Object Classification in Thermal Images using Convolutional Neural Networks for Search and Rescue Missions with Unmanned Aerial Systems

Christopher Dahlin Rodin[1,2], Luciano Netto de Lima[4], Fabio Augusto de Alcantara Andrade[1,3,4], Diego Barreto Haddad[4], Tor Arne Johansen[1] and Rune Storvold[3]

***Abstract*—In recent years, the use of Unmanned Aerial Systems (UAS) has become commonplace in a wide variety of tasks due to their relatively low cost and ease of operation. In this paper, we explore the use of UAS in maritime Search And Rescue (SAR) missions by using experimental data to detect and classify objects at the sea surface. The objects are chosen as common objects present in maritime SAR missions: a boat, a pallet, a human, and a buoy. The data consists of thermal images and Gaussian Mixture Model (GMM) is used to discriminate foreground objects from the background. Then, bounding boxes containing the object are defined and used to train a Convolutional Neural Network (CNN). The CNN achieves the average accuracy of 92.5% when evaluating a testing dataset.**

I. Introduction

Maritime Search and Rescue (SAR) operations are usually based on the drifting trajectory, which is influenced by the water streams and winds. In such operations, it is common to estimate the drift by deploying buoys with GPS sensors to transmit their positions [1]. Since changes in the environment at the search region are common, the search parameters might change many times during the mission, leading to the necessity of the reconfiguration of the mission itself. The search is usually performed using manned aircraft and vessels and is limited by the costs, the availability of human resources, and the mental and perception limitations of the human operators. All these limitations impose that a method for automatic classification of objects would be beneficial to the SAR mission as an additional assistance to the operators, due to its ability to process multiple inputs at higher speeds and with an invariable reliability rate, as it is not subject to exhaustion.

The use of Unmanned Aerial Systems (UAS) has grown rapidly, especially because of its high endurance, reduced cost, rapid deployment and flexibility. It also offers reduced risk for humans and impact on the environment compared to manned aircraft. Therefore, intelligent autonomous UAS equipped with image recognition capabilities to classify vessels, wrecks, people and objects pose as well suited tools to assist maritime SAR operations.

[1]Centre for Autonomous Marine Operations and Systems, Department of Engineering Cybernetics, Norwegian University of Science and Technology, Trondheim, Norway

[2]Maritime Robotics A/S, Trondheim, Norway

[3]NORUT Northern Research Institute, Tromsø, Norway

[4]CEFET/RJ Federal Center of Technological Education of Rio de Janeiro, Rio de Janeiro, Brazil

In these missions, it is fundamental to identifying key objects in aerial images using the techniques of object detection, classification and tracking. However, it might be more challenging to solve these classic computer vision problems when using UAS. Especially because of real-time requirements and top-down view angles. Moreover, running computationally intensive algorithms, such as deep neural networks with many filters and convolution layers, is an additional challenge due to UAS power consumption limitations, and space and weight constraint for embedded hardware.

Leira et al. [2] used thermal camera images captured by UAS to detect, classify, and track objects at the sea. The solution presented arises as a useful tool for SAR operations. The object detection algorithm used relies on static filter parameters and thresholds, which are determined manually a posteriori. The classifier used is based on the object area, the average object temperature, and its general shape. However, there are a number of scenarios where this classification would be challenging, e.g., when motion blur is present or when the object is moving across an image with varying sensor intensity, which can be caused by an uneven scene radiance or sensor noise. Therefore, a deep learning algorithm could be a more effective tool for the object classification, since it can handle variations on the images affected by environmental changes, as long as these effects are widely present in the dataset.

Convolutional Neural Networks (CNN) are the state-of-the-art deep learning tools for classification of images. Using convolution and pooling layers, it is possible to efficiently extract the most relevant features of the images. Some works were done with CNN and UAS, as in [3], where bounding boxes of images captured by a camera mounted on a UAS at a high altitude were classified in real-time into four classes: building, ground, tree and road. In [4], ground animals were detected using CNN in aerial images captured by a camera mounted on a low-cost UAS in Namibia and the step of object detection for bounding boxes prediction was also explored in the work. Sea animals were detected in aerial images in [5], where the bounding boxes were defined by the confidence of each pixel of being the center of a window containing a mammal and then a CNN is used to classify the images. Regarding the use of CNN to classify objects in aerial images taken by a UAS in maritime environments, a work was done by [6], where RGB images were used and bounding boxes

were classified into two classes: $boat$ or $notboat$. SAR, CNN and UAS are used together in [7], where near real-time object detection was performed by a UAS for avalanche SAR missions. A pre-trained CNN did the object detection and a Support Vector Machine (SVM) was used to classify the proposed human bodies. All of these works were done using datasets of RGB images, but there are also some works using CNN with thermal images, as in [8], to monitor machine health and in [9], to detect pedestrians. However, there were not found works using CNN to classify objects at the sea in aerial thermal imagery and this is particularly important in night time low visibility SAR operations.

In this paper, a CNN is trained to classify boats, buoys, people and pallets in images captured by a thermal camera mounted on a fixed-wing UAS. The foreground objects were detected by modeling the background as a mixture of Gaussian distributions and subtracting the foreground [10]. This method is computationally cheaper than other object proposal methods such as sliding windows [11] or selective search [12] because it is particularly suitable for thermal images at the sea, as there are two modes present in the distribution: the radiance reflected from the sky, and the heat emitted from the sea [13]. Subsequently a window was fitted around the objects and padded to ensure that the full objects were included in the window. One other novelty brought by this study is the use of the estimated observed area as an extra feature in the fully connected layer of the CNN.

II. Dataset

The dataset consists of images captured by a thermal camera mounted on a fixed-wing UAS. The thermal camera used is a FLIR Tau2, which provides analogue video data at a 640 $\times$ 512 pixels resolution. The lens has a focal length of 19 mm, which produces a 32° $\times$ 26° angle of view. The analogue video data is converted to digital using a 16 bit analogue-to-digital converter. In order to create 8 bit images, the 16 bit images are normalized between 0 and 255 for the smallest and largest intensity in the full dataset. The UAS was also equipped with an Inertial Measurement Unit (IMU) and Global Navigation Satellite System (GNSS) unit, in order to find the surface area of the objects in the images (see section II-C).

Four different objects were placed in the ocean: a 26 feet boat, an euro pallet, a human wearing an immersion suit, and a buoy with a 60 cm diameter. The objects are chosen as common objects present in maritime SAR missions, where e.g. pallets are a common object to search for when trying to locate fish aggregating devices. The objects can be seen in higher resolution visual light camera images in figure 1. The human varied between different actions during the experiment: floating horizontally on the surface (creating a large, long surface), swimming (creating a medium sized surface varying in shape), and standing vertically (creating a small surface, down to 20 cm across).

The total dataset consists of around 22,000 images were captured during a time span of 50 minutes. The objects were only fully inside in the camera field of view in a limited subset of the full dataset, leading to a smaller number of objects used in the CNN.

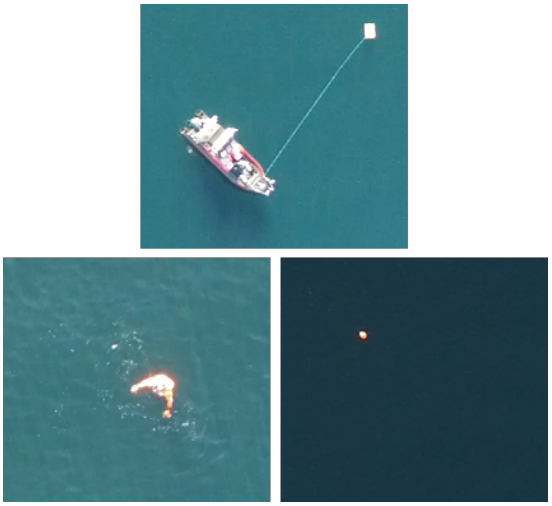

Fig. 1. The different objects present in the scene, as captured by a higher resolution, visual light camera. Top: boat and pallet, bottom left: human, bottom right: buoy. The images were captured at different altitudes.

Various imperfections were present in the images. Several images contain motion blur caused by the dynamics of the UAS. This effect is minor for larger objects such as the boat, however for smaller objects such as a human head sticking out from the water, it can greatly affect the shape, size, and intensity of the object. See figure 2 for an example of how motion blur changes the object size and dimensions. The pixel intensity is also varying throughout each image, which makes the same object take on intensities between 97 to 110 in an example 8 bit image sequence. This might be caused by noise in the uncooled thermal image sensor, internal camera intensity calibrations, or varying scene radiance. The background varies between 81 and 98 in the same image sequence. See figure 3 for an average of all images without objects, where the intensity variation can be seen.

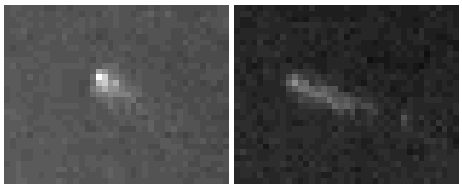

Fig. 2. The same object without motion blur (left) and with motion blur (right). The shape, size, and intensity is greatly affected.

In order to find the objects in the images and label them, their boundaries were first found (section II-A). The objects were then automatically labeled based on the physical area of the boundary (see section II-A for definition of the physical area) and finally manually corrected (section II-B). The number of labeled objects in the dataset used in the CNN is summarized in table I.

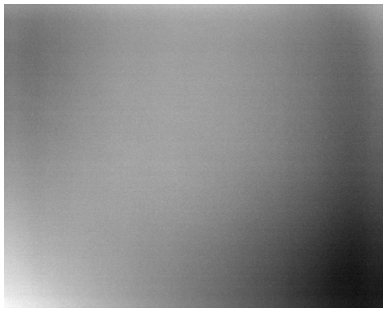

Fig. 3. The mean image without any objects, with its intensity stretched to show the varying image intensity.

TABLE I
NUMBER OF LABELED OBJECTS IN THE DATASET

Boats	620
Pallets	739
Humans	313
Buoys	276

A. Bounding Boxes

In order to discriminate the foreground objects from the background in the images, the background pixels were modeled using an adaptive background Gaussian Mixture Model (GMM) [10]. The GMM provides robust foreground segmentation and is suitable for thermal images at the sea since there are two modes present in the distribution: the radiance reflected from the sky, and the heat emitted from the sea. It can also model the sensor varying sensor noise, but might fail when the thermal camera is performing sudden noise corrections. The algorithm was implemented using the Background Subtraction Library [14]. A study by Borghgraef et al. [13] showed that more advanced algorithms, such as ViBe and the behaviour subtraction algorithm, outperformed the GMM for detecting objects at the sea surface in thermal images. However, this was for a static camera at a highly slant angle, which means that the study is not completely applicable to the scenario of this paper. For this project, the GMM is chosen as a good balance between robustness and simplicity. The bounding box was then defined as the smallest box that encloses the boundary of the object. The bounding boxes of all objects were then padded to the size of the largest bounding box found in the dataset. See figure 4 for a sample boundary and bounding box.

B. Labeling

In order to use the extracted foreground objects in the supervised learning algorithm, each object needs to be properly labeled. The objects were first assigned one of three labels based on their observed area in square meters (see section II-C) - boat, pallet, or human/buoy. Each label was then manually verified and adjusted if deemed incorrect.
Due to the low ground resolution and their similar dimensions, discriminating humans from buoys was not possible only using the size as a criterion or by looking at individual images due to the varying shapes of the human and other effects, e.g., motion blur. A manual classification was therefore done by analyzing the shape of each object appearing in a sequence of images, taking into consideration that the buoy is completely round while the human has a more elliptical and varying shape. See figure 5 and 6 for a comparison between a boat, a pallet, a human, and a buoy in the images.

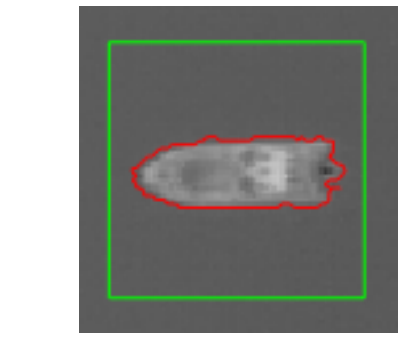

Fig. 4. The border around the extracted foreground object (red), and the bounding box (green).

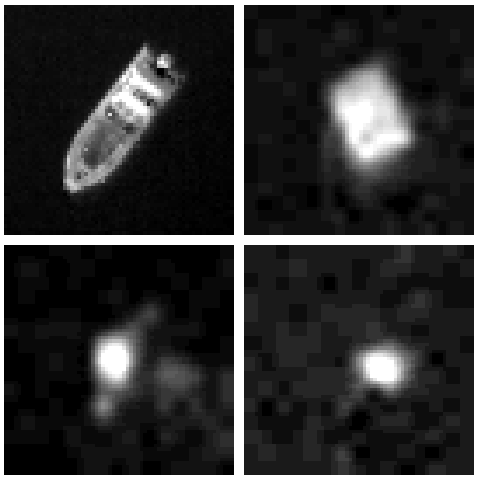

Fig. 5. The different objects which were labeled. Top left: boat, top right: pallet, bottom left: human, bottom right: buoy. The images are scaled to show more detail.

C. Object Area

The observed area of each object in square meters is used as an extra feature in the fully connected layer of the CNN. The real observed area is defined as the area of the object as seen by the camera, when projected at the plane spanning the North and East axes (NE-plane) at an altitude of zero ($D = 0$). See figure 7 for a visual description of the observed area of an object.

The pinhole camera model [15] is used to calculate the observed area of the boundary of each object. First, the observed area of the center pixel within the boundary is calculated, which is then calculated by the number of pixels within the boundary. In order to perform these calculations, it is necessary to know the attitude and altitude of the camera.

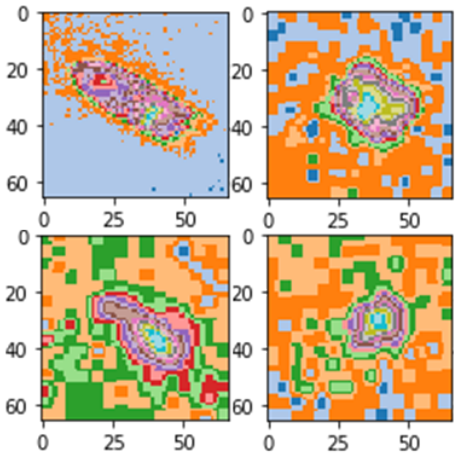

Fig. 6. Objects in a different color map, in order to aid in manually discriminating humans from buoys. Top left: boat, top right: pallet, bottom left: human, bottom right: buoy.

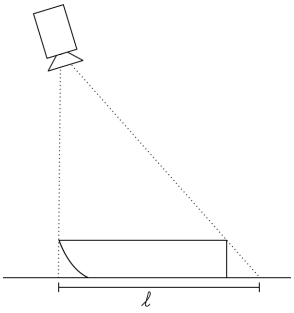

Fig. 7. The observed length, l, of an object. The observed area is the corresponding feature in two dimensions.

This data is obtained from the IMU and GNSS data, and is represented in the form of the extrinsic camera matrix. The intrinsic matrix is calculated from the camera specification. No lens distortion is considered due to the difficulty of performing distortion estimation for a thermal camera. Since the angle of view is relatively small, it is not likely to cause any major distortion.

The observed area distributions for each object is show in figure 8. It can be seen while boats and pallets can be almost completely classified based on their observed area (with minor overlap between pallets and humans), while there are major overlaps between humans and buoys. This is however an artifact of this dataset – in other datasets, buoys and boats can take on a variety of sizes. As previously mentioned, humans can take on a wide variety of sizes due to the different poses.

The real observed area of a buoy with a diameter of 60 cm should be 0.28 m^2. As can be seen in figure 8, the area is biased towards higher values. One reason for this is that a 60 cm circle can appear in 16 pixels (figure 9), when the observed area of each pixel is 17.9 cm – which is the case when flying at an altitude of 200 m with no roll or pitch using the camera system used in the experiment performed. This

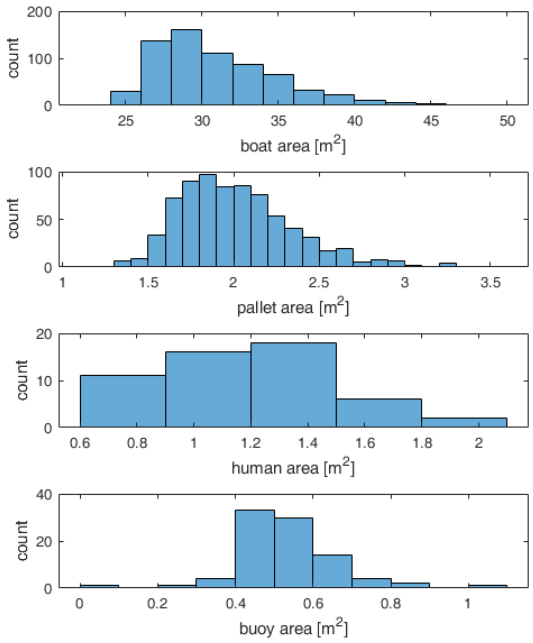

Fig. 8. Distribution of observed areas for the objects present. From top to bottom: boats, pallets, humans, buoys.

gives an observed area of 0.52 m^2. Additionally, the motion blur causes the object to appear larger than it really is.

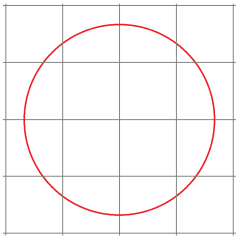

Fig. 9. A buoy with a diameter of 60 cm can appear in 16 pixels, when the observed width and height of each pixel is 17.9 cm.

III. Convolutional Neural Network

Traditionally, supervised learning based image analysis combines feature extraction with classical machine learning methods [16]. Convolutional Neural Network (CNN) is an alternative and recent trend for image classification that has been proven to produce high accuracy in image classification tasks [17] without requiring any task-specific feature engineering [18]. It is considered the most successful machine learning model in recent years [19] and the most eminent method in computer vision [20], in part because it consists of a powerful image features extractor [21].

A CNN is based on neuroscience researches about the processes that mammalian visual cortex uses to recognize images [22]. Several basic stages typically compose a CNN. Each stage consists of concatenation of convolution, normalization,

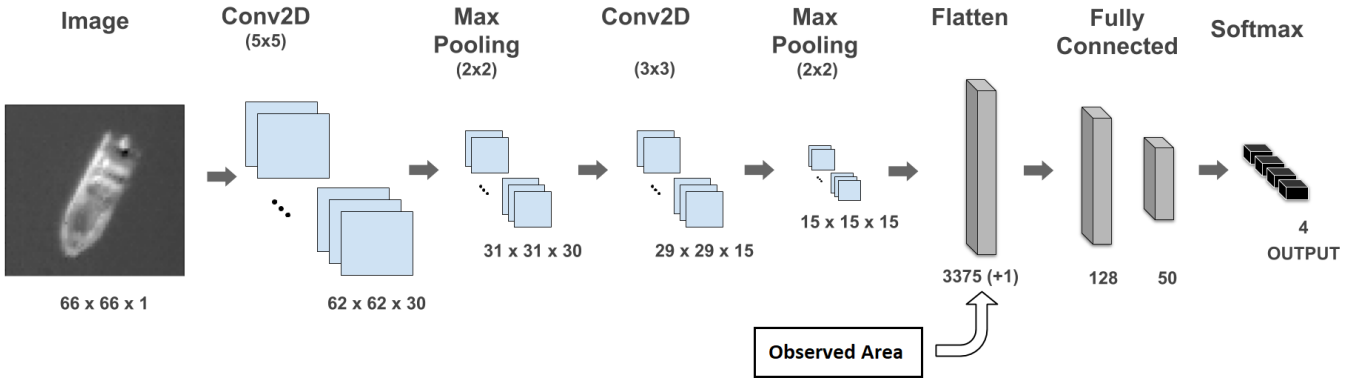

Fig. 10. CNN architecture. If the Estimated Object Area is used, one more element is added in the flatten layer, resulting in an output of 3.376 elements.

activation (nonlinear), and pooling layers [23]. In this work, two distinct architectures were used. The difference between them was the employment of the observed object area as an input of the fully connected layer.

A. Architecture

Regarding the architecture (see Figure 10), the proposed network starts with an input layer containing the image window. This layer is followed by the convolution layer which produces 30 feature maps from filters of size 5×5. The convolutional layer has a set of learnable filters called kernels. By the convolution between one kernel and a chunk of values from the layer, a feature map is generated, which consists of a presence representation of a specific feature in the image. The next layer is a max pooling with a filter of a size of 2×2, whose purpose consists of extracting the hierarchical features of the input image [24]. It works by mapping the bigger value from a 2×2 chunk in only one value in the next layer. Pooling helps to make the representation approximately invariant to small translations of the input [22]. This function is also responsible for reducing the width and the height of the feature map. Reducing this dimension, the computational demand is reduced due to the reduction of the number of parameters, which helps to avoid over-fitting. Then, it comes another convolution layer with 15 feature maps of size 3×3, and finally one more max pooling layer composed by 2×2 filters.

The next layer is a flatten layer used to adjust the tensor dimensions to the fully connected layers. At this point, the two architectures become different. One network has the estimated object size as an input and the other one does not. Then, it follows 2 fully connected layers composed respectively by 128 and 50 neurons using the rectifier activation function. The last fully connected layer is used to provide the predicted classification, using the softmax activation function. This is the most common solution for the regulation of the output values within the range from 0 to 1 [25], which assigns a multinomial probability distribution to the output vector [26]. It enhances the discriminative modeling power of the CNN, providing the probability of the input to belong to each possible class, namely: boat, buoys, human or pallet.

B. Dropout

One technique widely used to improve the performance and avoid the over-fitting (which is often a serious problem for a CNN [27]) is the dropout. The term "dropout" refers to dropping out units (hidden and visible) in a neural network during the training phase. By dropping a unit out, it means temporarily removing it from the network in the current epoch, along with all its incoming and outgoing connections [28].

The dropout parameter controlled in this paper was the independent probability of deactivating a neuron. This parameter was tested with 2 values: 0.2 and 0.5.

C. Cross-validation

In order to evaluate the generalization capacity of the classifiers, it is preferable that the set used in the evaluation process is different from the one used during the training process. Typically, the formation of the training and test set is based on non-repetitive sampling techniques, such as the k-fold cross validation method [29]. Cross-validation is a robust statistical technique for estimating the true risk function [30] (or the generalization error), the most important operational performance of a trained network [31].

In this paper, the database is divided into five sets of equal size. During each execution of the algorithm, one set is chosen to be out of the training phase, which will be the corresponding test set. This process is repeated five times, and the performance metric is inferred for each of the sets that were left out of the training process. The value of the overall performance will be defined by the average of the values obtained for each of the five executions.

D. Stopping criterion

In the neural network training phase, a stopping criterion has to be used to stop the training of the neural network. There are some stopping criterion, such as number of epochs, minimum mean square error, early stopping, among others. To ensure that the training was stopped in a way to provide an

appropriate generalization, a validation based early stopping is used in this work [32].

A small amount of the training dataset is sorted out to be used as a validation set. At each epoch, the performance index is evaluated for the new training set and for the validation set. When the performance of the validation set stops to decrease, i.e., when the training starts to over-fit, the training is stopped.

IV. Results

After testing the convergence for different parameters, the very first CNN was chosen and trained for 5-folds of images in 8 bit format, without taking into consideration the estimated size of the objects. The maximum number of epochs was 500 and the early stopping was set to stop the training after 50 validation evaluations without improvements. The validation split was 0.18 and dropout was 0.50. After doing 10 executions to get an indicative statistical performance, the average accuracy was 92.0% with 0.50% of standard deviation. This result shows that the configuration of the training algorithm was well set, so that the performances of all executions for all folds were similar.

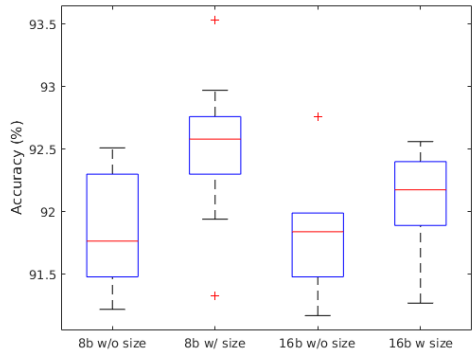

Fig. 11. Accuracy for different configurations.

When using the estimated objects size as an extra input of the fully connected layer, the resulted accuracy is higher as shown in figure 11 for 8 bit and 16 bit images, achieving 92.5% and 92.1% of accuracy, respectively. Regarding the classification, it is possible to notice that classifying between buoys and humans is a challenge as seen in the Confusion Matrix (table II). However, the use of the estimated object size helps the CNN to get better results (table III). When looking to the Confusion Matrix for buoys, there are fewer cases when the buoy is classified as a human. There is even a case of a boat being classified as a pallet when the estimated object size was not used (table II).

The ability of the CNN to classify humans vs. buoys is further investigated in figures 12 and 13, where the probability of each human and buoy test sample being either a human or a buoy are shown. In the humans samples, it is possible to notice that it is easier for the CNN to differentiate them from buoys. However, when analyzing the classification probabilities of the

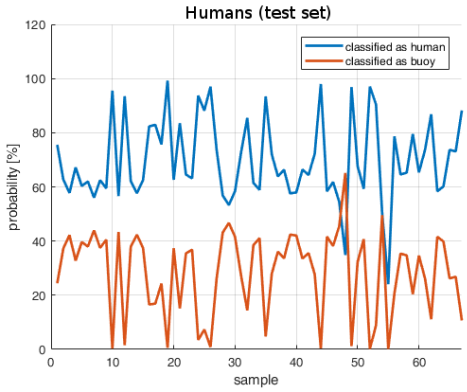

Fig. 12. Probability that a human is either a human (blue) or buoy (red).

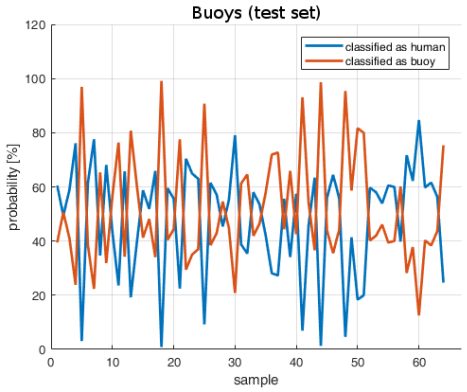

Fig. 13. Probability that a buoy is either a human (blue) or buoy (red).

buoys samples, it shows that it is very challenging to the CNN to decide if it is a buoy or a human.

Regarding the comparison of the performance between 8 and 16 bit images, even with the fact that the 16 bit images have more detail, the accuracy was sightly higher for the configuration with 8 bit images. This might be caused by the reduction of noise when reducing the bit depth, e.g., small intensity variations in the sea surface.

TABLE II
Confusion Matrix for 8 bit images without using object size.

		Predicted			
		Boat	Human	Buoy	Pallet
True	Boat	128	0	0	1
	Human	0	36	18	0
	Buoy	0	15	45	1
	Pallet	0	0	1	145

Images of three vessels from an external dataset (Figure 14) were used to evaluate the performance of the CNNs.

The observed areas of the vessels were estimated and chosen to be 200 m^2, 25 m^2 and 10 m^2 for the big, medium and small boat, respectively.

When using the CNN where the observed area is not used, all the three samples achieved 100% of probability of being a boat. Regarding the CNN where the observed area is used

TABLE III
CONFUSION MATRIX FOR 8 BIT IMAGES USING OBJECT SIZE.

		Predicted			
		Boat	Human	Buoy	Pallet
True	Boat	114	0	0	0
	Human	0	41	13	0
	Buoy	0	7	64	0
	Pallet	0	0	1	150

as an input, the big and medium vessel achieved 100% of probability of being a boat and the small vessel was classified as a pallet.

Fig. 14. Images of vessels from an external dataset. Top left: big vessel, top right: medium vessel, bottom: small vessel.

V. DISCUSSION

To obtain the images of the objects, the bounding boxes were defined as the smallest box that encloses the boundary of the object. Then, the bounding boxes of all objects were padded to the size of the largest bounding box found for the set of objects of the same class. However, in the SAR mission, it is not possible to know the class of the object a priori, therefore, another strategy need to be used to define the bounding boxes. One solution is to define a specific amount of pixels to be added to the boundaries of the detected object, to ensure that the whole object will be inside the bounding box, occupying as much of it as possible. This amount of pixels should be defined by the altitude of the UAS when the image is being captured and also the estimated size of the object. Thus, effects by the distance to the scene would also be mitigated.

The observed area, as well as its appearance in the thermal images, are greatly affected by motion blur. For larger objects this does not pose a major problem, but for objects just a few pixels in size, the difference can be of major concern. An actively stabilized gimbal and carefully chosen shutter times based on the UAS dynamics could prevent this. Another mitigating solution would be to collect a larger dataset in order to be able to properly classify objects even with motion blur.

The major difficulty of the CNN is to properly distinguish between humans and buoys, which is likely due to the low resolution of the thermal image sensor and relative high altitude, resulting in the objects being represented by very few pixels in the images. In real world maritime SAR missions, however, a buoy being classified as a human would not be a major issue, as the operator would still be notified, and could dismiss the notification from the CNN. Incorrectly classifying a human as a buoy could potentially cause a missed person, but could be solved by lowering the human probability threshold for notifying the operator.

In the dataset used in this work, all boat samples have similar observed areas. Therefore, when evaluating the classification performance for images of vessels from an external dataset, the result was superior when using the CNN where the object area was not considered as an input. However, the generalization power of the CNN containing the observed area can be improved by using a dataset with more samples of boats of different sizes. Also, in general, it is beneficial to have more data, especially at different angles and altitudes.

VI. FUTURE WORK

In the mission carried out to gather the data used for this work, an Electro-Optical (EO) camera was also equipped in the UAS to capture RGB images. However, the thermal and the RGB images were not obtained during the same flight, so it is not possible to use the two images together as inputs of the same CNN.

Therefore, one of the next steps is to develop a CNN to classify the objects in the RGB images, as the work done by [6]. Then, investigating a method to use both datasets together, for example, trying to use the results of each independent CNN multiplying the probability of each sample to be of each one of the classes.

For the CNN proposed by this work, the classification of images of vessels obtained in another mission in totally different conditions was evaluated. However, it is important to evaluate the classification for images of humans, buoys and pallets as well. Thus, it would be possible to estimate how well the CNN could perform in a real mission.

Another aspect that needs to be evaluated is how to improve the classification between humans and buoys, especially in the case of boys, where the calculated probability of one buoy sample being a buoys is very close to the probability of being a human. Examples of this approach would be to use an actively stabilized and sweeping gimbal together with a lens with higher focal length, in order to get a higher ground resolution.

VII. CONCLUSION

In this paper, the algorithm for detecting and classifying objects at the sea surface in thermal camera images taken by Unmanned Aerial Systems (UAS) has been discussed. The algorithm uses a Gaussian Mixture Model (GMM) in order to discriminate foreground objects from the background in the images. Then, bounding boxes around the objects are defined and used to train and test a Convolutional Neural Network (CNN). The observed area of the objects was also estimated and used as an input. The CNN was evaluated using the k-fold method with 5 folds and achieved an average of 92.5% of accuracy. Images of vessels from an external dataset were

also evaluated and all of them achieved 100% of probability of being a boat when using the CNN where the observed area was not used. The results and the robustness of the CNN algorithm prove it to be a useful tool to assist maritime SAR operations, and be a central part in a future fully autonomous UAS operation in SAR missions.

ACKNOWLEDGMENT

This project has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 642153. The information presented in this paper reflects only the author's view, and the Agency is not responsible for any use that may be made of the information it contains. The work was also supported by the Research Council of Norway through the Centre for Autonomous Marine Operations and Systems (NTNU-AMOS), grant number 223254. The authors would like to express their gratitude to the people from Maritime Robotics AS and NTNU UAV Lab involved in the experiment performed for this paper. This work was also partially developed with the support of CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico (National Council for Scientific and Technological Development, in Brazil), CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Coordination for Enhancement of Higher Education Personnel, in Brazil), and FAPERJ - Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (Rio de Janeiro Research Support Foundation, in Brazil).

REFERENCES

[1] International Maritime Organization and International Civil Aviation Organization, *IAMSAR Manual: Vol. 2: Mission Co-Ordination*. International Maritime Organization, 2007.

[2] F. S. Leira, T. A. Johansen, and T. I. Fossen, "Automatic detection, classification and tracking of objects in the ocean surface from uavs using a thermal camera," in *IEEE Aerospace Conference*, March 2015.

[3] C. Sheppard and M. Rahnemoonfar, "Real-time scene understanding for uav imagery based on deep convolutional neural networks," in *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017.

[4] B. Kellenberger, M. Volpi, and D. Tuia, "Fast animal detection in uav images using convolutional neural networks," in *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017.

[5] F. Maire, L. Mejias, and A. Hodgson, "A convolutional neural network for automatic analysis of aerial imagery," in *Digital lmage Computing: Techniques and Applications (DlCTA), 2014 International Conference on*, 2014.

[6] G. Cruz and A. Bernardino, "Aerial detection in maritime scenarios using convolutional neural networks," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2016, pp. 373–384.

[7] M. B. Bejiga, A. Zeggada, and F. Melgani, "Convolutional neural networks for near real-time object detection from uav imagery in avalanche search and rescue operations," in *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2016.

[8] O. Janssens, R. Van de Walle, M. Loccufier, and S. Van Hoecke, "Deep learning for infrared thermal image based machine health monitoring," *IEEE/ASME Transactions on Mechatronics*, 2017.

[9] V. John, S. Mita, Z. Liu, and B. Qi, "Pedestrian detection in thermal images using adaptive fuzzy c-means clustering and convolutional neural networks," in *Machine Vision Applications (MVA), 14th IAPR International Conference on*, 2015.

[10] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1999.

[11] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *International Journal of Computer Vision*, vol. 38, no. 1, pp. 15–33, 2000.

[12] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[13] A. Borghgraef, O. Barnich, F. Lapierre, M. Van Droogenbroeck, W. Philips, and M. Acheroy, "An evaluation of pixel-based methods for the detection of floating objects on the sea surface," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, no. 1, Feb 2010.

[14] A. Sobral, "BGSLibrary: An opencv c++ background subtraction library," in *IX Workshop de Viso Computacional (WVC'2013)*, Rio de Janeiro, Brazil, Jun 2013. [Online]. Available: https://github.com/andrewssobral/bgslibrary

[15] S. Prince, *Computer Vision: Models, Learning, and Inference*, ser. Computer Vision: Models, Learning, and Inference. Cambridge University Press, 2012.

[16] M. Valkonen, K. Kartasalo, K. Liimatainen, M. Nykter, L. Latonen, and P. Ruusuvuori, "Dual structured convolutional neural network with feature augmentation for quantitative characterization of tissue histology," in *International Conference on Computer Vision Workshops (ICCVW)*, Oct 2017.

[17] G. Forslid, H. Wieslander, E. Bengtsson, C. Whlby, J. M. Hirsch, C. R. Stark, and S. K. Sadanandan, "Deep convolutional neural networks for detecting cellular changes due to malignancy," in *International Conference on Computer Vision Workshops (ICCVW)*, Oct 2017.

[18] T. Nakazawa and D. V. Kulkarni, "Wafer map defect pattern classification and image retrieval using convolutional neural network," *IEEE Transactions on Semiconductor Manufacturing*, 2018.

[19] S. J. Lee, T. Chen, L. Yu, and C. H. Lai, "Image classification based on the boost convolutional neural network," *IEEE Access*, vol. 6, 2018.

[20] K. Chen, K. Fu, M. Yan, X. Gao, X. Sun, and X. Wei, "Semantic segmentation of aerial images with shuffling convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 2, pp. 173–177, Feb 2018.

[21] C. Bentes, D. Velotto, and B. Tings, "Ship classification in terrasar-x images with convolutional neural networks," *IEEE Journal of Oceanic Engineering*, vol. 43, no. 1, pp. 258–266, Jan 2018.

[22] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[23] Y. Wang, J. Lin, and Z. Wang, "An energy-efficient architecture for binary weight convolutional neural networks," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 26, no. 2, pp. 280–293, Feb 2018.

[24] N. Liu, J. Han, T. Liu, and X. Li, "Learning to predict eye fixations via multiresolution convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 2, pp. 392–404, Feb 2018.

[25] S. Lim and D. Lee, "Stable improved softmax using constant normalisation," *Electronics Letters*, vol. 53, no. 23, pp. 1504–1506, 2017.

[26] D. Yu and L. Deng, *Automatic speech recognition: A deep learning approach*. Springer, 2014.

[27] N. Yang, H. Tang, H. Sun, and X. Yang, "Dropband: A simple and effective method for promoting the scene classification accuracy of convolutional neural networks for vhr remote sensing imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 2, pp. 257–261, Feb 2018.

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[29] L. Rokach, *Pattern classification using ensemble methods*. World Scientific, 2010, vol. 75.

[30] S. Amari, N. Murata, K. R. Muller, M. Finke, and H. H. Yang, "Asymptotic statistical theory of overtraining and cross-validation," *IEEE Transactions on Neural Networks*, vol. 8, no. 5, pp. 985–996, Sep 1997.

[31] G. N. Karystinos and D. A. Pados, "On overfitting, generalization, and randomly expanded training sets," *IEEE Transactions on Neural Networks*, vol. 11, no. 5, pp. 1050–1057, Sep 2000.

[32] L. Prechelt, "Early stopping-but when?" *Neural Networks: Tricks of the trade*, pp. 553–553, 1998.