OBJECT DETECTION AND TRACKING

Amit Mohapatra
202218033
MSc Data Science

K. Sethu Srivatsa
202218021
MSc Data Science

Project Supervisor: - Prof. Srimanta Mandal

Abstract— In this project, we initiated the process by implementing YOLOv3 and SORT for robust object detection and tracking in underwater images. Subsequently, we employed autoencoder architecture to enhance image quality. Through an iterative approach, we further refined object representations and re-implemented YOLOv3 and SORT, resulting in improved detection scores. This holistic methodology showcases the effectiveness of our approach in enhancing underwater image analysis and advancing object tracking capabilities in dynamic environments.

I. INTRODUCTION

The surveillance in videos makes attempts to perform detection, tracking and recognition of object of interest from multiple frames, and interprets the behaviors of the object and the actions that they perform. These intelligent systems will surely replace the age-old traditional methods of surveillance. Starting with the lowest level of features of image to an extremely high-level approach of understanding, there exists three main methods of analysis: detecting the object, tracking of the given object from frame to frame and then finally evaluating the results obtained after tracking. The application of surveillance in videos extends beyond traditional methods, encompassing detection, tracking, and recognition of underwater objects across multiple frames. This intelligent approach, evolving from the lowest image features to a high-level understanding, revolutionizes underwater surveillance. This paradigm shift is not only limited to underwater scenarios; it also finds application in motion-based identification, access control, video indexing, communication between entities, and navigation of underwater vehicles, illustrating the versatility and broad impact of our methodology. Figure 1 depicts the block diagram of object detection and tracking.[1]

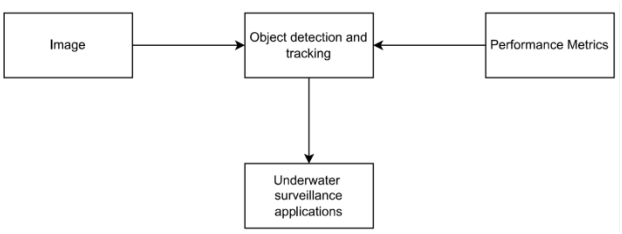

Fig.1. Block Diagram of object detection and Tracking

II. THEORY AND FUNDAMENTALS

A. YOLO v3

You Only Look Once (YOLO) is a state-of-the-art real-time object detection model that divides an image into a grid and predicts bounding boxes and class probabilities for each grid cell simultaneously. YOLOv3, an evolution of its predecessors, demonstrates superior accuracy and speed in detecting objects within an image. In the presented work, YOLOv3 is employed to identify and locate objects of interest in each frame of a video sequence. It works on the Darknet framework developed using C/Cuda. In this way, it shows a very high performance. Structurally, it is inspired by the GoogleNet structure, which consists of 24 convolutions and 2 dense layers. The third version of YOLO was released in April 2018[2]. The main innovation with this version is; it is the determination. of bounding boxes for different sizes.

Figure 2 Shows the basic architecture of YoloV3, it consists of 24 Convolution layer and 2 fully connected layer in a network architecture. Initial convolution layers extract features and fully connected layer predicts output probability and coordinates of object centers.

YOLO v3 uses darknet 53. It has fifty-three layers of convolution. Darknet-53 contains mainly 3x3 and 1x1 filters along with bypass links. The formulas given below explain the transformation of network output for obtaining bounding box predictions. Here, d_x and d_y, are center coordinates, dw is width and dh is the height of predicted result. Top left coordinates of the grid are m_x and m_y Network outputs are t_x, t_y, t_w, and t_h. Anchor dimensions for the box are P_h and P_w.

$$d_x = \sigma(t_x) + m_x \quad (1)$$
$$d_y = \sigma(t_y) + m_y \quad (2)$$
$$d_w = p_w e^{t_w} \quad (3)$$
$$d_h = P_h e^{t_h} \quad (4)$$

By using threshold value, a filter is applied to remove the box having class score less than the threshold value chosen. Because a score with less value represents that the box is insufficient for identifying the classes. Even after filtering by using a threshold value for the class scores, a lot of overlapping boxes still remain. When many boxes are overlapping with each other, then select only one box out of those overlapping boxes and identify the object. So, the second filter is used for choosing the desired boxes, which termed as nonmaximum suppression (NMS). It uses intersection over union (IOU) function.

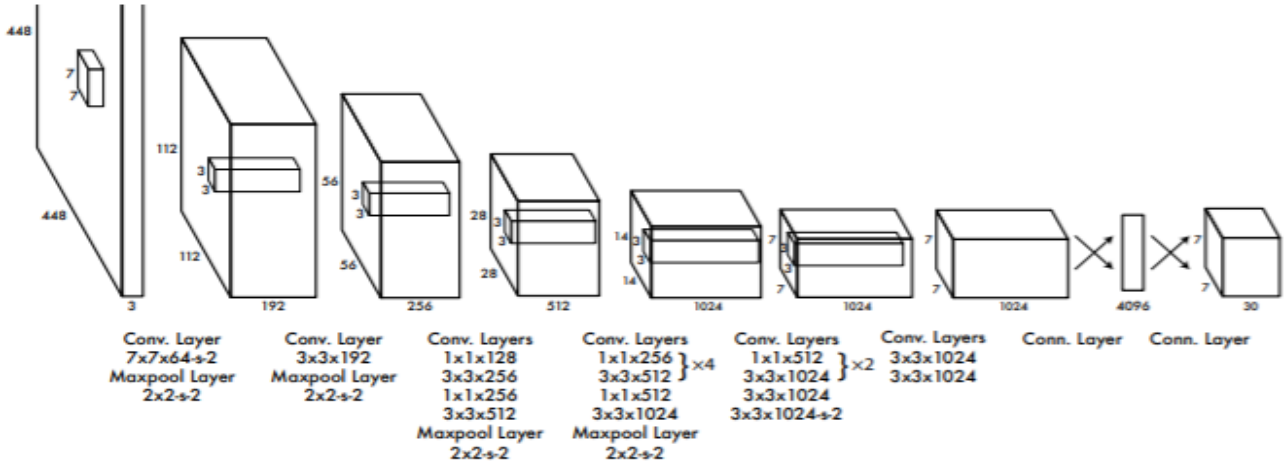

Fig.2. Architecture of YoloV3 algorithm

$$\text{IOU} = \frac{B_1 \cap B_2}{B_1 \cup B_2} \quad (5)$$

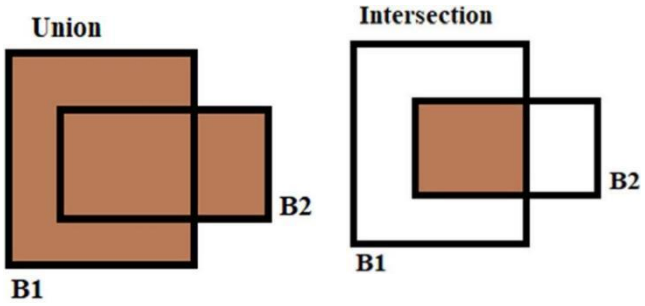

Fig. 3: Illustration depicting the definition of union and intersection

Above-left corner (a1, b1) and below right corner (a2, b2) are used to determine a box. For calculating the area of the rectangle, multiply its height (b2-b1) and its width (a2-a1). Then the coordinates (ai1, bi1, ai2, bi2) for the intersection of 2 boxes are obtained. Here, ai1 and bi1 are the highest value of a1 and b1 coordinate-position of the two boxes. Similarly, ai2 and bi2 is the lowest value of the a2 and b2 coordinate-position of the 2 boxes.

union(A, B) = A + B - inter(A, B) (6)

inter-area = (ai2 - ail) x (bi2 - bil) (7)

box1_area = (box1[3] - box1[l]) x (box1[2] - box1[0]) (8)

box2_area = (box2[3] - box2[l]) x (box2[2] - box2[0]) (9)

union-area = (boxl_area + box2_area) - inter-area (l0)

IOU = inter-area/union-area (ll)

The advantage of YOLO v3 is that some changes are included in error function and for objects of small to a considerable size detection occurs on three scales. The multiclass problem turned in a multilabel problem, and the performance improved over small size objects.[3]

B. SORT Algorithm for Object Tracking

The Simple Online and Realtime Tracking (SORT) algorithm is utilized for tracking the identified objects across consecutive frames. SORT employs a combination of Kalman filtering and Hungarian algorithm-based data association to provide robust tracking even in challenging scenarios, such as occlusions and temporary disappearance of objects. The algorithm maintains a set of active tracks and efficiently updates them as new detections are made.[4]

C. Integration of YOLOv3 and SORT

(

The integration of You Only Look Once version 3 (YOLOv3) and the Simple Online and Realtime Tracking (SORT) algorithm represents a synergistic approach to object detection and tracking. YOLOv3 excels in real-time object detection by dividing an image into a grid and simultaneously predicting bounding boxes and class probabilities for each grid cell. The detected objects, along with their confidence scores, serve as the input for SORT, a tracking algorithm that leverages Kalman filtering and the Hungarian algorithm for data association. This integration addresses the challenges posed by dynamic scenes, enabling the system to not only identify objects efficiently but also to track their trajectories robustly across frames. By fusing the strengths of YOLOv3 and SORT, the system achieves a comprehensive solution for real-world scenarios, ranging from crowded environments to scenarios with occlusions.[5]

D. Training Fish Object in YOLO:

To extend the capabilities of the YOLOv3 model for specific object detection, a fish object was incorporated into the pre-trained YOLOv3 framework. This involved augmenting the existing COCO dataset with annotated fish images and retraining the model using a combination of the new and original dataset. The addition of the "fish" class enhances the YOLOv3 model's versatility, allowing it to effectively detect and locate fish instances within the video frames. This fine-tuning process ensures that the YOLOv3

and SORT integration is tailored to the unique characteristics of the fish object, contributing to the system's adaptability and effectiveness in scenarios where fish tracking is essential.[6]

E. Object Tracking with SORT:

The proposed system seamlessly integrates the processed bounding boxes, filtered through our methodology, into the SORT (Simple Online and Realtime Tracking) algorithm, constituting a crucial phase in our object tracking approach. SORT efficiently manages object tracks in real-time, updating existing ones and generating new tracks for detected objects. Leveraging Kalman filtering for prediction and the Hungarian algorithm for association, SORT ensures the continuous and accurate tracking of objects, even in the face of challenges such as occlusions, object disappearance and re-emergence, and simultaneous tracking of multiple objects. The system's output includes the count of detected objects and active trackers in each frame, providing a comprehensive overview of the ongoing tracking process and affirming the system's proficiency in offering meaningful insights into dynamic underwater scenes. [7]

III. EXPERIMENTAL ANALYSIS AND RESULTS

In this section, we discuss the details of the dateset used for evaluating the proposed method. We describe the ablation study of the proposed approach. Later, we presented empirical evaluation and analyze the results.

A. Dataset Details

In real-time scenarios, the appearance of objects may vary across different time instances, necessitating the initiation of a new tracking process for each object entering the sequence. The track is terminated when an object fails to associate with any trajectory for a specified number of frames. The proposed method was evaluated using the GMOT-40 Dataset, which is the first publicly available dense Dataset for Generic Multiple Object Tracking (GMOT). This Dataset comprises 40 meticulously annotated sequences, evenly distributed among 10 object categories, featuring dense objects with over 80 instances of the same class possible in one frame. The Dataset is characterized by high-quality manual annotations, providing careful inspection in each frame. It introduces diversity in target classes, encompassing large variability both within and between sequences of the same class, while presenting real-world challenges such as occlusion, target entering/exiting, motion blur, and deformation. The challenging one-shot GMOT protocol is adopted for evaluation, and a series of baseline algorithms are introduced to address these complexities. Specifically, three fish sequences, namely fish-1, fish-2, and fish-3, were selected from the GMOT-40 Dataset for experimentation, demonstrating the adaptability and effectiveness of our proposed methodology in handling complex tracking scenarios.[8]

B. Performance Metrics

In order to demonstrate a fair comparison of the proposed tracker with the competing trackers available, we utilize the standard tracking performance metrics as defined below.

1) MOTA(↑): Multi-object tracking accuracy.

2) MOTP(↑): Multi-object tracking precision.

3) FAF(↓): number of false alarms per frame.

4) MT(↑): number of mostly tracked trajectories. I.e. target has the same label for at least 80% of its life span.

5) ML(↓): number of mostly lost trajectories. i.e. target is not tracked for at least 20% of its life span.

6) FP(↓): number of false detections.

7) FN(↓): number of missed detections.

8) ID sw(↓): number of times an ID switches to a different previously tracked object.

9) Frag(↓): number of fragmentations where a track is interrupted by miss detection.

Evaluation measures with (↑), higher scores denote better performance; while for evaluation measures with (↓), lower scores denote better performance.[9]

Table 1 : Performance of the approaches on original GMOT-40 benchmark sequences and its enhanced sequences.

	Original Fish-1 Vedio	Fish-1 Enhanced	Original Fish-2 Vedio	Fish-2 Enhanced	Original Fish-3 Vedio	Fish-3 Enhanced
MOTA(↑)	15.19	17.28	24.20	26.30	36.05	38.25
MOTP(↑)	35.42	40.30	34.24	48.34	32.38	39.22
FAF(↓)	4.75	4.51	5.99	6.03	4.53	4.51
MT(↑)	6.33	6.75	7.16	7.43	9.64	10.11
ML(↓)	34.67	33.65	49.84	45.87	40.36	41.87
FP(↓)	70	67	379	390	744	730
FN(↓)	771	760	4602	4200	1273	1200
ID sw(↓)	600	590	983	899	1267	1234
Frag(↓)	1700	1564	2698	2487	2631	2420

C. *Original and enhanced trajectories*

In this section, we will illustrate sample trajectories of fish-1, fish-2, and fish-3 in consecutive frames (a) to (e) original and (f) to (j) for enhanced videos. This comparative analysis aims to visually demonstrate the efficacy of our tracking methodology in capturing and enhancing object trajectories over time.

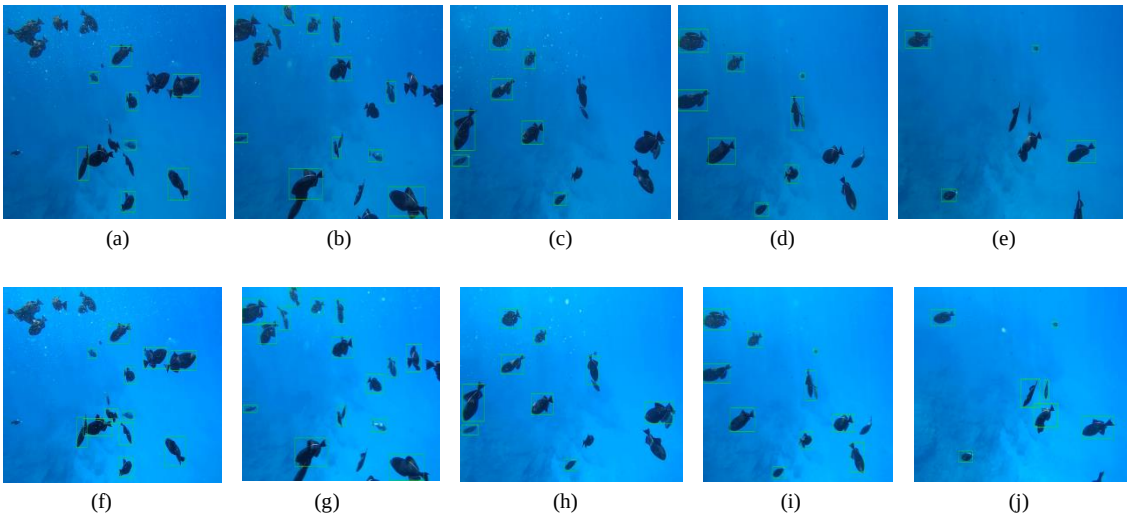

Fig. 4. Sample trajectories in (a) Frame t, (b) Frame t + 1, (c) Frame t + 2, (d) Frame t + 3, and (e) Frame t + 4 , original video of Fish -1, GMOT-40 Dataset. Sample trajectories in (f) Frame t, (g) Frame t + 1, (h) Frame t + 2, (i) Frame t + 3, and (j) Frame t + 4 , enhanced video.

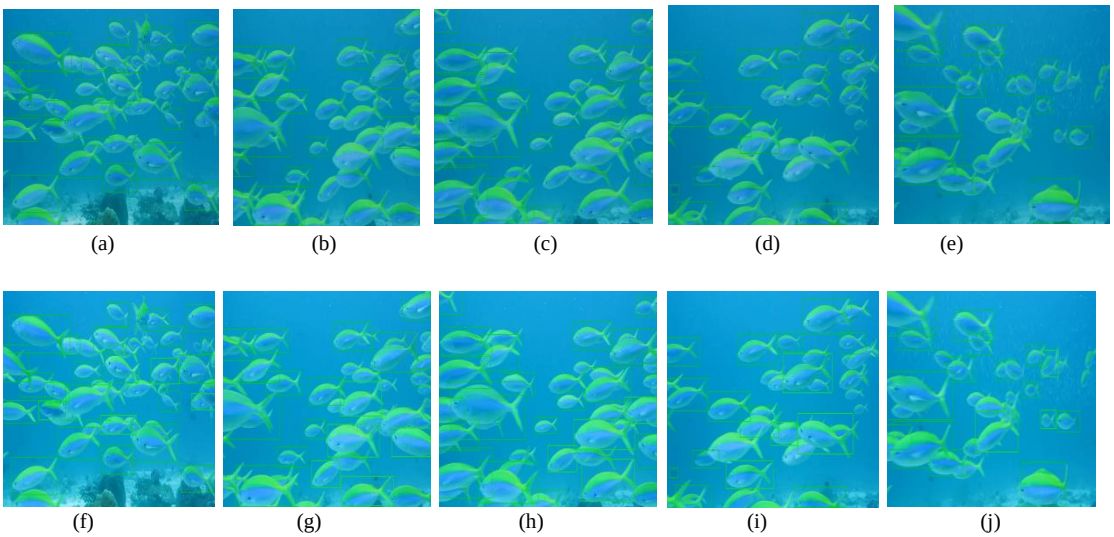

Fig. 5. Sample trajectories in (a) Frame t, (b) Frame t + 1, (c) Frame t + 2, (d) Frame t + 3, and (e) Frame t + 4 , original video of Fish -2, GMOT-40 Dataset. Sample trajectories in (f) Frame t, (g) Frame t + 1, (h) Frame t + 2, (i) Frame t + 3, and (j) Frame t + 4 , enhanced video.

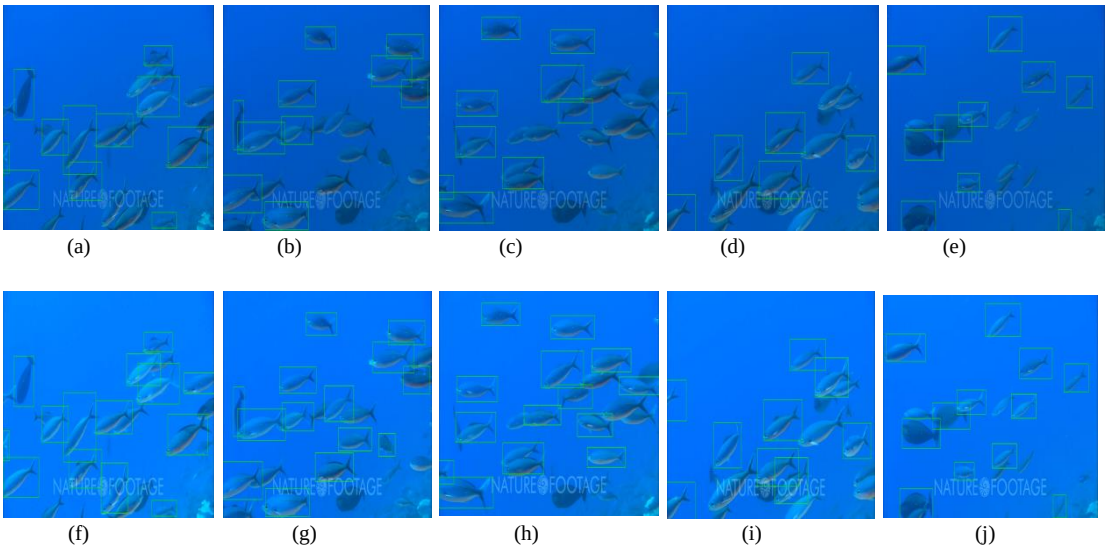

Fig. 6. Sample trajectories in (a) Frame t, (b) Frame t + 1, (c) Frame t + 2, (d) Frame t + 3, and (e) Frame t + 4 , original video of Fish -3, GMOT-40 Dataset. Sample trajectories in (f) Frame t, (g) Frame t + 1, (h) Frame t + 2, (i) Frame t + 3, and (j) Frame t + 4 , enhanced video.

IV. CONCLUSION

This research introduces a robust solution that seamlessly integrates YOLOv3 and SORT for comprehensive object detection and tracking. By harnessing the real-time capabilities of YOLOv3 for precise object detection and the tracking resilience of SORT for maintaining consistent object trajectories, our combined approach proves to be highly effective. Notably, the experimental results underscore the system's prowess in dynamic underwater environments. Crucially, our methodology goes a step further by applying YOLOv3 and SORT to both original and enhanced underwater images, the latter being enhanced using autoencoders. Strikingly, the metrics obtained for the enhanced images surpass those of the original ones, highlighting the significant improvement in object detection and tracking accuracy. This achievement underscores the potential of our proposed system for diverse applications in the realms of computer vision and video analysis, particularly in challenging underwater scenarios.

REFERENCES

[1] N. Jain, S. Yerragolla, T. Guha and Mohana, "Performance Analysis of Object Detection and Tracking Algorithms for Traffic Surveillance Applications using Neural Networks," 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2019, pp. 690-696, doi: 10.1109/I-SMAC47947.2019.9032502.

[2] J. Redmon, and A. Farhadi, “Yolov3: An incremental improvement,” arXiv preprint arXiv:1804.02767, 2018.

[3] P. Adarsh, P. Rathi and M. Kumar, "YOLO v3-Tiny: Object Detection and Recognition using one stage improved model," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 687-694, doi: 10.1109/ICACCS48705.2020.9074315.

[4] Simple Online and Realtime Tracking, Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, Ben Upcroft, arXiv:1602.00763 [cs.CV]

[5] K. O. Maung and T. Myint, "Performance Evaluation of Visual Object Tracking using YOLO Deep SORT with LCF," 2022 37th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), Phuket, Thailand, 2022, pp. 189-193, doi: 10.1109/ITC-CSCC55581.2022.9894855.

[6] Z. Chang, M. Wang, Z. Wei and J. Yu, "A Bionic Robotic Fish Detection Method by Using YOLOv3 Algorithm," 2020 Chinese Automation Congress (CAC), Shanghai, China, 2020, pp. 6045-6050, doi: 10.1109/CAC51589.2020.9326799.

[7] S. Manzoor, K. -H. Sung, Y. Zhang, Y. -C. An and T. -Y. Kuc, "Qualitative Analysis of Single Object and Multi Object Tracking Models," 2022 22nd International Conference on Control, Automation and Systems (ICCAS), Jeju, Korea, Republic of, 2022, pp. 1539-1545, doi: 10.23919/ICCAS55662.2022.10003784.

[8] H. Bai, W. Cheng, P. Chu, J. Liu, K. Zhang and H. Ling, "GMOT-40: A Benchmark for Generic Multiple Object Tracking," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 6715-6724, doi: 10.1109/CVPR46437.2021.00665.

[9] K. Bernardin and R. Stiefelhagen, “Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics,” Image and Video Processing, , no. May, 2008.