Embedded Online Fish Detection and Tracking System via YOLOv3 and Parallel Correlation Filter

Shasha Liu, Xiaoyu Li, Mingshan Gao, Yu Cai, RuiNian*
School of Information Science and Engineering
No. 238 Songling Road
Ocean University of China, Qingdao
E-Mails: liushasha117@163.com; 982502041@qq.com; 1239969217@qq.com; 347632365@qq.com; nianrui_80@163.com;
Peiliang Li
College of Physical and Environmental Oceanography
No. 238 Songling Road
Ocean University of China, Qingdao
E-Mails: Lpliang@ouc.edu.cn

Tianhong Yan
China Jiliang University, Hangzhou
E-Mails: thyan@163.com
Amaury Lendasse
Department of Information and Logistics Technology,
The University of Houston, Houston City, Texas
77204(713)743-2255, USA;
Arcada University of Applied Sciences, 00550 Helsinki, Finland;
E-Mails: alendass@central.uh.edu; lendasse@gmail.com

***Abstract*—Nowadays, ocean observatory networks, which gather and provide multidisciplinary, long-term, 3D continuous marine observations at multiple temporal spatial scales, play a more and more important role in ocean investigations. In this paper, we first perform image enhancement to produce depth information and benefit many vision algorithms and advanced image editing. We try to develop a novel underwater fish detection and tracking strategies combining you only look once(YOLO) latest detection algorithm YOLOv3 algorithm and parallel Correlation Filter. We demonstrated on the NVIDIA Jetson TX2 for online fish detection and tracking, enabling a fast system and rapid experimentation. It has been shown in the experiments that the developed scheme of this paper achieves consistent performance improvements on online fish detection and tracking for ocean observatory network.**

Keywords—Ocean Observatory Network; detection algorithm; parallel Correlation Filter; fish detection and tracking

I. INTRODUCTION

Among ocean observatory networks, those deployed for fish observation, which explores the species richness and activities of either wild or cultivated fish, has shown great prospect in aquaculture, fisheries, marine surveys and applications [1]. Underwater fish visual detection and tracking is one of the most essential and fundamental tasks to capture the visual cues in deeply exploring the fish ecosystem observation network [2]. However, it is still a challenging topic to conduct fish detection and tracking, due to the particular underwater visual properties in fish observation, such as poor visibility, non-rigid deformations, appearance variations in high frequency at diverse illumination and viewpoints, and so on [3, 4].

Dating back to the history, Considerable research has recently been focused on the analysis of underwater fish obtained from ocean observatory network. Among them, we try to develop a novel underwater fish detection and tracking strategies combining you only look once (YOLO) latest detection algorithm YOLOv3 algorithm and parallel Correlation Filter, a complete quantitative solution to fish detection and tracking, which integrated with marine knowledge, can analyze underwater objects and compose high level interpretations, like fish species distribution variation, and fish behavior patterns [5, 6, 7].

The structure of the paper is as follows: Section II describes the basic framework of underwater fish detection and tracking in our ocean observatory network. In Section III, object detection is introduced. Section IV introduces object tracking. Section V lists our experimental results and analysis. Conclusions are drawn in Section VI.

II. GENERAL FRAMEWORK

In this paper, we try to develop a novel underwater fish detection and tracking strategies combining you only look once (YOLO) latest detection algorithm YOLOv3 algorithm and parallel Correlation Filter. A brief flow chart of online underwater fish detection and tracking is shown in Fig. 1, which is made up of several steps, including image enhancement, object detection and object tracking. (1) Image enhancement: The performance of vision algorithms will inevitably suffers from the biased, low-contrast scene radiance. The haze removal can produce depth information and benefit many vision algorithms and advanced image editing. We adopt a basic underwater dehazing model to obtain the dehazed image. Each channel of the original image is pixel-stretched using two pixel thresholds that can be backpropagated to achieve image enhancement [8]. (2) Object detection: YOLOv3 Network predicts an objectness score for each bounding box using logistic regression. Each box predicts the classes the bounding box may contain using multilabel classification. During training it use binary cross-entropy loss for the class predictions. YOLOv3 predicts boxes at 3 different scales. The system extracts features from those scales using a similar concept to feature pyramid networks. The base feature extractor is added several

978-1-5386-4814-8/18/$31.00 ©2018 IEEE

convolutional layers. These predict a 3-d tensor encoding bounding box, objectness, and class predictions. Next it is taken the feature map from 2 layers previous and upsample it by 2×. It also is taken a feature map from earlier in the network and merge it with upsampled features using concatenation. Then, adding a few more convolutional layers to process this combined feature map, and eventually predict a similar tensor. It uses Darknet-53 network for performing feature extraction. This network uses successive 3×3 and 1×1 convolutional layers. We train on full images and we use multi-scale training, lots of data augmentation, batch normalization, all the standard stuff. We use the Darknet neural network framework for training and testing. YOLOv3 has significant benefits over other detection systems and it is faster and better [9]. (3) Object tracking: We propose to build realtime high accuracy trackers in a novel framework. It consists of two components: a fast tracker and an accurate verifier. The two components work in parallel on two separate threads. For tracker, we choose the algorithm which is correlation filter-based. We introduce a factorized convolution operator and a compact generative model of the training sample distribution to correlation filter, with the aim of simultaneously improving both speed and performance.

III. Object Detection

YOLO, an approach to object detection. Prior work on object detection repurposes classifiers to perform detection. Instead, YOLO frame object detection as a regression problem to spatially separated bounding boxes and associated class probabilities. A single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation. Since the whole detection pipeline is a single network, it can be optimized end-to-end directly on detection performance [10].

YOLO9000, a real-time object detection system that can detect over 9000 object categories [11]. The improved model, YOLOv2, is state-of-the-art on standard detection tasks. Using a novel, multi-scale training method the same YOLOv2 model can run at varying sizes, offering an easy tradeoff between speed and accuracy. YOLOv2 focus mainly on improving recall and localization while maintaining classification accuracy.

YOLOv3 is a little bigger than YOLOv2 but more accurate.

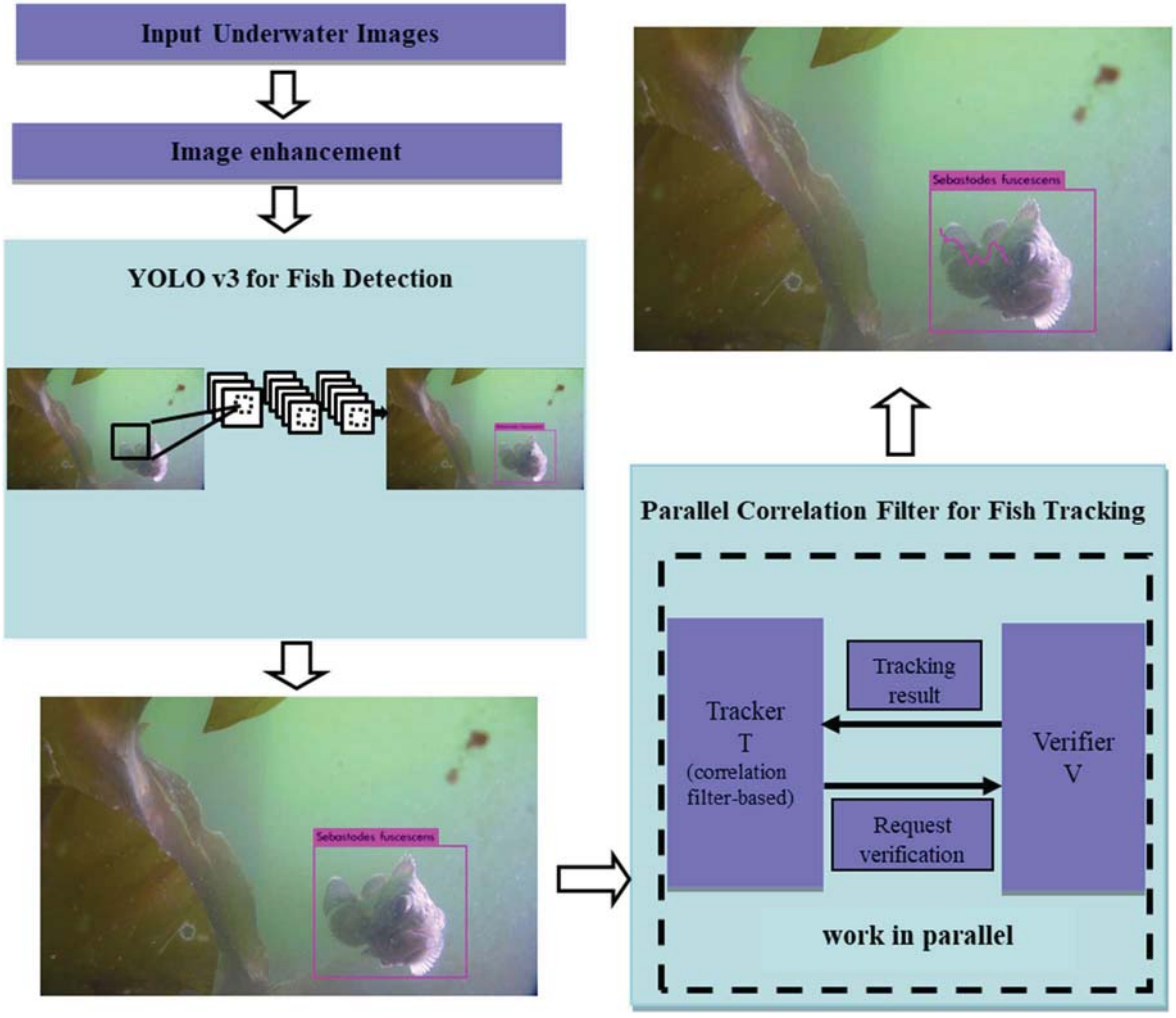


Fig. 1. The flow chart of system

A. Bounding Box Prediction

Following YOLO9000 the system predicts bounding boxes using dimension clusters as anchor boxes [11]. The network predicts 4 coordinates for each bounding box, t_x, t_y, t_w, t_h. If the cell is offset from the top left corner of the image by (c_x, c_y) and the bounding box prior has width and height p_w, p_h, then the predictions correspond to:

$$b_x = \sigma(t_x) + c_x \quad (1)$$

$$b_y = \sigma(t_y) + c_y \quad (2)$$

$$b_w = p_w e^{t_w} \quad (3)$$

$$b_h = p_h e^{t_h} \quad (4)$$

During training sum of squared error loss are used. If the ground truth for some coordinate prediction is $\hat{t}_*$ our gradient is the ground truth value (computed from the ground truth box) minus our prediction: $\hat{t}_* - t_*$. This ground truth value can be easily computed by inverting the equations above.

YOLOv3 predicts an objectness score for each bounding box using logistic regression. This should be 1 if the bounding box prior overlaps a ground truth object by more than any other bounding box prior. If the bounding box prior is not the best but does overlap a ground truth object by more than some threshold we ignore the prediction, following Faster R-CNN [12]. Unlike Faster R-CNN this system only assigns one bounding box prior for each ground truth object. If a bounding box prior is not assigned to a ground truth object it incurs no loss for coordinate or class predictions, only objectness.

B. Class Prediction

Each box predicts the classes the bounding box may contain using multilabel classification. It does not use a softmax as researchers have found it is unnecessary for good performance, instead it simply uses independent logistic classifiers. During training binary cross-entropy loss are used for the class predictions.

C. Prediction Across Scales

YOLOv3 predicts boxes at 3 different scales. This system extracts features from those scales using a similar concept to feature pyramid networks [13]. From base feature extractor several convolutional layers are added. The last of these predicts a 3-d tensor encoding bounding box, objectness, and class predictions.

Next the feature map from 2 layers previous are taken and upsample it by $2\times$. We also take a feature map from earlier in the network and merge it with our upsampled features using concatenation. This method allows to get more meaningful semantic information from the upsampled features and finer-grained information from the earlier feature map. Then a few more convolutional layers are added to process this combined feature map, and eventually predict a similar tensor, although now twice the size.

The same design are performed one more time to predict boxes for the final scale. Thus predictions for the 3rd scale benefit from all the prior computation as well as finegrained features from early on in the network.

D. Feature Extractor

Using a new network Darknet-53 for performing feature extraction. New network is a hybrid approach between the network used in YOLOv2, Darknet-19, and that newfangled residual network stuff. Network uses successive 3×3 and 1×1 convolutional layers but now has some shortcut connections as well and is significantly larger.

Darknet-53 is much more powerful than Darknet-19 but still more efficient than ResNet-101 or ResNet-152.

Experiments show that YOLOv3 is a good detector. It's fast, it's accurate. In terms of COCOs weird average mean AP metric it is on par with the SSD variants but is 3× faster. This indicates that YOLOv3 is a very strong detector that excels at producing decent boxes for objects. In the past YOLO struggled with small objects. However, now we see a reversal in that trend. With the new multi-scale predictions we see YOLOv3 has relatively high APS performance.

IV. OBJECT TRACKING

A novel parallel tracking and verifying (PTAV) framework is further taken as the object tracking tool in image sequences to focus on the motion trajectories of underwater fish all the time. The framework consists of two components: a fast tracker T and an accurate verifier V. The two components work in parallel on two separate threads. For tracker, we choose the algorithm which is correlation filter-based. The tracker T aims to provide a super real-time tracking inference and is expected to perform well most of the time; by contrast, the verifier V validates the tracking results and corrects T when needed. This way is to seek a trade-off between speed and accuracy [14].

The tracker T is responsible of the "real-time" requirement of PTAV, and needs to locate the target in each frame. Meanwhile, T sends verification request to V from time to time (though not every frame), and responds to feedback from V by adjusting tracking or updating models. To avoid heavy computation, T maintains a buffer of tracking information in recent frames to facilitate fast tracing back when needed. The verifier V is employed to pursue the "high accuracy" requirement of PTAV. Up on receiving a request from T, V tries the best to first validate the tracking result (e.g., comparing it with the template), and then provide feedback to T. To adapt V to object appearance variations over time, the tracking target template is not fixed. Instead, V collects a number of reliable tracking results, and then use k-means to cluster these results to obtain a target template pool for subsequent verification.

In PTAV, T and V run in parallel on two different threads with necessary interactions. The tracker T and verifier V are initialized in the first frame. After that, T starts to process each arriving frame and generates the result. In the meantime, V validates the tracking result every several frames. Because tracking is much faster than verifying, T and V work asynchronously. When V finds a tracking result unreliable, it searches the correct answer from a local region and sends it to

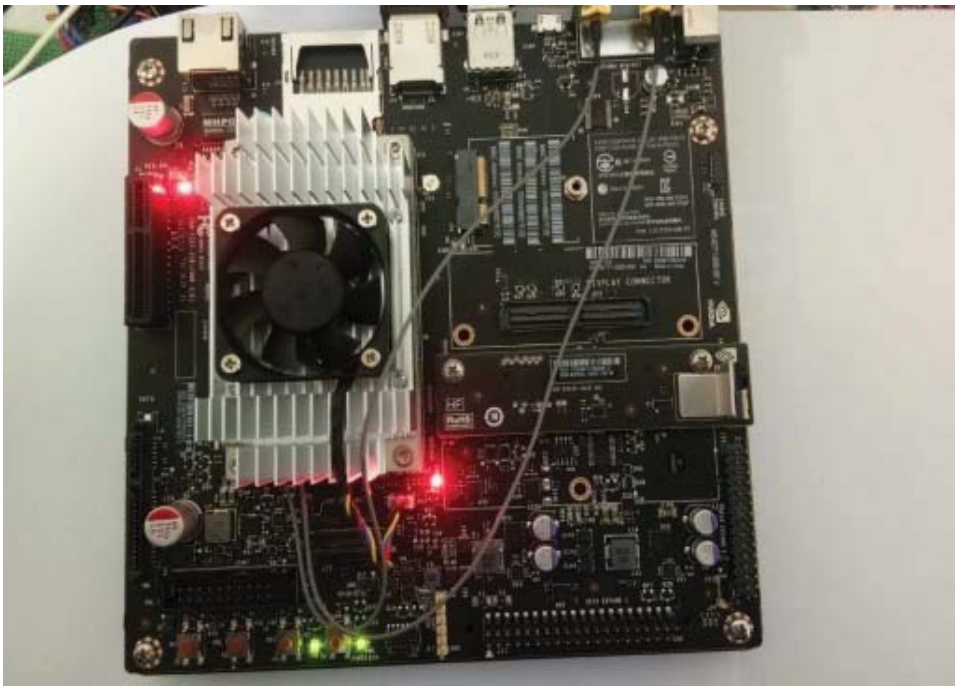
T. Upon the receipt of such feedback, T stops current tracking job and traces back to resume tracking with the correction provided by V.

It is worth noting that PTAV provides a very flexible framework, and some important designing choices are following. (1) The base algorithms for T and V may depend on specific applications and available computational resources. In addition, in practice one may use more than one verifiers or even base trackers. (2) The response of T to the feedback from V, either positive or negative, can be largely designed to adjust to specific requests. (3) The correction of unreliable tracking results can be implemented in various ways, and it can even be conducted purely by T (i.e., including target detection).

V. Simulation Experiment

In our simulation experiments, we used two data sets. One is the Fish4Knowledge dataset we collected. Their underwater video image of this datasets was taken from Orchid Island, Taiwan. Orchid Island stands on the western Pacific Ocean, and it located at 22°01'59"N, 121° 32'60"E. Another dataset was taken from the fish ecosystem in Shandong Province, China. The experimental data video has a total of 15 sea areas, including Ailian Bay, Beihai, and Liugong Island. We describe an underwater fish detection and recognition tracking visualization system, can be used for real-time visual environmental monitoring of fish ecosystem, and applied YOLOv3 and the Correlation Filter algorithm to establish an underwater fish detection and recognition tracking visualization system.

It can be said that Jetson TX2 is another breakthrough after the NVIDIA embedded computing series after Jetson TK1 and TX1.The Jetson TX2 processor has six CPU cores, four Cortex-A57, two self-developed Denver cores, a GPU-based Pascal architecture, 256 CUDA cores with 8GB of 128-bit LPDDR4 memory. Figure 2 shows the results on the NVIDIA Jetson TX2. Figure 3 shows the original image. Figure 4 shows the detection and tracking result by our proposed algorithm. The experiment results prove that this system can well realize the basic functions of fish real-time detection and multiple fish tracking. Experiments demonstrate the favorable performances of the proposed fish detection and tracking strategy in both efficiency and accuracy.



(a) NVIDIA Jetson TX2

(b) the result

Figure 2 The tracking results by NVIDIA Jetson TX2

Figure 3 The original image

Figure 4 The detection and tracking result

VI. CONCLUSION

In this paper, we have presented a novel underwater fish detection and tracking strategies combining YOLOv3 algorithm and parallel Correlation Filter. YOLOv3 is a fast and accurate detector. Parallel Correlation Filter achieves the high tracking accuracy among real-time trackers. The simulation results have shown the effectiveness and feasibility of our proposed method.

ACKNOWLEDGEMENTS

This work is partially supported by the National High-Tech R&D 863 Program (2014AA093410), the Natural Science Foundation of P. R. China (31202036), the National Program of International S&T Cooperation (2015DFG32180), the National Science & Technology Pillar Program (2012BAD28B05), and the Natural Science Foundation of P. R. China (41376140).

REFERENCES

[1] Chave, Alan D. "Cabled ocean observatory systems." Marine technology society journal 38.2: 30-43, 2004

[2] Clark, H. L. "New seafloor observatory networks in support of ocean science research." OCEANS. MTS/IEEE Conference and Exhibition. Vol. 1. IEEE, 2001

[3] R. Nian, F. Liu, B. He, “Early underwater artificial vision model in ocean investigations via independent component analysis”, Sensors, 13(7): 9104-9131, 2013.

[4] R. Nian, B. He, J. Yu, Z. M. Bao, Y.F Wang, “ROV-based underwater vision system for intelligent fish ethology research”, International Journal of Advanced Robotic Systems, 2013.

[5] Spampinato, Concetto. "Detecting, Tracking and Counting Fish in Low Quality Unconstrained Underwater Videos." 514-519, 2008.

[6] Xu Xiao, Che R, Nian R. "Underwater 3D object reconstruction with multiple views in video stream via structure from motion." OCEANS 2016-Shanghai. IEEE, 2016.

[7] Che R, Xu X, Nian R. Underwater non-rigid 3D shape reconstruction via structure from motion for fish ethology research OCEANS 2016 MTS/IEEE Monterey. IEEE, 2016.

[8] Zhang S, X.F Gong, R Nian, B. He. A depth estimation model from a single underwater image with non-uniform illumination correction OCEANS 2017-Aberdeen. IEEE, 2017.

[9] Redmon J, Farhadi A. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

[10] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition: 779-788, 2016.

[11] Redmon J, Farhadi A. YOLO9000: better, faster, stronger. arXiv preprint, 2017.

[12] Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems: 91-992015.

[13] Lin T Y, Dollár P, Girshick R B, He K, Hariharan B, Belongie S. Feature Pyramid Networks for Object Detection. CVPR. 1(2): 4, 2017.

[14] Fan H, Ling H. Parallel Tracking and Verifying. arXiv preprint arXiv:1801.10496, 2018.