Stereo Vision-Based Ship Detection and Positioning System Utilizing Deep-Learning

by

KOBAYASHI Mitsuru*, SATO Keiji*, NIWA Yasuyuki*, SAITO Eiko**, TOITA Keisuke***

Abstract

In recent years, research on Maritime Autonomous Surface Ships (MASS) and Remotely Operated Ships has grown significantly. For these ships to navigate safely, it is essential to compensate for the weaknesses of conventional sensors, such as X-band and S-band maritime radars, and the Automatic Identification System (AIS) with visual ship detection systems using cameras. In this study, the authors developed a stereo vision-based ship detection and positioning system incorporating the Faster R-CNN deep learning framework for object detection and recognition in maritime environments. The system consists of two general-purpose network cameras (1920 x 800-pixels resolution, 18.5° field of view), mounted on a test ship, which was a 15,000-gross-ton bulk carrier. The system is capable of sensing ships up to 8 km and navigation buoys up to 2 km ahead, provided that the object images exceed 16 pixels in length. The distance was determined using the principle of triangulation by comparing two simultaneously captured images. The distance estimation error rate was around 5% when compared with the AIS information of ships located 8 km away, and 1.2% when compared with the charted position of navigation buoys located 2 km away, aligning closely with theoretical estimates based on camera resolution. This paper describes the system architecture and presents the evaluation results for detection performance and range accuracy. In addition, a prototype monocular ranging algorithm was implemented, which estimates distance by detecting the horizon and the waterline of the target object in the image and calculating the vertical displacement. The accuracy of this method was then evaluated in comparison with the triangulation-based approach.

* National Maritime Research Institute, National Institute of Maritime, Port and Aviation Technology

** National Institute of Maritime, Port and Aviation Technology (At the time of research)

*** The University of Electro-Communications (At the time of research)

Received June 23rd, 2025

Accepted October 6th, 2025

Contents

1. Introduction

In recent years, research on autonomous navigation and remote ship operation has progressed across various domains [1]. Conventional ships typically rely on radar and the Automatic Identification System (AIS) for obstacle detection and positioning. However, radar has difficulty detecting ships with low radar reflectivity, whereas AIS is unable to detect ships that are not AIS-equipped and remains vulnerable to both jamming and spoofing. These limitations highlight the insufficiency of relying solely on these systems. Since conventional collision avoidance has depended on human visual watchkeeping, supplementing autonomous ships with visual detection systems based on visible light could significantly enhance safety.

As part of our previous research [2], one visible light camera and one far-infrared camera were installed on a coastal building to detect and estimate the positions of passing ships. The key findings were as follows:

- Deep learning-based ship detection significantly outperformed traditional image processing methods, reducing both false positives and missed detections.
- Ships with less than 19 pixels on the longer side of the image were difficult to detect.
- Distance estimation relied on the vertical position of ship images relative to the horizon. However, even with a camera installed 58 meters above sea level, the distance estimation error rate for a ship about 2,000 meters away was no less than 10%.

In the current study, we aim to advance the following areas:

- Ship Detection: Building on the success of deep learning methods in previous work, this study continues to employ AI-based techniques for ship image detection. The objective is to enhance the detection performance of small and distant ships, thereby enabling earlier collision avoidance. To achieve this, we select suitable detection algorithms and apply image preprocessing techniques.
- Distance Estimation: The previous method, which estimated ship distance continuously by detecting the waterline position of the ship and the horizon in a single image, lacks sufficient accuracy for estimating a ship’s course. In this study, we introduce a stereo vision approach using two cameras. By aligning the two ship captured images via template matching function, we aim to enhance the precision of distance estimation. On the other hand, the angular resolution of the direction estimation was sufficient compared to the distance resolution and there was little impact on the ship's position estimation.

2. System Configuration and Methodology

2.1 Hardware Configuration

For the purpose of this study, visible light cameras were installed on both port and starboard sides of the bridge of the 15,000 gross tonnage bulk carrier, referred to as the "own ship". The cameras were placed 25 meters apart and positioned approximately 15 meters above the water surface. Two images of ships are captured simultaneously every 10 seconds. The two camera's specification is as follows:

Model: SONY SNC-EP580　　Resolution: 1920×1080 pixels
Zoom Level: 3x　　Focus: Infinite
Horizontal Field of View (FOV) after zooming: 18.5°

2.2 Ship Image Detection

In this study, we focused on images of ships and navigation buoys, collectively referred to as "ship images." Our system aims to detect small ship images, as previously mentioned. However, due to the indistinct shape of small images, we do not perform discrimination between ships and navigation buoys.

Faster R-CNN [3] which is one of the representative frameworks for the Object Detection function in the field of Convolutional Neural Network (CNN) deep-learning network models is used to capture ship images. Specifically, this study examines the implementation of two notable frameworks: py-faster-rcnn [4] (a Faster R-CNN implementation) and its successor, Detectron [5]. Furthermore, performance varies depending on the choice of feature extraction layer such as VGG_CNN_M_1024 [6], ResNet [7] or AlexNet [8], which corresponds to the lower layers of the network model, and whether image expansion is applied as a preprocessing step.

Fig.1 shows the mean Average Precision (mAP) and VRAM requirements according to the implementation framework, feature extraction layer, and image expansion. As the results indicated in Fig.1, mAP can be improved by appropriately selecting feature extraction layers. Image expansion can slightly improve mAP although it requires larger VRAM.

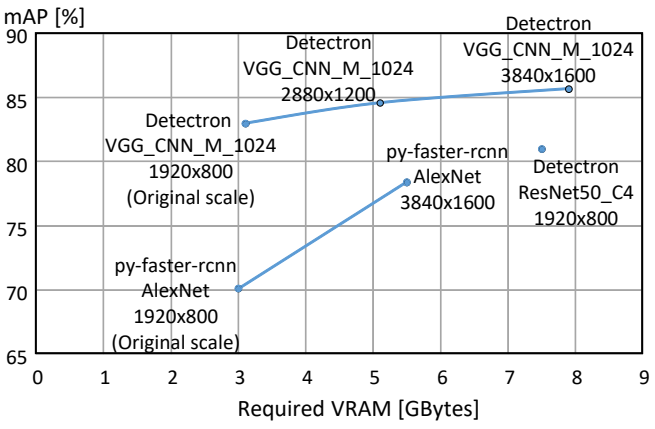


Fig. 1 Deep Learning Frameworks and Detection rate.

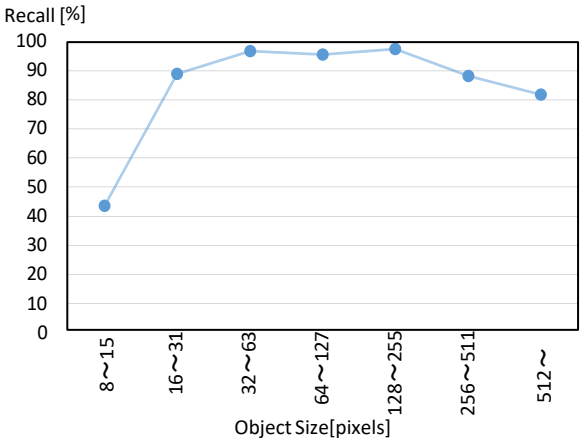


Fig. 2 Size of Objects and Recall rate of detection.

Fig.2 shows the recall, which is the proportion of ship images correctly detected, measured for different sizes on the long side of the image using the Detectron implementation with VGG_CNN_M_1024 feature extraction layer with no image expansion. For ship images with a long side of 32 pixels or more, the recall was around 95% (the decrease in recall for 256 pixels or more is due to the inclusion of ships docked at the quay and ships whose entire images do not fit in the photo). 32 pixels on the long side of the image corresponds to a 10-meter length object at 1 mile away.

Both Fig. 1 and Fig. 2 present comparisons of detection results on the same images captured under clear daytime conditions. Under rainy weather or low illumination, however, the detection rate decreases significantly.

2.3 Matching Ship images in Left and Right landscape pictures

In this process, identical ships are identified from multiple ship silhouettes in the left and right images, and their positional differences are measured.

First, candidates for the same ship are selected based on the bounding box positions and sizes obtained from captured ship image, along with the luminance histograms of the ship silhouettes.

Next, pairs of ship silhouettes corresponding to the same ship are matched using template matching to determine the positional differences in the images.

Specifically, the positional difference of the pair of ship silhouettes in the left and right images is determined by searching for the coordinate offset that results in the most similar images, shifting the offset 1 pixel at a time. While there are several types of cross-correlation functions, such as simply summing the differences in brightness, Normalized Cross Correlation (NCC) was adopted in this study. NCC is suitable even when the cameras have different sensitivity settings, as they operate independently. In this study, the cross-correlation calculations were performed on grayscale images derived from the original ship images.

If the cross-correlation does not exceed the threshold, the pair of ship silhouettes is removed from the candidate set, as they cannot be considered to correspond to the same ship.

2.4 Distance Estimation

A schematic diagram of the positions of the two cameras and the detected object is shown in Fig. 3 (note that the distance to the object P from the coordinate origin O is 100 ~ 400 times as long as the baseline L).

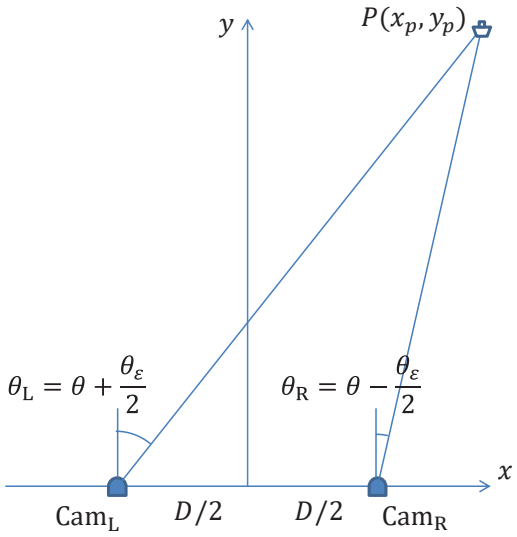


Fig. 3 Diagram for Position Estimation by Stereovision.

From the results of the template matching in the previous section, θ_L and θ_R, which are the relative direction from each camera to the target, are obtained as shown in Fig. 3. Furthermore, the relative direction θ from the center of the bridge of own ship to the target and the disparity θ_ε are obtained by the following equations.

$$\theta = \frac{\theta_L + \theta_R}{2} \tag{1}$$

$$\theta_\varepsilon = \theta_L - \theta_R \tag{2}$$

The relationship between the positions of the two cameras and an arbitrary point P (x, y) is expressed by the following equation.:

$$\begin{cases} x + L/2 = y \tan\theta_L \\ x - L/2 = y \tan\theta_R \end{cases} \tag{3}$$

The intersection point $P(x_p, y_p)$ of these two lines and the distance to the target D are calculated as follows:

$$y_{\mathrm{p}} \simeq \frac{L\cos^2\theta}{2\tan\frac{\theta_\varepsilon}{2}} \tag{4}$$

$$x_{\mathrm{p}} \simeq y_{\mathrm{p}}\tan\theta \tag{5}$$

$$D \simeq \frac{L\cos\theta}{2\tan\frac{\theta_\varepsilon}{2}} \tag{6}$$

The distance estimation error ΔD with respect to the disparity $\Delta\theta_\varepsilon$ is obtained by differentiating D with respect to θ_ε:

$$\frac{\Delta D}{D} \simeq -\frac{D}{L\cos\theta}\Delta\theta_\varepsilon \tag{7}$$

This shows that the distance estimation error rate $\Delta D/D$ for a detected target located directly in front is approximately equal to the estimation error of the disparity $\Delta\theta_\varepsilon$ multiplied by the ratio of the distance to the target to the distance between the cameras, D/L.

For example, under the configuration in this study (horizontal field of view of 18.5° and horizontal resolution of 1,920 pixels) and the template matching resolution set to 1 pixel, the estimation error of the disparity $\Delta\theta_{\varepsilon 1}$ is as follows:

$$\Delta\theta_{\varepsilon 1} = \frac{18.5\times\pi}{180\times 1920} = 1.68\times 10^{-4}\ [\mathrm{rad}] \tag{8}$$

Given $L = 25$ meters, the distance ratio D/L is 74 for a target 1 mile away and 296 for a target 4 miles away. Therefore, the estimated distance error rate $\Delta D/D$ for a target directly in front is 1.25% and 4.98%, respectively.

Additionally, assuming a yaw rate of 0.5 degrees per second and a time lag of 0.03 seconds between the left and right camera shots, the resulting error $\Delta\theta_{\varepsilon 2}$ is as follows:

$$\Delta\theta_{\varepsilon 2} = \frac{0.5\times 0.03\times\pi}{180} = 2.62\times 10^{-4}\ [\mathrm{rad}] \tag{9}$$

$$\Delta\theta_\varepsilon = \sqrt{\Delta\theta_{\varepsilon 1}^2 + \Delta\theta_{\varepsilon 2}^2} \tag{10}$$

This error leads to an estimated distance error rate of 1.94% for a target 1 mile away and 7.76% for a target 4 miles away. It is expected that relatively large errors occur when the orientation angle of the camera changes due to yaw or vibration. If the shooting timings of the left and right cameras could be perfectly synchronized, this error would not occur. However, with independent 30 frames per second general-purpose cameras, it is difficult to reduce the timing difference to less than 33 milliseconds even if the shooting commands are sent simultaneously. During this interval, ship motion can cause perspective transformations, which in turn introduce error $\Delta\theta_{\varepsilon 2}$. Moreover, since the cameras used in this study did not provide precise shooting timing information, it was not possible to apply corrective adjustments.

On the other hand, the estimation error rate $\Delta x_P/D$ in the relative direction to the target ship, approximately equal to $\Delta\theta$, is expected to be relatively small compared to the distance estimation error rate.

Other factors contributing to errors in measuring the relative position of targets include heading sensor errors, changes in the captured image due to ship movement, differences in the timing of image capture of two cameras, and estimation errors in other navigation equipment. These relative position estimation errors were larger than the error in measuring the relative position of targets using the stereo camera proposed in this study.

Fig. 4 Example Images of Detected Ships and Navigation Buoys. (Upper: Left Cam, Lower: Right Cam)

3. Results of Experiments

3.1 Overview

The proposed system is installed on the tested ship (the 15,000 gross tonnage bulk carrier) and conducted detection and position estimation of navigation buoys and oncoming ships during navigation. The weather was clear with a visibility of 20 kilometers. The network model used for detection was Faster R-CNN implemented with Detectron, utilizing VGG_CNN_M_1024 as the feature extraction layer. Input images were sized at 1920×800 pixels (original scale). Fig. 4 illustrates an example of a captured image and detected ship. The field of view of the two cameras is set to 18.5°. However, approximately 20 pixels on left and right side end cannot use for ship detection, resulting in an effective field of view of approximately 18°.

In detection, ships and buoys are not differentiated in the detection category and are displayed as "ship" within the detection frame, as described at the end of Chapter 1. The numbers represent the output of logistic regression.

3.2 Verification through Navigation Buoys detection

Fig. 5 shows the trajectory of the Own Ship and the positions of four navigation buoys captured by the camera. The red line represents the trajectory of the own ship. Circled number (such as ❶) in Fig 5 is buoy number of each buoy. Fig. 6 is the image of the navigation buoy ❶ from its initial detection position. The height of the navigation buoy is measured approximately 13 pixels.

Fig. 7 shows the relative distance estimation results for each buoy. In this case, the ship was approaching the four buoys and they were detected from about 2 km away from them. Note that the height of each buoy is 4.5 meters. However, estimation could not be made at a time when the buoys' or ship's images were obscured by the overlapping object such as the mast of own ship.

In the figures, the "Actual Distance" is measured based on the GPS positions of the buoys obtained from the nautical chart and the own ship.

Table 1 summarizes the position estimation errors for each buoy. In the Range measurement, the random error, represented by the Standard Deviation, is approximately 1.1% (about 20 m) which is consistent with the expected error in section 2.4. In contrast, the systematic error, represented by the Average, varies among buoys. The cause is unclear, but possible factors include

deviations of buoy positions from the nautical chart due to tidal currents, distortions in the camera images, or inaccuracies in the estimation formulas and parameters. In the Azimuth measurement, the random error was about 0.2 degrees, which corresponds to approximately 6 m when converted to linear displacement at the buoy's distance. This is about one-third of the error in the Range measurement.

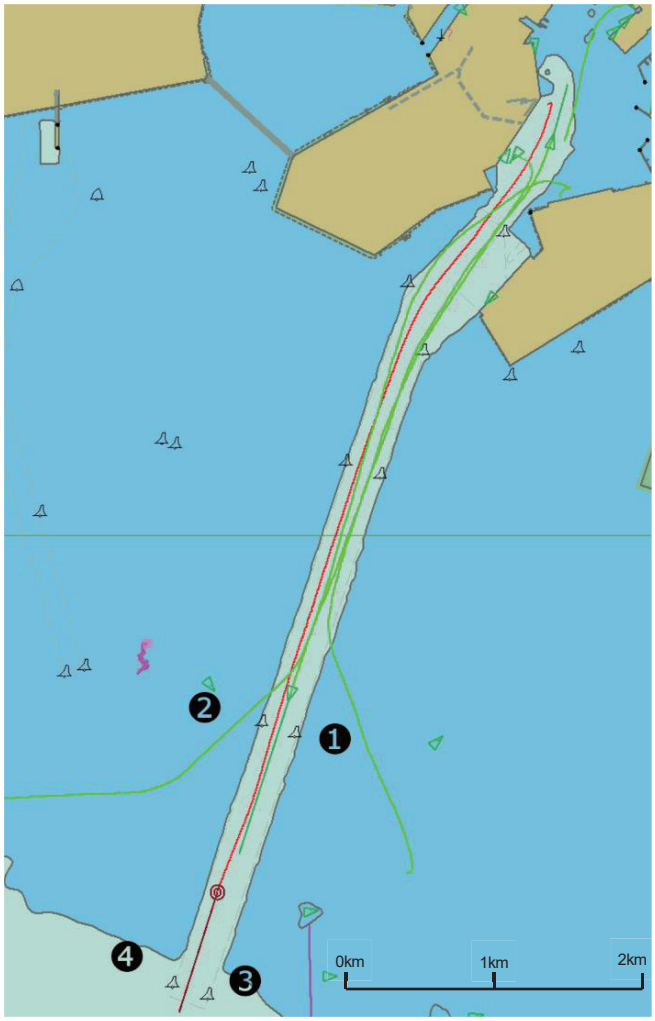


Fig. 5 Position of Navigation Buoys and Route of Own Ship (Red Line, downward).

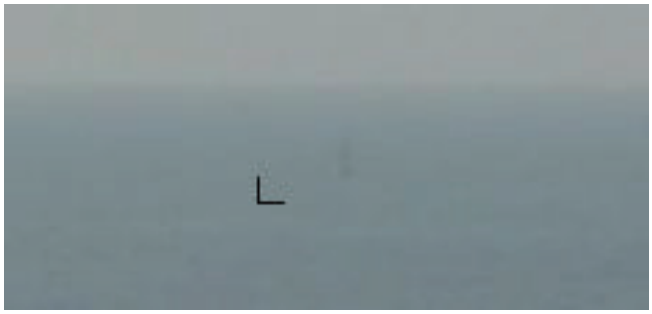


Fig. 6 Example image of Navigation Buoy (detected). (Black Line denotes 10 pixels)

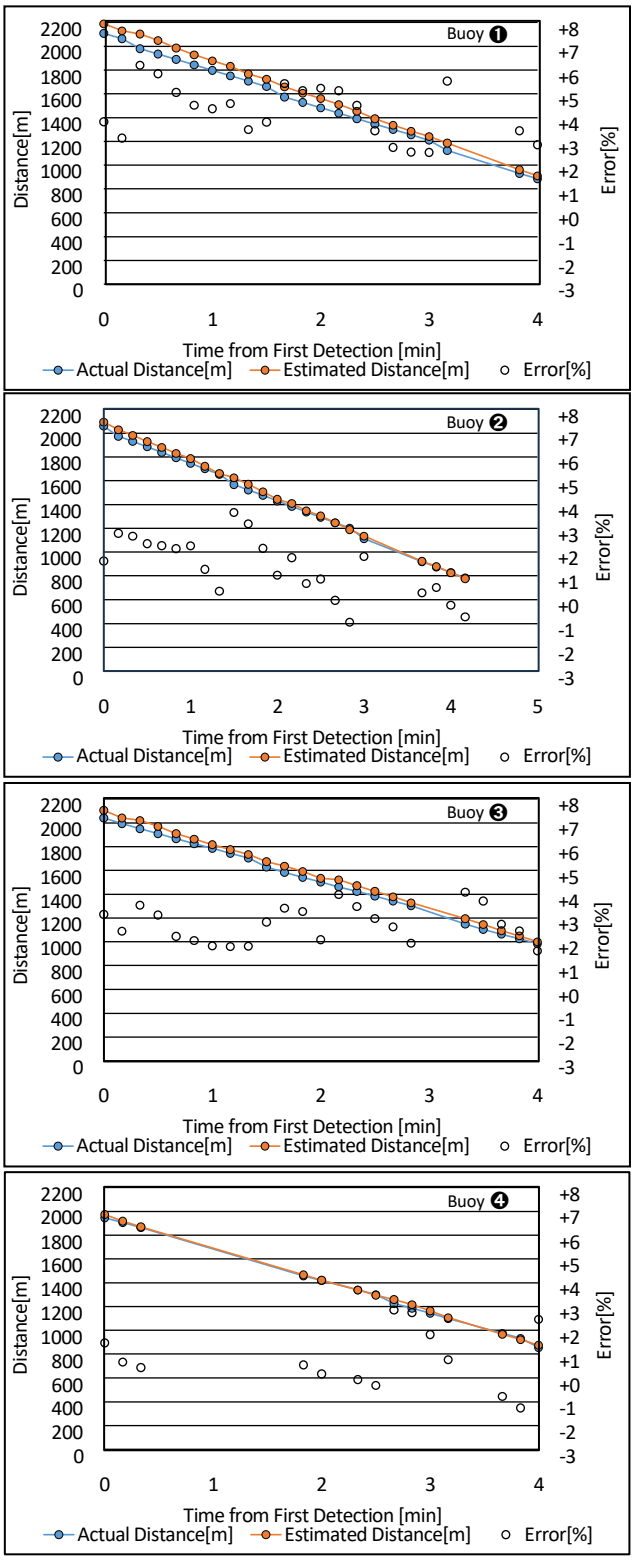


Fig. 7 Actual and Estimated Distances of Buoys. (Circled numbers correspond to those in Fig. 5)

Table 1 Position Estimation Errors for Navigation Buoys.

	Range error				Azimuth error			
	Average		St. Dev.		Average		St. Dev.	
	[%]	[m]	[%]	[m]	[deg]	[m]	[deg]	[m]
Buoy ❶	+4.25	+67.5	1.09	25.6	-0.21	-6.6	0.18	6.0
Buoy ❷	+1.39	+23.4	1.22	20.7	-0.57	-14.8	0.09	5.2
Buoy ❸	+2.75	+42.0	0.73	13.3	+0.28	+6.4	0.23	5.4
Buoy ❹	+0.82	+10.9	1.24	14.3	+0.34	+6.7	0.24	5.2

3.3 Verification through Oncoming ships detection

Fig. 8 illustrates the AIS positions and estimated positions of three oncoming ships along with the position of the own ship at the time of detection. Fig. 9 shows the relative distances to the oncoming ships based on the own ship's GPS and the oncoming ships' AIS, as well as the estimated distances and errors. Compared to the buoys, the larger size of the oncoming ships allows for detection even at longer distances. On the other hand, as shown in Table 2, the distance estimation error is relatively larger. The random error, represented by the standard deviation of the estimated error rate, is approximately 5% (for oncoming ship ③, the standard deviation is 5.1% when excluding two exceptional cases where the ship overlapped with the own ship's structures or the edges of the photo), which is similar to the expected error in section 2.5.

The specifications of each oncoming ship are as follows:

Oncoming ship ①: Container ship, 146 × 23 meters

Oncoming ship ②: Bulk carrier, 88 × 14 meters

Oncoming ship ③: Oil tanker, 112 × 19 meters

Fig. 10 shows the detection image of oncoming ship ③ when it is 8 km away. The main body of the ship measures approximately 22 pixels on its longest side.

Table 2 Errors of Estimated Position of Oncoming Ships.

	Range error				Azimuth error			
	Average		St. Dev.		Average		St. Dev.	
	[%]	[m]	[%]	[m]	[deg]	[m]	[deg]	[m]
Ship ①	-7.1	-383	3.7	198	+1.4	+134	0.7	62
Ship ②	-8.7	-524	4.6	277	+1.7	+173	0.6	60
Ship ③	-8.8	-727	11.7	963	+2.7	+391	0.3	47

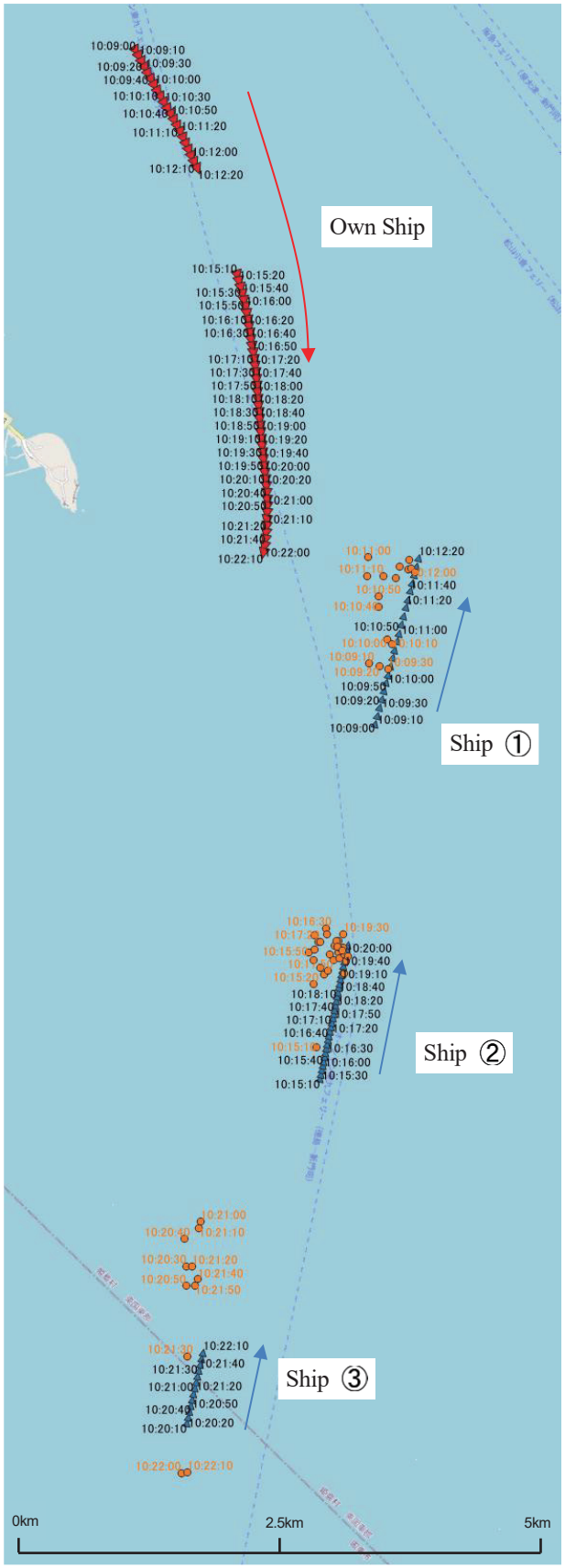


Fig. 8 Positions of Own Ship (Red) and
3 Oncoming Ships
(Blue: AIS position, Orange: Estimated position)

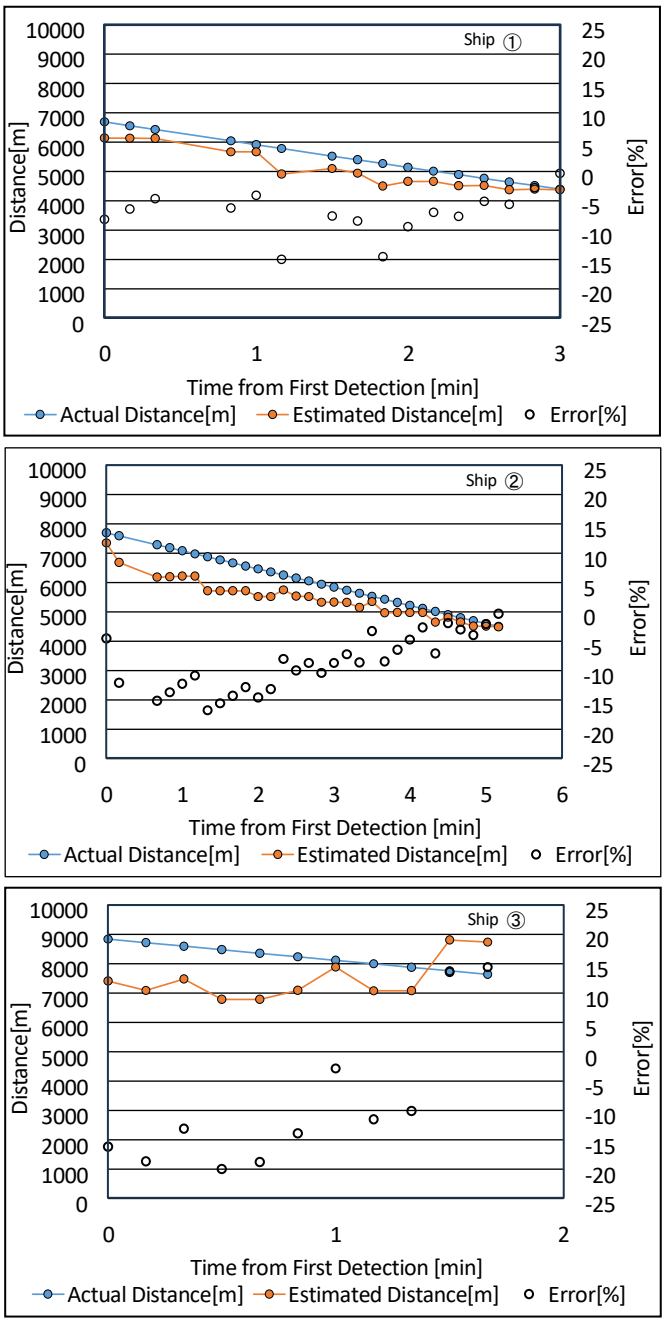


Fig. 9 Actual and Estimated Distances of Oncoming Ships.
(Circled numbers correspond to those in Fig. 8)

Fig. 10 Example image of Oncoming Ship (detected).
(Black line denotes 10 pixels)

3.4 Distance Estimation by Single Camera

To compare with the stereo-vision method, distance estimation method using a single camera was performed. As shown in Fig. 11, this method involves detecting the horizon and the waterline of the object from brightness changes in the scene image and estimating the distance based on the difference in vertical position in the image. Specifically, the average brightness was calculated both 30 pixels outside the left and right edges of the bounding box output by the detector, and within the bounding box edges. The horizon was defined as the first vertical position, when scanned downward from the top, where the brightness difference between the outside averages exceeded a threshold. Similarly, the waterline was defined as the first vertical position, when scanned upward from 5 pixels below the bottom edge of the bounding box, where the difference between the outside and inside averages exceeded a threshold.

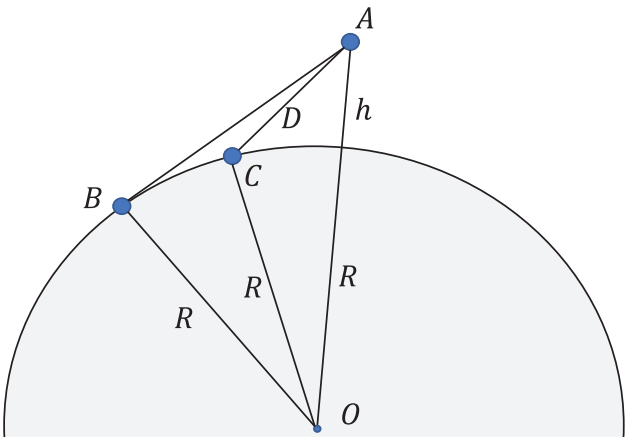
In Fig. 12, $\angle BAC$ is obtained from the vertical coordinate difference between the horizon and the waterline of the object in the scene image. The distance to the object is then calculated using the following equation.

$$\angle BAO = \sin^{-1}\frac{R}{R+h} \tag{11}$$

$$\angle CAO = \angle BAO - \angle BAC \tag{12}$$

$$D = (R+h)\cos\angle CAO - \sqrt{(R+h)^2\cos^2\angle CAO - (h^2+2Rh)} \tag{13}$$

Fig. 11 Detection of the Horizon and the Waterline of Navigation Buoys.



A : Camera
B : The Horizon
C : Object
O : Center of the Earth
R : Radius of the Earth
h : Height of the Camera
D : Distance to the Object

Fig. 12 Diagram for Distance Estimation by single camera

Table 3 Errors of Estimated Distance of Navigation Buoys with single camera.

	Range error			
	Average		St. Dev.	
	[%]	[m]	[%]	[m]
Buoy ❶	+5.05	+83.7	1.85	44.8
Buoy ❷	+4.91	+76.0	1.85	41.6
Buoy ❸	+2.87	+48.2	2.49	41.8
Buoy ❹	+3.87	+53.5	1.56	26.5

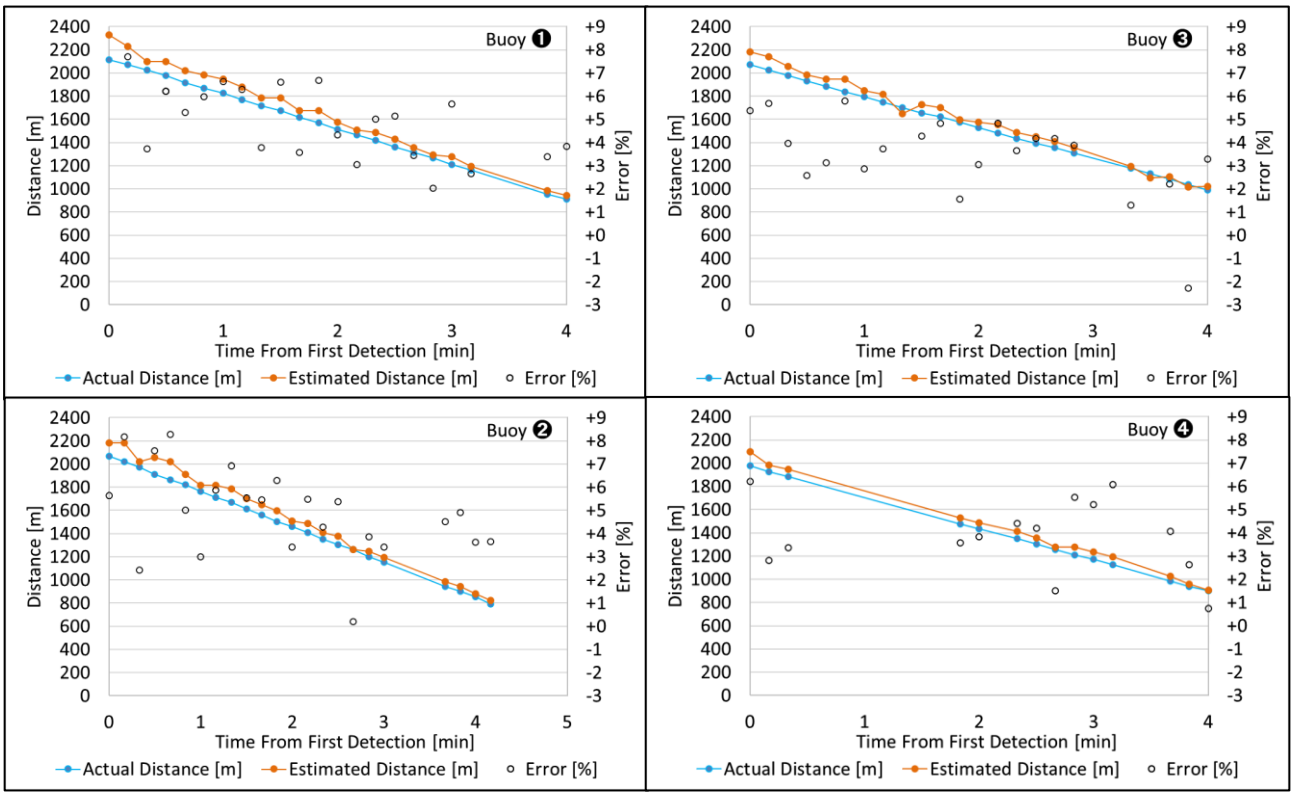


Fig. 13 Actual and Estimated Distances of Buoys by single camera.
(Numbers in a circle correspond to those in Fig. 5)

According to Table 3, the errors are slightly larger than those in Table 1. However, the most significant issue is not the magnitude of the error but the weakness in visibility. If visibility is poor, the horizon becomes blurred and cannot be detected. When the sea surface is not calm, the waterline of the target also cannot be detected. Depending on the weather, it is necessary to finely adjust the parameters for horizon and waterline detection. Additionally, as shown in Fig. 14, the distance per pixel increases rapidly as the object gets closer to the horizon.

The monocular camera distance estimation system of the subject ship, which had a camera installed 15 m above sea level, was unable to estimate the distance to the oncoming ship 8 km away, because the difference in height between the horizon and the waterline of the ship on the image was less than one pixel., as seen in Fig. 10. In stereovision, having visibility to the object is enough, and the left-right image matching is more robust compared to detecting the waterline.

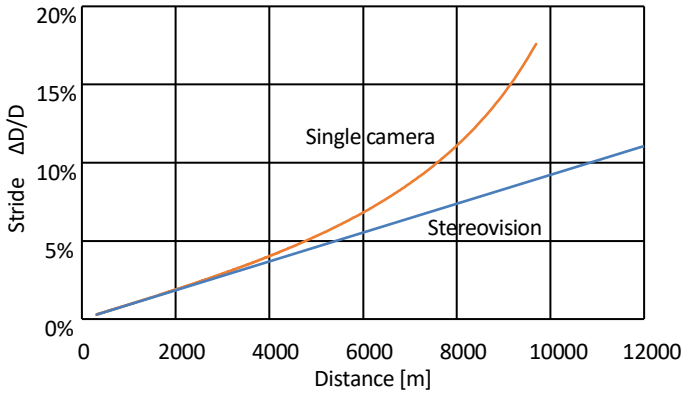


Fig. 14 Stride rate $\Delta D/D$ per 1 pixel
(Height 15 meters, Resolution 1920 pixel/15.2°)

4. Conclusion and Future Works

The authors developed a system that detects other ships and estimates their relative positions from two images simultaneously captured by two visible-light cameras mounted on a commercial ship facing forward. Deep learning was used for ship detection, and stereo-vision methods were employed for position estimation. As a result, the following achievements were obtained:

- By selecting appropriate feature extraction layers and implementations within the deep learning framework, the detection rate was improved. For ship images with a long side of 32 pixels (equivalent to a 10-meter object at 1 nautical mile away) or more, the recall rate was approximately 95%, excluding ships docked at the quay or those whose entire images were not captured in the photo.
- Navigation buoys with a height of 4.5 meters were reliably detected from approximately 2 km by the system.
- The random errors in the radial distance when estimating the relative position of navigation buoys and oncoming ships from the own ship are almost same as the error theoretically predicted based on the camera resolution.
- The random azimuthal distance error is smaller than the random radial distance error. It is likely that external factors such as ship motion have a significant impact for the azimuthal distance error.
- In the Distance Estimation by Single Camera method, the estimation error at a distance of 2 km was slightly larger than that of the stereo vision approach. However, due to its underlying principle, the error is expected to increase significantly in poor visibility conditions or for distant objects.

Using a general-purpose network camera, the random error in estimated distances was significant for ships 8 km away. This made it difficult to achieve sufficient accuracy for calculating stable ship speed vectors required for collision avoidance maneuvers. The error obtained through the experiment is in good agreement with error obtained by theory based on camera resolution. This suggests that employing higher-resolution cameras could reduce random errors.

As future challenges, detection of floating objects other than ships and buoys at sea in terms of obstacle detection is also important for safe navigation. In that case, the diversity of shapes makes it difficult to apply object detection through deep learning. A non-learning-based method is being developed in the ongoing work for detecting floating objects at sea, applying feature detection, feature matching, clustering, and object tracking techniques instead of deep learning [9].

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP20K04956. Parts of this manuscript are based on the author's prior conference proceedings paper [10], and are reproduced here with the publisher's authorization. We received tremendous

cooperation from the shipping company and the crew of the ship for conducting the experiments. We extend our gratitude to all those involved.

References

1) The Advanced Autonomous Waterborne Applications Initiative : Remote and Autonomous Ships - The next steps, https://www.rolls-royce.com/~/media/Files/R/Rolls-Royce/documents/customers/marine/ship-intel/aawa-whitepaper-210616.pdf (2024-07-01)
2) Mitsuru Kobayashi, et al.: Ship Detection Experiments in Uraga Channel using Visible Light / Far-Infrared Images and Deep Learning (in Japanese), Papers of National Maritime Research Institute, 20(4), 2020.
3) Shaoqing Ren, et al.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, arXiv:1506.01497
4) Ross Girshick, et al.: py-faster-rcnn, https://github.com/rbgirshick/py-faster-rcnn (2024-07-01)
5) facebookresearch, Detectron, https://github.com/facebookresearch/Detectron (2024-07-01)
6) Daniel Sonntag, et al.: Fine-tuning deep CNN models on specific MS COCO categories, arXiv:1709.01476
7) Kaiming He, et al.: Deep Residual Learning for Image Recognition, arXiv:1512.03385
8) Alex Krizhevsky, et al.: ImageNet Classification with Deep Convolutional Neural Networks, https://www.cs.toronto.edu/~kriz/imagenet_classification_with_deep_convolutional.pdf (2024-07-01)
9) Mitsuru Kobayashi, et al.: Development of Program for Detection of Unspecified Floating Obstacles from Camera Images (in Japanese), Journal of the Japan Institute of Navigation, 149, pp.48-55, 2023.
10) Mitsuru Kobayashi, et al.: Study for Binocular System for Ship Detection and Position Estimation (in Japanese), Conference Proceedings of the Japan Institute of Navigation, 7(2), 2019.