Deep Learning-Based Water Segmentation for Autonomous Surface Vessel

Muhammad Ammar Mohd Adam¹, Ahmad Imran Ibrahim¹,²,³, Zulkifli Zainal Abidin¹,² and Hasan Firdaus Mohd Zaki¹,²

¹ Mechatronics Engineering Department, Kulliyah of Engineering, International Islamic University Malaysia, Gombak KL 53100, Malaysia
² Centre of Unmanned Technologies (CUTe), Kulliyah of Engineering, International Islamic University Malaysia, Gombak KL 53100, Malaysia
³ Author to whom any correspondence should be addressed

imran@iium.edu.my

Abstract. Visual-based obstacle detection from an autonomous surface vessel (ASV) is a complex task due to high variance of scene properties such as different illumination and presence of reflections. One approach in implementing the task is through extracting waterlines to enable inferring of vessel orientation and obstacles presence. Classical computer vision algorithms for detection holds limitation in robustness and scalability. With recent breakthroughs in deep neural network architectures, vision-based object detection is seen to obtain high performance. In this work, the deep learning models based on Convolutional Neural Network (CNN) to implement binary semantic segmentation is studied. This architecture identifies each pixel to water and non-water classes. In purpose of benchmarking models, Fully Convolutional Network (FCN), SegNet and U-Net are trained on a publicly available dataset, IntCatch Vision Data Set (ICVDS), to evaluate the performance. From the experiments carried out, quantitative results show effectiveness of the models with accuracy all above 95.55% and lowest average speed of 11 frames per second. To improve, pre-trained networks (VGG 16, Resnet-50 and MobileNet) are used as a backbone, obtaining an improved accuracy above 98.14% with lowest inferring speed of 10 frame per second. Using the developed ASV, new dataset of 143 images called Malaysia ASV Dataset (MASVD) is collected, labelled and made publicly available. The trained models are tested with the newly collected dataset obtaining accuracy of 75%. The high accuracy performance shows potential for the models to be employed for collision avoidance algorithm in ASV navigation.

1. Introduction

Autonomous surface vessels (ASVs) are marine crafts that operates missions on the surface of water with minimal human intervention. Recent advancements in marine technology contributes to the development works of ASV to ensure a reliable navigation system and improve work efficiency. A fully autonomous system relies heavily on a good collision avoidance system to reduce risk of life or asset damage in the complex marine environment. On the other hand, with the recent breakthroughs in deep neural network (DNN), high accuracy object detection and classification is shown to be achievable particularly in the area of computer vision. This technology enhances safety and reliability of autonomous navigation. Currently, the common approach implemented in obstacle avoidance systems for marine vehicles are RADAR and LIDAR. A cheaper yet accurate and effective solution...
possible is by applying vision-based detection using DNN. Specifically, Deep Convolutional Neural Network (DCNN) is the common DNN approach for image-based applications.

2. Related work

2.1. ASV Development

Over the recent years, several commercial unmanned surface vehicles are available for hydrographic survey in inland waters from small- to large-sized class. The small-sized ASV includes Inception MK1 (Unmanned Survey Solutions), Z-Boat 1250 (Teledyne Marine) and SL20 (OceanAlpha). On the other hand, in the research sector, there are also several available vessels such as HydroDron (Marine Technology Ltd.) and Sonobot (EvoLogics GmbH). The vessels are integrated with variant sensors and components to be implemented for different marine applications. Stateczny et al. [1] demonstrates the possible use of HydroDron as a mobile platform to acquire bathymetric and sonar data in inland waters. While it is shown to be a successful development, the dimensions, however, are considered oversize for our application which are 4.23 m in length and 2.08 m in width. In another work, Kebkal et al. [2] demonstrates the development of a smaller sized boat called Sonobot, which is 1.2 m long, 0.92 m wide and 0.5 m high, with twin-hull design. The system architecture is used as a baseline, with improvement and modification extended from this work. One major difference in our ASV is in the hull design, where mono-hull structure is employed. The developed ASV is shown in Figure 1.

2.2. Obstacle avoidance

The recent main challenges in ASV development is in the autonomous system, especially in path planning and obstacle avoidance algorithms. There have been many works being done using different approaches and sensors. For obstacle detection, common sensors utilized in applications includes passive-ranging (monocular and stereo-vision) and active-ranging cameras (SONAR, RADAR and LIDAR) [3]. Passive-ranging has the advantage of better lateral and temporal resolution but relatively low depth resolution and accuracy. On the other hand, active-ranging provides higher resolution and accuracy in depth dimension but relies on combination of multiple factors including beam width, scan rates, pulse rates and effective ranges [3]. As a result, the complexity as well as operation cost in implementing active-ranging type of sensors is higher than passive-ranging sensing devices. In our work, passive-ranging sensors possess an extra advantage due to lower cost requirement in ASV development.

With the advancements in image processing algorithms and artificial intelligence techniques, vision-based approach is becoming more popular and reliable. Among the approach to implement visual-based obstacle avoidance in maritime applications is through detecting waterlines, or also commonly known as horizon line. There are several techniques in identifying the boundary line between water and non-water regions. An approach to deduce the boundary line is by first implementing segmentation for water and non-water region as done by Steccanella et al. [4]. This approach is explored in this work.

2.3. ASV datasets

Prior to implementation of object detection for obstacle avoidance, an approach to improve the detection performance is to first classify water against non-water regions. To evaluate the performance
of DCNN architectures for water segmentation, IntCatch Vision Data Set (ICVDS) is used for training, testing and validation phases. There are several publicly available maritime datasets for DCNN but mostly focusing on object detection instead of image segmentation.

ICVDS dataset is downloaded from IntCatch Vision Data Set website. The data consists of 318 annotated images which is split into three sets: 191 sample images used as training set, 87 sample images used as validation set and 40 sample images used as evaluation set. The images are frames extracted from videos taken at Lake Garda in Italy at different times of the day.

In relative to most other DCNN applications, the total sample available in the dataset is small. With the limited size of dataset, overfitting problem may occur due to constraint in generalization capacity as demonstrated by Steccanella et al.[4]. However, as presented by Ronneberger et al.[5], one of the common technique to overcome the small training set problem is through the implementation of data augmentation. In this experiment, the particular image transformations applied are flipping, rotation, shearing, cropping and filtering. For water environment, rolling motion of the platform is a normal condition which causes tilted views. To counter the effect, it is critical to consider this condition in training by including rotation augmentations. Apart from that, the dynamic range of brightness is also a vital concern as the ASV will be deployed at different times of the day. Brightness filter is also included in the augmentation process to enhance robustness of segmentation. The summarized detail of augmentation implemented are shown in Table 1.

| Augmentation | Detail                  |
|--------------|-------------------------|
| Flip         | Horizontal              |
| Rotation     | -20 to 20 degrees       |
| Shear        | -20 to +20              |
| Zoom         | 60% to 120%             |
| Filter       | Brightness, Motion blur, Saturation, Invert, Sigmoid |

2.4. Deep learning frameworks for segmentation

Deep learning is a representation learning through a multi-layer network, where each level of learning is transformed to a higher-level representation using non-linear elements. These layers extract important features without the need of specific human definition [6]. The significant achievements of deep learning in addressing cross-field problems, including image classification, object detection and object segmentation, have attracted interest among most computer vision researchers. One of the interesting applications utilizing this approach is in self-driving cars by implementing image segmentation to identify multiple objects of interest based on visual inputs, for instance detection of street lines, road signs and curbs. Similarly, for autonomous water surface vehicle applications, object detection is also required to implement visual-based obstacle avoidance. Image segmentation, a pixel-wise classification approach, plays a vital role in this approach, to identify objects in view, as well as localizing the objects within the image. In this work, semantic segmentation is applied as the extra computation for instance segmentation is not required and insignificant. Many architectures have been introduced for segmentation. Among the state-of-the-art models included in this work are Fully Convolutional Network (FCN) [7], SegNet [8] and U-Net [5].

3. Malaysia ASV Dataset

This new annotated dataset collected in Malaysia is called Malaysia ASV Dataset (MASVD) and is made publicly available. The visual images from perspective on a mini ASV in lake is extracted from a single video recorded from our developed ASV prototype. Specifically, the data collection is conducted in Maryam Lake in International Islamic University Malaysia. Figure 2 illustrates sample images in the dataset.
The total data obtained is 142 images. To enable supervised learning, these images are manually labelled using tools designed for segmentation on web-based annotation platform, Supervisely. The mask is defined by selected polygon coordinates using mouse clicks for connected points.

Figure 2. Sample original images and the corresponding annotated images for MASVD

As shown in Figure 2, the water condition is calm, with no moving obstacles. However, the scene is considered complex due to similar colour characteristics between water and non-water regions. The water is brownish in colour, which is similar to the surrounding trees. This will create difficulty for visual-based classification task if the model relies heavily on colour. Apart from that, there are aquatic plants on water surface, which will also create the complexity in classifying the regions. However, in general, the environment is possibly a good representation of lakes in Malaysia.

4. Experimental results and evaluation

In this paper, FCN, SegNet and U-Net are evaluated for water segmentation from ASV visuals. The starter-code is obtained from open-sourced code done by Steccanella et al. [4] which specifically implemented image segmentation using U-Net architecture. The models are developed using Keras library with TensorFlow as the backend engine. For this particular work, two-class output are considered: water and non-water. All the compared models are trained using the images from IntCatch Vision Data Set (ICVDS). To standardise evaluation results, the experiments are conducted using the same hyperparameters implemented by Steccanella et al. [4].

In order to evaluate the individual performances, accuracy, precision, recall and F1-score metrics are measured which best results are presented in Table 2. Despite Badrinarayanan et al. [8] demonstrated that more stored encoder feature maps and larger decoder size improves the performance
of image segmentation especially in F1-score, SegNet and U-Net have shown better score than FCN model on the tested dataset. However, the performance improved only in the range of 3-4% across all evaluation metrics. In overall, the results show high performances with metrics evaluated scoring above 93% for all the employed architectures. This demonstrates effective and successful pixel-wise classification for the specific data tested.

Figure 3 illustrates the qualitative results for each model tested on sample images from ICVDS. The ground truth segmentation results are shown as reference, where region coloured violet represents water and black represents non-water. For the implemented basic models, detected water region is coloured purple whereas non-water region is coloured yellow. All these segments are determined from predicting the class of individual pixels of the input image. As shown in Figure 3, pixel classification near boundaries between the two regions is very challenging and causing most of the false detections. FCN generates smoother segmentation as compared to SegNet and U-Net, missing tiny information near boundaries, hence obtaining lower F1-scores. In this particular implementation, it is critical to avoid false positive (FP), where pixels are misclassified as water for non-water regions and possibly causing collisions. Thus, apart from accuracy, precision is the next important metric to observe. U-Net classifies better in terms of accuracy, precision and F1-score, whereas SegNet outperforms other compared models in recall evaluation. Nevertheless, the segmentation models are able to detect more than 90% of the classes correctly, with issues mostly near boundaries. In general, the qualitative results are seen to correspond well to the quantitative evaluations.

In addition, the inference rate for the experiment is also presented in Table 2 to analyse if the models predict efficiently on the test data which is a common factor being considered in robotics application. Steccanella et al. [4] shown that the number of extracted features in layers has a negative relation to the speed of computation. On the other hand, Badrinarayanan et al. [8] demonstrated that SegNet, in comparison to FCN, has almost 6 times more feature maps in each decoder layer, resulting in higher training accuracy but slower inference computation. However, interestingly with the ICVDS, the opposite is observed for inference rate. By running the algorithms on the same machine, FCN obtained average computational speed of 11 frames per second for predicting a sample image, whereas for SegNet with larger overall network, the performance improvement is more than doubled.

| Network | Accuracy  | Precision | Recall  | F1-score | FPS |
|---------|-----------|-----------|---------|----------|-----|
| FCN-8   | 0.9555    | 0.9669    | 0.9330  | 0.9409   | 11  |
| SegNet  | 0.9686    | 0.9577    | 0.9757  | 0.9661   | 32  |
| U-Net   | **0.9802**| **0.9911**| 0.9612  | **0.9722**| 17  |

Table 2. Performance of FCN, SegNet and U-Net tested on IntCatch Vision Data Set. The highest score for each evaluated metrics is in **bold**.
To evaluate the effect of model weights pre-trained for image classification, three different pre-trained models, which are VGG 16 [9], Resnet-50 [10] and MobileNet [11], are introduced as backbone. The new integrated architectures are trained with the same dataset and evaluated with the same metrics to produce comparable results. As the resources for training are constrained, FCN is not applied for this training. The best result of each implementation is presented in Table 3. SegNet with Resnet as base model performs best in terms of precision and F1-score. However, the margin is small even comparing with the lowest performing model in this experiment, which is SegNet with VGG as backbone model. In overall, by introducing pre-trained model as encoder, the segmentation models are shown to improve. One of the reasons could be due to the relevant features extracted from water-related objects in ImageNet provide a good basis for the overall model.

Table 3. Performance of SegNet and U-Net with VGG, Resnet and MobileNet as backbone tested on IntCatch Vision Data Set. The highest score for each evaluated metrics is in bold.

| Segmentation Model | Base Model | Accuracy | Precision | Recall | F1-score | FPS |
|--------------------|------------|----------|-----------|--------|----------|-----|
| SegNet             | VGG 16     | 0.9841   | 0.9923    | 0.9689 | 0.9791   | 11  |
| U-Net              | VGG 16     | 0.9814   | 0.9962    | 0.9587 | 0.9752   | 11  |
| SegNet             | Resnet-50  | 0.9923   | 0.9984    | 0.9833 | 0.9906   | 11  |
| U-Net              | Resnet-50  | 0.9924   | 0.9978    | 0.9836 | 0.9904   | 10  |
| SegNet             | MobileNet  | 0.9922   | 0.9969    | 0.9838 | 0.9900   | 18  |
| U-Net              | MobileNet  | 0.9921   | 0.9959    | 0.9844 | 0.9899   | 17  |

To evaluate the segmentation performance for new unseen environment, the models are tested on MASVD without re-training process. Also, in this experiment, to simplify the evaluation, only two metrics are considered which are accuracy and inference speed. Each best quantitative result is presented in Table 4. Interestingly, the basic SegNet is able to outperform the corresponding models with this new dataset. This shows that the particular architecture has successfully extracted meaningful water features as well as important contextual information from ICVDS training dataset. As a result, the model is able to classify well despite similar colour appearance between objects in sample images.
The variance in accuracy metric between compared models is large, where the performance obtained is in the range of 65.17% to 96.21%.

Sample qualitative results of testing each model on MASVD are presented in Figure 4. From general evaluation, it is seen that SegNet tends to correctly extract water region, whereas U-Net has the tendency of extracting the trees and ground, possibly due to significant difference in colour of water. This shows that the architectures employed are insufficiently generalised for classifying the new dataset and will need further training to improve quality of collected features.

In terms of inference speed, the best performance is demonstrated by SegNet and U-Net networks with MobileNet as backbone, which is 6 frames per second. This is considerably a marginal improvement compared to corresponding basic models which possibly be due to smaller image as input. In general, the models pre-trained with ICVDS images are insufficiently generalised to be able to detect in new environment such as in the waters of Malaysia. Wider range of sample images or region-specific dataset is required to ensure a more robust and reliable water segmentation model.

Table 4. Performance of experimented models with pre-trained networks as backbone tested on Malaysia ASV Dataset. The highest score for each evaluated metrics is in bold.

| Segmentation Model | Base Model | Accuracy | FPS |
|--------------------|------------|----------|-----|
| FCN-8              | Vanilla    | 0.6517   | 4   |
| SegNet             | (Baseline) | 0.9621   | 5   |
| U-Net              |            | 0.8987   | 5   |
| SegNet             | VGG 16     | 0.8091   | 5   |
| U-Net              |            | 0.7712   | 5   |
| SegNet             | Resnet-50  | 0.7600   | 5   |
| U-Net              |            | 0.7570   | 5   |
| SegNet             | MobileNet  | 0.7486   | 6   |
| U-Net              |            | 0.7450   | 6   |
5. Conclusion
This research studies the potential use of deep learning architecture for waterline detection in ASV navigation. Feature extraction of water against non-water are trained using different segmentation models that are known to perform well in other applications such as autonomous cars. The performance of basic U-Net is shown to work best compared to other benchmarked models. Additionally, U-Net with Resnet-50 is seen to have best accuracy performance in overall.
An autonomous surface vessel of small-sized class is designed and developed as a platform to collect new visual datasets for waterline detection. The prototype is operated at Maryam Lake in Malaysia to capture different sequences. The dataset is annotated and made available for public use. The location environment is different from existing IntCatch Vision Data Set and is challenging due to smaller space for navigation and presence of small obstacles such as water plants.

Three popular DNN models for semantic segmentation applications (FCN, SegNet and U-Net) have been studied and implemented to classify water and non-water regions. It is shown that all three network architectures achieve good performance in terms of accuracy, precision, recall and F1-score, ranging from 93.3% to 99.11% with U-Net performing better than SegNet and FCN in average. For inference rate, the range obtained is from 11 fps to 32 fps. However, different trend is obtained, with SegNet outperforming FCN and U-Net by almost triple and double respectively.

It is also shown that using pre-trained networks (VGG, Resnet and MobileNet) as backbone model, achieves better performance ranging from 95.87% to 99.84%, championed by SegNet with Resnet-50 in average. For inference rate, it is shown that the speed is reduced, ranging from 10 fps to 18 fps. SegNet with MobileNet infers faster than other combined models with accuracy performance being among the highest.

Results show that ICVDS trained models obtains lower performance with MASVD achieving accuracy below 80.91%. While the prediction performance for this dataset is shown reduced, the performance is considerably as this shows that the pre-trained model on ICVDS generalises well for unseen environment as in MASVD which data augmentation has shown to contribute. With more datasets of different scenes trained for the model, a robust model can be achieved.

Acknowledgments
This research paper is supported by research initiative grant scheme with the number RIGS16-348-0512, International Islamic University Malaysia (IIUM) with equipment and additional financial support by Temasek Hidroteknik.

References
[1] Stateczny A 2018 Hydrodron – New Step for Professional Hydrography for Restricted Waters 2018 Balt. Geod. Congr. (BGC Geomatics) 226–30
[2] Kebkal K G, Glushko I, Tietz T, Bannasch R, Kebkal O G, Komar M and Yakovlev S G 2014 SONOBOT - an autonomous unmanned surface vehicle for hydrographic surveys, hydroacoustic communication and ... Int. Conf. Exhib. Underw. Acoust.
[3] Halterman R and Bruch M 2010 Velodyne HDL-64E lidar for unmanned surface vehicle obstacle detection Unmanned Syst. Technol. XII 7692 76920D
[4] Steccanella L, Bloisi D, Blum J and Farinelli A 2019 Deep learning waterline detection for low-cost autonomous boats Adv. Intell. Syst. Comput. 867 613–25
[5] Ronneberger O, Fischer P and Brox T 2015 U-Net: Convolutional Networks for Biomedical Image Segmentation Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015 ed N Navab, J Hornegger, W M Wells and A F Frangi (Cham: Springer International Publishing) pp 234–41
[6] Lecun Y, Bengio Y and Hinton G 2015 Deep learning Nature 521 436–44
[7] Long J, Shelhamer E and Darrell T 2014 Fully Convolutional Networks for Semantic Segmentation vol 79
[8] Badrinarayanan V, Kendall A and Cipolla R 2017 SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation IEEE Trans. Pattern Anal. Mach. Intell. 39 2481–95
[9] Simonyan K and Zisserman A 2014 Very Deep Convolutional Networks for Large-Scale Image Recognition arXiv 1409.1556
[10] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-Decem 770–8
[11] Howard A G, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M and Adam H 2017 MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications