National Physical Laboratory

NPL REPORT MS 33

GOOD PRACTICE IN TRAINING SET PREPARATION FOR MARINE NAVIGATION SYSTEMS

RAHUL KHATRY AND ANDREW THOMPSON

NOVEMBER 2021

Department for
Business, Energy
& Industrial Strategy

FUNDED BY BEIS

Good practice in training set preparation for marine navigation systems

Rahul Khatry and Andrew Thompson
Data Science

ABSTRACT

The automation of navigation systems for marine vessels has been gaining momentum recently. Such systems require the ability to perform object detection on images received by the vessel’s sensors, which in turn requires machine learning models and training data. This document addresses good practice in training set preparation for these object detection algorithms. We give an overview of a typical marine navigation system and describe the accompanying object detection task. We highlight three main areas of good practice: data transferability, data coverage and data accuracy, and we give concrete recommendations on good practice.

https://doi.org/10.47120/npl.MS33

ISSN 1754-2960

National Physical Laboratory
Hampton Road, Teddington, Middlesex, TW11 0LW

Approved on behalf of NPLML by
Louise Wright, Head of Science (Data Science)

CONTENTS

1 Introduction

This document addresses marine autonomous vehicles (marine AV) systems and the good practice issues involved in the preparation of training data for the machine learning components of such systems. In order to understand the requirements for training data, it is helpful to understand the marine navigation system in its entirety. Safety and validation of such systems is critical, and so it is also important to understand the operational domain and the system requirements. This document first discusses these aspects and then moves on to describe the machine learning task involved and give recommendations for good practice.

The structure of the rest of the document is as follows. Section 2 sets the context by giving a brief introduction to marine navigation systems. Machine learning is typically used for object detection and classification in marine AV, and so in Section 3 we describe the object detection and classification tasks encountered in marine AV and its training data requirements. Section 4 gives guidance concerning good practice in training set preparation for the object detection and classification tasks in marine AV.

2 Marine Navigation Systems

Marine vessels have been used to carry people and goods across the sea since early times, and these days they are an integral part of the transportation system. The automation wave has been gaining momentum in various industries, including manufacturing and the automotive sector. Experimentation with marine vessel automation has been ongoing for many years and validation of such systems is of crucial importance since accidents could cause loss of life and value. It hence becomes important that the highest safety standards are applied in every system operating these vessels.

Marine AV systems' functioning can be understood by tracking the transition of information from the input sensor data to the actuation commands (Figure 1). The first step is sensing, in which the system collects all the information from the different sensors including Cameras (images), Radar (distance information), GPS/IMU (location/position) and AIS (information about all other vessels). In the perception step, the images are used for object detection and classification, which is then fused with data from other sensors to infer locations of all obstacles. Sensor data fusion is then carried out to merge all data streams to obtain maps of the surroundings and obstacles. This is combined with the ship's location and position information to carry out a SLAM (Simultaneous Localization and Mapping) calculation which gives information about the overall environment, including where the ship is in that environment. This is also merged with the estimated trajectories of the moving obstacles to get the available and non-available movement regions. This information is used by path search algorithms to calculate the optimum trajectory that the ship needs to take while considering the moving and static objects. This information is then fed to the control system to create appropriate control commands which is passed to the actuation system to make appropriate movements in the vessel.

There are four input data streams that are typically relevant for marine object detection:

i) Camera: The camera arrays take pictures of the panoramic view and feed it to the Stereovision module.
ii) Radar: Radar reads the relative speed and azimuth of the surrounding objects further than the field of vision [1] and inputs into the sensor data fusion module.
iii) GPS/IMU: The GPS and IMU units measure the GPS location and the roll/pitch/yaw movement of the ship.
iv) AIS (Automatic Identification System): The AIS system is the communication that happens between ships, and between stations and ships. Broadcasts of these messages help the communication systems and position/trajectory estimation.

Since machine learning is typically used only for the object detection/classification component of the perception step, we will make this the focus of the rest of the report. The next section gives insight into the training data requirements of the object detection/classification component. Within the last decade, many object detection systems based on deep learning techniques have appeared which have shown great promise and have become state-of-the-art compared to classical computer vision-based methods (see Section 3), achieving higher classification accuracies than classical methods [2], and so in the rest of the report we focus primarily on guidelines for training images for use with deep learning techniques.

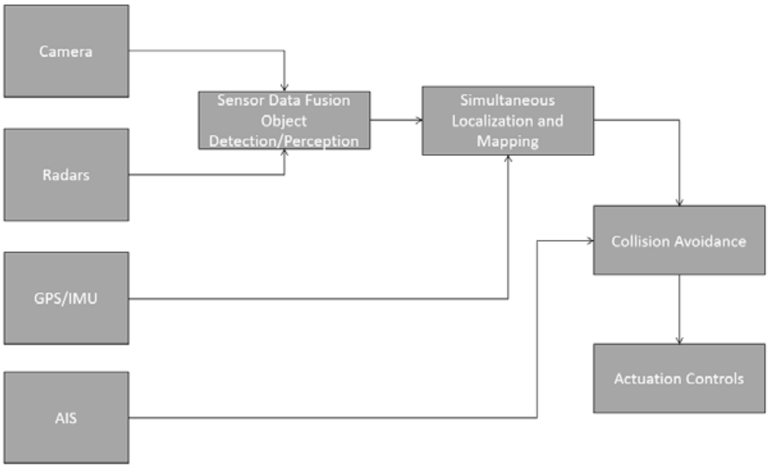

Camera
Radars
GPS/IMU
AIS
Sensor Data Fusion Object Detection/Perception
Simultaneous Localization and Mapping
Collision Avoidance
Actuation Controls

Figure 1 Simplified explanation of the AV navigation steps

3 Object Detection Systems for Marine AV

Since machine learning is used mainly in the object detection and classification system for marine AV, this section is focused on the training data that is applicable for the object detection/classification system. The object detection task involves identifying the presence of a vessel along with its location (typically using a bounding box), while the classification task involves subsequently determining the type of vessel. It is important to note that the two tasks should be distinguished, and the system requirements for detection and classification may be different. For example, it may be required to detect remote vessels and to perform low-granularity classification sufficient to identify extremely large vessels (for example mega-containerships), while it may be required to perform a higher-granularity classification once the vessels are closer. Similarly, the granularity of the classes must also be informed by the user requirements (for example it may be sufficient to distinguish between ships and buoys, or it may be necessary to distinguish between different types or sizes of ship).

The datasets required to train such a machine learning algorithm must consist of the following characteristics:

i) The domain images: The images which are taken in the operational design domain of the ship.
ii) The bounding boxes: Boxes to denote the ship within the image which is used for training the object detection system (as well as the output of the object detector).
iii) Object segmentation: Pixelwise location of the image (optional).

As an illustration of the type of data required, Figure 2 displays examples of different vessel types from SeaShips [3], a publicly available dataset captured by deploying a video monitoring system around the Hengqin Island, Zhuhai city, China. The dataset contains six principal ship types: ore carrier, bulk cargo carrier, general cargo ship, container ship, fishing boat and passenger ship. Figure 3 displays examples of the use of bounding boxes in SeaShips on the annotated images. The dataset consists of different vessel types and viewpoints (left of frame, middle of frame and right of frame) summarized in the Tables 4a and 4b. We present this example to illustrate some important aspects, namely coverage of classes, annotation and different viewpoints. However, for a marine navigation system, the training data would ideally be gathered *in situ*, using the same sensors and mast that are to be used in the deployment of the navigation system. Furthermore, side-on views alone are not adequate, and other viewpoints such as bow-on and bow-quarter which correspond to more directly approaching vessels are required.

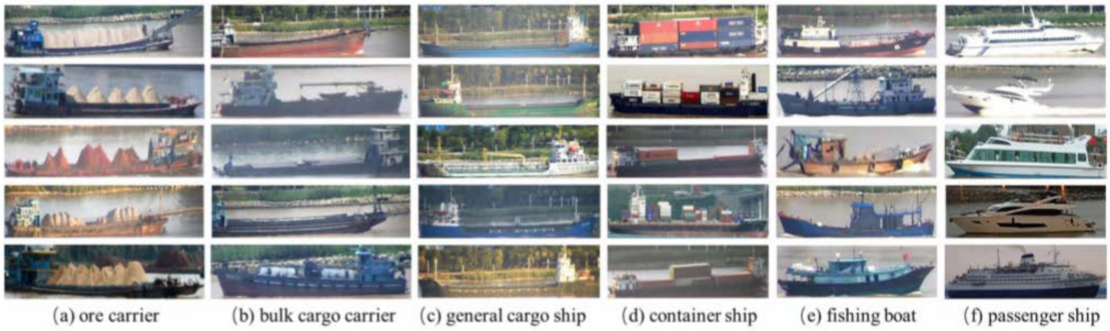

(a) ore carrier (b) bulk cargo carrier (c) general cargo ship (d) container ship (e) fishing boat (f) passenger ship

Figure 2 SeaShips classes example images [3]

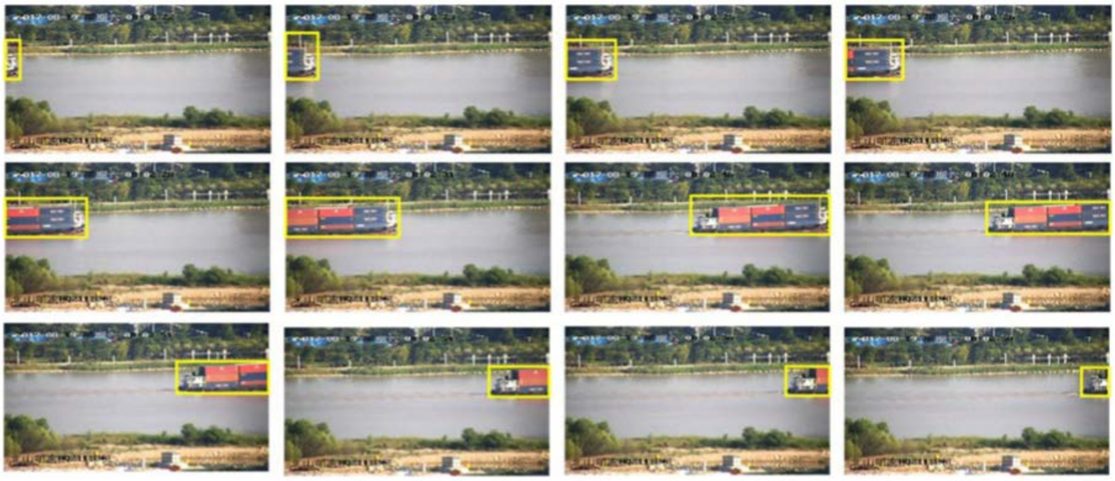
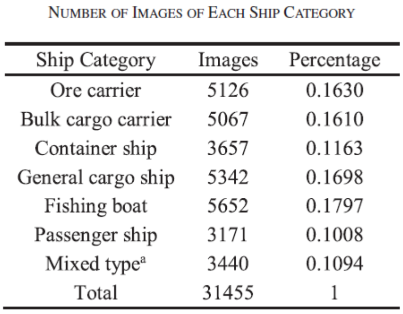
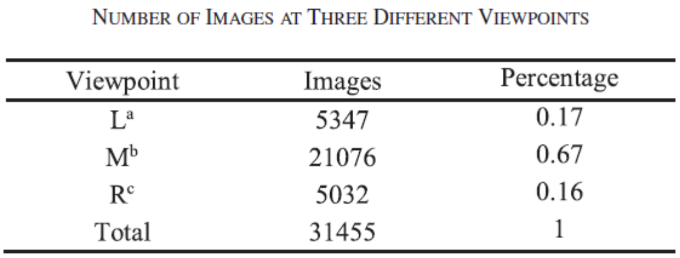


Figure 3 SeaShips based object detection example [3]


Number of Images of Each Ship Category

Ship Category	Images	Percentage
Ore carrier	5126	0.1630
Bulk cargo carrier	5067	0.1610
Container ship	3657	0.1163
General cargo ship	5342	0.1698
Fishing boat	5652	0.1797
Passenger ship	3171	0.1008
Mixed type[a]	3440	0.1094
Total	31455	1


Number of Images at Three Different Viewpoints

Viewpoint	Images	Percentage
L[a]	5347	0.17
M[b]	21076	0.67
R[c]	5032	0.16
Total	31455	1

Table 4 (a, b): Ship categories and image viewpoints in the SeaShips Dataset [3]

There are two main types of approach to ship detection in a maritime environment [4], [5], [6]:

a) **Classical computer vision-based methods.**

These include dynamic background modelling and shadow suppression to detect boats. Techniques such as background surveillance, background modelling for port surveillance and saliency detection to detect non-moving boats have also been applied. Appearance methods to model what a vessel looks like and classify it using edge detection have also been used [4], [5].

The technical tasks to be solved using traditional methods include horizon detection, background subtraction, foreground detection and tracking. These tasks are challenging due to the noise- and artefact-caused variations in the background, such as wakes, foams, clouds, etc. The effects of illumination conditions such as sunlight, twilight, night, rain, haze, fog can also make a model ineffective from one setting to another [5].

b) **Deep learning techniques like SSD, R-CNN and YOLO [6], [7], [8].**

These approaches can be thought of as specialized convolutional neural networks (CNNs), which combine feature detection and image classification.

Both deep learning and classical techniques are agnostic to the location of the object within their field of view or the relative motion of the camera and the ships since they both rely on template matching and detection [6]. However, in contrast to classical techniques, deep learning is potentially less sensitive to variations in background and lighting, though it still needs to be ensured that the model training process ensures that it is agnostic to such conditions.

Another advantage of deep learning systems is that they can be set up in such a way that they are continuously learning and improving over time [9]. Hence having a deep learning system could mean that the system continuously improves after re-training.

These advantages are compounded by the fact that, due to the increasing popularity of deep learning, it is becoming easier to hire people with skills in deep learning as opposed to the classical methods.

An example CNN architecture for object detection, namely the Single Shot MultiBox detector [6], is shown in Figure 5. In this architecture, an image is passed through a proprietary image classification network (VGG-16) of layers before passing through additional layers which predict the offsets to default bounding boxes of different scales and aspect ratios and their associated confidences.

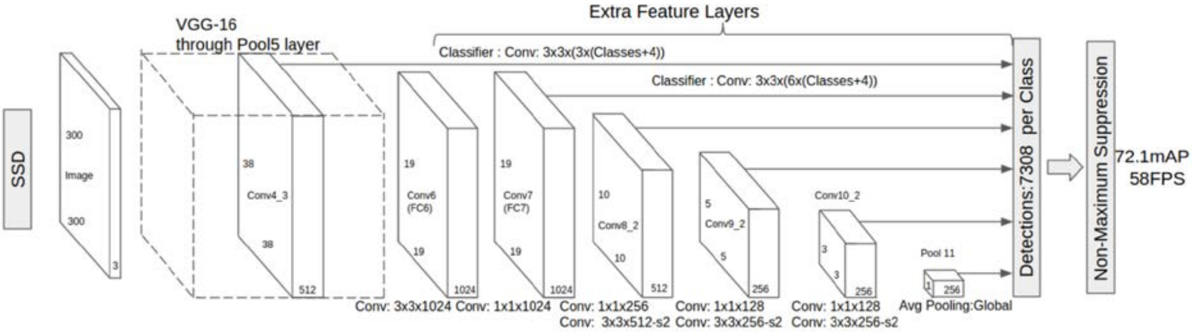


Figure 5 Architecture for a Single Shot Multibox detector [6]

Another approach often used in conjunction with deep learning is *transfer learning* [10], in which an existing model is adapted for a new dataset, rather than learning a new model from scratch. In this approach, the final layers of the CNN are typically stripped away to preserve the feature extraction capability, while the new classification layers are created and retrained on smaller volumes of new data. This accelerates the transfer of the pre-trained model onto a new domain (domain shift). Models built from existing marine vessel databases (such as SeaShips) could be used in this way, but it has also been observed that transfer learning can often be successful even where significant domain shift takes place, and so an alternative approach is to retrain an object detection system which is not specific to marine AV. For example, pretrained models for the SSD-mobilenet and YOLO architectures are publicly available [11] and can be deployed using open source deep learning frameworks such as PyTorch, Keras or TensorFlow [12].

4 Good Practice Guidance

In this section, we highlight three areas of good practice which are important to consider when collecting and preparing an image dataset for object detection and classification. The three areas are:

i) Data and model transferability (Section 4.1)
ii) Data coverage (Section 4.2)
iii) Data accuracy (Section 4.3).

In addition, an important general observation is that the ultimate measure of good practice in this context is whether the training data leads to a machine learning model which meets the system requirements. The areas of good practice are shown in Figure 6.

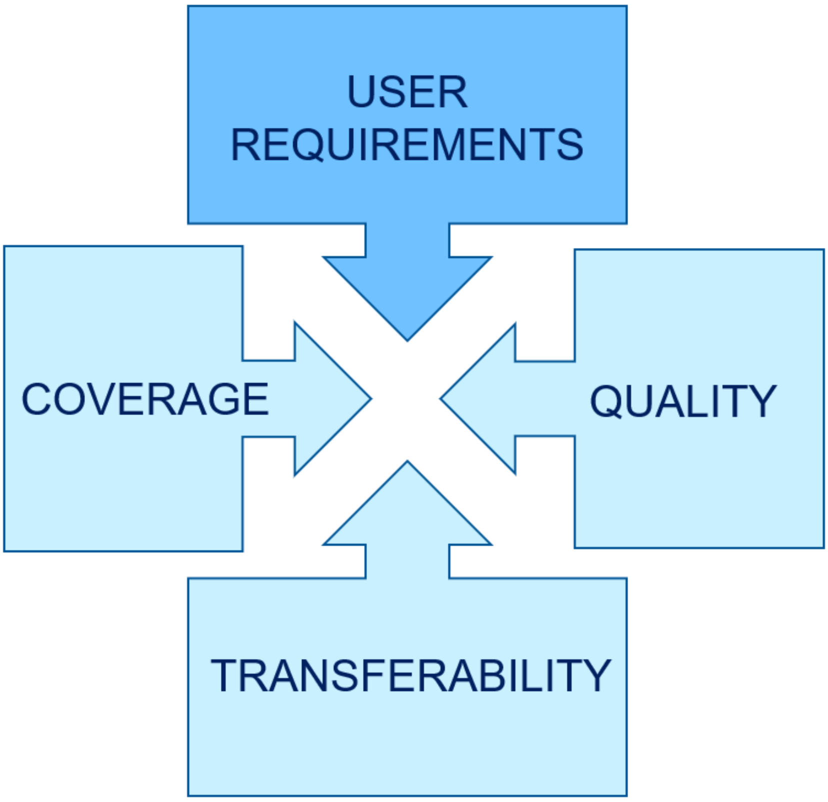


Figure 6 Areas of good practice for training data preparation

These areas of good practice align significantly with three of the popular four Vs of big data [13].

i) *Volume* (the number of examples in the dataset)
ii) *Variety* (the extent to which the dataset captures all possible scenarios)
iii) *Veracity* (the accuracy of the dataset).[1]

We point out some of these alignments in what follows.

4.1 DATA AND MODEL TRANSFERABILITY

Machine learning models are expensive to train, and it is difficult to collect and prepare the datasets. Wherever possible, it is therefore advantageous to reuse existing data or existing models.

a) **Existing data.**

Object detection and identification involves learning a complex and high-dimensional classification model. Statistical learning theory tells us that, in such a setting, a large training dataset is required to train an accurate model. As an example, the SeaShips database [3] contains 31,455 images. From this perspective, if there exists a dataset consisting of marine vessels, of a type relevant to the classification task at hand, it may make sense to use it. However, it is also important to be aware that an existing dataset may not be a good fit to the specific scenario of interest. For example, if two datasets were collected in the Antarctic and warm water regions respectively, one would expect the classes (types of vessels), lighting conditions and image quality covered by the two datasets to be different. Equally, image quality is likely to vary between two datasets which are not collected using the same sensors. Given such differences, a classification model built for one dataset may not be transferable to another dataset (a scenario often referred to as domain shift). Experience may be an initial guide to whether the benefits of reducing the need for data collection or augmenting the data set outweigh the negative impacts of domain shift. If some data is collected for which there is a higher confidence that it is relevant to the specific object detection task of interest, then numerical experiments can provide evidence. Two analytical approaches are likely to be helpful.

i) Unsupervised learning techniques, such as discriminant analysis [14, Section 4], can be used to identify pervasive differences between the two datasets. Unsupervised learning refers to the practice of discovering patterns and structure in an unlabelled database.
ii) A predictive model can be trained on the existing data set and then tested on both datasets. To achieve this, some of the existing data set must be set aside for testing. A comparison of prediction accuracy on the two test sets then gives insight into how transferable the model is.

If there is a financial cost to using existing data, then the benefits of using this data must be weighed against the financial costs.

If there is domain shift only in the sense that the existing dataset does not cover all types of object expected to be encountered, it is reasonable to supplement the existing

[1] The fourth V of big data is *velocity* (the speed of data acquisition), and this is less relevant in the context of training data. However, it certainly becomes important when it comes to deploying a marine navigation system where multiple data streams need to be evaluated within a narrow time window.

dataset with additional samples which extend coverage. If, on the other hand, the domain shift takes the form of different sensor specifications, it will be necessary to retrain the model on new data using the transfer learning approach described in Section 3.

b) **Existing models.**

In the fortunate situation that a model has already been trained on an existing dataset, then it may not be necessary to train a new model. However, the question of domain shift must be asked: is the model likely to be valid for the specific object detection task of interest? Experience can again be used as an initial guide, and the two analytical approaches mentioned above can provide evidence, provided that the data used to build the existing model is also available. It may be that the volume of data used to build the existing model is extremely large, in which case the unsupervised learning approach may be computationally challenging. While domain shift might mean that a model is not entirely transferable, transfer learning as described in Section 3 can be used to adapt an existing model. This approach still requires a modest volume of new data to be available, but it will likely reduce the amount of data that needs to be collected.

4.2 DATA COVERAGE

As was noted in the previous section, object detection involves learning a complex and high-dimensional classification model, and in such a context the requirements on the training set are that there is sufficient data to be representative of the range of different object detection and classification tasks expected to be encountered by the marine vessel. It is crucial therefore that the specific functions that the object detection system will be required to perform are used to inform the decisions about data collection from the outset. In addition, the dataset must have adequate *volume* and *variety* [13]. *Volume* in this context simply means the number of training samples, whereas *variety* is about how well the dataset captures the variations between the classes and within the same class.

In the context of training images for marine AV object detection, in order to sufficiently capture between-class variations, *class imbalance* should be avoided, in which the number of sample images of some types of vessels is much greater than others. When it comes to within-class variations, invariance to different viewing conditions (for example illumination, background and presence or otherwise of occlusions) is the main concern in a marine AV setting. It is described in [3], for example, how care was taken in the data acquisition of the SeaShips database by using panoramic video camera imaging, imaging in various locations, ensuring the presence of occlusions, and by careful selection from available data.

Data augmentation [15] refers to the practice of artificially generating additional training data to enhance an existing data set. Whether data augmentation should be performed will depend upon whether there is a need to correct for inadequacies in the collected data. There are three main reasons why data augmentation might be required.

i) To improve the robustness of a machine learning model to intra-class variations (such as differences in viewing conditions).
ii) To correct for class imbalance, namely that there are more data samples in some classes than others.
iii) To increase the size of a training set if the training set is originally too small.

Simple examples of data augmentation include rotation of images or parts of images to simulate different viewing angles, and the use of image adjustment techniques to simulate different lighting conditions, geometric distortions, obstructions and viewing angles.

There also exist algorithms for training deep neural networks which seek to improve the robustness of a machine learning model to perturbations by augmenting the training set with adversarial perturbations. Security of autonomous marine vessel machine learning models in the sense of robustness to adversarial attacks is an important concern, though this type of augmentation really belongs with algorithmic questions about how to train a machine learning model, and so is left outside the scope of the present document [16].

Data augmentation should be subjected to the same level of scrutiny as the original data: it is important that the augmented data is representative of real-world conditions. For example, if rotations are used, the rotations should be representative of the viewpoints that are expected to be encountered by the marine vessel during operation of the navigation system. The effect of different weather conditions on the resulting images should also be carefully modelled (see Section 4.3).

4.3 DATA ACCURACY

Inaccuracies in training data lead to inaccurate machine learning models, and consequently uncertainties in training data propagate through a model to give uncertainties in the classifications. With respect to training images for object classification, there are two aspects of data accuracy (or *veracity* [13]) that must be addressed: labelling quality and image quality.

a) Labelling quality.

Creating a labelled database of marine vessel images will need to be carried out manually, which can be a time-consuming process. Labelling can either be carried out in-house or by using crowd-sourcing services, e.g. reCAPTCHA. It is vital that labelling correctly captures ground truth, whilst at the same time identifying images which are known to the labeller because of other references but not identifiable through object features. There exist formal recommendations on custom labelling of datasets, for example the Visual Object Classes (VOC) Annotation Guidelines [17], which we reproduce below. This provides an illustrative example, highlighting some of the intricacies typically involved in labelling images for classification purposes.

i) What to label:
   - o All objects of the defined classes, unless
     - you are unsure what the object is.
     - the object is very small (at your discretion).
     - less than 10-20% of the object is visible, such that you cannot be sure what class it is. e.g., if only a tyre is visible it may belong to car or truck so cannot be labelled car, but feet/faces can only belong to a person.
   - o If this is not possible because too many objects, mark the image as bad.

ii) Viewpoint:
   - o Record the viewpoint of the 'bulk' of the object, e.g. the body rather than the head. Allow viewpoints within 10-20 degrees. If ambiguous, leave as 'Unspecified.' Unusually rotated objects, e.g. upside-down people should be left as 'Unspecified.'

iii) Bounding box:
   - o Mark the bounding box of the visible area of the object (not the estimated total extent of the object). Bounding box should contain all visible pixels,

except where the bounding box would have to be made excessively large to include a few additional pixels (less than 5%) e.g., a car aerial.

iv) Truncation:
- If more than 15-20% of the object lies outside the bounding box mark as Truncated. The flag indicates that the bounding box does not cover the total extent of the object.

v) Occlusion:
- If more than 5% of the object is occluded within the bounding box, mark as Occluded. The flag indicates that the object is not visible within the bounding box.

vi) Image quality/illumination:
- Images which are poor quality (e.g., excessive motion blur) should be marked bad. However, poor illumination (e.g., objects in silhouette) should not count as poor quality unless objects cannot be recognized. Images made up of multiple images (e.g., collages) should be marked bad.

vii) Clothing/mud/ snow etc.:
- If an object is ‘occluded’ by a close-fitting occluder e.g., clothing, mud, snow, etc., then the occluder should be treated as part of the object.

viii) Transparency:
- Do label objects visible through glass but treat reflections on the glass as occlusion.

ix) Mirrors:
- Do label objects in mirrors.

x) Pictures:
- Label objects in pictures/posters/signs only if they are photorealistic but not if cartoons, symbols, etc.

We highlight which of these considerations are especially apposite for detection and classification of marine vessels.

i) For detecting remote vessels, in order to capture the targeted scenario, training examples in which the vessel takes up relatively few pixels will be needed. However, for these examples, the resolution of the image may not be adequate to assign a classification label to the vessel.

ii) Both complete and incomplete parts of the ships must be annotated because we would expect to have to deal with both occlusions and ships entering and leaving the field of view of the camera, see for example [3].

iii) In a marine vessel context, expected viewpoint labels would be ‘side view’, ‘bow-on’ and ‘bow-quarter’.

Labelling and annotation are manual processes and are inherently subject to human error. It is therefore important that confidence in the quality of the labelling and annotation is established by assessing the labelling and annotation quality. This can be achieved either by assessment against ground truth if it is available, or by assessing inter-observer variability [18].

b) Image quality.

Images have an inherent accuracy limit determined by their resolution. The resolution of an image is determined by two factors: detector resolution (the number of detector elements) and the captured image resolution (number of pixels). The issue of resolution comes into play especially when designing specifications for datasets for classifiers for detecting ships near the horizon. In addition, weather effects can reduce image quality by introducing noise, blur and distortion. The performance of an object detection and classification system would be expected to degrade as image quality degrades. It is important to note that performance is impacted both by the quality of training and test images. If either of the two has low resolution or is subject to high levels of noise, the performance will degrade, even if the other is of higher quality.

Another aspect that needs to be kept in mind is that the precise way in which performance degrades with image quality will be highly dependent upon the detection/classification task. For example, in satellite imaging objects of 5 pixels in size can sometimes be localized with high confidence, whereas in face identification it has been reported that the object size needs to be 32 pixels to be identified with reasonable confidence [19], [20]. Hence it can be understood that acceptable resolution of an object is related to the kind of detection/classification task that needs to be performed. The same principle would apply to noise from weather conditions. In a marine AV context, it would be easier to distinguish between boats and islands but, in order to distinguish which kind of boat or boat model, higher quality images would be needed. Viewed another way, if the resolution is held fixed, distinguishing type of boat or boat model would only be possible at a relatively shorter distance compared to just detecting a boat etc.

Image quality assessment (IQA) can be classified into two kinds:

i) *Reference based IQA*: These metrics compare an image with a high-quality reference image, and therefore rely on such a high-quality image being present. These methods are particularly useful in quantifying the effect of weather on image quality. Popular metrics of this type are reviewed in [21] and are listed below:
   a. *Structural Similarity Index* uses a sliding window to compare the mean illumination and standard deviation of the two images.
   b. *Multiscale Structural Similarity* is more flexible for considering image details with different distortions and applies low-pass filters and down-sampling. The two images are compared on different down-sampling scales and the similarity index is calculated.
   c. *Most Apparent Distortion* calculates apparent distortions by comparing the luminescence of the two images and measuring local errors.

ii) *Non-Reference based IQA*: These metrics directly assess image quality for a given image based on image characteristics, without the presence of any reference image. Popular metrics of this type have been provided in [22]. Some interesting non-reference IQA metrics are listed below:
   a. *Blockiness* makes use of luminance masking, measured as luminance difference between neighbouring blocks, and texture masking, which is computed using neighbouring block properties.
   b. *Perceptual Blur and Ringing Metric* is based on edge detection using the Sobel operator. Noise and other subtle edges are removed by applying thresholds to gradient images. Local blur scores are aggregated to obtain a measure of the overall blur present in image.
   c. *Anisotropy* is used to measure image quality in the sense of the image having different qualities in different directions. In this approach, Generalized Renyi entropy and normalized pseudo-Wigner Distribution are used to calculate the directional entropy of the image.

We note in addition that, if the effect of different types of noise on image quality is carefully modelled, these models can then be used to perform data augmentation. For example, kernel filters, colour space transformation and random erasing can be used to mimic the effect of bad weather, stained lenses and partially covered objects respectively.

4.4 USER REQUIREMENTS

It is vital to emphasise that the ultimate measure of good practice in this context is whether the training data leads to a machine learning model which meets the system requirements. We have already seen in Section 3 that system requirements need to be determined carefully, and that there may be various modes of operation (for example far-field detection versus accurate classification of nearer vessels). Navigational safety regulations such as the International Regulations for Preventing Collisions at Sea (COLREGs) [23] provide operational requirements on the performance of marine AV navigation systems, and consequently upon its various components, including the object detection and classification. It is vital therefore that empirical testing is carried out to assess the performance of the machine learning models against these system requirements. These empirical results will reveal whether the machine learning model meets the system requirements.

There are two ways to improve a machine learning model: improve the model selection and improve the quality of the training dataset. Model uncertainty quantification tools [24] are useful for assessing to what extent the performance of a machine learning model can be enhanced by improving the data quality. If there is scope for improvement, then an iterative process of improvement of the training set based upon the points noted in Sections 4.1 to 4.3 will be needed to improve the model and achieve the required targets. Database management protocols which support the iterative testing of an evolving database will therefore be required, including adherence to metadata standards and version control.

It is important that principles of good practice are followed for empirical testing, such as use of appropriate evaluation metrics, train and test dataset splits, cross-validation, calibration, uncertainty quantification and evaluation of robustness. However, these considerations are beyond the scope of the present document, and we refer the reader to [25], [26] for more details.

5 Summary of recommendations

We conclude by presenting a checklist for auditing the training set collection and preparation of a database of images for object detection and classification. This checklist arises naturally from the principles of good practice outlined in Section 4. This checklist is suitable to be used either by a practitioner planning to create such a dataset or by an independent auditor assessing a dataset which is already in existence.

1. **The system requirements of the marine navigation system (for example COLREGs), and consequently of the object detection and classification component, should be identified.** In particular, the variability of system requirements — for example port versus open ocean and far-field detection versus nearby classification — should be clearly understood.
2. **The images should ideally be collected *in situ*, that is using the same sensing apparatus that will be used in the deployment of the navigation system.**
3. **The possibility of using transfer learning, namely using a model trained on a large existing image database and partially retraining the model with new data, should be explored (though not necessarily adopted).** If transfer learning has been used, suitable analysis should be performed to understand the benefits or otherwise of using it as opposed to training a full model from scratch.
4. **The number of original training images should be appropriate for the complexity of the detection or classification task.** As a rule of thumb, this number should be at least in the thousands and preferably larger.
5. **The dataset should include all types of objects and statuses reasonably expected to be encountered during deployment.**
6. **The training images should be balanced across the classes.** As a rule of thumb, no class should have an order of magnitude more samples than any other. To put the requirement another way, the number of samples in each class should be at least in the hundreds and preferably larger. If not, data augmentation should be used to artificially balance the dataset.

7. **The training data should give coverage of the different viewpoints expected to be encountered during deployment, including side-on, bow-on and bow-quarter.** The collected data should give such coverage, or else data augmentation should be used to artificially introduce the necessary rotations. These augmentations should be representative of the viewpoints that are expected to be encountered in deployment.
8. **The training set should give coverage of the different weather conditions expected to be encountered during deployment.** To enable this, the lighting and noise effects of different types of weather should be modelled. The collected data should give such coverage, or else data augmentation should be used to artificially produce images under various lighting conditions. These augmentations should be representative of the actual effects of weather conditions.
9. **The training data should be robust to varying backgrounds and occlusions.** To give robustness to occlusions, the training data should include a proportion of images in which the vessel is occluded. To give robustness to varying backgrounds, the training data should either give coverage of the different types of background expected to be encountered during deployment, or else background subtraction should be performed upon the training data (and during deployment).
10. **Formal labelling protocols should be followed, and there should be evidence that labelling accuracy meets some prescribed standard when compared against ground truth.** To this end, image resolution should be adequate so that manual object detection and classification can be carried out to the prescribed standard. Image quality metrics should also be used to understand the dependence of image quality and upon resolution and noise.
11. **Database management protocols which support the iterative testing of an evolving database, such as adherence to metadata standards and version control, should be followed.** Metadata concerning viewpoints, weather, lighting conditions, background and occlusions should be included.
12. **Empirical testing of the object detection and classification model should be performed, and good practice related to empirical testing of machine learning models, such as segregation of training and test datasets, should be followed.**
13. **Empirical testing should provide evidence that the system requirements have been met.**

6 Acknowledgment

This report is based upon an output of a collaboration between NPL and Lloyd's Register Group Limited.

7 References

[1] A Jarrah, M Jamali, J Ross, P Gorsevski, J Frizado, V Bingman. Sensitivity Analysis for Optimal Parameters for Marine Radar Data Processing. American Journal of Signal Processing 3(3) p78-83, 2013.
[2] Z Zou, Z Shi, Y Guo, J Ye. Object Detection in 20 Years: A Survey, arXiv:1905.05055v2, 2019.
[3] Z Shao, W Wu , Z Wang , W Du, C Li. SeaShips: A Large-Scale Precisely Annotated Dataset for Ship Detection. IEEE Transactions on Multimedia 20(10), 2018.
[4] E Jokioinen, J Poikonen, R Jalonen, J Saarni. Remote and autonomous ships-the next steps. AAWA Position Paper, Rolls Royce plc, London, 2016.
[5] S Grini. Object Detection in Maritime Environments. Master's thesis in Engineering Cybernetics, Norwegian University of Science and Technology, 2019.

[6] W Liu, D Anguelov, D Erhan, C Szegedy, S Reed, C Fu, A Berg. SSD: Single Shot MultiBox Detector. European Conference on Computer Vision p21-37, 2016.
[7] S Ren, K He, R Girshick, J Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Neural Information Processing Systems p91-99, 2015
[8] J Redmon, S Divvala, R Girshick, A Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In European Conference on Computer Vision and Pattern Recognition p779-788, 2016.
[9] Q Yao, M Wang, Y Chen, W Dai, Y Li, W Tu, Q Yang, Y Yu. Taking the Human out of Learning Applications: A Survey on Automated Machine Learning, arXiv:1810.13306v4, 2019.
[10] Q Yang, Y Zhang, W Dai, S Pan. Transfer Learning, Cambridge University Press, 2020.
[11] GLUON MXNET Model Zoo, https://gluon-cv.mxnet.io/model_zoo/index.html
[12] N Ketkar, E Santana. Deep Learning with Python. Apress, 2017.
[13] R Patgiri, A Ahmed. Big Data: The V's of the Game Changer Paradigm. IEEE 18th International Conference on High Performance
[14] T Hastie, R Tibshirani, J Friedman. The Elements of Statistical Learning (2nd Edition). Springer Series in Statistics, 2001.
[15] B Zoph, E Cubuk, G Ghiasi, T Lin, J Shlens, Q Le. Learning Data Augmentation Strategies for Object Detection, arXiv:1906.11172v1, 2019.
[16] I Goodfellow, J Shlens, C Szegedy. Explaining and harnessing adversarial examples. Stat 1050:20, 2015.
[17] http://host.robots.ox.ac.uk/pascal/VOC/voc2007/guidelines.html
[18] L Joskowicz, D Cohen, N Caplan, J Sosna. Inter-observer variability of manual contour delineation of structures in CT. European Radiology 29(3) p1391-1399, 2019.
[19] A Torralba. How many pixels make an image? Visual neuroscience 26(1) p123-131, 2009.
[20] A Mansour, A Hassan, W Hussein and E Said. Automated vehicle detection in satellite images using deep learning. IOP Conference Series: Materials Science and Engineering 610(1), IOP Publishing, 2019.
[21] A George, S Livingston. A Survey on Full Reference Image Quality Assessment Algorithms. International Journal of Research in Engineering and Technology 2(12) p303-307, 2013.
[22] V Kamble, K Bhurchandi. No-Reference Image Quality Assessment Algorithms: A Survey. Optik 126:11-12 p1090-1097, 2015.
Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems, p17-24, 2016.
[23] http://www.imo.org/en/About/Conventions/ListOfConventions/Pages/COLREG.aspx
[24] Y Gal. Uncertainty in deep learning. University of Cambridge, 2016.
[25] P Flach. Performance Evaluation in Machine Learning: The Good, the Bad, the Ugly, and the Way Forward. In AAAI Conference on Artificial Intelligence 33 p9808-9814, 2019.
[26] S Raschka. Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:1811.12808, 2018.