SynDHN: Multi-Object Fish Tracker Trained on Synthetic Underwater Videos

Mygel Andrei M. Martija, Prospero C. Naval, Jr., Ph.D.

Computer Vision and Machine Intelligence Laboratory, University of the Philippines

Abstract

In this paper, we seek to extend multi-object tracking research on a relatively less explored domain, that of, underwater multi-object tracking in the wild. Multi-object fish tracking is an important task because it can provide fish monitoring systems with richer information (e.g. multiple views of the same fish) as compared to detections and it can be an invaluable input to fish behavior analysis. However, there is a lack of an annotated benchmark dataset with enough samples for this task. To circumvent the need for manual ground truth tracking annotation, we craft a synthetic dataset. Using this synthetic dataset, we train an integrated detector and tracker called SynDHN. SynDHN uses the Deep Hungarian Network (DHN), which is a differentiable approximation of the Hungarian assignment algorithm. We repurpose DHN to become the tracking component of our algorithm by performing the task of affinity estimation between detector predictions. We consider both spatial and appearance features for affinity estimation. Our results show that despite being trained on a synthetic dataset, SynDHN generalizes well to real underwater video tracking and performs better against our baseline algorithms.

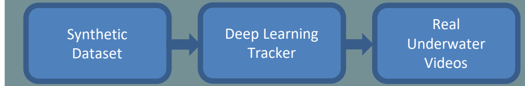
Introduction

Multi-object tracking (MOT) serves as a foundation in many video understanding tasks. Its application spans surveillance, autonomous driving, sports video analysis, and behavior analysis. Just like in other deep learning subfields, part of the reason why MOT research has accelerated is due to the availability of large-scale datasets. However, the domains covered by these datasets remain limited.

This paper seeks to extend MOT research on a relatively less explored domain – underwater multi-fish tracking in the wild. Generating tracking annotations, however, is not a trivial task. This is an expensive endeavor both in terms of manpower and costs. Hence, instead of generating manual annotations for our use case, we propose to train on a synthetic dataset.

Finally, we also propose SynDHN, an end-to-end tracker that can perform both fish detection and association, regardless of the fish species. SynDHN is a multi-component tracker that uses Faster-RCNN as its detector and the Deep Hungarian Network (Xu et at., 2019) as its association module. Furthermore, we explore the merit of using multiple cues (i.e. spatial and appearance cues) and show that SynDHN trained on both cues perform best on a real underwater test set.

Methodology

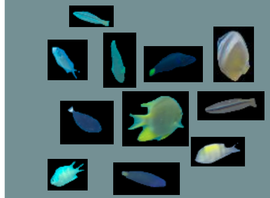


Synthetic Dataset → Deep Learning Tracker → Real Underwater Videos

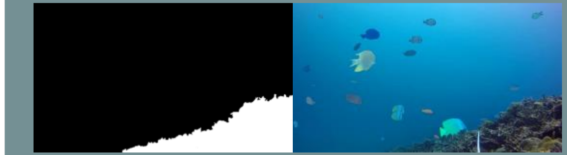
Synthetic Dataset

We propose an end-to-end framework, starting from dataset preparation, designing the proposed tracker, and finally testing it on actual data to evaluate model performance. To circumvent the lack of a large-scale training set in the context of multi-fish tracking, we design a synthetic dataset.

The synthetic dataset is constructed by extracting fish blobs from real underwater video captures. An off-the-shelf detection and segmentation model called PANet (Liu et al., 2019) is used to extract these blobs.



We patch these blobs on to the background and we simulate fish motion mainly by applying translation and rotation operations on each fish blob. Furthermore, we attempt to mimic a major challenge for MOT systems—occlusion—by extracting non-fish foreground objects such as corals and rocks. We then let some fish move behind these objects so that they get partially or fully occluded.

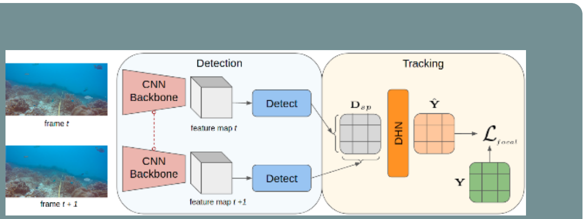


Deep Learning Tracker

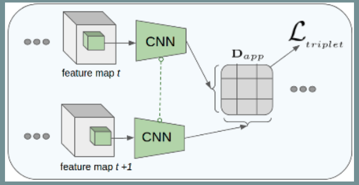
For our tracker, we propose an end-to-end setup that performs both fish detection and tracking. Detection is performed by a Faster-RCNN network (Ren et al., 2015). In its simplest form, our tracker uses only spatial features. Specifically, we take the pairwise Intersection-over-Union between detections in the previous frame (t) and detections in the next frame ($t+1$).

$$d_{ij} = IoU(x_i^t, x_j^{t+1})$$

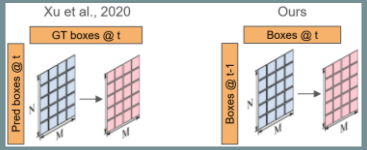
Hence, if we have N detections from frame t and M detections from frame $t+1$, we obtain a spatial affinity matrix of size NxM. This matrix is then fed through a Deep Hungarian Network or DHN (Xu et al., 2020)



An enhanced version of SynDHN incorporates appearance features as input to the tracker. For every detection, its appearance embedding is extracted from the region of the Faster-RCNN's feature map output that corresponds to the position of the said detection. This is then passed through a lightweight CNN. Similar to spatial affinities, an appearance affinity matrix is also obtained, which then gets passed through its own DHN.

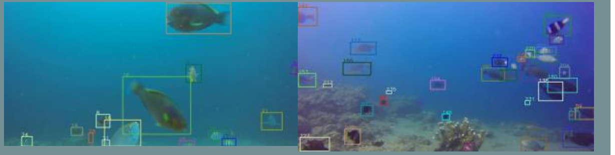


The Deep Hungarian Network is a series of bi-directional gated recurrent units that mimic the Hungarian matching algorithm (Kuhn et al., 1955). It is designed to be a differentiable approximation of the latter but DHN is not designed to be a tracker per se. To fit our purposes, we repurpose DHN by making a simple but fundamental change to the inputs being passed to it. Unlike in the original paper of Xu et al., the rows of the DHN's inputs will be the detections from the previous frame and the columns will be the detections at the next frame. Hence, making associations between row and column entries essentially accomplishes the process of building object tracks.



Test Dataset: Real Underwater Videos

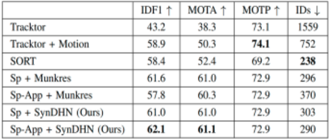
Finally, to test the efficacy of our proposed training scheme and model, we annotate a small scale dataset containing real underwater video sequences. In total, we annotated 4 short video sequences from different sites in the Philippines. Sample sites and predicted tracks are shown below.



Results

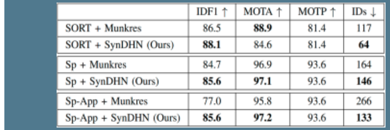
We obtain the following findings from the test results on the real underwater dataset:

- SynDHN that uses appearance cues perform best
- Weighing mechanism is essential when using multiple cues



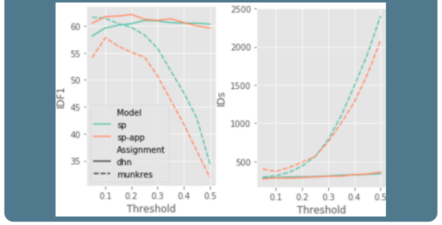
	IDF1 ↑	MOTA ↑	MOTP ↑	IDs ↓
Tracktor	43.2	38.3	73.1	1559
Tracktor + Motion	58.9	50.3	**74.1**	752
SORT	58.4	52.4	69.2	**238**
Sp + Munkres	61.6	61.0	72.9	296
Sp-App + Munkres	57.8	60.3	72.9	370
Sp + SynDHN (Ours)	61.0	61.0	72.9	303
Sp-App + SynDHN (Ours)	**62.1**	**61.1**	72.9	290

- Advantage of SynDHN is more evident when using better detections (e.g. ground truth detections)
- The DHN module can be applied on other tracking frameworks and still perform well



	IDF1 ↑	MOTA ↑	MOTP ↑	IDs ↓
SORT + Munkres	86.5	**88.9**	81.4	117
SORT + SynDHN (Ours)	**88.1**	84.6	81.4	**64**
Sp + Munkres	84.7	96.9	93.6	164
Sp + SynDHN (Ours)	**85.6**	**97.1**	93.6	**146**
Sp-App + Munkres	77.0	95.8	93.6	266
Sp-App + SynDHN (Ours)	**85.6**	**97.2**	93.6	**133**

- The DHN module makes trackers more robust against the matching threshold setting



Conclusion

- It is possible to circumvent the absence of ground truth tracking labels when training deep learning trackers
- A synthetic dataset can substitute for an actual training set
- SynDHN can perform well on real multi-fish tracking in the wild despite being trained on

Acknowledgements

Philippine Council for Industry, Energy and Emerging Technology Research and Development; Philippine Council for Agriculture, Aquatic, and Natural Resources Research and Development; Laura T. David, Ph.D. and Mark John Manalo from the Marine Science Institute.