IMAGE FEATURE EXTRACTION METHODS FOR STRUCTURE DETECTION FROM UNDERWATER IMAGERY

P. Roberts[1], P. Helmholz[2*], I. Parnum[2], A. Krishna[1]

[1]School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Australia
[2]School of Earth and Planetary Sciences, Curtin University, Australia

Commission II, WG II/1

KEY WORDS: Autonomous Underwater Vehicles; Image Feature Extraction; Unsupervised Learning; Underwater Object Detection

ABSTRACT:
The use of autonomous underwater vehicles (AUVs) for surveying underwater infrastructure presents a potential cost saving in comparison to remotely operated vehicles (ROVs). One of the challenges when processing images of underwater structures captured by an AUV, is that vast number of images captured during the mission usually do not show the structure. For instance, images captured during the dive to the structure or of the sea floor, or of the deep sea facing away from the structure. Too many images captured, without relevant information for a 3D reconstruction of the structure, leads to increased processing time and issues during the reconstruction process. There are two solutions to reduce the images to only images showing the structure. Firstly, only images of the structure are captured in the first place or remove images that are not useful after the capture and before further processing. This study developed and evaluated techniques that would enable the first strategy to be applied in an AUV. To apply this strategy in an AUV, would require an on-board structure detection system to ensure that they are correctly orientated for capturing useful footage during a survey mission. However, the marine environment poses several challenges to image-based object detection. Furthermore, small AUVs have limited power and computational resources available while deployed on a mission. To investigate the suitability of creating a lightweight structure detection model for the purpose of image evaluation, three computationally efficient image feature extraction methods (colour moments, local binary patterns (LBP), and Haar wavelet decomposition) were evaluated for their ability to distinguish underwater structures from background areas using unsupervised k-means models. LBP was found to be an effective method for identifying underwater structures in open water conditions. For identifying a structure against the seabed, colour moments were identified as the most effective method.

1. INTRODUCTION

Developments in autonomous underwater vehicles (AUVs) will potentially revolutionize surveys and inspections of subsea infrastructure in the near future (Rumson, 2021). Small portable AUVs will soon be available that can be launched and retrieved by a single operator. This will provide a significant reduction in cost and complexity in comparison to the current use of remotely operated vehicles (ROVs), which require a trained pilot and specialized equipment to operate. However, moving from a tethered ROV to a self-contained AUV raises a number of technical challenges. Firstly, an AUV is limited to its on-board power supply: It must operate as efficiently as possible to extend operating times before it needs to be recharged. Secondly, communication with remote underwater electronic devices is generally limited to acoustic signals, which have a severely constrained transmission bandwidth. Aerial drones can utilise 4G or 5G networks to operate in fleets, transfer video feeds to human operators, or offload information for near real-time processing on servers. This is not possible for a wireless underwater vehicle, so the AUV must be able to intelligently determine if it is capturing useful information during its survey mission. If the navigation path has not been set correctly, there is a risk that an AUV will complete its mission without capturing useful images of the target. To circumvent this risk, a possible solution is to equip the AUV with object detection capabilities, which would mean that only images of the structure are captured. In the context of this paper, we will use the word structure as a general term for any parts belonging to a subsea infrastructure such as a drilling rig.

Only capturing images of the target structure, has several advantages when further processing the images for 3D reconstruction. Firstly, the number of images to be processed is reduced, which will decrease the processing time. Secondly, the risk of false matches of the feature-based matching algorithms, which are part of a 3D reconstruction workflow, will be decreased. For instance, the potential of mismatches on the seafloor are reduced. Finally, the object detection algorithm can be used to mask the images of interest within the 3D reconstruction processing pipeline.

Image classification and image-based object detection has experienced remarkable developments over the past decade. In some applications, state of the art algorithms are able to rapidly detect and classify objects in images with better accuracy than humans (Buetti-Dinh et al., 2019; De Man et al., 2019). The best performing models use deep convolutional neural network (CNN) architecture (Zaidi et al., 2022). They generally require a large amount of labelled data to train for accurate performance and have traditionally required powerful computing hardware to operate at practically useful speeds (Capra et al., 2020). Recent developments have made it easier to incorporate neural network-based object detection in lightweight electronics devices (Zaidi et al., 2022; Zhao et al., 2020). Some examples include object tracking features in unmanned aerial vehicles, and advanced autofocus systems for digital cameras (Ramachandran and Sangaiah, 2021; Herrmann et al., 2020). However, there is a lack of available labelled data for underwater imaging applications (Wang et al., 2019), which creates a significant barrier to implementing deep neural network-based detection methods for AUVs.

Where large, labelled data sets are not available for training predictive models, traditional machine learning models can be used in conjunction with numerical features extracted from images (Tiwari et al., 2013). Studies have shown that manual

https://doi.org/10.5194/isprs-archives-XLVIII-1-W2-2023-1067-2023

feature extraction methods can outperform CNN based architecture when small training sets are used (Lin et al., 2020). This study investigated the classification efficacy of available feature detection methods when applied to subsea structure detection.

The aim of this study was to identify image feature extraction methods that can be utilized to create a lightweight structure detection model for use in AUV surveys. The structure of the paper is as follows, Section 2 (Related works) provides a summary of notable research that has been completed in the field of underwater monocular vision-based object detection and describes the contribution of this study. Section 3 outlines the candidate feature extraction methods and the clustering method used Section 4 presents an evaluation of each candidate feature extraction method's performance, as well as the prediction results from the models created. Section 5 contains a discussion of the results and potential alternative approaches. Finally, Section 6 provides the conclusions of the study and outlines future research directions.

2. RELATED WORK

Monocular vision-based object detection in the underwater environment remains a challenging task, mainly due to the reduced clarity of underwater images. Most of the recent developments in object detection have come from deep learning methods (Gomes et al., 2020). However, these studies have generally been focussed on the detection and classification of marine organisms, such as fish and plankton, for which a number of labelled datasets are available to train networks (Fayaz et al., 2022). Studies concerned with the detection of man-made or miscellaneous objects, have generally used traditional computer vision approaches to identify regions of interest within a frame.

Some of the object detection methods are applied to finding moving objects from a static camera. This problem becomes complicated in underwater scenes due to the presence of moving particulates or seaweed, and the increased sensor noise due to low light conditions. (Seese et al., 2016) used a combination of a Gaussian mixture model and a Kalman filter to estimate the background area of an underwater video scene. The method was used to isolate moving fish against complicated backgrounds. The initial background segmentation utilised data from consecutive video frames to train the Gaussian mixture models and Kalman filters, so this approach required significant parallelisable computing resources. (Vasamsetti et al., 2018) used a combination of colour and texture information across three frames of a video to detect moving objects in underwater scenes. Texture information is extracted for objects using a novel multi-frame triplet feature that compares neighbouring pixel intensity values across consecutive frames to segment moving objects. Colour information is then used to refine the prediction by comparing RGB colour channel intensities of the video frames to a temporary background image. The detection of moving objects is more suitable for fish detection than structure detection. However, given a moving camera, a static structure may be interpreted as a moving object if it is captured against a uniform featureless background. Therefore, these techniques could potentially be used for structure detection against a watery background during an AUV survey.

(Hou et al., 2016) used colour information in the YUV colour space to extract regions of interest containing man-made objects. To overcome the image quality limitations of underwater scenes, the images were pre-processed to equalize illumination, boost colour contrast and, reduce noise. Once regions of interest were detected, shape signals were used to identify the basic geometry of the man-made object. The identification stage of this method is more suited to simple geometries of small man-made objects than larger complex structures. However, the initial region of interest detection highlights the effectiveness of colour-based background segmentation in underwater scenes.

Dark channel information was used by both Zhu et al. (2016) and Chen et al. (2017), as part of an imaging pipeline to detect regions of interest in underwater scenes. Zhu et al. (2016) used the dark channel prior method for initial haze removal of underwater images. The refined images were then processed with a discriminative regional feature integration algorithm to produce a saliency map, and a mean-shift over segmentation algorithm to segment the images. The results of both algorithms are then combined to show only the segmented regions with highly salient objects as the regions of interest in the image. Chen et al. (2017) used dark channel information to estimate light transmission across an image frame in an underwater scene. The contrast in light transmission, combined with colour contrast and pixel intensity contrast, is then used to calculate a region of interest from the scene.

The use of deep learning methods has become prominent in recent underwater object detection research. Villon et al. (2016) highlighted the performance benefits of deep learning methods by comparing classifications from a convolutional neural network (CNN) to a support vector machine (SVM) classifier using histogram of gradients (HOG) for image feature extraction. The CNN based model returned more accurate results and faster detection performance than the SVM model. Notable deep learning architectures have also been applied to underwater detection; for instance, the performance of Fast-RCNN, Faster-RCNN, and YOLO-V3 were evaluated by Fayaz et el. (2022) for their performance at detecting sea-cucumbers, seaurchins, and scallops. The YOLO v3 algorithm was noted for its detection accuracy and rapid performance (Fayaz *et al.,* 2022). Mahmood et al. (2016) applied the VGG network to coral classification. Thum et al. (2020) applied transfer learning techniques to pre-trained CNN based classification models, in order to classify images as either containing or not containing underwater cables amongst complicated backgrounds. By utilising transfer learning, Thum et al. (2020) were able to obtain accurate classification results without the need for massive underwater datasets, which are usually required to create robust deep learning models. However, this method still required the extraction and manual collation of 2000 images.

Deep learning methods have also been used to apply image segmentation to underwater datasets (Liu and Fang, 2020; Drews Jr et al., 2021; Nezla et al., 2021). However, a lack of properly labelled datasets for underwater imaging applications has been a notable challenge in this area. To address this problem, Drews-Jr et al. (2021) created synthetic images by applying contrast reduction and colour adjustments to pre-labelled images to simulate the colour cast and turbidity of underwater environments. The simulated images were then used to augment training sets of real underwater images to improve the segmentation accuracy of Segnet (Chen et al., 2018) and Deeplav3+ (Badrinarayanan et al., 2017) models. Unfortunately, the results indicated that augmenting the training dataset with simulated images led to a slight reduction in accuracy for segmenting underwater scenes.

In conclusion, while the difficulties of capturing high-quality images in underwater environments, and the subsequent challenges of performing image classification and object detection in dark, low-contrast scenes are well understood, to our knowledge there has been limited research into methods that would allow an AUV to autonomously verify when it is collecting useful images during a survey mission. Thum et al. (2020) suggest that small neural networks designed for edge devices, such as the Mobilenet family (Howard et al., 2017), could be used to classify images as either containing or not containing a structure. However, the transfer learning methodology employed, still requires a large amount of manual image labelling. Furthermore, the classification models only provide a binary present/absent response. It does not account for scenarios where the target structure is poorly framed in a small margin of the camera's view. This study introduces a method that is computationally lightweight, so it can be run efficiently on many sub-sections of the original image to provide coarse localization of any detected structures.

3. METHODS

3.1. Feature Extraction Methods

Initially the common feature extraction methods (surf, sift, orb) have been tested but were excluded due to extremely variable results based on how close the structure is. Hence, the three candidate feature extraction methods considered as part of the study are: colour moments, local binary patterns (LBP), and Haar wavelet decomposition. Colour moments were used to extract high-level statistical information from the colour channels of the video images (Tiwari et al., 2013). Local binary patterns and Haar wavelet decomposition were considered for texture information extraction because of the reported efficacy and low computational cost (Tiwari et al., 2013).

3.1.1 Colour Moments: Colour moments is the term used to describe summary statistics of the separate colour channels of an image. When used as a feature extraction method for machine learning, the first four statistical moments are used i.e. mean, standard deviation, skewness, and kurtosis. These values are calculated from the distribution of pixel intensity values for red, green, and blue colour channel matrices, and then combined into a feature vector (Tiwari et al., 2013). The attenuation of red-light frequencies in water is much higher than green or blue frequencies, so there was no significant red channel information in the survey footage (Figure 1). Therefore, the general colour moments feature extraction method was modified to only include moments from the blue and green image channels.

To extract colour moment features, the original image was split into individual colour channels. The blue and green channels were then split into 20 x 20 pixel arrays. The mean, standard deviation, skewness, and kurtosis of pixel values for both channels were then combined to form a 1 x 8 length vector.

3.1.2 Haar Wavelet Decomposition: Haar transform is one of the simplest discrete wavelet decomposition methods. It can be applied to a two-dimensional array. On larger matrices, each of the four decomposition products is a new matrix with both height and width dimensions at half the size of the original matrix. When the Haar transform is applied to a single channel image, the four product matrices can be cast as separate images, each a quarter of the size of the original.

The Haar wavelet decomposition can be applied recursively to the product of successive transforms. Feature vectors are extracted through Haar transforms by calculating the mean and standard deviation of the four matrices produced by each successive decomposition (Tiwari et al., 2013). The length of the feature vector is determined by the number of decomposition steps utilized: The total length of the vector for k decomposition steps will be k ×6+2.

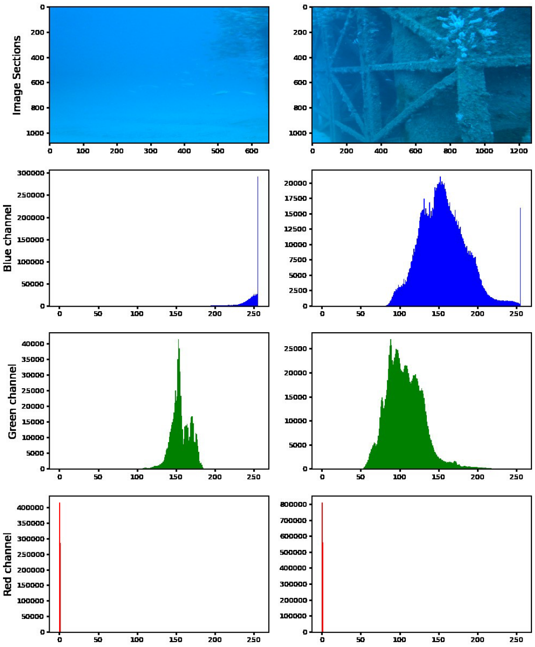


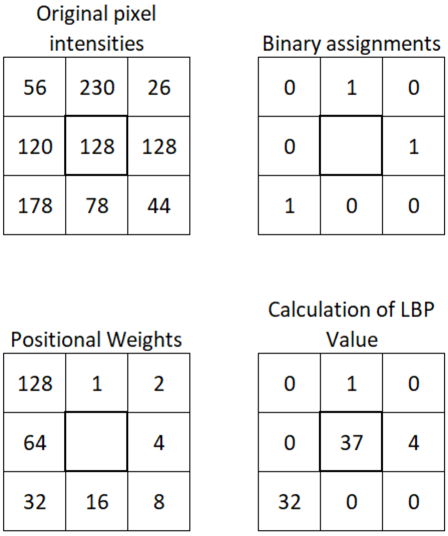
Figure 1. Input images (row 1) with their blue (row 2), green (row 3), and red (row 4) channel histograms f.

The feature extraction process was slightly more involved for the wavelet extraction method. Each Haar wavelet decomposition of an image reduces the size of the resultant product matrices by 50% along both its width and height. Therefore, the tile sizes had to be modified to produce an equivalent number of feature vectors. The entire image was first converted to grey-scale, then two consecutive decomposition calculations were performed. This produced three 960 x 540 sized arrays, and four 480 x 270 sized arrays. To maintain consistency with the other methods, the mean and standard deviation from 10 x 10 pixel-tiles on the larger arrays, and 5 x 5 pixel-tiles on the smaller arrays were combined into a 1 x 14 length feature vector.

3.1.3 Local Binary Patterns: Local binary patterns (LBP) is a texture-based feature detection method, which has been used successfully in various object detection applications (Karis et al., 2016). The fundamental concept of LBP is based on comparing a given pixel's intensity value to the intensity value of each surrounding pixel in a circular pattern (Figure 2). In the simplest implementation, each pixel is compared to its eight surrounding pixels, however, a larger radius and circumference can also be used. Each pixel in the circular pattern is assigned a threshold value of either 0 or 1 based on the difference in intensity of the surrounding pixel g_p and the intensity of the central pixel g_c. A single value for the central pixel is then calculated by multiplying each threshold value by the weight assigned to its position in the circular pattern, then taking the sum of all threshold and weight products.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-1-W2-2023-1067-2023 |

Feature vectors are created by forming a histogram from each pixel's LBP values. For a circular pattern with a P of 8 pixels, there are 256 unique LBP values that can be set as bins for the histogram. However, for image texture analysis, there is a subset of LBP values that provide a greater amount of information. These are the "uniform" LBP patterns, where the threshold value does not change more than twice when traversing the circular pattern. The uniform LBP values correspond to edges, corners, and flat areas in the image. When P = 8, the LBP values corresponding to uniform patterns can be grouped into 58 histogram bins. All the remaining non-uniform LBP values are grouped into one extra bin, resulting in a feature vector that is 59 values long.



Original pixel intensities

56	230	26
120	128	128
178	78	44

Binary assignments

0	1	0
0		1
1	0	0

Positional Weights

128	1	2
64		4
32	16	8

Calculation of LBP Value

0	1	0
0	37	4
32	0	0

Figure 2. LBP calculation process.

For LBP features, an LBP value array was calculated from the grey-scale conversion of the entire image. Grayscale conversion was completed as per CCIR 601 (Y = 0.299R+0.587G+ 0.114B). The edges of the image were extended by reflection, so the calculated LBP array maintained the same dimensions as the original image. For each 20 x 20 pixel tile of the LBP array, a 59 bin histogram was created, and the counts of each bin were stored as 1 x 59 length feature vector.

3.2 K-means Unsupervised Learning

Unsupervised learning, in the form of k-means clustering, was used to evaluate the structure distinguishing efficacy of each feature extraction method. The aim of the clustering is to detect images which contain structures relevant for 3D reconstruction (foreground) compared to images showing ocean or sea floor (background).

Traditional clustering methods do not perform well with large feature vectors due to the curse of dimensionality. So, the candidate methods were each normalized, and then reduced through principal component analysis (PCA). The first three principal components of the colour moments, LBP, and Haar wavelet feature vectors accounted for 73%, 37%, and 67% of the respective variances.

Clustering was performed on the first three principal components of each candidate method. A k-means algorithm was used, with the number of clusters set as two to differentiate the background from the structure. The performance of each feature extraction method was evaluated by inspecting the clustering results on the data. Finally, the k-means models were applied to a separate data set from another underwater survey of similar structures.

4. VALIDATION

4.1 Datasets

Dataset 1: The first dataset consisted of video files from an ROV survey of one of the two purpose-built artificial reefs on the Western Australian (WA) coast, called the Rottnest Island fish towers. The fish tower is a large structure designed to act as an artificial fish habitat. The tower features a truss-based structure with large cylindrical forms enclosed at its corners. It is visually similar to underwater infrastructure used in industries such as hydrocarbon production or offshore wind power generation. The tower is located at an approximate water depth of 40 m near Rottnest Island, Western Australia (Mufti et al., 2019). Footage of the survey was captured by a Sony RX0 digital camera at a resolution of 1920 x 1080 pixels and a frame rate of 50 fps. The survey was completed in fine weather conditions and all the footage was captured with natural light only. Underwater visibility is very clear, with only a small amount of particulate matter present. Three dives were performed on the day of the survey; the ROV was able to record the structure from short and long range on all sides.

Reference Dataset 1: To evaluate the candidate feature extraction methods, a subset of video frames was extracted as lossless png images. Each frame was manually chosen from across the video timeline so that the subset provided a representative sample of the entire survey. Care was taken to ensure that the sample dataset included close, medium, and long-range views of the structure. Images of the structure in shadow, as well as in well-lit conditions were also included. To reduce the effect of red channel interference on colour moments feature vectors, the footage collected before and after the dive was excluded, as were any frames from when the ROV was very near the water surface. In total 29 images were extracted.

Each image of the reference datasets was split into a grid of 20 x 20 pixel-sized tiles for feature extraction The size of 20 x 20 pixel were found empirically across the used images. At this size, a 1920 x 1080 image produces 5184 observations, totalling 150,336 observations for the 29 image of the reference dataset 1.

Dataset 2: The second dataset consisted of footage that was captured by a GoPro camera mounted on a small ROV for a survey of artificial reefs off the coast of Bunbury, Western Australia (Rofallski et al., 2020). The artificial reefs are similar to the Rottnest fish tower, in that they are a truss like structure located on the seabed, but in shallower water. They are covered in marine growth and visually distinct from the surrounding ocean floor. The artificial reef survey was captured at 4000 x 3000 pixel resolution, so it contains considerably more detail than the 1920 x 1080 pixel footage of the Rottnest fish towers. There is a noticeable difference in colour cast between the two data sets; the Rottnest fish tower data set has a strong blue cast, whereas the artificial reef footage has a stronger green channel. Finally, the Rottnest fish tower survey was captured predominately from a horizontal camera angle; Whereas the Bunbury artificial reefs were captured from an angled vertical orientation, which meant the background consisted predominately of seabed, not open ocean, in most of the survey footage.

Reference Dataset 2: A smaller validation set was also created by manually selecting 7 further individual frames from the video file. These frames were similarly selected to provide a representative subset of the entire survey. Each of the validation set images was used to create a ground-truth mask for the location of the structure within the frame. The primary interest of the study was to segment the background from the rest of the image; fish were generally considered to be part of the structure creating the ground truth image masks. A 20 x 20 pixel tile was determined to be part of the structure if the corresponding ground-truth mask contained at least 20% white pixels. Otherwise, it was designated as background.

4.2 Evaluation

4.2.1 Visual Evaluation: Clustering results from the three candidate feature extraction methods were visually evaluated by inspecting the classes against the images of dataset 1. Figure 3 shows a representative image with the clustering classes overlaid (foreground is highlighted in red). The k-means output is naive, but as a binary signal, it is easy to interpret as structure and background.

From visual inspection of the overlaid results, LBP appeared to provide the most consistent performance across all scenes depicted in the images. Colour moment features performed well on a selection of the images, with the clustered results identifying the almost all the structure in most cases. Unfortunately, the colour moments-based segmentation also tended to identify darker areas of the background as part of the structure Figure 3).

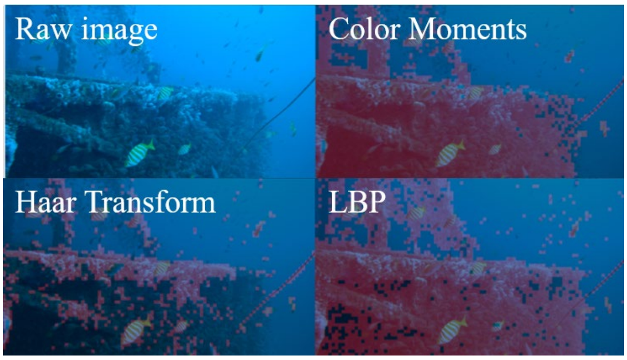


Figure 3. K-means Clustering results for a representative image using of dataset 1 for different features. Foreground is highlighted in red. All other parts belong to the background.

The features extracted by Haar wavelet decomposition did not perform well. The clustering results indicate that the only the most detailed sections of the structure were separated from the background. The benefit of the Haar based features is that there are very few false positive structure detections in the background areas.

LBP features led to clustering results that appear to be somewhere between the colour moments and Haar decomposition methods. The LBP results show generally good identification of the pelagic tower, although darker areas of the structure are not identified as well as with colour moments. The LBP based clustering also resulted in false detections among the background areas. These tended to be as random single points, possibly triggered by small, suspended particles.

4.2.2 Numerical Evaluation: Numerical evaluation of the candidate methods was performed using dataset 1. Classification performance metrics of the clustering of the different image processing methods are shown in Table 1, and they generally agree with the visual analysis of each feature extraction methods performance. The f1 score shown in the table is a weighted average of precision and recall value. Using LBP, returns the highest accuracy, while also providing a good balance between precision and recall. The colour moment features show a similar true structure detection rate to LBP, but almost four times as many false structure predictions. The Haar features have a very low false detection rate (0.99%) for the structure, although the overall low number of structural predictions leads to significantly lower accuracy and recall scores. The only significant difference between the test set and validation set is that the colour moment based model did not identify more true structure points. This may be due to the specific frames selected for the validation set. It highlights the fact that the validation set does not provide a true objective evaluation of the candidate methods.

Metric	Colour Moments	Haar	LBP
Accuracy	0.833	0.767	0.916
Precision	0.771	0.990	0.933
Recall	0.856	0.449	0.862
F1	0.812	0.618	0.895

Table 1. Performance Metrics for Predictions using dataset 1.

It is worth noting, it takes significantly longer to process a 1920 x 1080 video frame with Colour Moments than with Haar wavelets or LBP. This is because the Haar and LBP methods used a single grayscale array, whereas the Colour Moments method accessed the three colour channel arrays.

4.3 Structure Predictions

4.3.1 Dataset 1: K-means clustering based on the LBP-derived feature vectors produced the best structure detection results (Table 1), so the LBP k-means model was used to analyse the entire dataset. The foreground prediction for each frame of the video can be cast as a binary image where each 20 x 20 pixel-tile from the original image is represented as either a single black or white pixel in a 96 x 54 pixel binary image. Figure 4 shows a selection of video frames with their associated prediction images; white being foreground (structures predicted) and black being background. The prediction images were combined into a video file and the playback was observed at the original data-set's native 50 fps. It became apparent that the false positive predictions of the structure caused a significant amount of temporal noise in the video. It was observed that most of the noise appeared as dispersed single white pixels in the black areas of the frame, so a simple erosion-dilation convolution was performed to denoise the image (Bradski, 2000). This resulted in a significant reduction in the temporal noise, along with a minor loss of detail in each frame.

To determining whether an underwater structure is visible in the frame, preliminary analysis found the most informative metric was the total number of tiles that returned a positive prediction. Figure 5 placed at the end of the paper shows the number of structure predictions for each frame of the video file (the chart has been limited to footage from the descent, survey, and ascent). Both the raw prediction values and the denoised signal track well with manual observations from the video file. The ROV's descent from the surface down to the structure (frames 27900 to 33800) at the start of the deployment, and ascent back up at the end of the deployment (frames 133200 to 141900) were characterized in the dataset by the video frames containing predominately background. Therefore, the descent and ascent periods were clearly visible as sections of low total predictions.

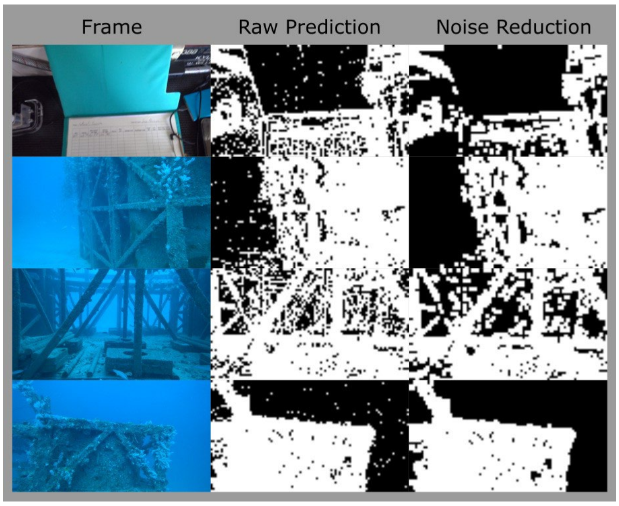


Figure 4. Images without (raw prediction) and with noise reduction applied. Foreground is shown as white and background as black.

Given that the prediction for each frame was cast as an image, it was also possible to provide some basic metrics about the location of the structure within the frame. While the structure was being surveyed, the number of positive predictions generally remained greater than 2000 tiles (approximately 40% of the frame area). There were several points during deployment where the number of structure predictions dropped below 1000. These points correspond to instances where the ROV rotated, and the structure was not visible in the camera's field of view. At close examination of these sections of the data set, the horizontal and vertical median pixel values of the prediction images reflect the location of the structure within the frame. Figure 5 provides a detailed view of the structure dropping out of the image frame between the 89000th and the 89500th video frames. The horizontal and vertical median plots indicate that the structure falls out of view as the ROV rotates laterally to the right. The structure comes back into view as the ROV then banks to the left.

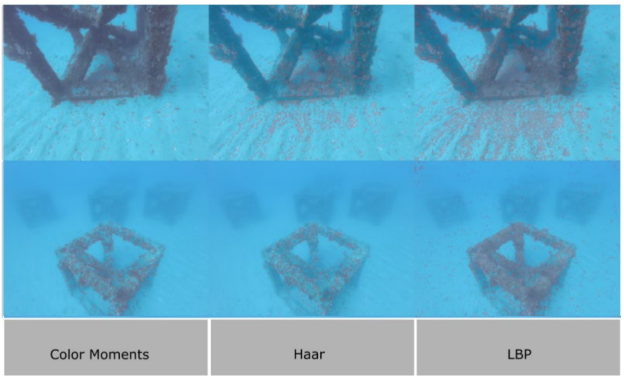


Figure 6. Re-trained k-means prediction results applied to dataset 2.

4.3.2 Dataset 2: As a final evaluation of the feature extraction methods, k-means were applied to dataset 2 (Figure 6). When trained on the actual data, the colour moments-based k-means model gave good results, with accuracy at 0.89 (Table 2). The colour moments model identified the structures cleanly at close range, while rejecting almost all the seabed or background. The LBP model also showed notable improvements. It produced a less noisy prediction image and was slightly better at identifying the reefs at medium to short range than the colour moments model. However, the LBP based model still tended to identify areas of the seabed as part of the structure at close range and the overall accuracy remained low at 0.724. The Haar based model performance showed little improvement to the pre-trained model, with the accuracy dropping slightly from 0.763 to 0.757.

Metric	Colour Moments	Haar	LBP
Accuracy	0.890	0.757	0.724
Precision	0.713	0.391	0.398
Recall	0.791	0.300	0.633
F1	0.750	0.340	0.489

Table 2. Performance Metrics for Re-Trained Models on Bunbury Artificial Reef Data

5. DISCUSSION

Considering the study was completed with unsupervised learning only, both colour moment and LBP based feature extraction methods showed very promising performance. The k-means model was used mainly as a tool to evaluate the candidate methods' abilities to extract useful feature data. However, using k-means data to make predictions on dataset 2 gave reliably accurate results. In the context of assisting AUV missions, a similar approach could be used to determine whether the vehicle is collecting useful images during a mission. By casting the prediction results as a low-resolution image, it is also possible to detect when the structure of interest is moving in or out of the camera's field of view.

Of the three candidate feature extraction methods considered, LBP appeared to be the most effective. LBP features gave the best predictive performance on dataset 2 showing reasonably robust performance when determining structure with water and no seafloor in the background. LBP feature extraction probably worked well because it is very effective at finding areas of low detail. Both underwater surveys used small form digital cameras to capture the data. The Sony RXO has a 1-inch sensor, and the GoPro has a 1/2.3 inch sensor. High ISO values are needed to effectively record images in low-light underwater settings, so both cameras are likely to produce image noise at a pixel level. Therefore, the LBP calculations will predominately find non-unform patterns in the areas without sufficient texture detail to override the noise patterns. In the context of binary separation of a structure from a featureless background, this approach works very well. However, when a detailed seabed is also visible in the frame, the LBP method would need to extract information that is specific to the texture of objects. Further work is required to determine if LBP can be effectively applied for this application.

Colour moments also appeared to be effective feature extraction method for the purpose of basic subsea object detection. The method was slightly less robust than LBP when being applied to different datasets. This may be due, at least in part, to the noticeably different colourcast between the two datasets. Therefore, colour moments model that was developed for a fixed camera system on an AUV would not have that challenge. However, different underwater environments can produce natural colour casts, so data would be required from a variety of underwater environments to ensure that a robust model is produced. During the project, the only significant detected fault of the colour moments method was its tendency to classify darker areas of purely background frames as structure. This was probably an effect of using unsupervised learning. It may be possible to distinguish between the dark background and dark structure colour channel information when using the full feature vector with a supervised learning approach.

Based on the observed results, the Haar wavelet decomposition feature extraction method was the least effective. Although, this may be a consequence of this study's implementation of the method. The image tiles were initially selected at 20 x 20 pixels, so only two decomposition steps were performed when extracting features from images. Any further decompositions would have required halving the dimensions of a 5 x 5 pixel tile. Therefore, the Haar method was only extracting small-scale texture information from the images. This was probably why the models trained with Haar features only tended to identify structure in highly detailed areas of the image foregrounds. To better evaluate the Haar method, the study could be repeated with a tile size of 32 x 32 pixels, which would allow the extraction of texture information from at least four scale ranges. Nevertheless, the dimension of the tile also depends on the resolution of the camera and further investigations are required.

The LBP calculation used in this study is the most basic version of the LBP method. Since its inception, there have been several modifications proposed that have been found to produce improved results in various applications (Liu et al., 2016). When developing a trained model for application with an AUV, a more advanced LBP method, such as median robust extended local binary patterns (MRELBP), should be considered (Liu et al., 2016).

Deep neural networks were not considered for this project due to the large amounts of data required for accurate performance. However, in decent visibility conditions, the underwater structure identification problem is significantly simpler than multi-object classification problems where deep CNN models are usually applied. It is possible that a shallow CNN architecture could be trained to produce useful results with a modestly sized labelled dataset.

6. CONCLUSION AND FUTURE WORK

The results of this investigation indicate that LBP and colour moments are effective feature extraction methods for identifying structures in underwater survey footage. A simple k-means model, trained with three principal components from LBP feature vectors was able to reliably detect structures in open water surveys at over 90% accuracy. Unfortunately, the LBP method was less successful when distinguishing a man-made structure from the seabed in high-resolution footage. This is comparable accuracy to the results from Thum et al. (2020), who used CNN-based classifiers to identify man-made cables in underwater locations. T

While the models created for evaluating the feature extraction methods were simple, their prediction performance confirms that lightweight models can be an effective solution for AUV surveys. By splitting the image into a grid of tiles and making a prediction for each tile, the contents of the image can be summarized with a single value metric, which is easily communicable to a human operator through acoustic signals. Furthermore, it is easy to detect when the structure is moving out of the AUV camera's field of view. The AUV's control system can use this information to adjust its pose or position in order to continuously capture high quality survey footage.

While this study showed promising results, several research questions require further investigations. This includes defining the tiles to calculate the features based on the camera's solution and the approximate distance of the camera to the object, which is mostly defined by visibility. Then, the investigation of further features such as ORB or LAB as well as the combination of features. Finally, a validation using ConvNet pretrained on ImageNet is planned.

REFERENCES

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE PAMI*, 39(12), 2481–2495.

Bradski, G., 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Buetti-Dinh, A., Galli, V., Bellenberg, S., Ilie, O., Herold, M., Christel, S., Boretska, M., Pivkin, I. V., Wilmes, P., Sand, W. et al., 2019. Deep neural networks outperform human expert's capacity in characterizing bioleaching bacterial biofilm composition. *Biotechnology Reports*, 22, e00321.

Capra, M., Bussolino, B., Marchisio, A., Shafique, M., Masera, G., Martina, M., 2020. An updated survey of efficient hardware architectures for accelerating deep convolutional neural networks. *Future Internet*, 12(7), 113.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of ECCV*, 801–818.

Chen, Z., Zhang, Z., Dai, F., Bu, Y., Wang, H., 2017. Monocular vision-based underwater object detection. *Sensors*, 17(8), 1784.

De Man, R., Gang, G. J., Li, X., Wang, G., 2019. Comparison of deep learning and human observer performance for detection and characterization of simulated lesions. *Journal of Medical Imaging*, 6(2), 025503–025503.

Drews-Jr, P., Souza, I. d., Maurell, I. P., Protas, E. V., C. Botelho, S. S., 2021. Underwater image segmentation in the wild using deep learning. *Journal of the Brazilian Computer Society*, 27, 1–14.

Fayaz, S., Parah, S. A., Qureshi, G., 2022. Underwater object detection: architectures and algorithms–a comprehensive review. *Multimedia Tools and Applications*, 81(15), 20871–20916.

Gomes, D., Saif, A. S., Nandi, D., 2020. Robust underwater object detection with autonomous underwater vehicle: A comprehensive study. *Proceedings of the International Conference on Computing Advancements*, 1–10.

Herrmann, C., Bowen, R. S., Wadhwa, N., Garg, R., He, Q., Barron, J. T., Zabih, R., 2020. Learning to autofocus. *Proceedings of the IEEE/CVF*, 2230–2239.

Hou, G.-J., Luan, X., Song, D.-L., Ma, X.-Y., 2016. Underwater man-made object recognition on the basis of colour and shape features. *Journal of Coastal Research*, 32(5), 1135–1141.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Karis, M. S., Razif, N. R. A., Ali, N. M., Rosli, M. A., Aras, M. S. M., Ghazaly, M. M., 2016. Local binary pattern (lbp) with application to variant object detection: A survey and method. *2016 IEEE 12th International Colloquium on Signal Processing & Its Applications (CSPA)*, IEEE, 221–226.

Lin, W., Hasenstab, K., Moura Cunha, G., Schwartzman, A., 2020. Comparison of handcrafted features and convolutional neural networks for liver MR image adequacy assessment. *Scientific Reports*, 10(1), 1–11.

Liu, F., Fang, M., 2020. Semantic segmentation of underwater images based on improved Deeplab. *Journal of Marine Science and Engineering*, 8(3), 188.

Liu, L., Fieguth, P., Wang, X., Pietikainen, M., Hu, D., 2016.¨ Evaluation of lbp and deep texture descriptors with a new robustness benchmark. *ECCV 2016: Proceedings, Part III 14*, Springer, 69–86.

Mahmood, A., Bennamoun, M., An, S., Sohel, F., Boussaid, F., Hovey, R., Kendrick, G., Fisher, R. B., 2016. Coral classification with hybrid feature representations. *2016 IEEE ICIP*, IEEE, 519–523.

Mufti, A., Parnum, I., Belton, D., Helmholz, P., 2019. Investigation of in-field devices for underwater surveying of reef structures. Publication pending

Nezla, N., Haridas, T. M., Supriya, M., 2021. Semantic segmentation of underwater images using unet architecture based deep convolutional encoder decoder model. *2021 ICACCS*, 1, IEEE, 28–33.

Ramachandran, A., Sangaiah, A. K., 2021. A review on object detection in unmanned aerial vehicle surveillance. *International Journal of Cognitive Computing in Engineering*, 2, 215–228.

Rofallski, R., Tholen, C., Helmholz, P., Parnum, I., Luhmann, T., 2020. Measuring Artificial Reefs using a MultiCamera System for Unmanned Underwater Vehicles. *ISPRS Archives*, 43(B2), 999– 1008.

Rumson, A. G., 2021. The application of fully unmanned robotic systems for inspection of subsea pipelines. *Ocean Engineering*, 235, 109214.

Seese, N., Myers, A., Smith, K., Smith, A. O., 2016. Adaptive foreground extraction for deep fish classification. *2016 ICPR CVAUI*, IEEE, 19–24.

Thum, G. W., Tang, S. H., Ahmad, S. A., Alrifaey, M., 2020. Toward a highly accurate classification of underwater cable images via deep convolutional neural network. *Journal of Marine Science and Engineering*, 8(11), 924.

Tiwari, A., Kumar, A., Saraswat, G. M., 2013. Feature extraction for object recognition and image classification. *International Journal of Engineering Research & Technology (IJERT)*, 2(10), 2278–0181.

Vasamsetti, S., Setia, S., Mittal, N., Sardana, H. K., Babbar, G., 2018. Automatic underwater moving object detection using multi-feature integration framework in complex backgrounds. *IET Computer Vision*, 12(6), 770–778

Villon, S., Chaumont, M., Subsol, G., Villeger, S., Claverie,´ T., Mouillot, D., 2016. Coral reef fish detection and recognition in underwater videos by supervised machine learning: Comparison between deep learning and hog+ svm methods. *ACIVS 2016,* Proceedings 17, Springer, 160–171.

Wang, Y., Song, W., Fortino, G., Qi, L.-Z., Zhang, W., Liotta, A., 2019. An experimental-based review of image *enhancement* and image restoration methods for underwater imaging. IEEE access, 7, 140233–140251.

Zaidi, S. S. A., Ansari, M. S., Aslam, A., Kanwal, N., Asghar, M., Lee, B., 2022. A survey of modern deep learning based object detection models. *Digital Signal Processing,* 103514.

Zhao, H., Zhou, Y., Zhang, L., Peng, Y., Hu, X., Peng, H., Cai, X., 2020. Mixed YOLOv3-LITE: A lightweight real-time object detection method. *Sensors*, 20(7), 1861.

Zhu, Y., Chang, L., Dai, J., Zheng, H., Zheng, B., 2016. Automatic object detection and segmentation from underwater images via saliency-based region merging. *OCEANS* 2016Shanghai, IEEE, 1–4

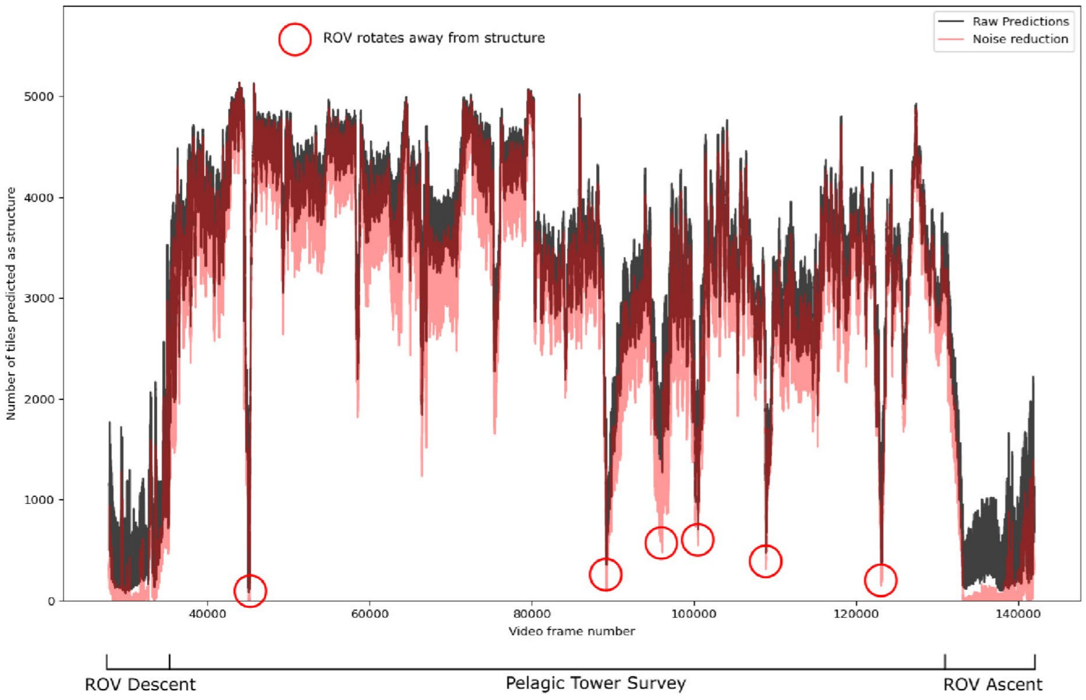


Figure 5. Number of tiles classified as structure in each video frame.