Comparing Spectral Bands for Object Detection at Sea using Convolutional Neural Networks
Paper

Stets, Jonathan Dyssel; Schöller, Frederik Emil Thorsson; Plenge-Feidenhans'l, Martin K.; Andersen, Rasmus Hjorth; Hansen, Søren; Blanke, Mogens

Published in:
Journal of Physics: Conference Series (Online)

Link to article, DOI:
10.1088/1742-6596/1357/1/012036

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Stets, J. D., Schöller, F. E. T., Plenge-Feidenhans'l, M. K., Andersen, R. H., Hansen, S., & Blanke, M. (2019). Comparing Spectral Bands for Object Detection at Sea using Convolutional Neural Networks: Paper. Journal of Physics: Conference Series (Online), 1357(1), [012036]. https://doi.org/10.1088/1742-6596/1357/1/012036

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Comparing Spectral Bands for Object Detection at Sea using Convolutional Neural Networks

To cite this article: Jonathan D. Stets et al 2019 J. Phys.: Conf. Ser. 1357 012036

View the article online for updates and enhancements.
Comparing Spectral Bands for Object Detection at Sea using Convolutional Neural Networks

Jonathan D. Stets, Frederik E. T. Schöller, Martin K. Plenge-Feidenhans’l, Rasmus H. Andersen, Søren Hansen and Mogens Blanke
Technical University of Denmark, DK 2800 Kongens Lyngby, Denmark
E-mail: stet@dtu.dk, \{fets, mkpl, rhjan, sh, mb\}@elektro.dtu.dk

Abstract. This study compares spectral bands for object detection at sea using a convolutional neural network (CNN). Specifically, images in three spectral bands are targeted: long wavelength infrared (LWIR), near-infrared (NIR) and visible range. Using a calibrated camera setup, a large set of images for each of the spectral bands are captured with the same field of view. The image sets are then used to train and validate a CNN for object detection to evaluate the performance in the different bands. Prediction performance is employed as a quality assessment and is put in a navigational perspective. The result is a quantitative evaluation that reveals the strengths and weaknesses of using different spectral bands individually or in combination for autonomous navigation at sea. The analysis covers two object classes of particular importance for safe navigation.

1. Introduction
Autonomous marine systems require the ability to accurately detect and classify objects at sea to navigate safely. A vision system can, in combination with other sensors, be an effective way to detect objects on the sea surface. However, changing weather conditions and illumination changes throughout the day cause great variations in visibility. This means that a vision system cannot solely rely on the human visible spectrum.

A recent study comparing electronic and human outlook \[1\] demonstrated how a learning-based approach identified objects on the water, from images in the visible range, using...

Figure 1: Images of Ærøskøbing harbour in the South Funen Archipelago in three spectral bands. These images have not been transformed, so their view is similar but not identical.
Table 1: Acronyms and Abbreviations

| acronym | explanation                      | acronym | explanation                      |
|---------|----------------------------------|---------|----------------------------------|
| Adam    | Adaptive Moment Estimation       | HFOV    | Horizontal Field Of View         |
| AIS     | Automatic Identification System  | IoU     | Intersection over Union          |
| CNN     | Convolutional Neural Network      | LWIR    | Long Wave Infrared 8 – 14µm     |
| CPA     | Closest point of approach         | NIR     | Near-Infrared 0.8 – 1.0µm       |
| ECDIS   | Electronic Chart Display          | nmi     | Nautical miles                   |
| FN      | False Negatives                   | TCPA    | Time to CPA                      |
| FP      | False Positives                   | TN      | True Negatives                   |
| FPN     | Feature Pyramid Network           | TP      | True Positives                   |

This study investigates how different spectral bands can contribute to the detection of objects at sea. Specifically, three types of images are targeted: long wavelength infrared (LWIR) images, near-infrared (NIR) images and colour images in the human visible spectrum. The vision system consists of three cameras, one for each spectral band, mounted close together and pointing in the same direction. Figure 1 shows images of the same scene from each spectral band. To get a base for comparison, the images from the three cameras are transformed and cropped so they have a similar field of view. Data were recorded from two different ferries operating in the South Funen Archipelago with images taken from morning to after sunset.

A RetinaNet [6] convolutional neural network (CNN) is used to detect and classify objects. This architecture was chosen as it has proved to work well for object detection and classification in a marine environment [13]. A set of images for training and evaluation for each spectral band has been annotated into two object categories *ship* and *buoy*, which are of particular importance for safe navigation in confined waters. The intention is to use one set of cameras for coarse object detection while object specific details are to be found subsequently with a different classifier and potentially with a dedicated camera system with a narrower field of view. The classification of object details is beyond the scope of this paper. This study hence aims to compare detection using a CNN for three different spectral bands. We wish to investigate how detectable the selected classes are in different illumination conditions and distances from the
camera. The performance in a navigational setting is assessed using Time to Closest Point of Approach (TCPA).

2. Requirements
In a maritime navigational setting, two of the most common object classes to recognise is buoys, which indicate limits for safe navigation, and other vessels or objects in the vicinity of own ship, which could influence navigation decisions for own vessel. Buoys provide information about areas with shallow water, dangers or wreckages. Buoys are marked with colours and geometry markers that convey their meaning. These are visible during daylight conditions. At night, selected buoys are marked with lights that blink in specific patterns to relay information. Ships’ conditions are indicated by flags or geometric markers in the mast with shapes that show intent or condition. These are visible during daytime. At night, these markers are replaced by lights arranged in special patterns on the mast. If a camera-based system is used to supplement or temporally replace human outlook, it needs to be able to determine the specific type of ships and buoys and interpret navigation information made by markers on these objects.

To minimize the risk of collision with other vessels, master mariners, see [8], estimate measures of commonly accepted Closest Point of Approach (CPA) for the safe passage of two vessels in open waters and the TCPA required for a vessel to react if course or speed alterations should be needed. Such manoeuvring action will be needed if the estimated CPA is below the safe distance of passage. This paper will assess vision system performance system against requirements:

- TCPA > 10 min for larger ships.
- TCPA > 6 min for smaller vessels and stationary objects.

Information about relative speed and course of nearby vessels are essential for this computation, and it is required to assess whether they have conditions of reduced manoeuvrability. Such special conditions include being anchored, fishing or not under command. A vision system is not a sole sensor for object detection, which will rely on the fusion of data from Radar, AIS if this is sent by the object, and Electronic Chart Display and Information System (ECDIS).

3. Method
This section presents the vision system and outlines the method used to evaluate the spectral bands. Images are acquired from three different types of cameras and aligned before they are annotated and sent to the object detection network. A transformation of the images is needed such that the same objects are visible in all images at approximately the same spatial location. The steps to achieve this are presented in the following paragraphs.

3.1. Vision System
The images are acquired using the following sensors:

- Teledyne Dalsa Calibir 640, 8 – 14µm Long Wavelength Infrared camera. 42.5° horizontal field of view (HFOV). Raw resolution: 640x480 pixels, 14 bit.
- JAI GO-5000M 5M pixel, monochrome camera with a near-infrared low pass filter attached. 55° HFOV. Raw resolution: 2560x2048 pixels, 12 bit.
- JAI GO-5000C 5M pixel, visible range colour camera with a polarizing filter attached. 55° HFOV. Raw resolution: 2560x2048 pixels, 12 bit.

The cameras are fixed on a rigid mount to ensure a common baseline and positioned approximately 10 cm apart with parallel principal axis. This means that ideally, the extrinsic parameters of the camera systems consist of only a translation on one axis and no rotation. Figure 2 shows a photo of the camera setup.
3.2. Image Alignment

With cameras slightly offset and lenses not identical, the acquired images do not have the exact same field of view. To achieve an almost similar view, the images need to be transformed, cropped and scaled. This is achieved by first correcting the image for lens distortion and then identifying a transformation from the NIR and visible range images to the LWIR images.

The cameras are calibrated with a planar checkerboard pattern as a calibration target using the method described in [15]. A checkerboard printed on paper is used for the visible range and NIR cameras, and a checkerboard pattern cut from a thin plywood sheet as foreground and a solid aluminium plate as background, is used for the LWIR camera. The aluminium plate is heated to increase the contrast between plywood and aluminium in the LWIR calibration images. The parallax occurring due to the thickness of the plywood sheet is neglected in the calibration process. Calibration of the cameras provides intrinsic parameters, which are used to correct for lens distortion effects.

While most of the viewing frustums for the cameras are overlapping, it is not possible to achieve an overlap of the images with perfectly similar field of view (FOV). But with a relatively small translation between cameras compared to the distance of the class objects, a 2D transformation between the images suffices for the purpose of this study. Eight matching points across the images are selected and these are used to find a similarity transform consisting of rotation, scaling and translation between the camera images.

3.3. Data Set

Data was acquired on-board two different ferries operating in the Southern Funen Archipelago. An image was captured every 1s from each of the cameras and the set was recorded at springtime from midday till dawn, ensuring a broad range of illumination. A set of 9229 images were selected and annotated with bounding boxes of the object classes. Annotations were shared between the different cameras, i.e. if an object was annotated on the LWIR image, the same annotation would be placed in the corresponding images of the two other spectral ranges, regardless of visibility in these images.

Two classes are annotated, \textit{buoy} and \textit{ship}, and Table 2 outlines the total number of annotations and images in the data set. The slight discrepancies in the number of each class for the different camera types are due to objects leaving the frames at slightly different times, as a consequence of an offset in FOV near the cameras. The data set is split into a training and validation set by selecting 10\% of the images in the data set at random, and assigning them to the validation set. This yields a training set containing 8306 images and a validation set containing 923 images. An annotated image from the data set is seen in Figure 3.
3.4. Network Architecture

The detector used for the comparison in this paper is the RetinaNet CNN proposed by Lin et. al [6]. It predicts both a class of an object and its containing bounding box. This architecture was chosen as it was found to be the best in terms of performance and speed [13].

RetinaNet is a one stage detector, meaning that an image needs only to be inferred once through the network. This gives a significant speed-up compared to the popular two-stage detectors, such as the R-CNN family of networks [10, 3, 11]. It shares many features with other one-stage detectors, such as the use of anchors, as introduced in the RPN network in [11], and feature pyramids as used in the SSD [7] and FPN [5] networks. However, the reason that RetinaNet performs better than other one-stage detectors is in its novel loss function, called the Focal Loss, a modified version of the widely used Cross Entropy function. The cross entropy loss function is defined as:

$$CE(p, y) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1 - p) & \text{otherwise} \end{cases}$$

where \( y \in \{0, 1\} \) is the ground truth class, and \( p \) is the network’s probability for the class \( y = 1 \). The problem with this loss function is that it yields a non-trivial loss value even in the case where \( p \gg 0.5 \), i.e. for well-classified cases. This can lead to inefficient training or degenerate models if the class imbalance is high, i.e. too many of the training examples contain no meaningful information. One stage detectors of the same type as RetinaNet, are vulnerable to this effect when predictions containing background outnumber those containing relevant objects. RetinaNet modified in the cross entropy introducing a tuneable modulating factor (2) in the focal loss,

$$FL(p) = -(1 - p_t)^{\gamma}\log(p_t)$$

where \( p_t \) is defined as:
\[ p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise.} \end{cases} \] (3)

The \( \gamma \) in (2) is tuned to reduce the contribution from well-classified examples. This increases the weighting of hard examples contained in training data, and in turn, increases the classification performance. Furthermore, RetinaNet uses a weighting parameter, \( \alpha \in [0, 1] \), for the two classes, which can also reduce the effect of class imbalance. This parameter is used as another factor on the loss, yielding the \( \alpha \)-balanced focal loss,

\[ \text{FL}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t), \] (4)

where

\[ \alpha_t(y) = \begin{cases} \alpha & \text{if } y = 1 \\ 1 - \alpha & \text{otherwise.} \end{cases} \] (5)

The backbone of RetinaNet is a version of the Feature Pyramid Network (FPN) [5], built upon the ResNet-50 architecture. Furthermore, two sub-networks are attached to the feature pyramid generated by the FPN, a classification and a regression network, respectively.

The version of RetinaNet used in this study takes images with a resolution of 1440×1080 pixels as input. Higher resolution could technically be used, but the on-board setup used for inference is limited by memory. As a consequence, some information loss occurs in the NIR and visible range images, due to downsizing from their native resolution of 2560×2048 pixels. The LWIR range images are up-scaled to match the network input resolution. The anchor sizes used are found by k-means clustering of the ground truth boxes, and set to \( 14^2, 21^2, 34^2, 60^2 \) and \( 132^2 \). For each anchor size, the following aspect ratios are used in the network: 0.76, 1.14, 1.57, 2.18 and 3.2. The modulating and weighting parameters were chosen as \( \gamma = 2 \) and \( \alpha = 0.25 \), which are the values suggested in [6]. The Adam [4] optimizer was initialized with a learning rate of \( 1 \cdot 10^{-6} \).

4. Results
The networks were tested on the randomly selected validation set of images to compare network performance in the different spectral ranges. The number of images and classes in the validation set are seen in Table 3. Predictions on a set of images acquired at daytime and after sunset are shown in Figure 4.

| Images | Buoy | Ship |
|--------|------|------|
| 923    | 1431 | 624  |

Performance is calculated based on statistics for whether or not labeled objects in the validation data are detected. An object is considered detected if the predicted bounding box by the network and the ground truth box have an intersection over union (IoU) of above 0.3, together with the predicted class being correct. The low IoU threshold is chosen because annotations are shared between images from the different cameras. As the cameras’ FOV are not perfectly aligned, the bounding boxes need to be large enough to encompass an object in images by all cameras, which leads to larger bounding boxes than needed for the individual cameras and
Figure 4: Sample predictions on images in the dataset. First row of images are captured at noon, the second 30 min after sunset. The visual and NIR images outperform the LWIR during daylight conditions, due to the lower LWIR resolution. The LWIR is superior in low light conditions.

hence to lower IoU’s measures for the ensemble of cameras. This behaviour is caused by the ensemble labelling method used in the study. Better labelling could make it possible to raise the IoU threshold. The performance is evaluated using the standard measures, precision and recall, defined in (6).

\[
\text{precision} = \frac{TP}{TP + FP}, \quad \text{recall} = \frac{TP}{TP + FN},
\]

where \( TP \) (true positives) is the amount of correctly classified objects in the validation set, \( FN \) (false negatives) is missed detections and \( FP \) is false positives. From a navigational perspective, the detection range is an important metric to evaluate. Assuming perfect envelopment of the object, the detection range is related to the pixel area of the associated bounding box. The precision and recall are therefore computed as functions of the pixel area of the bounding boxes noting that an object is only a fraction of the bounding box area. Overall precision and recall for the spectral ranges are presented in Table 4.

Table 4: Overall precision and recall for the three camera types using a IoU threshold of 0.3 and a confidence threshold of 0.7

| Type    | Precision | Recall |
|---------|-----------|--------|
| Visible range | 0.96 | 0.9 |
| NIR     | 0.96      | 0.94   |
| LWIR    | 0.95      | 0.86   |

As precision and recall varies with the used IoU and confidence thresholds, one threshold could favour one model over the others, making the result biased. Therefore, the mean average precision (mAP) is used. Average precision \( AP(q) \) for a given class \( q \) is found from precision and recall for confidence thresholds in the range \( c = [0, 1] \) and then interpolating precision, \( p(r|q) \) such that \( p_{\text{inter}}(r|q) = \max_{\tilde{r} \geq r} p(\tilde{r}|q) \), where \( p_{\text{inter}}(0|q) \) and \( p_{\text{inter}}(1|q) \) are extrapolated from data.
**AP(q)** is computed by averaging \( p_{\text{inter}}(r|q) \) over a subset of recalls and mAP is the mean of \( AP(q) \) over \( Q \) object classes:

\[
AP(q) = \int_0^1 p(r|q)dr \approx \frac{1}{11} \sum_{r \in 0, 0.1, \ldots, 1} p_{\text{inter}}(r|q), \quad mAP = \frac{1}{Q} \sum_{q=1}^{Q} AP(q).
\]  

(7)

The mAP was calculated for three IoU thresholds, with presented in Table 5. Looking at the mAP, it can be seen that the visual range and NIR spectral ranges perform better than LWIR.

**Table 5: Mean average precision**

|                  | Visual range | NIR  | LWIR |
|------------------|--------------|------|------|
| mAP@0.25 IoU     | 0.98         | 0.98 | 0.96 |
| mAP@0.50 IoU     | 0.95         | 0.96 | 0.89 |
| mAP@0.75 IoU     | **0.45**     | 0.42 | 0.37 |

The plots for evaluation of the precision and recall for the different spectral ranges, along with a histogram of detections, is presented in Figure 5. Precision and recall are shown as function of area and are calculated as follows,

\[
P(a) = \frac{\sum_0^a TP(a)}{\sum_0^a (TP(a) + FP(a))}, \quad R(a) = \frac{\sum_0^a TP(a)}{\sum_0^a (TP(a) + FN(a))}
\]  

(8)

### 4.1. Viability in a navigational setting

From the navigational requirements outlined in Section 2, the maximum detection range needs to be assessed. Figure 6 shows the earliest frame in the validation set, where the M/F Marstal ferry was detected. Correlating with AIS data, which all larger vessels must transmit, showed the distance to the ferry was 5.5 Nautical Miles (nmi) in this frame. The ferry travelled with a speed of 12 knots, while own vessel was travelling at 11.6 knots. Accordingly, the estimated TCPA is 14 minutes, which is sufficient time for route planning.

Using the statistics from Figure 5, we can calculate the minimum distance from the camera to an object, based on its pixel area. The smallest pixel area where the networks obtain a reliable detection rate, i.e. a recall of at least 0.5, was chosen. The expression used for calculating the distance to an object, \( d_{\text{phys}}^{\text{object}} \), is described in (9).

\[
d_{\text{phys}}^{\text{object}} = \frac{s_{\text{obj}}}{\tan(\alpha_{\text{obj}} H)}
\]  

(9)

where \( s_{\text{phys}}^{\text{obj}} \) is the object’s physical size and \( \alpha_{\text{obj}} H \) is the horizontal angle occupied by the object. This angle is calculated as:

\[
\alpha_{\text{obj}} H = \frac{s_{\text{cam}}^{\text{H}}}{s_{\text{img}}^{\text{H}}} \cdot \text{FOV}_{\text{H}}^{\text{cam}},
\]  

(10)

where \( s_{\text{cam}}^{\text{H}} \) is the pixel width of the object in the image, \( s_{\text{img}}^{\text{H}} \) is the width of the image in pixels and FOV_{\text{H}}^{\text{cam}} is the camera’s horizontal FOV. As image alignment is used, the images all have FOV_{\text{H}}^{\text{cam}} = 42.5^\circ \), and \( s_{\text{img}}^{\text{H}} = 1440 \) pixels. From Figure 5, the pixel areas resulting in a recall of 0.5 or above was found and denoted in Table 6.
Figure 5: Detection histograms, precision and recall curves for the three spectral ranges, as function of area. The input images to the networks for each spectral band are scaled to have the same resolution: 1440 × 1080 pixels. The ripple in the precision curve at low values of area is due to low amount of false positives with low area.

Figure 6: Cropped visible range image with augmented bounding box detections from the network. The class name, classification certainty and intersection over union (IoU) are denoted for each prediction. The ship-detection on the left is the M/F Marstal ferry.

Assuming square bounding boxes perfectly enveloping the object, and using the physical dimensions of the M/F Marstal ferry for our calculations, i.e. $s_{phys}^{obj} = 13m$, the minimum detection distances with a recall of 0.5 is shown in (11).
Table 6: Pixel areas corresponding to recall = 0.5 for the three spectral ranges.

| spectral range | Visible | NIR | LWIR |
|----------------|--------|-----|------|
| Pixels         | 70     | 60  | 70   |

\[
\begin{align*}
\text{d}_{\text{phys,visible}}^{\text{object}} &= 1.27 \text{ nmi} \\
\text{d}_{\text{phys,NIR}}^{\text{object}} &= 1.27 \text{ nmi} \\
\text{d}_{\text{phys,LWIR}}^{\text{object}} &= 1.06 \text{ nmi} \\
\end{align*}
\]

These distances are the absolute minimum detection ranges, assuming a perfectly enveloping bounding box, which is rarely the case on objects far away from the camera when using shared annotations. Using Figure 6 as an example, the M/F Marstal ferry takes up approximately a sixth of the horizontal width of the bounding box, i.e. 2.5 pixels. Using this width in (9) yields:

\[
\text{d}_{\text{phys}}^{\text{object}} = 5.45 \text{ nmi},
\]

which is in line with the AIS measurements, and will yield a TCPA of more than 10 minutes for a vessel similar to the M/F Marstal ferry.

5. Discussion

The overall precision and recall in Table 4 shows that recall across all bounding box pixel areas is better in the visible and NIR ranges when going for an approximately equivalent precision of 90%. Objects generally need to be closer to the LWIR camera before their heat signature is detected by the CNN. As we are using shared annotations in this investigation, some annotations in the validation set will not correspond to visible objects on the LWIR images. This leads to a slightly lower recall for the LWIR range. From a navigational point of view, the LWIR range will be worse for long-range detection of objects, which is evident from the recall curves in Figure 5. Although performing worse than the visible range and NIR cameras, an LWIR camera can still have merit in a navigational setting. In a night-time setting, an LWIR camera can have more information in the frame from the heat signatures, than the other spectral ranges, as these will only be able to detect the lights placed on the ships and buoys. Figure 7 shows how the appearance of a buoy is more significant in the LWIR image compared to visible range and NIR.

When comparing the visible range colour and NIR cameras, results show a slightly higher recall in the NIR range for objects for bounding box pixel area of less than 250 pixels. From a navigational perspective, this argues for the use of a NIR camera for long-range detection of objects. Colour information is needed to decode navigation information from buoys and signal lights during night time. Therefore, the results suggest that cameras in all ranges investigated are relevant. The differences in image resolution used must be taken into account. The 640 × 480 resolution of the LWIR images have a negative impact on the amount of information extracted from the environment, compared to the NIR and visible range cameras. Consequently, worse performance in the LWIR range should be expected. Camera elevation also plays a role. The dataset in this paper was recorded from different ferries which results in images acquired from distinct heights in the range 6-9 m above sea surface. Sensor elevation will mainly be of significance for objects at close range.
Figure 7: Evening photo from the three bands. Motion blur, due to a slow shutter speed, is affecting the buoy in the visible range and NIR images. The silhouette is however very clear in the LWIR image.

5.1. Perspectives
The goal of this research was to obtain object detection based on two main classes. Further classification into subclasses, such as specific types of ships and buoys, is expected to be supported by a dedicated network for reasons of performance, plus a pan-tilt-zoom camera. The latter would be employed to obtain greater detail of an object, e.g. to read day signs that show vessel status. Further research will also expand the image database with nighttime images, a range of different weather conditions and all representative boat types. A pertinent issue with safety critical use of neural networks occurs when training has been insufficient such that detection fails in particular cases. This concern is a hot topic in industries, including marine, where strict safety standards and approval procedures apply. Techniques for anomaly detection, classical image analysis, object tracking and multi mode sensor fusion are envisaged, in combination, to achieve the resilience and safety integrity needed but significant research is necessary.

6. Conclusion
Based on requirements formulated by master mariners from SIMAC, this paper assessed the performance of object detection using a RetinaNet CNN on images obtained in the visible, NIR and LWIR spectral ranges. Unifying to 1440 pixels, i.e. down-scaling visible range and NIR images from 2560 to 1440 and up-scaling LWIR from 640 to 1440 pixels horizontally, object detection precision of > 90% was achieved along with recall in the 80 – 90% range. Detection distance was investigated using pixel areas, and calibrated by AIS transmissions. Detection ranges in daylight were found sufficient to obtain TCPA which meets the requirements for safe navigation. Performance of the CNN was best in the visible and NIR ranges, in daylight conditions, while the LWIR range would be necessary for detecting objects without lights at nighttime.

Acknowledgments
This research was conducted as parts of the ASAN and ShippingLab projects. The first and fourth authors are sponsored by ASAN, the second and third by the ShippingLab Autonomy Workpackage. ASAN is sponsored by The Danish Maritime Foundation, Orients Fund and LauritzenFonden. ShippingLab is sponsored by these and also by The Danish Innovation Fund. Requirements were obtained in workshops with SIMAC, who is a partner in both projects. GPUs donated by the NVIDIA Corporation were used to train the networks. The support and the professional input from SIMAC are gratefully acknowledged.
References

[1] M. Blanke, S. Hansen, J. D. Stets, T. Koester, J. E. Brøsted, A. Llopart Maurin, N. Nykvist, and J. Bang. Outlook for navigation – comparing human performance with a robotic solution. In Proc. of ICMASS’2018. SINTEF Academic Press, May 2019.

[2] R. Dulski, S. Milewski, M. Kastek, P. Trzaskawka, M. Szustakowski, W. Ciurapiński, and M. Zyczkowski. Detection of small surface vessels in near, medium and far infrared spectral bands. In Proceedings of SPIE - The International Society for Optical Engineering, volume 8185, 2011.

[3] R. Girshick. Fast R-CNN. In Proc. IEEE Int. Conference on Computer Vision, 2015.

[4] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. CoRR, abs/1412.6980, 2014.

[5] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. CoRR, abs/1612.03144, 2016.

[6] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In Proc. 2017 IEEE Int. Conf. on Computer Vision (ICCV). IEEE, 2017.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In Lecture Notes in Computer Science, page 21–37. Springer Int. Publishing, 2016.

[8] D. Papageorgiou, M. Blanke, M. Lützen, M. Bennedsen, J. Mogensen, and S. Hansen. Parallel automaton representation of marine crafts’ COLREGS based maneuvering behaviours. In Proc. IFAC Conference on Control Applications in Marine Systems, Robotics and Vehicles, CAMS’19. Submitted.

[9] D. K. Prasad, D. Rajan, L. Rachmawati, E. Rajabally, and C. Quek. Video processing from electro-optical sensors for object detection and tracking in a maritime environment: A survey. IEEE Transactions on Intelligent Transportation Systems, 18(7812788):1993–2016, Aug. 2017.

[10] T. D. R. Girshick, J. Donahue and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proc. Ieee Conf on Computer Vision and Pattern Recognition, 2014.

[11] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39, 2017.

[12] C. D. Rodin and T. A. Johansen. Detectability of objects at the sea surface in visible light and thermal camera images. In Proc. of Oceans 2018, 2018.

[13] F. E. T. Schöller, M. K. Plenge-Feidenhans'l, J. D. Stets, and M. Blanke. Assessing deep-learning methods for object detection at sea from LWIR images. In Proc. IFAC Conference on Control Applications in Marine Systems, Robotics and Vehicles, CAMS’19. Elsevier, 2019. Submitted.

[14] T. A. Tolga Can, A. Onur Karali. Detection and tracking of sea-surface targets in infrared and visual band videos using the bag-of-features technique with scale-invariant feature transform. Applied Optics, 50(33):6302–6312, 2011.

[15] Z. Zhang. A flexible new technique for camera calibration. IEEE Transactions on pattern analysis and machine intelligence, 22, 2000.