A Camera-based Deep-Learning Solution for Visual Attention Zone Recognition in Maritime Navigational Operations

Baiheng Wu, Peihua Han, Motoyasu Kanazawa,
Hans Petter Hildre, Luman Zhao, Houxiang Zhang, and Guoyuan Li[1]

***Abstract*— The visual attention of navigators is imperative to understand the logic of navigation as well as the surveillance of navigators' status and operation. Current studies are implemented with the help of wearable eye-tracker glasses; yet, the high expenditure demanded by such equipment and service and its limitations on usability have impeded related research from further development. In this letter, the authors propose a framework, which is the first attempt in the maritime domain, to provide a camera-based deep-learning (CaBDeeL) visual attention recognition solution that outperforms the intrusive eye tracker regarding its shortcomings. A wide-angle camera is configured in front of the navigator in the advanced ship-bridge simulator in a way that visual attention reflected by their facial and head movements is captured in the front view. A pair of eye-tracker glasses is used to classify the captured visual attention images, which then form the primary database. During the process of classifying camera-captured images, a convolutional neural network (CNN) is built as an automatic classifier. The CNN is applied to two scenarios, and it shows an overall 95 % accuracy.**

***Index Terms*—Maritime, maritime autonomous surface ships (MASS), human-in-the-loop (HITL), attention tracking, computer vision (CV), deep learning.**

I. Introduction

The development of maritime autonomous surface ships (MASS) has triggered interest from both industry and academia in recent years. According to different institutes and organizations, some necessary stages must be performed before ultimate fully-autonomous ships come into being [1] [2] [3]. One common ground on which these stages are built is the scale of human-in-the-loop levels of ship autonomy [4]. In another respect, human-related errors play a dominant role (75 % - 96 %) in causing marine accidents, according to the statistics [5]. Therefore, how to effectively provide practical guidance and decision supports to both the staff onboard and those at remote control centers to enhance navigational efficiency and safety will continue to be one of the most crucial issues in developing MASS [6].

Research on visual attention and eye movement is of great relevance to achieving the goal. Human visual system reflects navigators' concentration [7], fatigue status [8], maneuvering interests [9], among other factors that are closely related to sailing safety and efficiency. Conducting research on navigators' visual attention and eye movement, therefore, enables us to understand the logic and mechanism of how navigational decisions and commands are made. It also helps to learn sailing patterns and detect anomalies, which in return provides practical information for both on-board navigators and remote surveillance personnel.

Nevertheless, the feasibility and cost of equipment for recording visual attention have hindered collecting data from navigators on either real ships or simulators. Some common feedback from the participants after using the eye-tracker glasses is listed below:

- most of them claim that wearing the eye-tracker glasses places them in an unusual situation, especially for those who do not have ordinary experience of wearing glasses (for example, short/long-sight glasses);
- some of them complain that the frame edge of the glasses sometimes obstructs their sights when they squint sideways. And this is also affirmed by previous research [10].

In addition to the feedback from the navigators, the glasses may also fail to track saccadic eye movement [11]. Although the manufacturer also provides screen-mounted eye-tracker products, it is no doubt that the expenditure to fully equip a ship bridge is far beyond the budget of most stakeholders because navigators need to pay attention to multiple screens, including the Electronic Chart Display and Information System (ECDIS), the automatic radar plotting aid (APRA) (usually two ARPAs in an actual situation), the dashboard, the vast 180 ° front window among some other specific screens. These issues are considered to have an influence on either navigators' performance or the researchers' accessibility to data. Therefore, developing a low-cost solution that enables visual attention data collection without requiring wearable eye-tracker equipment is necessary.

With the rapid development of image technology and artificial intelligence (especially computer vision and pattern recognition algorithms), we have an opportunity to develop a camera-based deep-learning (CaBDeeL) solution specifically for maritime navigational operations on ship bridges. Consumer-level system/sports cameras are capable of providing high-resolution (1080P-8K) videos with a high

*The research is supported in part by the MAROFF KPN project "Digital Twins for Vessel Life Cycle Service" (Project No.: 280703), in part by the IKTPLUSS Project "Remote Control Centre for Autonomous Ship Support" (Project No.: 309323)

*NSD (Norsk Senter for Forskningsdata, Norwegian Centre for Research Data) has assessed and approved that the processing of personal data in the IKTPLUSS Project "Remote Control Centre for Autonomous Ship Support" (Project No.: 309323) will comply with data protection legislation, so long as it is carried out in accordance with what is documented in the Notification Form and attachment, 20. 12. 2021, as well as in correspondence with NSD.

*All authors are with the Department of Ocean Operations and Civil Engineering (IHB).

[1] Corresponding author: guoyuan.li@ntnu.no

frame rate (60-240 frames per second) at an affordable price. The camera can clearly capture head movement, orientation, and visual gaze direction, which all reflect navigators' visual attention. From the image recognition and classification perspective, mature algorithms with different architectures claiming the state-of-the-art performance have been updated year by year, for example the principal component analysis [12], the local binary pattern histograms [13], convolutional neural networks (CNN) [14] [15], the long short term memory [16], the transformer [17], and etc. Benefiting from both sides, we consider it is worth integrating the techniques and implementing them in the maritime navigation domain.

Similar integrated technologies have emerged in other transportation branches including automotive [18] and aviation [19] (more details are explored in Section II). In these branches, whether it is a pilot in the cockpit or a driver on a seat, the mutual point is that they usually have fixed sitting positions. Nevertheless, the situation may vary depending on different types of ships in the maritime domain. Considering our attempt in this letter is the pioneer in the maritime domain, we carry out this project on an advanced ship bridge system as shown in Fig. 1 and the top-right embedded chart in Fig. 2. This type of ship bridge is often instrumented on the high-speed cruising catamarans, commonly used to execute passenger commuter routes for public transportation. In this type of ship bridge, the navigator operates and maneuvers on a fixed seat which is similar to the operational environments of pilots and drivers.

In general, the authors are committed to providing a low-cost and non-intrusive visual attention recognition solution by developing an integrated system that consists of a consumer sports camera and a CNN deep-learning algorithm. This framework enables to trace navigators' visual attention at a high frequency (up to 120 Hz) to specific visual attention zones (VAZs). As the data can be exported and transmitted in real-time, they have the potential to provide online decision support (such as fatigue monitor and alarm, and operation prediction). The data also can be received by a remote control center to evaluate onboard situations.

The solution proposed by the authors mainly contains two parts:

- database establishment: videos are recorded by the wall-mounted consumer camera to capture the head and eye movement of navigators; then, we split videos into single frames and sort the images into different classes in terms of the visual attention to the VAZs;
- deep-learning model training: a CNN model is architectured by selecting a proper filter, kernel, activation functions, and layer depth; the CNN model is trained with the database established in the first step by setting appropriate epochs and batch sizes to achieve a satisfying classification ability.

While the model is satisfactorily trained, it is applied to two scenarios for a test to verify the classifier's performance on visual attention recognition.

The letter is organized as follows: Section II explores related research items from other transportation branches and progress on specific computer-vision-based visual attention recognition; Section III depicts the workflow of this letter, the algorithm architecture, and the experimental setup in details; Section IV demonstrates the results of training process and model performance in two testing scenarios; At last, a conclusion is given as a summary of the contribution, limitation, and future works.

II. Literature Review

This section explores the relevant research from two aspects: visual-attention-recognition-related research, with a primary focus on computer-vision-based studies, and the current progress in the transportation industry.

A. Algorithm-based visual attention recognition

Computer vision has been extensively studied using cameras to realize visual attention recognition. The approaches to achieve it can be sorted into mainly two classes: eye-gaze tracking [20] and head orientation detection [21], and the two ways are often combined in recent mainstream. Lee *et al.* carried out an interesting study by using eye gaze as a remote controller to a TV, and 2D geometric transform was used to achieve eye gaze mapping in the process [22]; Cheung *et al.* developed a low-cost solution by applying a simulated 3D head model based on a web camera to achieve eye gaze tracking [23]; Chi *et al.* designed a global calibration method on a multi-camera structure, which solved the problem when calibration reference was not within camera's range [24]; Gudi *et al.* applied CNN algorithm to evaluate different types of inputs, including the whole face, two eyes, and single eye based on a webcam as well [25]; Dai *et al.* integrated binocular features and spatial attention mechanism into the CNN algorithm [26]. The fore-mentioned algorithms and applications indicate that current computer vision techniques can perform precise and robust visual attention recognition with low-cost solutions.

B. Applications in transportation industry

As autonomous automotive projects trigger the hot spot earlier than MASS, camera-based torso and eye-tracking frameworks and algorithms have been studied extensively in the past decades. Smith *et al.* developed an automatic face tracing system with functions to detect eye blinking and closing for visual attention assessment with one camera [27]; Jiménez *et al.* used a stereo camera to realize face pose and gaze estimation to evaluate drivers' distractions caused by in-vehicle information systems [28]; Vora *et al.* compared four different CNN architectures in driver gaze zone estimation, among which SqueezeNet recorded the highest accuracy [29]. Yang *et al.* used a dual-camera system to capture drivers' behavior in order to detect non-driving activities [30]; Li *et al.* performed a field investigation by using two cameras to record the environmental conditions and drivers' scanning patterns separately to synthetically analyze the behavioral difference between signalized and unsignalized intersections [31]; Wu *et al.* utilized infrared cameras to estimate gaze points to anticipate drivers' intention in semi-autonomous vehicles [32]. Rangesh *et al.* performed generative adversarial networks to remove the effects brought by eyeglass [33].

Similar research in the aviation department is not as prosperous as in the automotive one. Glahot not only summarized the relationship between eye movements and aviators' cognitive status but also reviewed different types of eye-tracking system architectures used in the cockpit [34]; Ellis used a three-camera system to estimate aviators' workload based on eye-tracking metrics [35]; Pavelková *et al.* developed the OptiTrack system consisted of six cameras to enhance the communication flow and in-cockpit operational safety [19]; Murthy *et al.* revealed the limitation of using wearable eye-tracker in the cockpit and developed a screen-mounted camera system to solve issues happened in critical situations, such as when against the strong illumination [36].

The existing research items in other transportation departments provide a paradigm for reference when developing similar systems and applying related technologies in the maritime domain [37]. Up to date, no research item has been implemented with non-intrusive solutions. Commercial-off-the-shelf apparatus are utilized to perform various onboard investigations over multiple operations such as crane lifting [38], and navigation [39], during which wearable eye trackers are often used. Eye-tracking data are also used for mental load assessment [40], situational awareness analysis in terms of sailing safety promotion [41], improvement of navigational apprentice training [42] [43] in the maritime domain. As the eye tracking and visual attention are significant in related research and there is not yet a low-cost, non-intrusive flexible solution, the authors are triggered to develop a feasible, low-cost, and non-intrusive solution specifically for the maritime domain so that marine experiments can be designed and carried out more independently.

III. Methodology & Setup

In this section, we introduce the whole workflow (depicted in Fig. 2) of the methodology by dividing it into two parts: training flow which includes the database formation and the details of the designed CNN-based deep-learning algorithm; and the testing flow which includes how the trained model is applied in trial sailings and how the performance is evaluated.

A. Training flow

Training flow forms the base of the developed solution (the upper in Fig. 2). The first step is to invite navigators to perform random trials in the ship-bridge simulator, from which the data build up the primary database. A wall-mounted sports camera is used to capture navigators' head and eye movements from their front view (as shown in Fig. 2). In this step, the eye-tracker glasses are used only to sort the collected images into different classes in the database. According to the configuration of the ship-bridge simulator, as shown in the top-right embedded chart in Fig. 2, there are three VAZs including:

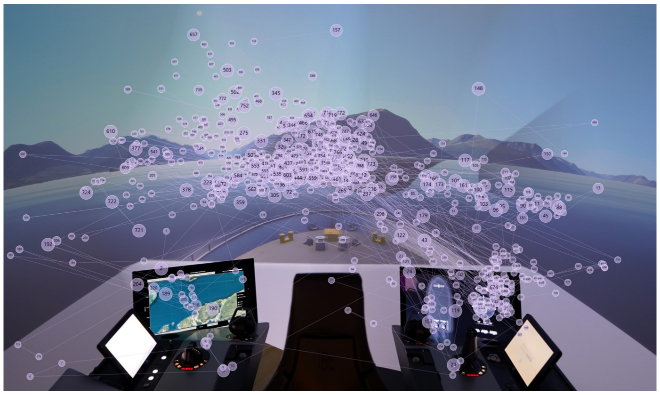


Fig. 1. Binocular movement trajectory and fixation points.

- the projected-scene wall (VAZ-I), which displays the simulated scenario's outside environment;
- the ECDIS screen (VAZ-II), which shows the traffic situation on the map and also has the function of providing decision support;
- the dashboard screen (VAZ-III), which contains machinery information.

The database continues to expand as trials accumulate. The second step is then to train the deep-learning model with such a database. After the model is well-trained, it is able to classify navigators' visual attention into corresponding VAZs correctly.

1) Database formation: Fig. 1 shows the eye movement trajectory captured by the glasses when collecting data from a trial sailing that lasted approximately 10 minutes. It shows that the navigator focused on all three VAZs, but the time was not evenly distributed.

We collected around 40 minutes of video to establish the primary database, and the video was recorded with a resolution of 1080P and at a rate of 60 frames per second. Applying the eye-tracker glasses' data as reference and after necessary data pruning, the distribution of the collected data in each class is 50.0 %, 16.7 %, and 33.3 % for VAZ-I, II, and III, respectively.

2) Algorithm: In this study, we developed a CNN deep-learning model whose structure is shown in Fig. 3. In general, it contains three convolutional layers, three max-pooling layers, one flatten layer, and two fully-connected layers.

Each convolutional layer in this network includes two operations:

$$\begin{aligned}\mathbf{s} &= Conv(\mathbf{x}) \\ \mathbf{s} &= ReLU(\mathbf{s})\end{aligned} \tag{1}$$

where $\mathbf{x}$ is the input; *Conv* denotes the convolutional operation and its weight is to be learned by training; *ReLU* is selected as the activation function to solve this image classification problem [44]. In the three convolutional layers, their corresponding kernels are selected with sizes of 5×5, 3×3, and 3×3.

Max pooling layers (sub-sampling layers) are performed after each convolutional layer to compress the image data,

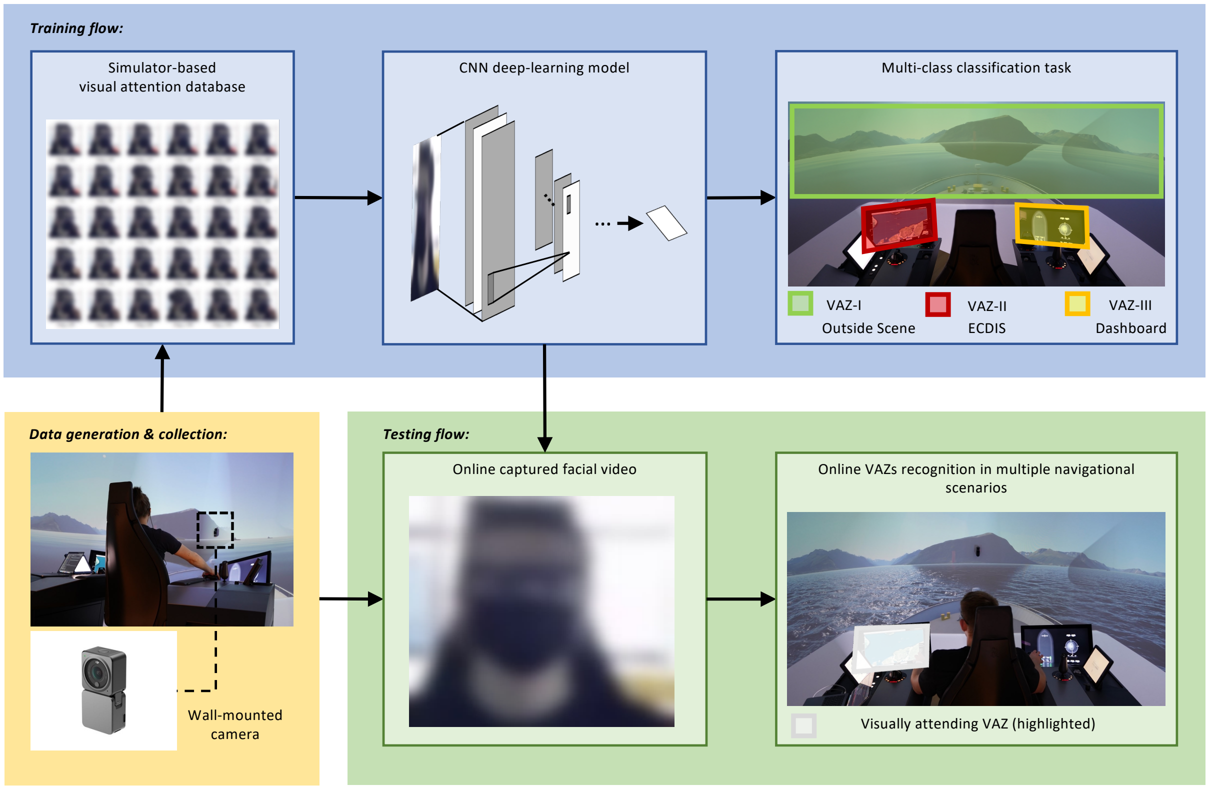

Training flow:
Simulator-based
visual attention database
CNN deep-learning model
...
Multi-class classification task
VAZ-I
Outside Scene
VAZ-II
ECDIS
VAZ-III
Dashboard
Data generation & collection:
Wall-mounted
camera
Testing flow:
Online captured facial video
Online VAZs recognition in multiple navigational
scenarios
Visually attending VAZ (highlighted)

Fig. 2. Workflow of the CaBDeeL solution (images containing facial identity have been blurred for demonstration in the figure according to the General Data Protection Regulation (GDPR)).

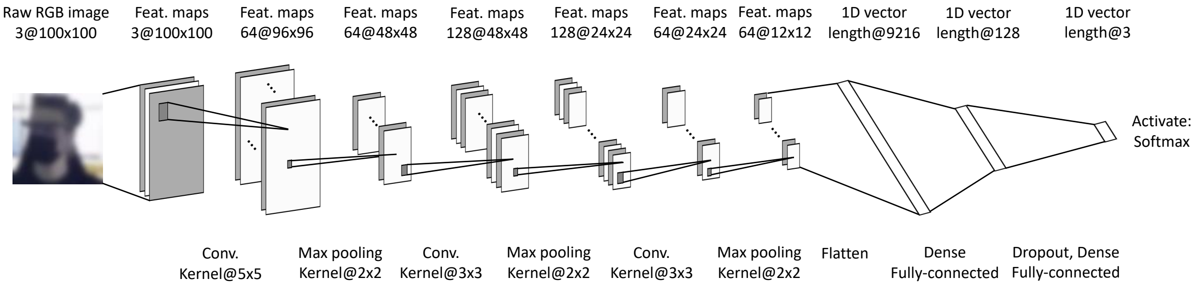


Fig. 3. Structure of CNN-based deep-learning algorithm.

reduce the number of the weights and avoid over-fitting. In this network, all max-pooling layers have the same kernel size at 2×2. After the convolutional and max-pooling layers, the multi-channel maps are flattened to a 1D vector. Then two fully-connected layers follow up to weigh and rectify features. The dropout operation in the last fully-connected layer also prevents over-fitting. Finally, the features map is dense to a 1D vector with a length of 3; and as we are about to solve a multi-class single-label problem, *softmax* is chosen as the activation function to produce the probability of each class.

B. *Testing flow*

Testing flow is downstream when the model is trained and ready to use. Navigators are invited to maneuver on the ship-bridge system again to generate videos for testing. The videos are exported and decomposed into frames, and then the trained deep-learning model is applied to the frames to recognize the VAZ of the navigator in the image. When collecting data in this flow, navigators are also asked to wear the eye-tracker glasses, and it is only to verify the results and performance of the deep-learning model.

IV. Results & Discusion

This section lists the results from two aspects: the training process and verification of the trained model in some trial sailings. We also illustrate the comparison between CaBDeeL VAZs recognition method and the eye-tracker glasses to demonstrate the significance of maritime application.

A. *Model training*

1) Training process: The algorithm is written and realized in the Keras framework. When training the model, the

TABLE I

Comparison on sampling rate/accuracy

No.	Duration (mm:ss)	Sampling rate/accuracy
Eye-tracker glasses		
Trial 1	16:44	93 %
Trial 2	5:47	90 %
Trial 3	12:02	84 %
Trial 4	5:16	83 %
CaBDeeL		
Overall	-	95 %

database is split 80 % - 20 % into train and validation sets. The learning rate is set as 0.01, and the Stochastic gradient descent (SGD) optimizer is selected [45]. The batch size is set to 128. The model is trained for 500 epochs.

Fig. 4 shows the loss and accuracy changes along the 500 epochs. In the first 20 epochs, the accuracy increases rapidly while the loss decreases at the same pace. From the 50th epoch, the change rates of accuracy and loss become small and steady. At the 500th epoch, the accuracy has stabilized over 95 %, and the loss is lower than 0.13. The trend in Fig. 4 implies that more epochs may continue to improve the performance, but to keep a balance between computational efficiency and accuracy, we stop at the 500th and take it as the trained model for further verification.

2) Training result: Table I compares the sampling rate of the eye-tracker glasses and the accuracy of CaBDeeL. The glasses can precisely locate the gaze zone when eye movement is sampled effectively. However, the sampling process can be unstable, especially when navigators squint over the edge of the glasses frame. The glasses cannot properly predict the gaze estimation if the eye movement fails to be sampled. Compared to the glasses, CaBDeeL shows more robustness as long as the camera functions normally. Accuracy over 95 % has already outperformed the eye-tracker glasses in this VAZs recognition task on the ship-bridge simulator.

B. Test in two trials

1) Scenario setup: Two different scenarios are designed (as shown in Fig. 5) to test the performance of the trained model, including in heavy traffic conditions where collision avoidance operations are needed as well as in light traffic conditions where the navigator freely maneuvers the ship to cruise on the sea. In Fig. 5(a), according to the Convention on the International Regulations for Preventing Collisions at Sea (COLREGs), the own ship (OS) yields to give ways to target ships (TSs) coming from its starboard side (colored in red), and shall not give way to the TS coming from its port-board side (colored in green). In this case, the operator has to ensure proper sailing and maneuvering to avoid potential risks. In this process, we assume the navigator has to meticulously monitor the situation by inspecting the visual sight outlook, information on ECDIS, and the maneuvering commands on the dashboard. Although the scenario in Fig. 5(b) is relatively simple as there is little traffic, we expect to observe the behavioral difference of the navigator reflected by the visual attention data from the two scenarios.

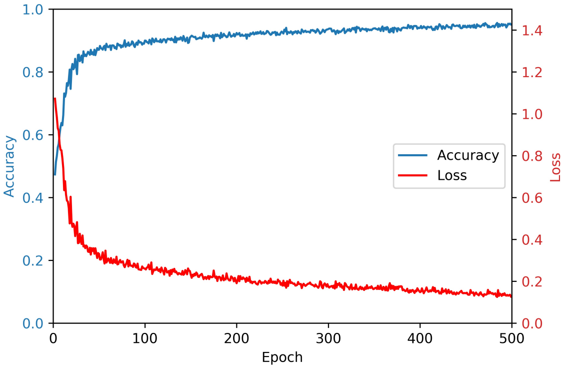


Fig. 4. Performance development along the training process.

TABLE II

Sampling rate/accuracy comparison in the two scenarios

Scenario	CaBDeeL-250	CaBDeeL-500	Eye-tracker
(a)	89.5 %	93.5 %	69.0 %
(b)	92.1 %	95.9 %	95.0 %

2) Classification accuracy: The classification accuracy is plotted in Fig. 6. In Fig. 6(a), the prediction accuracy has overall satisfactory performance. It is the most accurate when predicting the VAZ-III while the least accurate when predicting the VAZ-I. The fact that VAZ-I is the scene screen with broad coverage within eyesight may explain this. In addition, it is relatively in the middle among the other two VAZs, so it is likely to have more failures in prediction on this VAZ. VAZ-II and VAZ-III are distinguishable as one is to the navigator's left, and another is to the right, so these two VAZs are never confused (both wrong prediction rates are 0). In Fig. 6(b), it is interesting to find that VAZ-II is never visually visited by the navigator, which implies that the navigator prefers to use pure visual sight to observe the environment on VAZ-I when the traffic situation is less demanding. In this scenario, the prediction accuracy is even higher when training the model (95 %). A reason to explain it is that since one of the VAZs is never paid attention to by the navigator, it reduces the probability of incorrectly sorting the frame into that class.

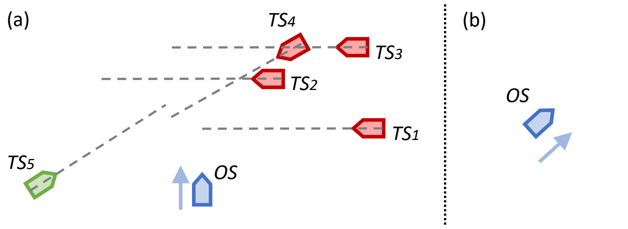


Fig. 5. Two scenarios: (a) heavy traffic with collision avoidance demand; (b) cruise in light traffic.

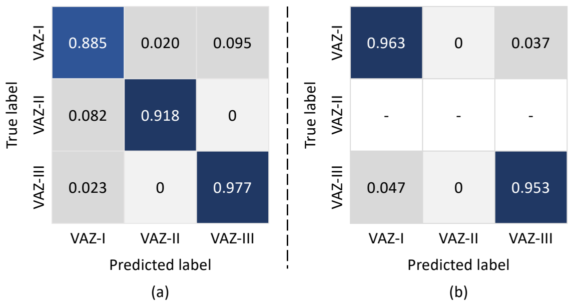


Fig. 6. Confusion matrices in the two scenarios of CaBDeeL (500 epochs).

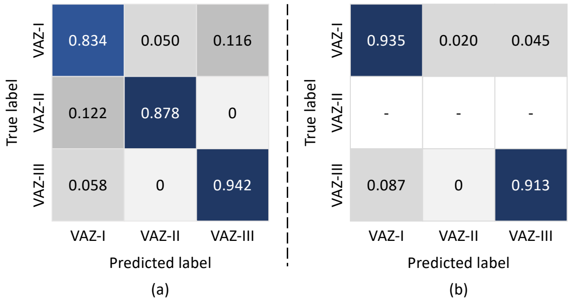


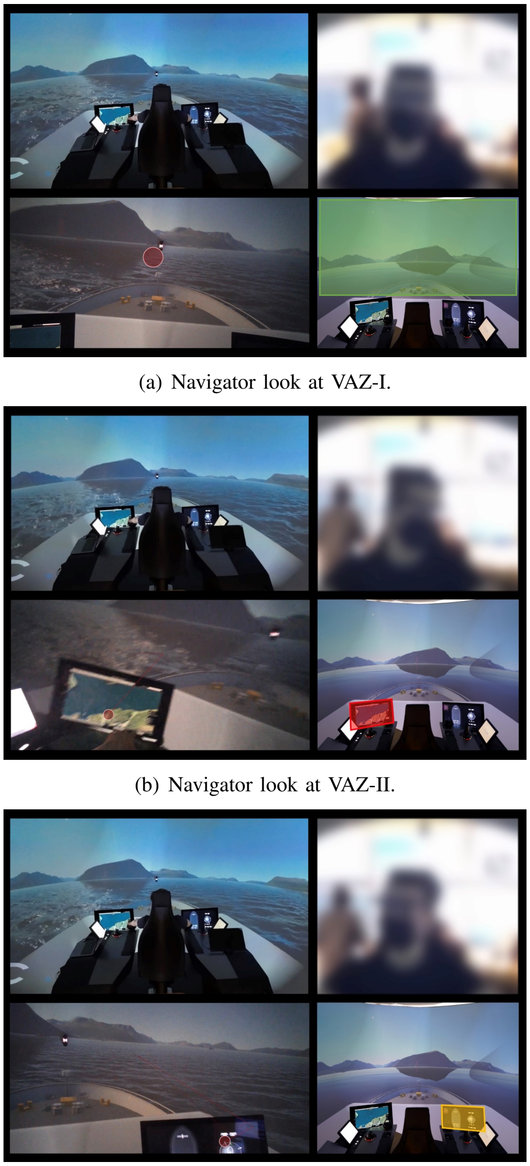
Fig. 7. Confusion matrices in the two scenarios of CaBDeeL (250 epochs).

In general, CaBDeeL (500 epochs) scores overall accuracy at 93.5 % and 95.9 % as in Table II for the Scenario (a) and b respectively. While the sampling rate of the eye-tracker glasses meets some critical issues which result in a low rate in Scenario (a), it can be caused by an improper way of wearing the glasses, failure in eyes location calibration, swift eye sweeping, and extreme glare. In contrast, the rates are close to each other in Scenario (b). Table II also shows the accuracies of CaBDeeL trained at the 250th epoch, and compared with the accuracies at the 500th epoch in both scenarios, it corroborates that the performance of CaBDeeL model is promoted uniformly as the amount of epochs increases. The detailed accuracies of each class can be found in Fig. 7. It is noticed that when the model is trained at the 250th epoch, it may still falsely sort VAZ-I as VAZ-II, but this phenomenon disappears when CaBDeeL is trained after 500 epochs.

3) Visualization & Comparison: Fig. 8 shows three subfigures when the navigator looks at different VAZs. The bottom-left of each subfigure in Fig. 8 shows that eye-tracker glasses can locate the gaze to an exact point, although they fail to track eye movement in some cases. While CaBDeeL can recognize different VAZs, which means an approximate area of visual attention. Fig. 8 proves the accuracy of the CaBDeeL in the trial sailings in the designed scenarios.

C. Applicable function

Like the eye-tracker glasses, CaBDeeL is capable of realizing pragmatic visual attention analysis by providing commonly used metrics, such as transition frequency, duration of fixation, and the total duration time of fixation. Here we illustrate how it performs analysis on the two testing sailings in Section IV-B.



(a) Navigator look at VAZ-I.

(b) Navigator look at VAZ-II.

(c) Navigator look at VAZ-III.

Fig. 8. Matts plot of (1) top-left: back-view (only used for demonstration, not relevant to CaBDeeL); (2) top-right: front-view (input image to CaBDeeL; the facial image shown here is blurred according to GDPR); (3) bottom-left: eye-tracker glasses marked video; (4) bottom-right: CaBDeeL recognized VAZ is highlighted.

1) Transition frequency: Fig. 9 depicts the transition frequency between every two VAZs. In Scenario (a), the navigator transits their visual attention between VAZ-I and III at the highest frequency, although it is hard to distinguish direct transits between VAZ-II and III. To handle collision risks, the navigator needs to gather information from the ECDIS, for example, the distance/time to the closet point of approach, speeds of TSs, route prediction, and other traffic information, and they have to assess the situation from the ECDIS and with direct observation of the situation and traffic. While In Scenario (b), VAZ-II never captures any

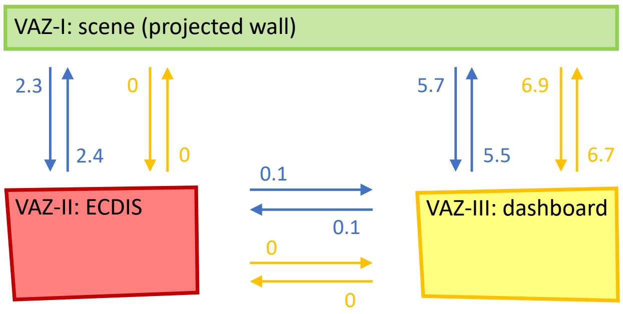


Fig. 9. Transition frequency between every two VAZs in the two scenarios (transitions per minute). The blue lines depicts for Scenario (a) and the yellow lines depicts for Scenario (b).

TABLE III
DURATION OF FIXATION FEATURES

No.	Scale (s)	Median (s)	Mean (s)
Scenario (a): heavy traffic			
VAZ-I	[0.3, 13.8]	1.5	2.0
VAZ-II	[0.2, 10.3]	1.8	2.6
VAZ-III	[0.1, 8.5]	3.0	4.0
Scenario (b): light traffic			
VAZ-I	[0.4, 24.0]	4.0	5.6
VAZ-II	-	-	-
VAZ-III	[0.6, 8.4]	2.2	2.8

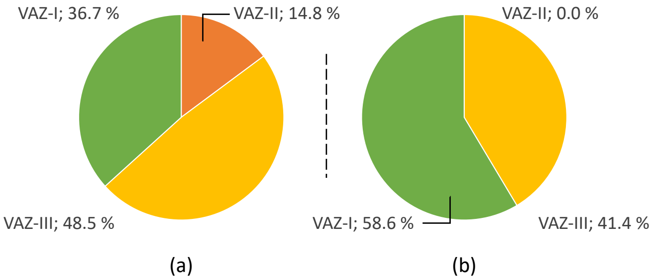


Fig. 10. Total duration proportion of each VAZ in the two scenarios.

attention of the navigator, which implies that the navigator tends to fully rely on his own observation when sailing in light traffic. Moreover, the total transition frequencies in Scenario (a) and (b) are 15.9 and 13.6 transitions per minute. This difference also demonstrates that attention activeness is lower when sailing under ordinary circumstances than in complicated ones.

2) Duration of fixation: Table III lists the features in the duration of fixation. In Scenario (a), VAZ-III has the highest mean and median, i.e., more information on the dashboard to be read at each time. When the navigator receives such information, their attention shifts away only shortly as it is dangerous to leave the situation unattended. Both scenarios yield interesting results when comparing their features. In Scenario (b), the maximum on VAZ-I reaches 24.0 seconds which is much higher than any maximums in Scenarios (a). Light traffic requires fewer operations and consequently demands less attention on the dashboard that provides maneuvering and machinery information. In addition, the lower threshold of the scale in Scenario (b) is much higher, which also proves that the navigator is less visually active.

3) Total duration proportion: Fig. 10 plots the total duration time spent on each VAZ in the two scenarios. In Scenario (a), according to Fig. 10(a), VAZ-III dominates the navigator's attention. From the proportion distribution, it can be inferred that navigators are more sensitive to the machinery commands (such as propulsion rate, speed, and rudder angle) to achieve fine maneuvering when sailing in a congested water channel. Meanwhile, the navigator also needs to obtain information and receive decision support from the ECDIS system to select a collision avoidance scheme. Different from the heavy traffic situation, Fig. 10(b) reveals that the navigator places their visual sight over the window to observe the environment and become aware of the situation. This is in accordance with to the discovery in Section IV-B.2.

D. *Summary*

In summary, CaBDeeL proves its capability in terms of prediction accuracy and robustness when compared with the traditional solution in Section IV-B. Concerning usability, CaBDeeL can generate and export ordinary visual attention metrics in a case study and also in statistics for analysis (demonstrated in Section IV-C). We convey that CaBDeeL meets the research goal in the scope of visual attention in maritime navigation and outperforms the traditional solution in specific features such as robustness in continuous working (eye-tracker host usually gets overheated) and accuracy (against the sampling rate of eye-tracker glasses).

V. CONCLUSION

The authors attempt to develop a camera-based deep-learning (CaBDeeL) solution that solves the VAZs recognition problem in this letter. The developed framework attains excellent results in achieving the goal. When CaBDeeL is applied to trial sailings, it scores overall accuracy beyond 95 %. This solution outperforms the traditional visual attention tracking method given its high robustness, low cost, and non-intrusive feasibility. However, CaBDeeL, at this stage, is only able to recognize navigators' visual attention around an approximate zone, which is not as precise as the traditional solution. Limitations of this framework will be further discussed in another study that is coming soon. To sum up, as the first attempt in the maritime domain to develop a single-camera-based visual attention tracking solution, the proposed framework has the potential to lower the threshold for maritime researchers to dive into studies via visual attention of maritime operators and navigators.

ACKNOWLEDGMENT

Acknowledgment is extended to the engineers at Offshore Simulator Centre AS for providing technical support, the experiment participants for providing their expert skills and knowledge.

References

[1] "Regulatory scoping exercise on maritime autonomous surface ships," in *Maritime Safety Committee, 100th session*. International Maritime Organization, 2018.

[2] "*DNVGL-CG-0264* class guideline: Autonomous and remotely operated ships," *DNVGL*, 2018.

[3] "*LR* code for unmanned marine systems," in *ShipRight Design and Construction, Additional Design Procedures*. Lloyd's Register, 2017.

[4] "Review of maritime transport 2021." UNCTD (United Nations Conference on Trade and Development), United Nations Publications, New York, USA, 2021.

[5] A. Galieriková, "The human factor and maritime safety," *Transportation research procedia*, vol. 40, pp. 1319–1326, 2019.

[6] B. Wu, G. Li, L. Zhao, H.-I. J. Aandahl, H. P. Hildre, and H. Zhang, "Navigating patterns analysis for onboard guidance support in crossing collision-avoidance operations," *IEEE Intelligent Transportation Systems Magazine*, 2021. [Online]. Available: doi.org/10.1109/MITS.2021.3108473

[7] O. Arslan, O. Atik, and S. Kahraman, "Eye tracking in usability of electronic chart display and information system," *The Journal of Navigation*, vol. 74, no. 3, pp. 594–604, 2021.

[8] M. Sant'Ana, G. Li, and H. Zhang, "A decentralized sensor fusion approach to human fatigue monitoring in maritime operations," in *2019 IEEE 15th International Conference on Control and Automation (ICCA)*. IEEE, 2019, pp. 1569–1574.

[9] O. S. Hareide, "The use of eye tracking technology in maritime high-speed craft navigation," *Doktoravhandlinger ved NTNU*, 2019.

[10] O. S. Hareide and R. Ostnes, "Maritime usability study by analysing eye tracking data," *The Journal of Navigation*, vol. 70, no. 5, pp. 927–943, 2017.

[11] K. Ooms, L. Dupont, L. Lapon, and S. Popelka, "Accuracy and precision of fixation locations recorded with the low-cost eye tribe tracker in different experimental setups," *Journal of eye movement research*, vol. 8, no. 1, 2015.

[12] I. Bacivarov, M. Ionita, and P. Corcoran, "Statistical models of appearance for eye tracking and eye-blink detection and measurement," *IEEE transactions on consumer electronics*, vol. 54, no. 3, pp. 1312–1320, 2008.

[13] X. Feng, M. Pietikäinen, and A. Hadid, "Facial expression recognition based on local binary patterns," *Pattern Recognition and Image Analysis*, vol. 17, no. 4, pp. 592–598, 2007.

[14] M. Selim, A. Firintepe, A. Pagani, and D. Stricker, "Autopose: Large-scale automotive driver head pose and gaze dataset with deep head orientation baseline." in *VISIGRAPP (4: VISAPP)*, 2020, pp. 599–606.

[15] B. Ahn, J. Park, and I. S. Kweon, "Real-time head orientation from a monocular camera using deep neural network," in *Asian conference on computer vision*. Springer, 2014, pp. 82–96.

[16] P. L. Mazzeo, D. D'Amico, P. Spagnolo, and C. Distante, "Deep learning based eye gaze estimation and prediction," in *2021 6th International Conference on Smart and Sustainable Technologies (SpliTech)*. IEEE, 2021, pp. 1–6.

[17] N. Wang, W. Zhou, J. Wang, and H. Li, "Transformer meets tracker: Exploiting temporal context for robust visual tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1571–1580.

[18] J. Wang, W. Chai, A. Venkatachalapathy, K. L. Tan, A. Haghighat, S. Velipasalar, Y. Adu-Gyamfi, and A. Sharma, "A survey on driver behavior analysis from in-vehicle cameras," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[19] A. Pavelková, A. Herout, and K. Behún, "Usability of pilot's gaze in aeronautic cockpit for safer aircraft," in *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. IEEE, 2015, pp. 1545–1550.

[20] H. Chennamma and X. Yuan, "A survey on eye-gaze tracking techniques," *arXiv preprint arXiv:1312.6410*, 2013.

[21] A. Al-Rahayfeh and M. Faezipour, "Eye tracking and head movement detection: A state-of-art survey," *IEEE journal of translational engineering in health and medicine*, vol. 1, pp. 2 100 212–2 100 212, 2013.

[22] H. C. Lee, D. T. Luong, C. W. Cho, E. C. Lee, and K. R. Park, "Gaze tracking system at a distance for controlling iptv," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 4, pp. 2577–2583, 2010.

[23] Y.-m. Cheung and Q. Peng, "Eye gaze tracking with a web camera in a desktop environment," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 4, pp. 419–430, 2015.

[24] J. Chi, Z. Yang, G. Zhang, T. Liu, and Z. Wang, "A novel multi-camera global calibration method for gaze tracking system," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 5, pp. 2093–2104, 2019.

[25] A. Gudi, X. Li, and J. v. Gemert, "Efficiency in real-time webcam gaze tracking," in *European Conference on Computer Vision*. Springer, 2020, pp. 529–543.

[26] L. Dai, J. Liu, and Z. Ju, "Binocular feature fusion and spatial attention mechanism based gaze tracking," *IEEE Transactions on Human-Machine Systems*, 2022.

[27] P. Smith, M. Shah, and N. da Vitoria Lobo, "Determining driver visual attention with one camera," *IEEE transactions on intelligent transportation systems*, vol. 4, no. 4, pp. 205–218, 2003.

[28] P. Jiménez, L. M. Bergasa, J. Nuevo, N. Hernández, and I. G. Daza, "Gaze fixation system for the evaluation of driver distractions induced by ivis," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 3, pp. 1167–1178, 2012.

[29] S. Vora, A. Rangesh, and M. M. Trivedi, "Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis," *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 3, pp. 254–265, 2018.

[30] L. Yang, K. Dong, A. J. Dmitruk, J. Brighton, and Y. Zhao, "A dual-cameras-based driver gaze mapping system with an application on non-driving activities monitoring," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 10, pp. 4318–4327, 2019.

[31] G. Li, Y. Wang, F. Zhu, X. Sui, N. Wang, X. Qu, and P. Green, "Drivers' visual scanning behavior at signalized and unsignalized intersections: A naturalistic driving study in china," *Journal of safety research*, vol. 71, pp. 219–229, 2019.

[32] M. Wu, T. Louw, M. Lahijanian, W. Ruan, X. Huang, N. Merat, and M. Kwiatkowska, "Gaze-based intention anticipation over driving manoeuvres in semi-autonomous vehicles," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 6210–6216.

[33] A. Rangesh, B. Zhang, and M. M. Trivedi, "Driver gaze estimation in the real world: Overcoming the eyeglass challenge," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020, pp. 1054–1059.

[34] M. G. Glaholt, "Eye tracking in the cockpit: a review of the relationships between eye movements and the aviators cognitive state," 2014.

[35] K. K. E. Ellis, *Eye tracking metrics for workload estimation in flight deck operations*. The University of Iowa, 2009.

[36] L. Murthy, A. Mukhopadhyay, S. Arjun, V. Yelleti, P. Thomas, D. B. Mohan, and P. Biswas, "Eye-gaze-controlled hmds and mfd for military aircraft," *Journal of Aviation Technology and Engineering*, vol. 10, no. 2, p. 34, 2021.

[37] R. Mao, G. Li, H. P. Hildre, and H. Zhang, "A survey of eye tracking in automobile and aviation studies: Implications for eye-tracking studies in marine operations," *IEEE Transactions on Human-Machine Systems*, vol. 51, no. 2, pp. 87–98, 2021.

[38] G. Li, R. Mao, H. P. Hildre, and H. Zhang, "Visual attention assessment for expert-in-the-loop training in a maritime operation simulator," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 522–531, 2019.

[39] O. S. Hareide and R. Ostnes, "Scan pattern for the maritime navigator," *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation*, vol. 11, no. 1, 2017.

[40] F. Di Nocera, S. Mastrangelo, S. P. Colonna, A. Steinhage, M. Baldauf, and A. Kataria, "Mental workload assessment using eye-tracking glasses in a simulated maritime scenario," *Proceedings of the human factors and ergonomics society europe*, pp. 235–248, 2016.

[41] O. Atik, "Eye tracking for assessment of situational awareness in bridge resource management training," *Journal of Eye Movement Research*, vol. 12, no. 3, 2019.

[42] J. L. Rosch and J. J. Vogel-Walcutt, "A review of eye-tracking applications as tools for training," *Cognition, technology & work*, vol. 15, no. 3, pp. 313–327, 2013.

[43] F. Sanfilippo, "A multi-sensor fusion framework for improving situational awareness in demanding maritime training," *Reliability Engineering & System Safety*, vol. 161, pp. 12–24, 2017.

[44] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Icml*, 2010.

[45] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.