagged data was available, it was necessary to use a few-shot learner for the task.

4 Data

We were given about 45 different videos of a single beetle swimming in an arena, filmed with a static camera (static background) by a researcher of the Zoological faculty of TAU. The videos have a spatial resolution of 832×832 pixels, a length of about one minute each, and contain 120/125 FPS. The videos do differ in scale, brightness, and arena type. When processing the video we encountered lack of RAM resources so we batched each video to 200 frames at a time, whereas for the prediction itself we lowered it even more to 50 frames. As mentioned above to utilize a few shot learner we created 30 ground truth masks of the waves from a sample of these videos. An example of frames from these videos and the ground truth masks are shown in Figure 2, in the "Wave Segmentation" block.

5 Methods

In this section, we will describe our method step by step to both segments and track the swimming beetle and the waves it creates on the water's surface. As mentioned above, our proposed framework consists of two parts, beetles segmentation, and waves segmentation, as shown in Figure 2.

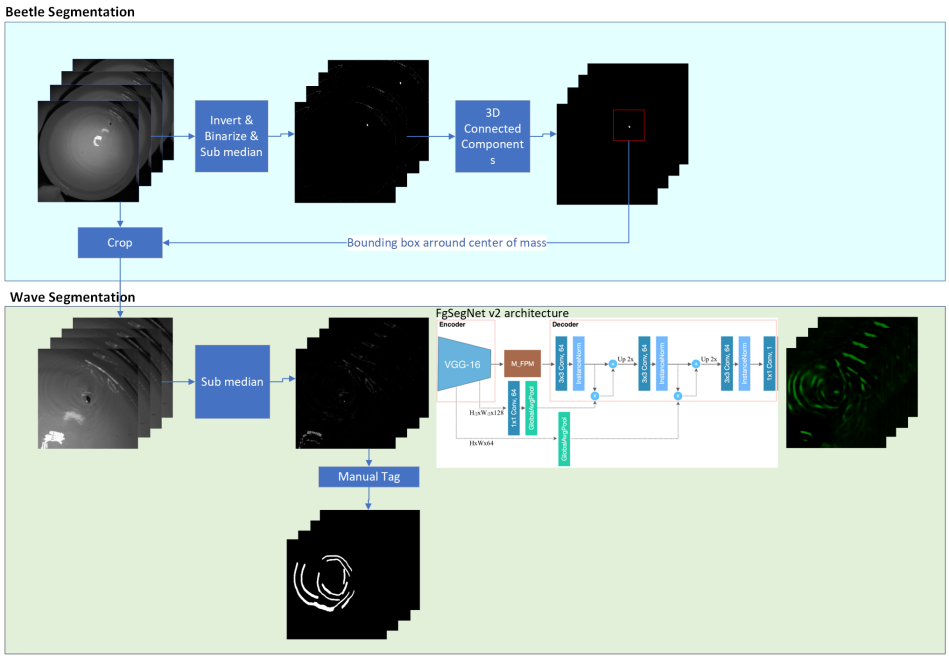

Figure 2: End-To-End solution

5.1 Beetles Segmentation

Deep learning has been a game changer in the field of computer vision and it is used for a wide variety of tasks in this field. However, often to realize its potential tagged data will be needed. As mentioned before, our data does not contain ground truth, which made the problem much harder. Some state-of-the-art methods for unsupervised segmentation were tried such as [1, STEGO] and [2, COSNet], neither of them was able to segment the beetle (nor the waves). While trying to process the image the prepare it for training we noticed it would be better to solve this part with a classical approach.

First, for each batch, we make noisy foreground masks with a simple preprocess and background subtraction. The preprocess includes changing the frames to grayscale, then inverting them such that the beetle color will be white instead of black, and lastly binarizing the images with a simple threshold. For the background subtraction, we used the median (along the time axis) of the preprocessed batch as the background. Finally, we binarize it with a simple threshold again. In the last step, to extract only the beetle from the noisy foreground masks, the largest connected component (in all of

the batch dimensions) was taken. This method exploits the time continuity of the beetle's movement as the beetle is connected to a large component along all the frames, unlike blobs of noise, even if they were larger than the beetle in some of the frames. The process is demonstrated in the "Beetle Segmentation" block in Figure 2.

5.2 Waves Segmentation

Unlike the beetles, the waves have many different patterns and their appearance was often subtle. Therefore, a classical solution for their segmentation was much more complicated to produce. Similar to the beetles, the performances of the unsupervised segmentation networks for the wave segmentation were still insufficient. Hence, a small number of ground truth masks were tagged manually, and a few shot learner for foreground segmentation was used [3, FgSegNetV2].

Before using the network, another preprocessing was preformed. Similarly to the preprocessing for the beetles, binary foreground masks of each batch was made for the waves. Though this time without the color inversion, since the beetle is not wanted in this part and should be removed from the mask, and without the image binarization (as it was done for the connected components). Finally, only a small bounding box of the image around the beetle was taken as input for the network. These images were fed into the proposed network shown in Figure 2 inside the "FgSegNet v2 architecture" block, generating a background which is common across time, and foreground - the waves, which represents temporal changes along time.

6 Experiments

To test our approach, we examined it on a sample of beetle videos. We ran it on Google Colab with GPU/CPU as devices. The training is fast and converged after only 18 epochs which took a few minutes. The segment an entire video is much slower, with a ratio of about 30[sec] of processing for 1 second of video (as the video has a large FPS). For the beetle segmentation we used a binarize threshold of 215 and after the background subtraction a threshold of 255 (else is zero). For the training of the waves segmentation, we used the architecture and configuration of FgSegnet-v2 for the UCSD data set. A RMSprop ($\rho = 0.9$ and decay of 0) as an optimizer, a learning

rate of $lr = 1 \cdot 10^{-4}$, and the binary cross-entropy for the loss function,

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p(y_i) + (1 - y_i)\log(1 - p(y_i)) \tag{1}$$

Where y_i is the label and $p(y_i)$ is the predicted probability. Due to a lack of memory, we used a batch of size 1 frame for the training.

6.1 Results

Figure 3 shows the performance of the beetle segmentation (on the right) by visualizing the results for different examples. Although our approach seems to be robust, there are a few assumptions that it is based on, and will fail if they are not satisfied. First, the background subtraction will eliminate the beetle if it is static through out most of the batch. Secondly, if there is more than one moving objects in the water, the algorithm might segment them instead.

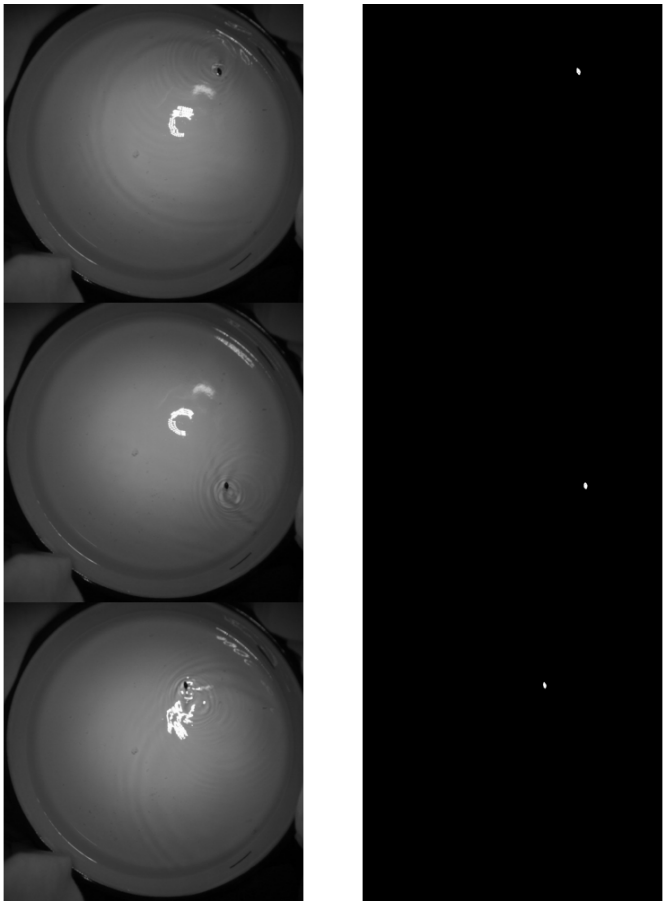

Figure 3: Beetle segmentation sample

Figure 4 shows the performance of the wave segmentation (marked with green lines on the right) by visualizing the results of the network for different examples. Although we tested this approach on our data, there is always the risk in supervised learning to overfit. If major changes will happen like different scales or more moving objects, it might be needed to train the network again with more and new relevant tags. To allow easier manual tagging we created a *tagging tool*.

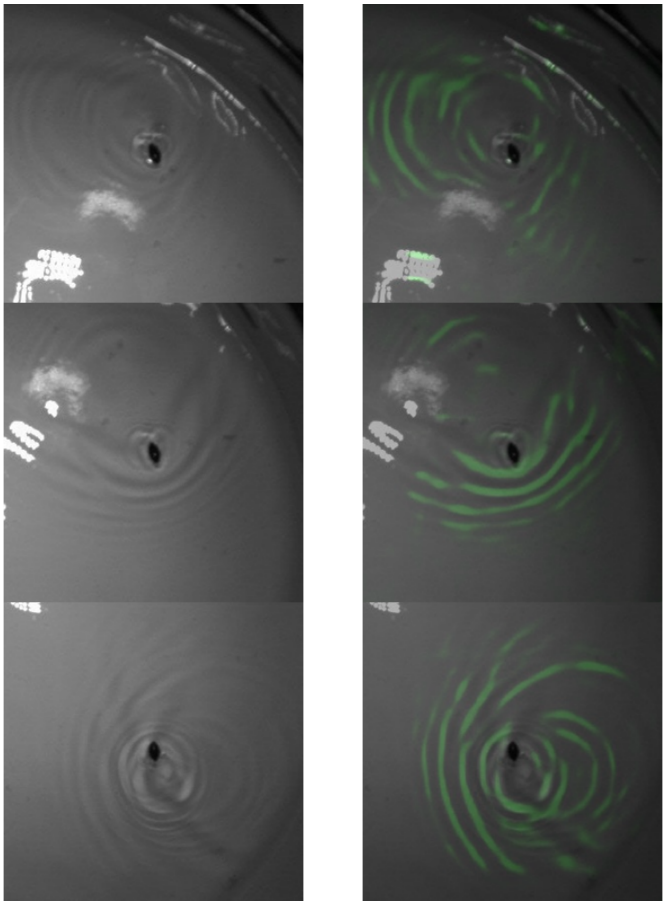

Figure 4: Wave segmentation sample

To numerically evaluate the results of the wave segmentation we left 20% of our tagged data for validation. As mentioned before, the manually tagging did not include the subtle and small waves in the frame. Therefore, although the visualized results seem to be good, the precision of the segmentation is not as good as expected. For the evaluation, we binarized the probability map predicted by the network with a threshold of 0.2.

The results yielded an average recall of 0.597, average precision of 0.339, and average F-Measure of 0.421, as shown in the results Table 1.

	Recall	Precision	F-Measure
frame1	0.62	0.29	0.39
frame2	0.58	0.17	0.26
frame3	0.64	0.38	0.48
frame4	0.66	0.30	0.41
frame5	0.64	0.48	0.55
frame6	0.43	0.39	0.41
Average	**0.597**	**0.339**	**0.421**

Table 1: Recall, Precision and F-Measure metrics

To check for overfitting and convergence, we plotted our loss of the train and validation for each epoch, as shown in Figure 5. It looks like there is a small overfit. However, due to the imperfection of the manual tags, we will not expect and want the validation loss to decrease to zero, in order to still be able to generalize to the subtle small untagged waves in the frame, which indeed occurred. So as long as the validation loss is converges, and the visual results look good, we can say that the network has been successful with the task.

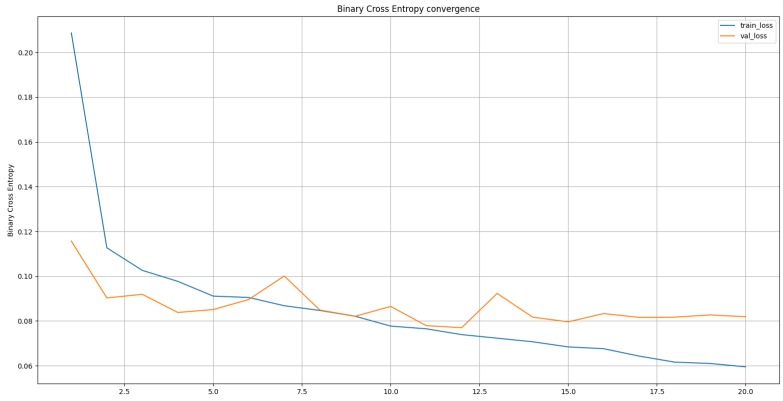

Figure 5: Binary Cross Entropy Loss convergence

7 Conclusion

In this work, we propose a robust method to segment and track a video of swimming whirligig beetles, and the waves they produce. Our method includes a robust classical approach combined with a network with pre-trained weights ready for the task, that also can be retrained for different future use using more ground truth labels. From the visualization of the results, it looks like both the beetle and the waves segmentation matched well visually. Through this work, we came across many different ideas and methods which might have been helpful to solve our problem or parts of it. We learned the strengths and weaknesses of these methods and in the end, could choose the best for us. As future work, our method can be extended to find a number of beetles that swim simultaneously in the pool instead of only one, by leaving all the largest components instead of only the largest one. Another important improvement can be paralleling the video segmentation to dramatically decrease its run-time.

8 References

[1] M. Hamilton, Z. Zhang, B. Hariharan, N. Snavely, and W. T. Freeman, "Unsupervised semantic segmentation by distilling feature correspondences," in *International Conference on Learning Representations*, 2022.

[2] X. Lu, W. Wang, J. Shen, D. Crandall, and J. Luo, "Zero-shot video object segmentation with co-attention siamese networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[3] L. A. Lim and H. Y. Keles, "Learning multi-scale features for foreground segmentation," *arXiv preprint arXiv:1808.01477*, 2018.

[4] V. A. Tucker, "Wave-making by whirligig beetles (gyrinidae)," *Science*, vol. 166, no. 3907, pp. 897–899, 1969.

[5] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," vol. 106, p. 107404, 2020.

[6] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "VoxelMorph: A learning framework for deformable medical image registration," *IEEE Transactions on Medical Imaging*, vol. 38, pp. 1788–1800, aug 2019.

[7] J. Kim, J. Kim, T. Kim, D. Huh, and S. Caires, "Wave-tracking in the surf zone using coastal video imagery with deep neural networks," *Atmosphere*, vol. 11, no. 3, 2020.

9 Appendix

9.1 Video and Code

We include a short video description of our work at:
https://youtube.com/shorts/6SamWSabx9A?feature=share
We also provide training and evaluation code at:
https://github.com/nadavmisgav/whirl_beetle_segmentation