Deep Dive: On Estimating Fish Biomass Using Monocular Unconstrained Underwater Videos

Final Year Project

Report by

Usman Jalil (346172)
Syed M. Abubakar (337385)
Muhmmad Saad (33414)

In Partial Fulfillment

Of the Requirements for the degree

Bachelor of Electrical Engineering

(BEE)

School of Electrical Engineering and Computer Science
National University of Sciences and Technology
Islamabad, Pakistan

(2024)

DECLARATION

We hereby declare that this project report entitled "***Deep Dive: On Estimating Fish Biomass Using Monocular Unconstrained Underwater Videos***" submitted to the "School of Electrical Engineering and Computer Sciences (SEECS)", is a record of an original work done by us under the guidance of Supervisor "Dr. Ahmad Salman" and that no part has been plagiarized without citations. Also, this project work is submitted in the partial fulfillment of the requirements for the degree of Bachelor of Electrical Engineering

Team Members

Usman Jalil

Syed M. Abubakar

Muhammad Saad

Supervisor

Dr. Ahmad Salman

Date:

24-04-2024

Place:

National University of Science and Technology, H-12 Islamabad

DEDICATION

We would like to dedicate our work to our teachers and parents, without whom this could not have been possible.

ACKNOWLEDGEMENTS

We would like to wholeheartedly thank our advisor Dr. Ahmad Salman and the co-advisor Ma'am Hafsa Iqbal for helping us throughout the final year project. Without their keen guidance and support, it would not have been possible for us to meet the scope of the project in time.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Thousands of terabytes of underwater imagery or video data is available today in a pursuit of rapid sampling through AI and computer vision techniques. Such methods provide support for laborious manual sampling to estimate relative abundance and biomass of fish fauna by marine scientists and conservationists. The fact that most of the video data is available in monocular form poses a great challenge in automatic fish biomass estimation through modern AI and machine learning tools which need stereo vision to calculate depth and consequently size and mass of fish. Furthermore, challenges such as poor luminosity, dynamic backgrounds, water murkiness and fish camouflage persist, hampering effective detection and classification of fishes which ultimately leads to poor biomass estimate. To bridge this gap, this study proposes a hierarchical deep learning pipeline to accurately detect and classify fish species followed by specially designed monocular RGB to depth imagery translation module in the domain of generative AI. To achieve this, a co-existing Gaussian mixture models and YOLOv8 deep neural network yield accurate detection and classification using temporal and spatial fish-dependent features. Then a neural autoencoding algorithm supported by fish tracking and segmentation calculate size and mass of fish ending up in a species-wise biomass estimation. A benchmark Fish4Knowledge Life CLEF 2015 monocular dataset is used to estimate biomass of 15 species in LifeClef-15 covering Taiwan coral reefs holding a rich fish biodiversity. By gathering standard average size data of species and comparing the figures with our estimates by the proposed model yields 5.44% error in mean square terms with the standard deviation of 8.809%. This novel research encourages deployment of state-of-the-art computer vision

and machine learning approaches in fish detection, classification and biomass estimation for the available huge underwater monocular video data, a big step towards conservation and habitat monitoring.

Chapter 1

INTRODUCTION WITH LITERATURE REVIEW

The automated estimation of fish populations through the analysis of underwater video footage holds significant importance for marine scientists as it enables them to gauge the relative abundance and biomass of different species across various marine ecosystems. This critical information is instrumental in safeguarding endangered species from the perils of overfishing and environmental shifts. Furthermore, determining the maximum count of a specific fish species aids marine scientists in establishing environments conducive to fostering greater fish biodiversity in specific regions.

The advent of rapid data acquisition through underwater camera systems has made thousands of hours of video data available from diverse regions around the world. However, manually observing and sampling such vast volumes of data presents a labor-intensive and cost-prohibitive challenge for marine biologists and conservationists, despite the undeniable merits of its non-destructive nature. In contrast, the automatic sampling of fish using modern machine learning and computer vision tools is increasingly garnering attention from marine and fisheries communities as an essential requirement. Video-based fish detection is an essential precursor to their species classification and then mass estimation, as each video frame may contain multiple fish belonging to various species. To address this challenge, numerous machine learning (ML) algorithms are available; however, they grapple with the considerable variability within species, non-rigid deformations, alterations in orientation, reduced visibility, and intricate lighting conditions.

In the past, marine scientists have relied on textural features to detect and classify fish, utilizing classical methods such as Principal Component Analysis (PCA), hierarchical decision trees with Support Vector Machines (SVM), and Gaussian Mixture Models (GMM). These techniques aimed to automate fish sampling [16, 38, 42, 15]. However, given the intricate nature of the problem, recent research endeavors have shifted towards Deep Neural Networks (DNN) for

tasks like fish detection, tracking, and classification. For example, fish classification based on Convolutional Neural Networks (CNN) is applied to the LifeCLEF 2014 and 2015 datasets, which featured an extensive class distribution [47]. Similarly, YOLO, a renowned object detector, is used to achieve a remarkable 93% detection accuracy for a dataset comprising 839 fish samples [51]. Another study [58] also leveraged YOLO for fish detection across multiple datasets, achieving an impressive mean average precision score of 54%. In a different approach, a multi-class SVM on AlexNet CNN features are utilized for the LifeCLEF 2015 fish detection task, securing a 74% F-score [28]. In another work, image enhancement strategy tailored for the task of coral reef fish detection is introduced for the LifeCLEF 2015 dataset [50]. Their method employ saliency maps generated through a Siamese network to effectively reconstruct input images, markedly improving visual clarity despite challenges such as variations in luminosity, the presence of moving background objects, and image blurriness. Subsequently, they applied Cascade-RCNN to the refined dataset, attaining an F-score of 81.7%.

Taking a distinct route, authors [59] propose pre- and post-processing techniques applied to deep learning to extract fish patches for detection. GoogleNet [5] for fish detection and classification achieves an F-score of 84% for 15 species on the same dataset. On the other hand, a combination of GMM features and pixel-wise posterior analysis for fish detection in complex backgrounds is proposed, attaining an average F-score of 84.28% in the LifeCLEF 2014 fish detection challenge. Furthermore, a study presents promising results on LifeCLEF 2015 dataset by adopting a hybrid approach combining temporal and CNN features. They utilize GMM and optical flow features over raw sequential images and applied the Resnet-50 fish classifier for fish identification. Moreover, they employed the YOLOv3 architecture on raw images in parallel, ultimately achieving F-scores of 95.47% and 91.64% for fish detection and classification, respectively. Further advancements have involved the incorporation of

dense optical flow as temporal features, along with CNN analysis of fish grazing behavior on specific fish species. This integration has demonstrated improvements in classification performance, a finding endorsed by in a controlled environment, achieving over 95% accuracy in a dataset of more than 1,000 videos. Additionally, applied optical flow and 3D CNN for fish feeding intensity estimation, achieving 95% accuracy in their collected dataset of 24 videos under constrained conditions.

In a recent contribution, a novel application of semi-supervised learning for fish detection has been introduced, specifically focusing on the FishInTurbid Water dataset. Their approach harnesses the potential of weakly-labeled datasets through an ensemble of two deep neural networks, resulting in a high level of accuracy with reduced turnaround time. The methodology involves training a self-supervised model on unlabeled data, followed by fully supervised incremental learning using weakly-labeled data. Furthermore, the researchers implement a pioneering weakly-supervised XGBoost ensemble, incorporating pre-trained DNNs such as EfficientNet and Vision Transformer, with fine-tuning on the FishInTurbidWater dataset. This multifaceted approach yields an impressive 93.6\% F-score, offering an innovative perspective on semi-supervised learning techniques for fish detection.

The application of unsupervised learning for fish tracking and segmentation has been explored using optical flow and CNN on the DeepFish video dataset, with an average precision and recall of 50% and 72%, respectively. This approach, which combines temporal features with learning-based solutions, extends beyond fish fauna detection in underwater imagery, showcasing the benefits of feature combination in both temporal and spatial contexts.

Fish biomass estimation traditionally involves multiplying the total count of fish within a designated aquatic region by the average weight of sampled fish. In the last decade, many biomass estimation studies relied on taking the fish out of water for measuring its size or

other parameters like length or area. However, this method has been critiqued for its reliance on manual weighing techniques, which are not only labor-intensive and time-consuming but also prone to inaccuracies. Additionally, manual handling of fish during weighing procedures has been shown to induce stress, adversely impacting their health and well-being. Consequently, the conventional approach to biomass calculation presents notable limitations.

In recent years, considerable attention has been devoted to the exploration of automatic and non-invasive methodologies for estimating underwater fish biomass. These methods include Machine Vision, Acoustics, and environmental DNA. Using advanced Computer Vision techniques can help in the calculation of fish mass without the need for human intervention. Many of these methodologies rely on the utilization of 2D computer vision systems and external devices, such as polystyrene boards, conveyor belts, tanks, and boxes, to facilitate the estimation of biomass alongside other parameters such as length, width, area, and weight. The utilization of stereo-vision technology has gained popularity in addressing the challenges posed by the variability of distance and angle in the bodies of free-swimming fish, which cannot be effectively managed by 2D computer vision systems. This approach has been highlighted in recent studies.

Our primary focus in this research centers on enhancing contemporary state-of-the-art to estimate the biomass of fish. This enhancement is achieved through a meticulously crafted algorithm designed to capture fish motion-related features while concurrently eliminating irrelevant noise and background interference. Our evaluation draws upon the LifeCLEF 2015 dataset, which pose challenges stemming from the underwater environment's inherent variability, characterized by poor visibility and aquatic background ambiguity. Out contributions are summarized as follows:

- A spatio-temporal hierarchical pipeline is proposed for fish detection where GMM in conjunction with ResNet-152 is

utilised to capture fast moving fish while YOLOv8 extract static fish candidates.

- The detected fish candidates are classified by the YOLOv8 and then tracked across the video frames to omit the missed detection and enhance the recall rate yielding better F-Score.
- A CNN-based autoencoder network is devised with residual inter-block connection enriched with multi-head attention mechanism to project monocular underwater imagery to depth maps. These depth maps are help in estimating the fish size and consequently biomass of the classified species by keeping the the account of fish tracks.

1.1 OUTLINE OF THE REPORT

This report comprises 4 major chapters, each of which focuses on a specific portion of the final year project.

- **Chapter 1 gives the literature overview to highlight the limitations of traditional methods and elucidates the shift towards using advanced computer vision techniques for biomass estimation.**
- **Chapter 2 explains the in-depth information about dataset and complete methodology of our approach for biomass estimation.**
- **Chapter 3 validates our approach by showcasing the results of our methodology and cross papers analysis.**
- **Chapter 4 provides information about Graphical User Interface of our Biomass detector along with its real time performance.**

Chapter Summary

The introduction outlines the shift towards automated and non-invasive methods for estimating underwater fish biomass, highlighting the limitations of traditional manual weighing techniques and the emergence of advanced Computer Vision and stereo-vision technologies.

Chapter 2

METHODOLOGY AND DATASET

In this section we will discuss comprehensive details about the datasets used in this article. After that we will provide details about the architecture and Methodology proposed in our design. We will also talk about the motivation and reasoning. behind our Methodology and how it is beneficial.

2.1 DATASET

The dataset we used is LifeClef-2015 that has a set of videos. The LifeClef-2015 dataset includes a total of 93 videos of flv format, each containing instances of 15 distinct fish species. This dataset is a subset of a more extensive collection of underwater videos known as Fish 4Knowledge, as documented by Joly et al. in 2016. The dataset has a split of 21.5% training data and 78% test data. The resolution of the videos is 640 x 480 for training and 320 x 240 for the test set. The data is manually annotated with fish locations (as bounding boxes) and species. In addition to this, there is a provided set of samples for each species. The dataset contains a total of more than 14,000 annotations and 20,000 sample images. The dataset is unbalanced in terms of the number of appearances of fish species in frames as shown in the table.

The LifeClef-15 dataset comprises of all the environmental variations and video distortions that we stated above as challenges. To determine the robustness of our algorithm, it is required that our data set consists of such diversities. Some of these variabilities are discussed below:

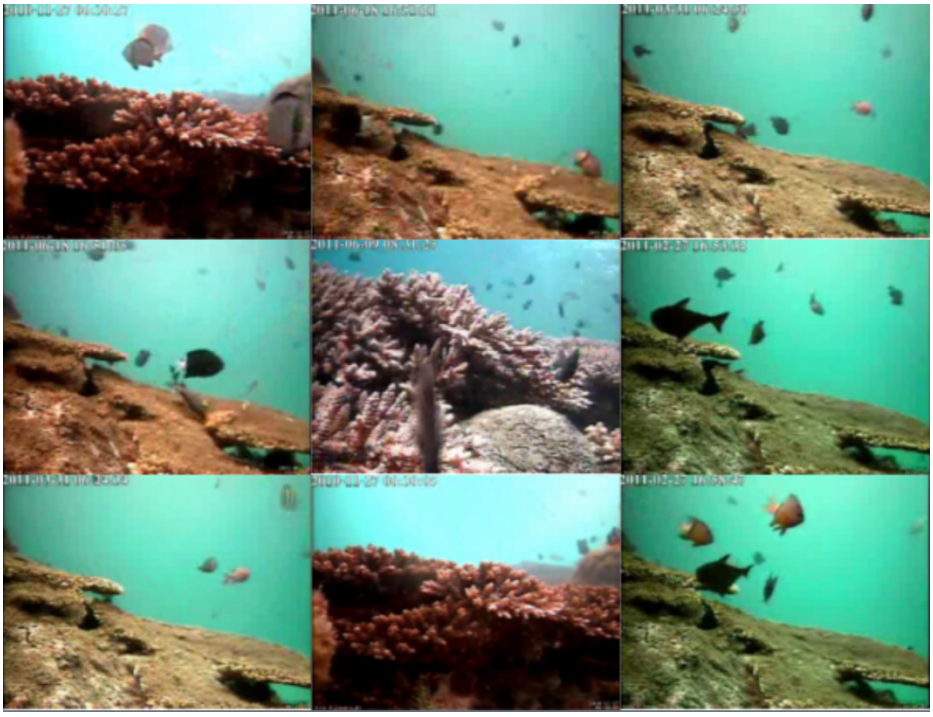

Figure 1 Overview of Life Clef 15 Dataset

1. **Blurry Vision:** Two of the videos consists blurry vision.
2. **Background Complexity**: Six of the videos included with complex background making the fish disguised and difficult to extract foreground from background.
3. **Low Resolution:** All of the videos have a very low resolution.
4. **Crowded**: Seven of the videos comprise of high density of movement.
5. **Noisy**: One video with intensive noise.

The total instances of fish in training set is 13647 whereas the test set includes 5936 instances. The fish **Hemigynous Melapterus** has 0 occurrences in the Ground Truth. Hence, it is not mentioned in the table.

LCF-15 Species	No. of Images
Abudefduf vaigiensis	434
Acanthurus nigrofuscus	2770
Amphiprion clarkii	3265
Chaetodon lunulatus	3544
Chaetodon speculum	162
Chaetodon trifascialis	704
Chromis chrysura	3859
Dascyllus aruanus	1749
Dascyllus reticulatus	5327
Hemigymnus melapterus	361
Myripristis kuntee	3231
Neoglyphidodon nigroris	213
Pempheris vanicolensis	906
Plectroglyphidodon dickii	3102
Zebrasoma scopas	343

Table 1 Species Instances

2.2 METHODOLOGY

Utilizing deep neural networks (DNN) remains the primary strategy for fish detection and species classification in both images and videos. However, improving DNN capabilities can be achieved by integrating traditional computer vision algorithms. Given recent advancements in spatio-temporal feature extraction, our proposed methodology is inherently hybrid. We incorporate temporal features obtained from segmenting moving fish and eliminating static background using a direction-based clustering technique applied to the magnitude of motion within video frames. Moreover, a specialized branch, powered by a modified YOLOv8 fish detector, is further reinforced by merging the outcomes of our motion-based segmentation algorithm. This integrated approach aims to enhance fish detection performance, particularly in addressing the distinctive challenges presented by dataset like LifeCLEF 2015.

2.2.1 Static Image Based Detection

Our pipeline includes two branches. The first branch is based on static image-based detection system. Here We relied on the capabilities of YOLOv8. YOLOv8 is an advanced real-time object detection model known for its efficiency and accuracy. It operates by dividing the input image into a grid and predicting bounding boxes and class probabilities for each grid cell. Unlike traditional object detection models that apply a sliding window approach, YOLOv8 predicts objects in a single pass through the network, making it significantly faster. Similarly in our case, Yolov8 predicted the bounding box coordinates (x, y, width, height) along with confidence scores for each detected object in an image and classified them from the 15 annotated species in the dataset.

2.2.2 Motion Based Detection

The second branch is based on motion-based detection where we utilized Gaussian Mixture Model, Gaussian mixture models, or GMMs, represent an unsupervised learning technique employing probabilistic models to discern normally distributed subpopulations within a larger population. For detecting fish within videos, GMMs rely on temporal variations. The RGB input frames were converted to binary images with a certain threshold. The binary images were processed by GMMs to model the background, effectively segmenting out foreground elements. In this process, each pixel value is treated as a feature at fixed background locations, forming feature vectors across multiple frames, corresponding to the total pixel count in a video frame. To assign a bounding box for each segmented foreground output within binary images, we computed its coordinates and mapped these bounding boxes onto the original RGB frames.

Effective utilization of GMMs necessitates a substantial dataset for training and estimating parameters such as mean and covariance of the background. This enables the model to distinguish fish from non-fish entities, encompassing diverse elements like kelp, coral reefs, sea

grass beds, and other aquatic flora, as well as sessile invertebrates such as sponges, gorgonians, and ascidians, along with the physical seabed structure. Statistical patterns exhibited by stationary structures, such as coral reefs and seabed formations, differ distinctly from the movement patterns of fish. Additionally, objects with constrained movement, like swaying kelp and aquatic plants, possess statistical signatures divergent from fish movement. Leveraging temporal changes in frames, GMMs can also identify unannotated fish instances. However, subsequent filtering based on Intersection over Union (IoU) criteria ensures retention of only annotated fish detections. Once the fish are detected using techniques such as Gaussian mixture models (GMMs) to extract motion-based features from video data, they are passed through a classification stage. This stage employs a deep learning model called ResNet-152. Using ResNet-152 model we classified the species in our motion based detection branch.

The GMM can becomes disadvantageous at some situations if we solely rely on it. The fish with small temporal changes, such as the stationary fish in our video sequence are failed to be detected by GMM. GMMs are effective for modeling temporal changes in pixel distributions, making them suitable for detecting moving fish. However, they were unable to detect stationary fish due to the lack of noticeable changes in pixel intensities over time. However, YOLOv8 effectively addresses the detection of stationary fishes with significant accuracy.

In numerous instances, both the YOLO and GMM models produced bounding boxes for the same fish, leading to ambiguity in selecting the correct bounding box. Furthermore, there were scenarios where GMM detected a fish missed by Yolo or GMM detected an unannotated fish. To address these cases, we employed Intersection over Union (IoU) as a metric to compare the overlapping areas between the bounding boxes generated by both models and the ground truth boxes. IoU quantifies the extent of overlap between two

bounding boxes by calculating the ratio of their intersection to their union.

$$IoU = \frac{|\, A \cap B \,|}{|\, \mathrm{A} \cup \mathrm{B} \,|}$$

Here, A represents the bounding box obtained from either YOLO or GMM, while B denotes the ground truth box. In cases where both models detected the same fish, we selected the bounding box with the higher IoU value. For instances where GMM detected fish missed by YOLO, we set a threshold for the IoU value. Bounding box pairs with IoU values exceeding this threshold were deemed overlapping detections between GMM and the ground truth, thus retaining them. However, for detections by GMM that were unannotated, no IoU value was applicable, leading to their exclusion. Finally, the retained detections from both GMM and YOLO were merged to create the final set of fish detections. The unified output image is fed into the tracking module explained next.

2.2.3 Tacking

In the detection pipeline, we implemented a tracking module in order to make sure that specie is being detected throughout the frames. This is a necessary step to precisely estimate the biomass of fish in the later phase. We used the methodology to track species along with detections. It involves two stages (i) The Detection Stage and (ii) Association Stage. The detection stage is already implemented using Yolo which gives the bounding box. In the association stage, ByteTrack associates every detection box in each frame to create object tracks. It introduces a novel association algorithm based on bipartite graph matching, which considers both motion and appearance information for association. Through this way we will know that this specie is already detected and will not increase the number of counts in Biomass Calculations.

2.2.4 Biomass Estimation

Fish biomass is a critical measure in aquatic ecology, representing the total weight of all fish in a particular ecosystem, habitat, or population at a specific time. It helps scientists understand how many fish are present and where they are located within water environments. This information is crucial for managing fisheries effectively and protecting aquatic ecosystems. To calculate fish biomass, researchers typically use the following formula which is widely used and accepted in fisheries science:

$$Fish\ Biomass = Number\ of\ fishes \times Total\ mass\ observed$$

Our primary objective is to determine the amount of fish biomass in an underwater setting. To achieve this, we require both the quantity and weight of the observed fish species. Utilizing the YOLO + GMM model, we identify and document the count of fish from 15 distinct species in the dataset.

2.2.5 Mass Calculations

Previous studies extensively explored the connection between the size in square centimeters, length in centimeters, and other factors such as fish height in centimeters, with its mass in kilograms. Among the historical precedents, Fulton (1904) proposed a influential method rooted in the relationship between fish length and mass, which was aimed to establish mathematical relationships among these factors, enabling the estimation of fish biomass using any of the mentioned parameters.

$$Fish\ Weight = a \times Length^{b}$$

where a & b are constants derived from regression techniques.

This study examines the relationship between the length and weight of 15 fish species in the LifeClef Dataset to gauge their combined biomass. This approach maintains uniformity in our computations. In this article, the terms "mass" and "weight" are used interchangeably when discussing the fish. They both indicate the amount of matter the fish contains, measured in grams.

The formula shows that we need the fish's length to estimate its weight. Estimating length is tough, especially on a monocular dataset. To address this, we created a scale that translates the pixel length of a detected fish into real lengths using depth data. We devised a simple scale that uses the fish's pixel length and depth (which we'll discuss later) to calculate the fish's length in centimeters.

$$scale = \frac{\text{Pixel Length x depth from camera}}{\text{actual length [cm]}}$$

$$actual\ length\ [cm] = \frac{\text{pixel length x depth from camera}}{scale}$$

To construct our scale, we began by determining the species of fish depicted in a given video using known factual information. Next, we calculated the average length of this fish species using data from FishBase, an online database. Our sole assumption was that the fish in our dataset were healthy and represented an average size for their respective species. Using this methodology, we derived the average lengths for each fish species, which can be verified by referencing a corresponding table.

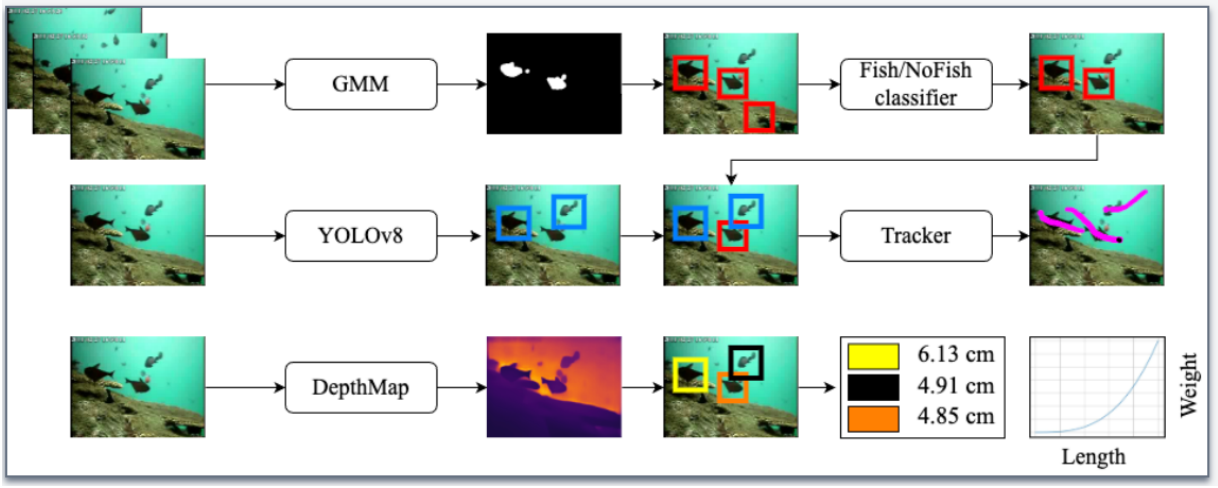

Figure 2 Overall Methodology

The fish length is determined by measuring from its head tip to the middle of its tail. Our study utilized the YOLOv8 + GMM model, which demonstrated high performance. Consequently, we opted not to employ segmentation for fish length estimation. Instead, we relied on YOLOv8 bounding boxes and their pixel length to approximate the fish's size from head to tail midpoint, demonstrated in Figure a.

Due to the continuous movement of the fish in our dataset, accurately measuring its length is challenging, as it aligns with the camera for only a brief period (1-2 frames). To address this issue, we maintain a record of fish lengths corresponding to their unique tracker IDs throughout the entire video. Subsequently, we use the maximum recorded length to estimate the fish's actual length in centimeters.

Moreover, due to uncertainty about the fish's orientation and to prevent including its width when it moves vertically in the video, we compute the mean value of the length of the bounding box and its vertex.

2.2.6 Depth

In our study, we employed a blend of transformers and a U-net structure to gauge depth in underwater settings. Our model underwent training using more than 1.5 million annotated images from the Kitti + NYU dataset, resulting in superior RMSE performance

compared to prior methods. This model outperformed existing models such as Zoe-Depth in metric depth estimation, where each pixel's depth is expressed in meters. Due to its exceptional performance, it was considered an optimal choice. Additionally, we applied transfer learning techniques to adapt this model to our LifeClef dataset for further analysis.

To determine the depth of a solitary fish, we utilize the bounding box provided by YOLOv8 and obtain the depth value from the depth array generated by the Depth-anything model at the precise center of this bounding box. It is reasonably assumed that the fish will consistently occupy the precise center of the bounding box. Consequently, this allows for the calculation of the fish's depth.

Chapter Summary

This Chapter give us the detailed information of Dataset and complete methodology of our approach to estimate biomass of fish in monocular dataset.

Chapter 3

RESULTS & DISCUSSIONS

We evaluated the performance of our architecture using the F1 score metric, a widely recognized measure in the field of machine learning and classification tasks. The F1 score takes into account both precision and recall, providing a balanced assessment of a model's ability to correctly classify instances across multiple classes. We gained insights into the effectiveness of our architecture in handling classification tasks, capturing both the model's ability to correctly identify relevant instances and its capability to minimize false positives and false negatives.

$$F1 = \frac{2 \text{ x (Precision x Recall)}}{\text{Precision} + Recall}$$

3.1 FISH DETECTION & CLASSIFICATION

Our fish detection system comprises two branches, with one employing YOLOv8 for detecting fish instances due to its exceptional accuracy and computational efficiency compared touch as Single-Shot Detector (SSD), Region Proposed Network, and Feature Pyramid Network (FPN). YOLOv8, pretrained on the extensive ImageNet dataset serves as a versatile image feature extractor, encompassing a broad spectrum of maritime categories such as goldfish, ray fish, jellyfish, and more. Leveraging this pretrained model significantly reduces the need for extensive fine-tuning on our fish dataset.

For training YOLOv8, we utilized the used of labels. The label file contains information about fish candidate locations and their corresponding labels, where legitimate fish instances are labeled as 1 along with their bounding box coordinates. Images without fish instances are represented by empty vectors. During training, YOLOv8's detector branch focuses on regression tasks for coordinates using mean square error, while the classifier branch is trained on

labels using cross-entropy loss.

The training process involves various parameters: 500 epochs, a batch size of 16, an image size of 640 × 480 pixels, a learning rate of 0.01, optimizer weight decay of 0.0005, and warm-up momentum set to 0.8. To enhance model generalization, we apply various data augmentation techniques including hue-saturation-value adjustments (HSV = [0.015, 0.7, 0.4]), translation by 0.1, mosaic augmentation at a factor of 1.0, and horizontal flipping with a probability of 0.5.

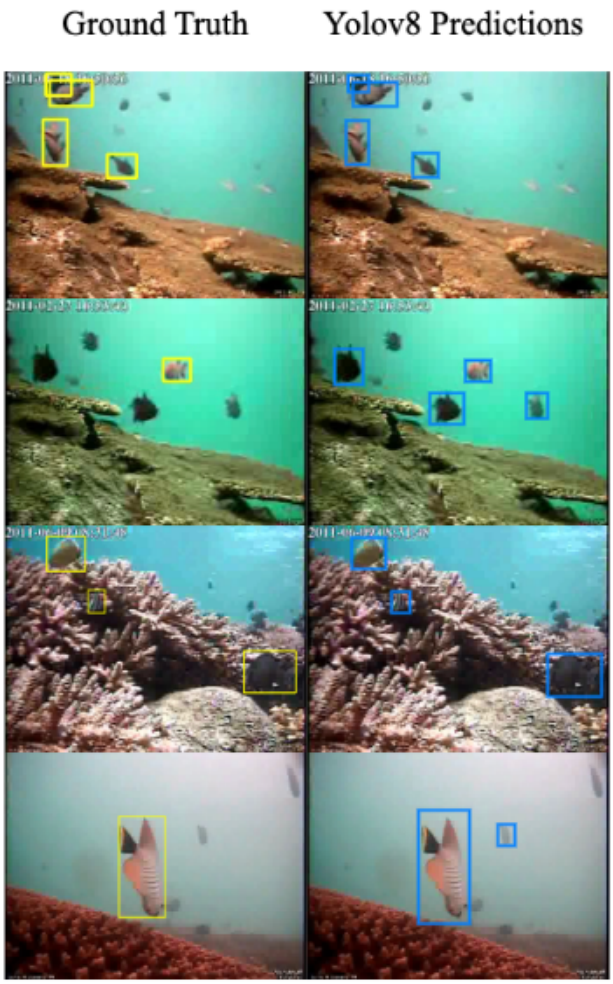

Figure 3 Yolov8 Results

The training and inference are conducted on images with dimensions of 640 × 480 and 320 x 240 pixels respectively, with detections filtered out below a confidence score of 0.6 during the inference stage. It's important to note that both the original images and the fish/non-fish crops maintain a resolution of 640 × 480 pixels. The training process for YOLOv8 took approximately 18 hours, with the

larger amount of training data provided by the Detection Set contributing to a more stable training curve and resulting in a significant improvement in the F-score on the test set. These results are specific to frame-based detection. We obtained an accuracy of 96.3% for Yolov8.

Species	Precision	Recall	F1 Score	Accuracy
Abudefduf vaigiensis	0.949	0.833	0.914	0.877
Acanthurus nigrofuscus	0.967	0.983	0.994	0.975
Amphiprion clarkii	0.954	0.985	0.985	0.969
Chaetodon lunulatus	0.988	0.999	0.992	0.993
Chaetodon speculum	0.988	1.000	0.995	0.994
Chaetodon trifascialis	0.989	0.981	0.993	0.985
Chromis chrysura	0.964	0.906	0.946	0.934
Dascyllus aruanus	0.988	0.999	0.992	0.993
Dascyllus reticulatus	0.960	0.945	0.980	0.952
Myripristis kuntee	0.943	0.981	0.992	0.962
Neoglyphidodon nigroris	0.937	0.971	0.989	0.954
Pempheris vanicolensis	0.930	0.938	0.971	0.934
Plectroglyphidodon dickii	0.942	0.920	0.953	0.931
Zebrasoma scopas	0.985	1.000	0.995	0.992
Overall	**0.965**	**0.959**	**0.975**	**0.963**

Table 2 Detection & Classification Results

We trained Gaussian Mixture Models (GMM) for background subtraction in videos. The network's training process involves the following parameters. Firstly, the threshold of variance in pixel intensities is fixed with a value of 127 indicating the minimum variance required for a pixel to be considered part of the foreground. Secondly, the background ratio that determines the model's sensitivity to classify pixels as foreground is set to a value of 0.7 indicating that a pixel must have at least 70 % probability of belonging to the background to be classified as such. Additionally, the number of Gaussian mixture models is set to 20 in this case, enabling the model

to capture complex background distributions effectively. These hyperparameters are crucial for fine-tuning the background subtraction process, balancing sensitivity to changes with computational efficiency, and ensuring accurate extraction of foreground objects from video sequences. Adjusting these parameters allows for optimized performance tailored to specific video data characteristics and application requirements. Our overall accuracy after combining Yolo and GMM came out to be 98.2% with an f-score of 97.5%.

3.3 BIOMASS CALCULATIONS

We did our computations and training on an intel core i5 processor with 32GB RAM and Nvidia 1080Ti 11GB GPU.

Specie	Avg. Length (cm)	Avg. Weight (grams)	Biomass (grams)
Abudefduf vaigiensis	7.7342	13.107	2806.297
Acanthurus nigrofuscus	13.377	54.205	22166.045
Amphiprion clarkii	6.6431	6.743	5677.834
Chaetodon lunulatus	4.4639	2.038	3781.290
Chaetodon speculum	25.87	403.588	55243.356
Chaetodon trifascialis	9.6074	16.931	27699.947
Chromis chrysura	9.4866	25.699	7416.312
Dascyllus aruanus	6.2532	4.995	14215.755
Dascyllus reticulatus	7.1927	10.766	74409.924
Myripristis kuntee	5.7681	5.182	1758.798
Neoglyphidodon nigroris	6.7393	13.107	8984.170
Pempheris vanicolensis	6.4616	4.995	3185.570
Plectroglyphidodon dickii	5.3535	2.960	2057.840
Zebrasoma scopas	6.5364	6.757	1726.235

Table 3 Biomass Calculation of Each Specie

3.4 DISCUSSION

In recent times, many computer vision and machine learning algorithms have surfaced, displaying potential across various fish-related tasks such as automatic fish detection, species classification, tracking, and biomass estimation. These innovative techniques predominantly lean on sophisticated deep learning algorithms. However, a notable research gap remains in understanding the influence of environmental factors on the performance of these systems. While state-of-the-art object detection algorithms excel at pinpointing static fish based on textural patterns, they often falter in detecting moving fish, particularly in challenging underwater settings with poor visibility. Our recent study has addressed this gap by achieving state-of-the-art performance in the fish detection task. Our approach integrates a novel algorithm for temporal or motion-based feature extraction seamlessly with spatial features through the YOLOv8 deep architecture in our hybrid machine vision pipeline (illustrated in Figure 2). The temporal features crucial for fish detection are derived through simultaneous motion segmentation and background subtraction, allowing us to filter out irrelevant background pixels with lower motion magnitudes while retaining potential fish candidates.

The YOLOv8, in its standard configuration, struggles to accurately detect fast-moving fish instances, particularly when lacking clear texture or shape due to factors like water murkiness or camouflage in the background. Our temporal pipeline effectively addresses this limitation, significantly boosting the overall F-score of our algorithm and paving the way for potential applications such as assemblage and biomass estimation with tracking, which we plan to explore in future work. Nonetheless, our proposed algorithm may fail to detect fish within the frame in some instances due to challenges like camouflage, a confusing background, and the static posture of fish. The yellow boxes represent ground truth fish locations that our algorithm missed, some of which were inaccurately labeled as fish in

the original ground truth files provided by AIMS, adding further complexity to the detection task.

Chapter Summary

This chapter presents the validation of our proposed approach via results. Being the first ones to estimate biomass on monocular dataset having underwater videos dataset with a lot of noise, murkiness, blurriness, and background complexity we achieved pretty good results.

Chapter 4

GRAPHICAL USER INTERFACE

4.1 HOW TO USE

Here is the User Guide to our Graphical User Interface of Biomass Estimation. It consists of two 3 stage algorithms. First frames get passed through the YOLO Darknet and static fish species around the corals get detected.

After that, detected frames get passed through the Gaussian Mixture Model for temporal detection of fish species. Overall, these detected frames get passed through encoder-decoder based U-NET architecture for monocular depth estimation along with length mass and biomass estimation of entire scene.

4.1.1 Video Upload

The first step in using the application is to upload a video file containing underwater footage. This could be footage captured from a camera placed underwater, showcasing various marine environments.

The interface provides a button allowing you to select the desired video file from your device. Once the video is uploaded, the application is ready to analyze its contents. Supported formats are flv, avi, mp4 with maximum size of 200 mb.

4.1.2 Start Detection

After uploading the video, you initiate the detection process by clicking the "Start Detection" button. This action signals the application to begin analyzing each frame of the video in real-time. Behind the scenes, sophisticated algorithms process the video frames, identifying any fish present within them. This detection process is performed swiftly, enabling you to see results almost instantly.

4.1.3 Real-time Processing

As the video is being processed, the application diligently scans each frame for the presence of fish. Advanced computer vision techniques, such as object detection, are employed to identify and localize fish within the underwater scene.

Additionally, the application estimates key attributes of the detected fish, such as their length and mass. These estimations are based on various factors, including the size of the fish in the frame and predefined parameters specific to each fish species.

4.1.4 Visual Feedback

Throughout the detection process, the application provides visual feedback to aid in understanding the analysis. Each frame of the video is displayed in the interface, with annotations overlaid to indicate the presence of detected fish.

These annotations highlight the location of each fish within the frame and may include additional information, such as the estimated length and mass of the fish. This real-time feedback allows users to observe the progress of the analysis and gain insights into the underwater environment.

4.1.5 Detection Results

Upon completion of the analysis, the application presents a summary of the detection results. This summary includes detailed information about the fish detected in the video, such as the number of fish belonging to each species, their average length, and the total biomass of each species. By reviewing these results, users can gain

valuable insights into the composition and characteristics of the fish population within the analyzed video.

4.1.6 User Interaction

Throughout the entire process, the interface allows for seamless interaction with the application. Users can easily upload different video files for analysis or stop the detection process if necessary. This flexibility empowers users to tailor the analysis to their specific needs and preferences, ensuring a smooth and user-friendly experience.

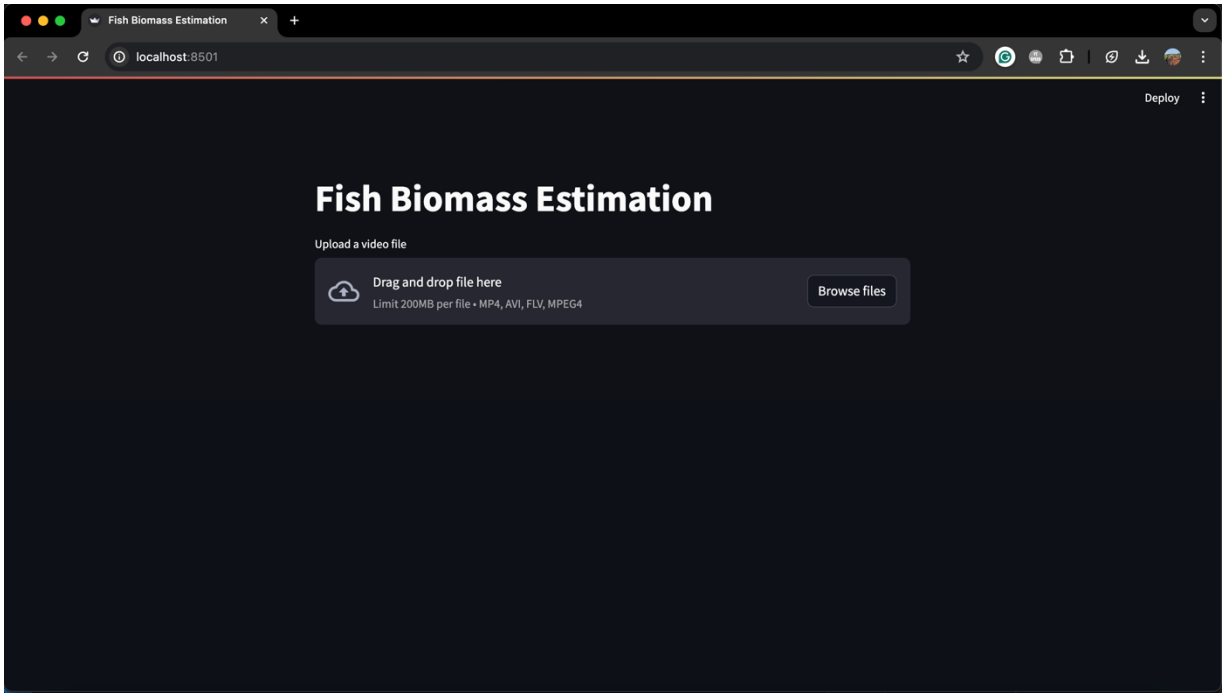

Figure 4 GUI

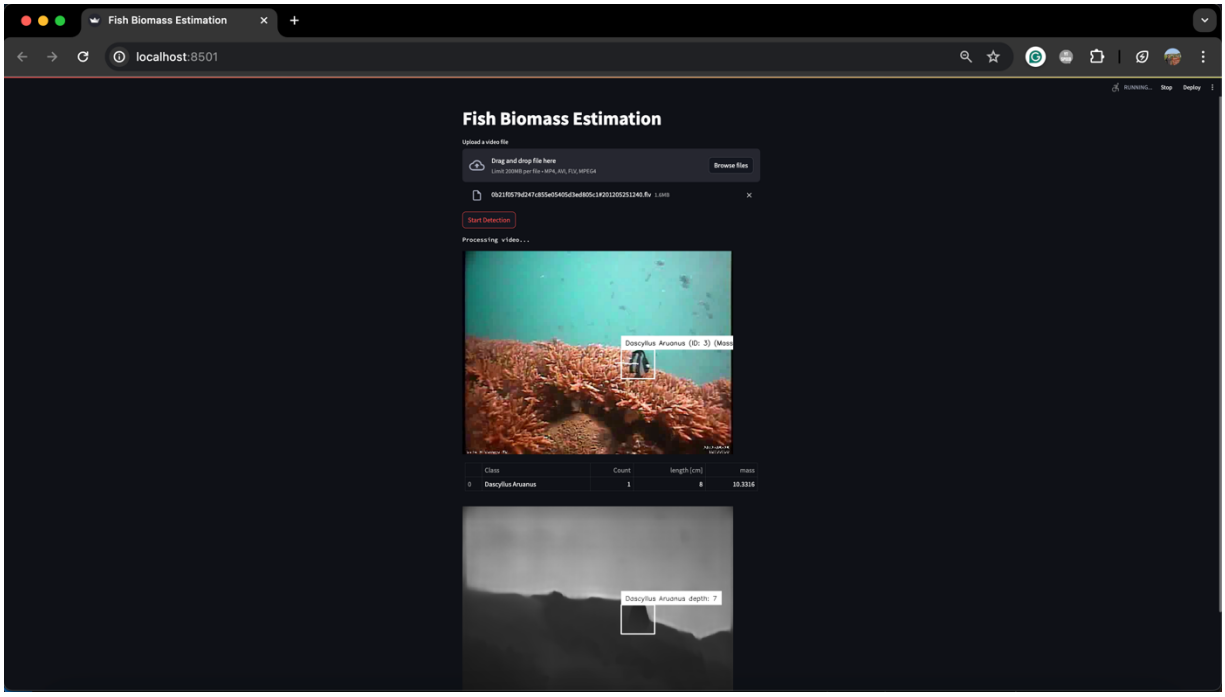

Figure 5 Real-time Processing

Chapter Summary

This chapter presents the detailed user guide of graphical user interface for biomass estimation on videos with real-time feedback.

Chapter 5

CONCLUSION AND FUTURE WORK

8.1 Conclusion

In conclusion, our study highlights the significant advancements and challenges in automated fish population estimation and biomass calculation using underwater video analysis. Leveraging datasets like LifeClef-2015, which provides a comprehensive collection of annotated videos, we've explored the efficacy of machine learning and computer vision techniques in detecting and classifying fish species. Despite the progress made, challenges such as dataset imbalance and variability in underwater conditions persist, underscoring the need for continued research and innovation in this field. Moving forward, the integration of advanced technologies like monocular and deep learning holds promise for enhancing the accuracy and efficiency of fish biomass estimation, ultimately contributing to better conservation efforts and sustainable management of marine ecosystems.

8.2 Future work

In the realm of underwater video analysis for fish biomass estimation and population monitoring, several avenues present themselves for future exploration and improvement. One promising direction involves enhancing the modalities used for biomass estimation, particularly by refining methods that incorporate depth maps derived from stereo-vision technology or other depth-sensing techniques. The inclusion of depth information can lead to more accurate estimates of fish size and biomass, thereby providing valuable insights for marine management decisions. Additionally, enriching existing datasets with depth maps can serve as a valuable resource for training and evaluating algorithms, enabling researchers to develop more robust and generalizable models. To further enhance the effectiveness of machine learning algorithms, efforts should be directed towards improving their generalization capabilities to accommodate variations in underwater conditions, such as changes in

lighting, water clarity, and fish behavior. This may involve exploring transfer learning techniques or domain adaptation strategies to enable models trained on one dataset to perform effectively on others. Furthermore, the integration of multimodal data sources, including acoustic data, environmental sensors, and satellite imagery, holds promise for enhancing the context-awareness and accuracy of fish biomass estimation systems. Collaboration between research institutions, industry partners, and government agencies will be instrumental in driving progress in underwater video analysis, fostering interdisciplinary approaches that can lead to more effective and sustainable marine management practices. Through these collective efforts, the potential for impactful contributions to marine science and conservation can be maximized.

Chapter Summary

This chapter concludes the report by summarizing the major features of our project. It presents recommendations for future work which can improve the overall system by a great extent.

Chapter 6

REFERENCES

References

[16] Huang, X. and Huang, P.X., 2014. Balance-Guaranteed Optimized Tree with Reject option for live fish recognition.

[38] Lawson, G.L., Barange, M. and Fr´eon, P., 2001. Species identification of pelagic fish schools on the South African continental shelf using acoustic descriptors and ancillary information. ICES Journal of Marine Science, 58(1), pp.275-287

[42] Palazzo, S. and Murabito, F., 2014, November. Fish species identification in real-life underwater images. In Proceedings of the 3rd ACM international workshop on multimedia analysis for ecological data (pp. 13-18).

[15] Huang, P.X., Boom, B.J. and Fisher, R.B., 2015. Hierarchical classification with reject option for live fish recognition. Machine Vision and Applications, 26(1), pp.89-102.

[28] J¨ager, J., Rodner, E., Denzler, J., Wolff, V. and Fricke-Neuderth, K., 2016, September. SeaCLEF 2016: Object Proposal Classification for Fish Detection in Underwater Videos. In CLEF (working notes) (pp. 481-489).

[50] Sun, H., Yue, J. and Li, H., 2022. An image enhancement approach for coral reef fish detection in underwater videos. Ecological Informatics, 72, p.101-862.

[51] Sung, M., Yu, S.C. and Girdhar, Y., 2017, June. Vision based real-time fish detection using convolutional neural network. In OCEANS 2017-Aberdeen (pp. 1-6). IEEE

[58] Xu, W. and Matzner, S., 2018, December. Underwater fish detection using deep learning for waterpower applications. In 2018 International conference on computational science and computational intelligence (CSCI) (pp. 313-318). IEEE.

[59] Zhuang, P., Xing, L., Liu, Y., Guo, S. and Qiao, Y., 2017, September. Marine Animal Detection and Recognition with Advanced Deep Learning Models. In CLEF (Working Notes) (pp. 166-177).

[5] Choi, S., 2015, September. Fish Identification in Underwater Video with Deep Convolutional Neural Network: SNUMedinfo at LifeCLEF Fish task 2015. In CLEF (Working Notes) (pp. 1-10).