DOI: 10.21917/ijivp.2023.0417

AN ENHANCED LIGHTWEIGHT TRANSFORMER METHOD USING STRONG ENCODER TECHNIQUES IN UNDERWATER OBJECT DETECTION

N. Sivakumar[1], A. Sumalatha[2] and K. Prabhu[3]

[1]*Department of Electrical and Electronics Engineering, Varuvan Vadivelan Institute of Technology, India*
[2]*Department of Computer Science, Kristu Jayanti College, India*
[3]*Department of Civil Engineering, GRT Institute of Engineering and Technology, India*

Abstract

The two-stage lightweight transformer methodology can now be used for underwater object detection thanks to some small adjustments that have been made to it. These adjustments were made to accommodate the needs of the underwater environment. By utilising multi-scale training, making enhancements to the backbone network of the Faster RCNN, and optimising the model response to both positive and negative examples, with an emphasis on the latter, we can achieve significant gains. The enhanced lightweight transformer method has been shown to be effective through the utilisation of comparative experiments which demonstrate that the network module may be utilised in the method to serve as the feature extraction structure. This has been shown to be the case thanks to the fact that the enhanced lightweight transformer method has been shown to be effective. We first disassemble the entire system to break it down into its component components before running the detection algorithm and the ablation tests. We can perform additional tests regarding the efficiency of the object detection algorithm for lightweight transformers. When compared to the unimproved lightweight transformer technique, the F1 result is rapidly approaching 99%, which represents a substantial leap forward in terms of quality. This contributes to demonstrating that the methods that were recommended are effective.

Keywords:

Lightweight Transformer, Strong Encoder, Underwater Object Detection

1. INTRODUCTION

Sonar methods such as sector scan sonar, side scan sonar, and synthetic aperture sonar (SAS) are examples of signal processing methods that can be used to keep track of objects that are underwater in environments that are dynamic [1]. The 1960s saw the development of a sonar technique called sector sweep.

Acoustic formation is a term that refers to echoes that are reflected by the target and then analysed by the receiver. These techniques utilise acoustic formation to determine whether the object is present. An absolute value technique, which is also known as an amplitude method, is required for the purpose of this acoustic analysis to get an approximation of the time-of-flight of the reflected signal in water [2].

This method is also known as an amplitude method. On the other hand, amplitude methods have a number of drawbacks that decrease their effectiveness when it comes to tracking [3]. The current emphasis of research that is still being carried out is on the numerous applications that can be found underwater. Object detection, imaging of wrecks, and underwater monitoring are just a few instances of the many possible implementations of this technology [4].

However, because there are so many different applications of underwater acoustics, it is unavoidable that noise will be introduced into the region of interest, which reduces the effectiveness of these applications. The interference processes and formation methods are to blame for the introduction of this cacophony, which lowers the efficiency of the applications [5].

The precision with which coherent interpretation of underwater images can be achieved is reduced; edge separation, image segmentation, target identification and classification are hampered; confusion is introduced into submerged navigation and texture parametric inversion [6]; and edge separation, image segmentation, target identification and classification are hampered.

The precision with which coherent interpretation of underwater images can be achieved is reduced, edge separation, image segmentation, target identification and classification. On the other hand, the approach that is proposed in this paper includes the processing of images that were taken underwater to determine the images.

1.1 BACKGROUND

The broadcast recordings have been analysed by researchers from a wide variety of fields, including marine biology, to determine the total number of objects that can be seen at this time. It is essential to have the ability to differentiate between a variety of objects. This decision has the greatest influence on how accurately the counting and detection systems function, it was the most challenging part of the process to decide which object identification algorithm to put into place.

In [6] developed a vision system that can recognise and counting the number of fish in movies captured in real-time. The analysis of video textures, as well as object identification and tracking, are all utilised by this system. Approaches such as those described above depend heavily on manually crafted features, which places a limit on the amount of information that can be represented. Other approaches, such as those described below, do not have this limitation.

Convolutional neural networks (CNNs) [7] to correctly identify the species of fish that were present in the image. This was achieved with the assistance of a foreground detection technique so that potential windows for fish could be extracted. The results of experimental research support the conclusion that a deep learning approach to detecting underwater objects is superior to the Histogram of Oriented Gradients (HOG) method and the Support Vector Machine (SVM) method when it comes to locating coral reef fish. A deep learning approach to underwater object detection was compared with these two methods.

The recognition of different species of fish was made possible with the assistance of Rapid RCNN [9]. On the other hand, these fast RCNN techniques are dependent on the features that are

extracted from the neural network final convolution layer. These features are much too coarse to be utilised for the purpose of detecting small things because those would require a much more precise analysis. The lack of sufficient databases for the identification of underwater objects is another obstacle in the way of the development of techniques for the detection of underwater objects.

An underwater dataset for underwater saliency detection that includes annotations at the object level was proposed [3] for the purpose of evaluating various underwater object detection algorithms. This dataset would be intended for use in underwater environments. This dataset would be utilised while sailing on the ocean.

The insufficiency of these algorithms is likely to blame for the excessive amount of processing time that is required [2]. This excessive quantity of processing time limits the flexibility of the algorithm at the places where the application of filters and parameter sequences needs to be followed up. We are investigating an algorithm for moving averages that involves extracting the foreground from an already-obtained reference image to produce more accurate results.

Using point-based comparisons between a reference image and a target image, it is possible to identify a specific object in the image being analysed.

There are several benefits to comparing the target image to the reference image, and these benefits exist even if the comparison is carried out in a haphazard manner. One of these benefits is the capability to recognise objects although they may have been shrunk, enlarged, or shifted in plane without losing their original form.

It is noticeable when there is even a slight amount of rotation or occlusion that is occurring out of plane. The primary focus of this article is on non-repetitive texture pattern objects because these types of objects produce one-of-a-kind feature matches and benefit the most from the application of this method.

2. PROPOSED METHOD

As a preliminary step in the research endeavour, a total of 2000 unique samples of the marine environment were gathered. On the other hand, the experimental collection does not include a large enough number of individual data points to make the construction of a deep learning network model possible. On the images that were initially a part of the dataset, the authors of this article apply a variety of data enhancement strategies and then analyse the results.

Flipping, local trimming, colour correction, Gaussian noise, and salt and pepper noise are some of the techniques that fall under this category. These procedures were developed with the intention of resolving a problem involving a limited number of samples. Image flipping involves inverting the original image of the sample of experimental data that was taken, whereas image cropping involves making a local cut in the image, choosing the position of the object, and erasing the portions of the image that are distracting or unnecessary. Image flipping is distinguished from image cropping by the fact that image flipping involves inverting the original image of the sample of experimental data that was taken.

During the process of data enhancement, this article intentionally incorporates Gaussian noise as well as salt and pepper noise into some of the images. This is done to improve the resilience of the model and to protect it from becoming overfit. Underwater images typically have an unbalance in the colour primary since the colour components of underwater images are distinct from those of land-based images. to improve the quality of the features captured in the underwater images, colour compensation was applied to the images that were taken with the subject under water.

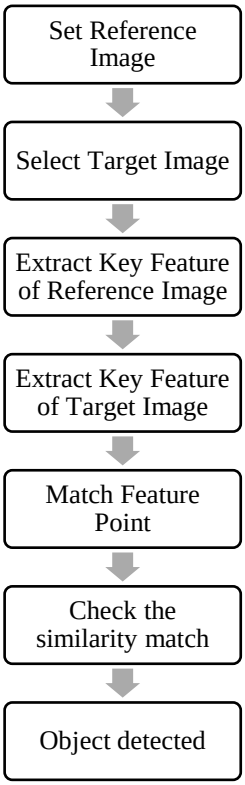

Fig.1. Object Detection

The red component of light is the one that is diminished the most noticeably when it travels through the ocean. As a direct consequence of this, the red component in the images that were collected of the underwater environment is relatively insignificant, whereas the information in the blue and green channels is generally preserved well.

2.1 FEATURE EXTRACTION

The Fig.2 show the outcomes of a process that visualised the 10 most prominent features of both the target image and the reference image. This process was performed so that the results could be compared. Because of carrying out these steps, we now have a reasonable comprehension of the degree to which the two images share common characteristics.

The Detect Surf procedure is utilised to accomplish the task of feature extraction. Despite this, there is a great deal of leeway in the approaches that can be taken to arrive at the same destinations to achieve the same objectives.

2.2 FEATURE POINT MATCH

The geometric transform function is used to match the descriptors of each image with the inliers and outliers of the contours of the other image. This is done so that the two images have a consistent appearance. This stage is completed once it has been determined which aspects of the images constitute the most crucial elements.

By utilising the transformation of the matched points that exist between the target and the reference, this sub-loop can determine the position of the target object within the reference image. If each of the feature points coincides at the same moment, it indicates that the object recognition was successful, and the algorithm will extract the target object from the scene that is being used as a reference to complete the process. However, if the reference image and the target image are not identical, the algorithm will not be able to identify any of the objects in the target image.

2.3 OBJECT TRACKING

The first step in the process of object tracking is known as simple object recognition; however, object tracking is still considered to be its own separate process. Object tracking is the process of finding and following an object across multiple frames or throughout an entire video stream.

Object recognition is the process of locating an object within a single frame. Object tracking is the process of locating and locating an object over the course of an entire video stream. Components such as the optimization of noise dilation and occlusion methods are included in the comprehensive framework that is used for tracking a single object under dynamic circumstances.

These components can be implemented as an autonomous vision-based heuristic tracking system or as the extraction of features followed by the application of the Kalman filter. Both methods are examples of possible implementations. The recommended methodology includes the following three primary components: data association, prediction, and discovery. The Fig.1 is a flowchart that illustrates the detection process, as well as the prediction and data correlation processes.

Image processing gives us the ability to understand the position and orientation of a linear object, but it takes some time to find that object within a massive image. Image processing also gives us the ability to understand its position and orientation. The height of the image is also a consideration.

Moving in a straight line behind an object will help minimise the strain that is placed on the computational resources, so it is required that they do so. Keeping up with a linear object that has parameters that change consistently from one image to the next can be difficult if you are trying to do so because of the complexity of the object.

If a prediction is made during the phase of image processing regarding the location of a linear object, then the search area can be reduced in size while still providing an acceptable range. This is possible if the size of the search area is reduced. A linear Kalman filter is utilised to make projections about the linear object parameters. This is necessary since the movement of forward-scan sonar through ocean disturbances and the motion of its carrier are not completely understood.

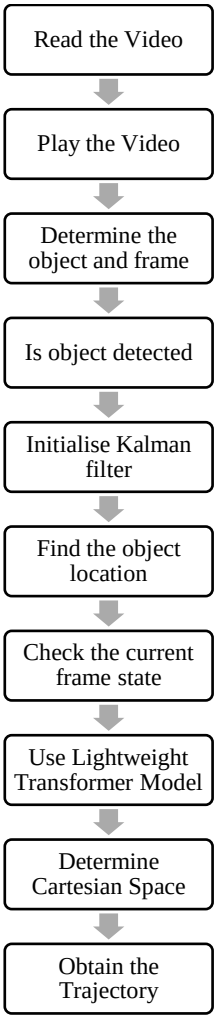

Fig.2. Object Tracking

The Kalman filter model is provided for any two different sources of error, *v* (systematic), and *w*. This is something that can be done for any two distinct kinds of observations.

$$X(t+1) = X(t) + v \tag{1}$$

$$X = (\rho, \theta) \tag{2}$$

$$Z(t+1) = Z(t+1) + w \tag{3}$$

where,

v - systematical error and

w - observation error.

The ability to acquire an accurate ability to forecast outcomes relies heavily on the use of calculation. When tolerance ΔX is considered, the following formula is used to determine the new search radius for linear objects:

$$R(t+1) = X(t+1) + \Delta X \tag{4}$$

The only thing that is necessary is the tracking of a linear object in the image at $R(t+1)$, which will be shown below. Both the capability of determining velocity and the range of image processing that is necessary will improve because of this.

2.4 OBJECT EXTRACTION

The primary objective of this research is to identify, isolate, and keep track of linear objects found in underwater environments. After employing a Gabor filter, one can obtain the binary image because of their efforts. The line is reconstructed by making use of the Hough transform in conjunction with the contour function of the linear object.

The Hough transform is a fundamental and efficient algorithm for line extraction that performs exceptionally well in the field of border-discontinuous straight-line extraction. Its name comes from the fact that it was developed to extract lines that were discontinuous. This is because it is unaffected by background noise and can function normally despite occasional disruptions.

When the Hough transform is applied, the angle that is present between the normal vector of a straight line and the x-axis as well as the distance that the straight line is from the centre do not change.

$$\rho = x_i \cos\theta + y_i \sin\theta \quad (5)$$

It is possible, in principle, to construct any point on any line. The point on the line indicates where the position is on the line (x_i,y_i). The equation of the line is the only question that needs to be resolved at this time (ρ,θ).

3. RESULTS AND DISCUSSION

We develop a methodology for locating three-dimensional pipelines that are buried at great depths below the surface of the ocean. These pipelines are buried at great depths below the surface. The strategy that was recommended maintains what about our earlier efforts was successful and builds upon those successes in novel way. The research uses UOT32 (Underwater Object Tracking) Dataset for training, testing and validation.

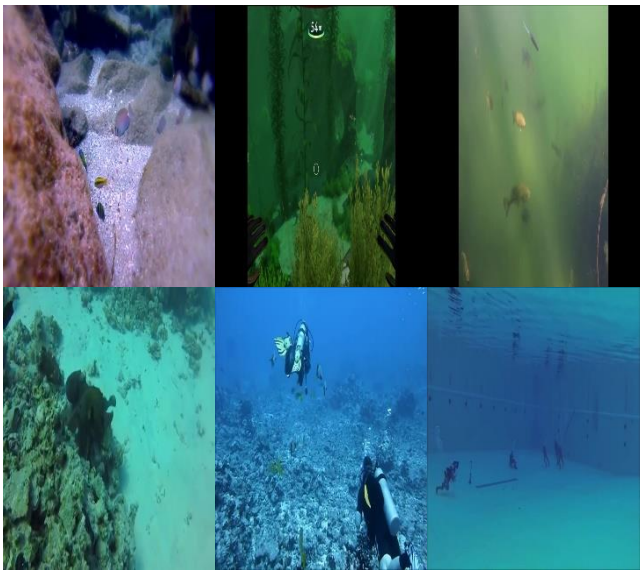

Fig.3. Training Images

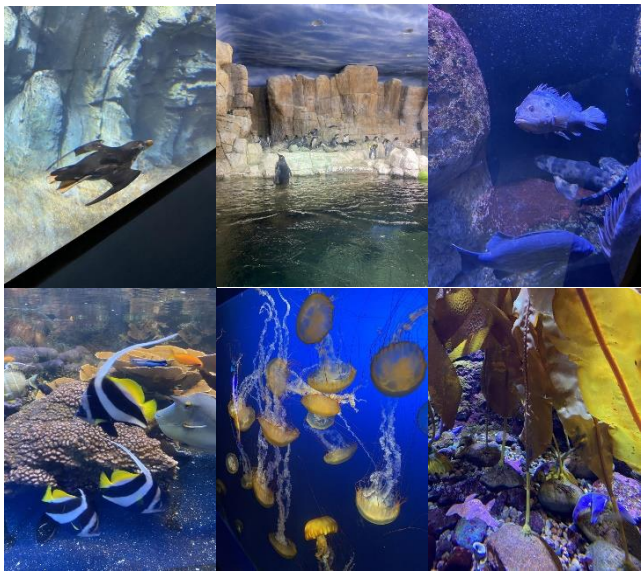

Fig.4. Test Images

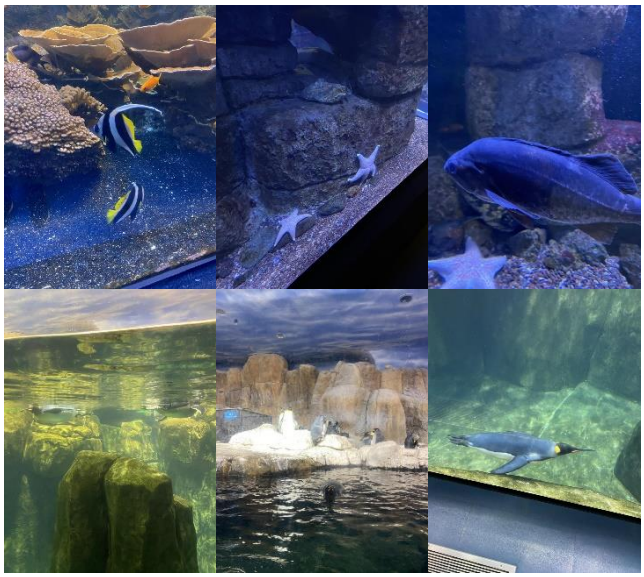

Fig.5. Validation Images

Table.1. Object Detection for Image 1

Model	Experiment	Image 1		
		Accuracy	Recall	Precision
CNN	Training	0.635	0.713	0.586
	Testing	0.899	0.977	0.830
	Validation	0.967	0.977	0.957
Proposed	Training	0.625	0.479	0.684
	Testing	0.762	0.586	0.899
	Validation	0.899	0.821	0.967

Table.2. Object Detection for Image 2

Model	Experiment	Image 2		
		Accuracy	Recall	Precision
CNN	Training	0.860	0.850	0.420
	Testing	0.967	0.879	0.977
	Validation	0.977	0.948	0.977
Proposed	Training	0.655	0.850	0.176
	Testing	0.733	0.909	0.215
	Validation	0.909	0.948	0.479

Table.3. Object Detection for Image 3

Model	Experiment	Image 3		
		Accuracy	Recall	Precision
CNN	Training	0.684	0.410	0.928
	Testing	0.811	0.557	0.977
	Validation	0.811	0.557	0.977
Proposed	Training	0.664	0.401	0.957
	Testing	0.733	0.459	0.977
	Validation	0.801	0.557	0.948

Table.4. Object Detection for Image 4

Model	Experiment	Image 4		
		Accuracy	Recall	Precision
CNN	Training	0.782	0.244	0.576
	Testing	0.938	0.938	0.821
	Validation	0.967	0.938	0.938
Proposed	Training	0.860	0.576	0.879
	Testing	0.938	0.967	0.850
	Validation	0.977	0.967	0.977

Table.5. Object Detection for Image 5

Model	Experiment	Image 5		
		Accuracy	Recall	Precision
CNN	Training	0.713	0.283	0.049
	Testing	0.821	0.000	0.000
	Validation	0.821	0.977	0.166
Proposed	Training	0.684	0.176	0.029
	Testing	0.743	0.000	0.000
	Validation	0.801	0.860	0.147

Table.6. Object Detection for Image 6

Model	Experiment	Image 6		
		Accuracy	Recall	Precision
CNN	Training	0.733	0.498	0.508
	Testing	0.889	0.674	0.723
	Validation	0.909	0.879	0.801
Proposed	Training	0.694	0.489	0.547
	Testing	0.782	0.586	0.586
	Validation	0.879	0.830	0.703

This article presents a method that, while it is comparable to other approaches that have been taken in the past, is significantly more accurate when detecting multiple categories simultaneously and has the potential to significantly reduce the occurrence of both missed detections and false detections. This method can be found in this article. The underwater landscape images that are utilised for the purposes of training, validating, and evaluating.

The learning rate will initially begin at the value of 0.001, which will serve as the beginning point. It is recommended that changes in the performance of the model be monitored at intervals of three iterations to obtain a more in-depth understanding of the model performance and to ensure that adjustments are made in a convenient and timely manner. Monitoring the changes in the model performance at intervals of three iterations.

If performance does not show any signs of improvement, the learning rate for the subsequent training exercise should be adjusted to 90 percent of the value it was at the beginning. If you want to wring every bit of worth out of your model, you should consider utilising the Adam planner. The parameters of the model that is utilised for recognising objects that are submerged in water coincide after 500 epochs of training and learning at a learning rate of 0.0001.

The regular NMS is switched out for the soft NMS, which optimises the model bounding box mechanism. As a result, the general performance of the model sees a slight bump because of this change. In the end, the model of the improved algorithm for detecting objects underwater is achieved through the integration of several different enhanced schemes. This makes it possible to have a representation of the procedure that is more accurate. Because of this, the technique can be helpful when attempting to locate things that are buried or hidden beneath the surface.

Sea urchins can be located, according to the data displayed in Table.1-Table.6, anywhere on the memory curve. When it comes to locating aquatic plants, sea cucumbers, starfish, and scallops, the accuracy of the detection diminishes as the area being searched becomes more constrained. On the other hand, accuracy is maintained at its highest degree throughout the entirety of the process of locating underwater targets. When applied to a single map, the speed of detection that is provided by the method that is discussed in this article is preferable to the speed of detection.

4. CONCLUSION

The time complexity of our proposed method is *M* times greater than that of a single model since it is a lightweight transformer methodology. This is the situation even though it achieves performance that is considered state-of-the-art on challenging datasets. It will be essential for upcoming work to discover methods that can reduce the computational complexity of the strategy that we have recommended. Specifically, this will be essential for the following reasons. In addition, modern deep models incorporate attention mechanisms and innovative loss functions to tackle the problem of detecting noise and small objects.

REFERENCES

[1] L. Chen and L. Wu, "Underwater Target Detection Lightweight Algorithm Based on Multi-Scale Feature Fusion", *Journal of Marine Science and Engineering*, Vol. 11, No. 2, pp. 320-334, 2023.

[2] Y. Sun, "Underwater Object Detection with Swin Transformer", *Proceedings of International Conference on Data Intelligence and Security*, p. 422-427, 2022.

[3] K. Liu and S. Tang, "Underwater Object Detection Using TC-YOLO with Attention Mechanisms", *Sensors*, Vol. 23, No. 5, pp. 2567-2577, 2023.

[4] K. Ali and N. Mahmood, "Marine Object Detection using Transformers", *Proceedings of International Bhurban Conference on Applied Sciences and Technology*, pp. 951-957, 2022.

[5] Z. Li and Y. Gao, "Deep Learning-Based Object Detection Techniques for Remote Sensing Images: A Survey", *Remote Sensing*, Vol. 14, No. 10, pp. 2385-2397, 2022.

[6] Y. Tang, R. Sagawa and R. Furukawa, "AutoEnhancer: Transformer on U-Net Architecture search for Underwater Image Enhancement", *Proceedings of the Asian Conference on Computer Vision*, pp. 1403-1420, 2022.

[7] X. Xu and J. Xia, "MulTNet: A Multi-Scale Transformer Network for Marine Image Segmentation toward Fishing", *Sensors*, Vol. 22, No. 19, pp. 7224-7234, 2022.

[8] F. Qingyun and W. Zhaokui, "Cross-Modality Attentive Feature Fusion for Object Detection in Multispectral Remote Sensing Imagery", *Pattern Recognition*, Vol. 130, pp. 108786-108798, 2022.

[9] S. Han and S. Zhang, "Former-CR: A Transformer-Based Thick Cloud Removal Method with Optical and SAR Imagery", *Remote Sensing*, Vol. 15, No. 5, pp. 1196-1209, 2023.