TARGET CLASSIFICATION USING MULTI-VIEW SYNTHETIC APERTURE SONAR IMAGERY

Benoît d'Alès de Corbet, David P. Williams and Samantha Dugelay

Science and Technology Organization - Centre for Maritime Research and Experimentation
La Spezia, SP, ITALY

Benoît d'Alès de Corbet, STO-CMRE, Viale San Bartolomeo 400, 19126 La Spezia, email: benoit.dales@cmre.nato.int

Abstract: *Synthetic Aperture Sonar (SAS) can be used commonly in many different underwater applications such as mine countermeasures, habitat mapping and archeology. It offers high resolution images over wide swath areas. A single-view SAS image however may lack critical information for an object classification task (a mine hidden by a rock for example, or a partial image). Instead, multi-view images of the same scene could provide much richer information. In this context, Thales developed a sonar capable of processing three views under different angles simultaneously. CMRE and Thales have teamed up to investigate deep learning applications for multi-view. This paper demonstrates the potential benefits of such a technology in the matter of target classification. The data used for this study are real SAS data collected at sea trials by the MUSCLE. The preliminary work compares different ways of classifying with Convolutional Neural Network (CNN) architectures. Transfer learning is also performed from pre-trained models.*

Keywords: *synthetic aperture sonar, multi-view, convolutional neural network, classification.*

1. INTRODUCTION

Synthetic Aperture Sonar (SAS) is commonly used in underwater applications such as mine countermeasures [1], habitat mapping [2] and archeology [3]. The principle is based on the coherent combination of successive pings synthesized in a large array. Compared to real aperture, the technique delivers a much higher resolution independent of range and frequency. Indeed, it is able to image the seabed with centimeter resolution up to hundreds of meters range. To perform, the sensor needs to move at a constant speed following a known path. In the context of mine countermeasures, a SAS payload is embedded on an Autonomous Underwater Vehicle (AUV). This sensor platform has the particularity of incorporating an Automatic Target Recognition (ATR) system [4]. During missions, large quantities of data are collected and processed in order to detect a Region of Interest (ROI) and classify it. In recent years, Convolutional Neural Network (CNN) has led to very good performance on sonar image classification [5]. However, this stage is largely dependent on the image quality and on the view-angle of the object. Indeed, if the latter is hidden behind a rock, it could mislead the process. Consequently, the ATR system adjusts the AUV's mission path to revisit potential target locations. The method is nevertheless time-consuming and uncertain.

A new approach is proposed by Thales [6]. They built a sonar able to simultaneously capture three high-resolution views under different angles. This system provides a new way of classifying mines. Rather than a single-view, the classification stage can exploit much richer information, either about the mine-like object or its environment. With the great expertise of the CMRE in ATR, Thales and the Centre have teamed up to explore the possibilities in the matter of target classification.

In the present study, Section 2 describes the multi-view classification process, from the proposed CNN model architectures to data processing. Section 3 provides the analysis of the results and Section 4 finally concludes the study.

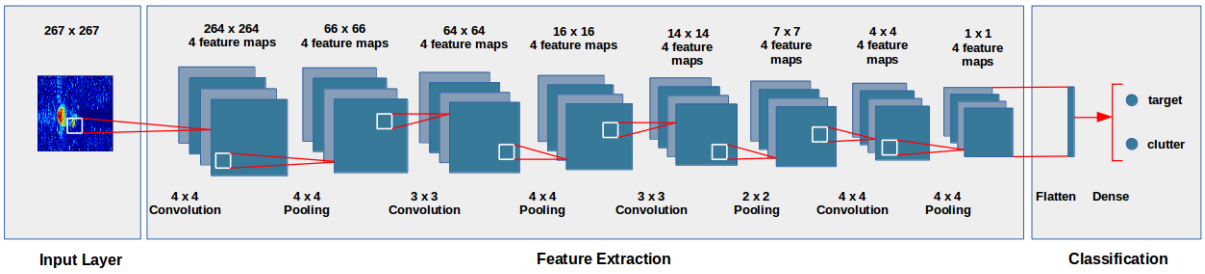
2. MULTI-VIEW CLASSIFICATION

2.1. Convolutional Neural Networks

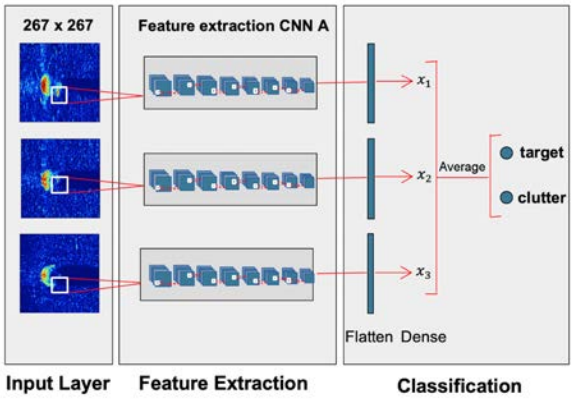
A CNN is a specific type of neural network which is specially used for image analysis. The CNN model architecture in Fig. 1a is used for the single-view classification (CNN A). It is taken from [7] where it has already proven its efficiency. It consists of a stack of convolutional and average pooling layers. On one hand, the convolutional layer applies a linear convolutional filter followed by a nonlinear activation function (ReLU). Thus, feature maps are generated from the input data. On the other hand, the pooling layer reduces the size of the image by only keeping the most important pixels. It distorts the image by losing the precise pixel positions. This helps to limit the risks of overfitting. After the feature extraction stage, the neural network finishes via a fully connected layer (also called dense layer). It combines all the specific characteristics detected by the previous layers and gives as result a binary prediction (i.e. clutter or target) with a sigmoid activation. This model serves as a basis for the building of our multi-view classifiers.

We propose two ways of classifying multi-view images. The first one consists in classifying each image independently with the previous model (CNN A). The resulting predictions are then averaged to give the final result. Let us call this model CNN B (*cf.* Fig. 1b).

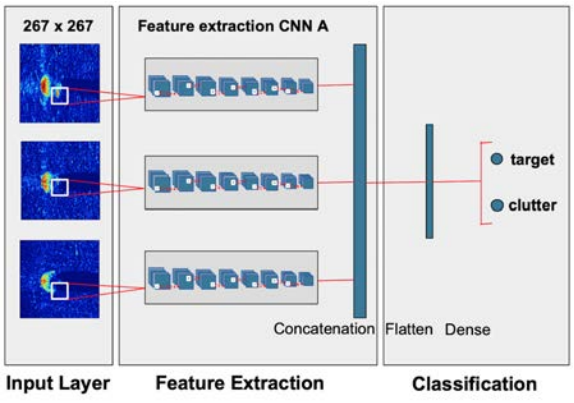
For the second one (*cf.* Fig. 1c), we have 3 inputs that follow the same feature extraction process on 3 independent branches. Then, the output branches are concatenated, just before the classification stage. Finally, the fully connected layer gives a prediction according to a sigmoid activation. Let us call this model CNN C.



(a) Single-view CNN architecture (CNN A)



(b) CNN architecture with averaged predictions (CNN B). x_1, x_2 and x_3 represent the prediction values of each image



(c) CNN architecture with the concatenation of the three branches (CNN C)

Fig. 1: CNN architectures

2.2. Datasets

Since 2007, CMRE has been conducting many sea trials. The Centre has collected a significant amount of SAS data with the SAS-equiped AUV called MUSCLE. We find there various seafloor compositions and mine-like objects with high-resolution imagery. The system has a center frequency of 300 kHz and a bandwidth of 60 kHz. It provides a resolution cell of 2.5 cm in the along-track dimension and 1.25 cm in the range dimension.

The training of the single-view CNN model in Fig. 1a is already performed with data from 8 expeditions conducted between 2008 and 2013 (*cf.* Table 1). Therefore, only weights resulting from the training are used. The data we have focuses on object detection from a specific trial conducted in 2014, in Italy. The detection process is achieved by the algorithm detailed in [8]. It delivers a rich database content such as muds, ripples, posidonias and various shapes of targets (symmetric and asymmetric). However, we do not use all the database. In order to build a multi-view dataset, only object detections with three different views are kept. It implies

that we are looking for the detections for which the AUV revisited the scene. The original dataset is consequently reduced. In addition, by grouping the three views together, the number of classifications is then divided by three. All images are labeled, either as clutter (class 0) or as target (class 1).

By building this multi-view dataset, we try to get closer to what the Thales sonar imagery could provide. However, we observe that the sonar from Thales processes SAS images under three fixed orientations. On our side, the angle view of the object is linked to the way the AUV revisits the scene. In order to deal with this important feature, we propose the following dataset (*cf.* Fig. 2). It consists in dividing the image wavenumber spectrum in three equal parts where each part corresponds to a view. The process is based on [9]. It allows to obtain three different orientations with a known deviation of a few degrees (approximately 5°). This generated dataset is denoted DS2.

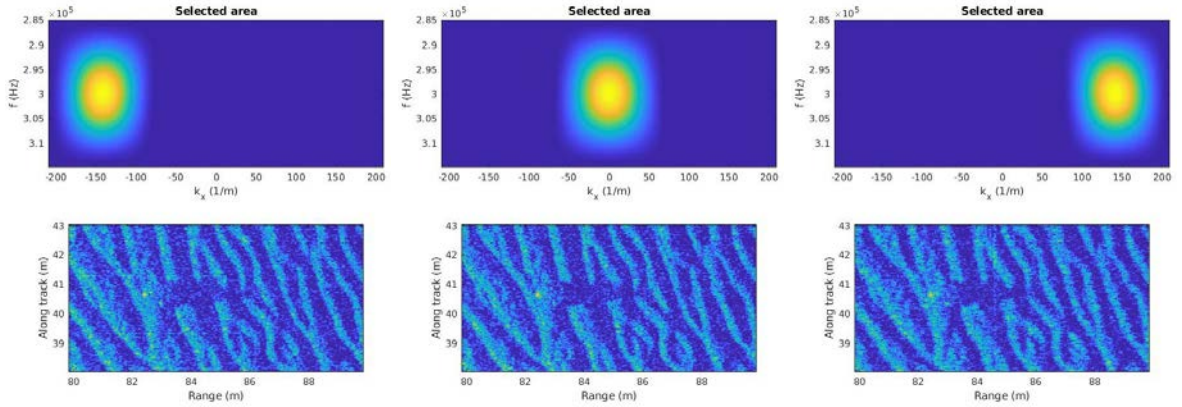


Fig. 2: Generated multi-view images

Table 1: SAS datasets

Sea trials	Type of dataset	Targets	Clutter
2008-2013 - Latvia, Italy, Spain	Single-view images	2912	29280
2014 - Italy	Multi-view images (DS1)	116×3	40×3
	Generated multi-view images (DS2)	348×3	120×3

Before input images are used by CNNs, a preprocessing stage is performed. The complex SAS images are first interpolated by converting pixels into squares covering 1.5 cm in each direction. Then, they are normalized so that pixel values fall between a fixed range, $[-1,1]$. The normalized image x'_{ij} is:

$$x'_{ij} = 2 \times \frac{\min\left(40,\ \max\left(0,\ x_{ij} - \frac{1}{N}\sum_{m,n} x_{mn}\right)\right)}{40} - 1, \tag{1}$$

where x_{ij} represent the unnormalized pixel values and N is the total number of pixels in the image.

Finally, the image is centered around the object highlight and cropped to an area of approximately 4.005 m $\times$ 4.005 m (corresponding to a size of 267 pixels $\times$ 267 pixels).

3. EXPERIMENTAL RESULTS

With these multi-view datasets (DS1 and DS2), we aim to experiment the benefits of the multi-view CNN architectures compare to the single-view CNN architecture. It is necessary for that to re-train CNN C, once with DS1 and once with DS2. The weights associated with the convolutional layers from the feature extraction stage in CNN A (*cf.* Fig. 1a) are kept for the three branches. After the concatenation of the independent branches, the parameters from the fully connected layer needs to be re-trained, according to both multi-view datasets (DS1 and DS2). Thus, we freeze weights from the feature extraction stage during the training. The other parameters are trained by minimizing the binary cross-entropy loss using back-propagation through RMSprop with a learning rate of 10^{-3}. One third of the respective datasets are used for the training. It represents a small amount of data but only 13 parameters have to be trained. Data augmentation is accomplished by randomly mixing the order of the input images in the 3 branches of the CNN model. To compare fairly the different datasets DS1 and DS2, the train-sets is based on the same original database imagery. The same applies for the test-sets.

In order to find out how accurate are the predictions, we use different performance metrics, such as precision, recall, F1-score and the Area Under the Curve (AUC). These measures are expressed below and the parameters are detailed in Table 2. As for the classification results, they are shown in Table 3.

Table 2: Definition of TP, FN, FP and TN

	Actual positive	Actual negative
Predicted positive	True Positive (TP)	False Positive (FP)
Predicted negative	False Negative (FN)	True Negative (TN)

On one hand, the precision metric is a ratio that shows us the correctly predicted targets to the total predicted targets. On the other hand, the recall metric is used as a ratio between the correctly predicted targets and the total true targets. Finally, F1-score is the harmonic mean of precision and recall.

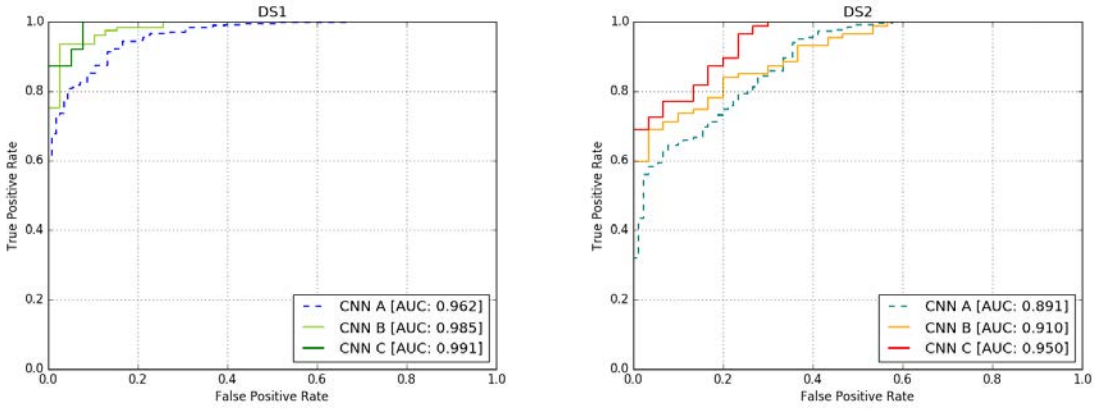
$$precision = \frac{TP}{TP+FP} \quad , \quad recall = \frac{TP}{TP+FN} \quad \text{and} \quad F_1 = 2 \times \frac{precision \times recall}{precision + recall}. \tag{2}$$

Table 3: Classification performance

Dataset	Model	Precision	Recall	F_1
DS1	CNN A	0.967	0.833	0.895
	CNN B	0.992	0.936	0.963
	CNN C	0.947	1	0.972
DS2	CNN A	0.919	0.736	0.817
	CNN B	0.942	0.747	0.833
	CNN C	0.95	0.874	0.91

Fig. 3 shows the Receiver Operating Characteristic (ROC) curves by plotting the probability of false alarm over the probability of detection, while threshold is varied. The calculated area under each curve is also indicated. With both datasets, CNN C delivers the most interesting

results. This can be explained by CNN B that does not take full advantages of the different object points of view. Because it averages predictions, it is not able to give more prominence to a specific image with valuable information. The drop in performance with the DS2 is justified because of its degraded resolution and because the feature extraction stage was pre-trained with orginal images. However, even with a small difference in orientation between images, the classification results of DS2 with CNN C demonstrated a much better performance than a single-view classification.



(a) CNNs performance with DS1 *(b) CNNs performance with DS2*

Fig. 3: Classification performance

4. CONCLUSION

In this preliminary work, we wanted to investigate the potential of multi-view SAS image classification. We described two ways to enhance SAS classifier performance. One independently classifies three images and averages the predictions. The other one fuses the data resulting from three independent feature extraction branches and delivers the prediction. We compared them with two types of multi-view dataset. The different classification tests showed very encouraging results. Fusing data at the end of the feature level demonstrated a significant gain in terms of performance. Indeed, it makes the best use of additional information given by the complementary views.

Future work will aim to enlarge our datasets by gathering more multi-view images. This would allow to reinforce the training of the last layer of CNN C. It could also allow to re-train the entire network for the generated dataset. Other types of CNN could also be explored and compared with those studied in this paper.

ACKNOWLEDGEMENT

This work was supported by the NATO STO Centre for Maritime Research and Experimentation (CMRE) and Thales. Data used in this paper were acquired by the CMRE with funding provided by the NATO's Allied Command Transformation (ACT). This manuscript has been approved for public release.

REFERENCES

[1] J. Groen, E. Coiras, J. Del Rio Vera, and B. Evans. Model-based sea mine classification with synthetic aperture sonar. *IET Radar, Sonar Navigation*, 4(1):62–73, February 2010.

[2] T. Thorsnes and L. R. Bjarnadóttir. Experiences from using autonomous underwater vehicles and synthetic aperture sonar for sediment and habitat mapping. *AGU Fall Meeting Abstracts*, 2017.

[3] Øyvind Ødegård, Roy E. Hansen, Hanumant Singh, and Thijs J. Maarleveld. Archaeological use of synthetic aperture sonar on deepwater wreck sites in skagerrak. *Journal of Archaeological Science*, 89:1 – 13, 2018.

[4] D. P. Williams, M. Couillard, and S. Dugelay. On human perception and automatic target recognition: Strategies for human-computer cooperation. In *2014 22nd International Conference on Pattern Recognition*, pages 4690–4695, August 2014.

[5] D. P. Williams. Underwater target classification in synthetic aperture sonar imagery using deep convolutional neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2497–2502, December 2017.

[6] M. Chabah, N. Burlet, J.-P. Malkasse, G. Le Bihan, and B. Quellec. *SAMDIS: A New SAS Imaging System for AUV*, volume 6. Springer International Publishing, Cham, 2016.

[7] D. P. Williams, R. Hamon, and I. Gerg. On the benefit of multiple representations with convolutional neural networks for improved target classification using sonar data. In *Proceedings of the Underwater Acoustics Conference*, July 2019.

[8] D. P. Williams. Fast target detection in synthetic aperture sonar imagery: A new algorithm and large-scale performance analysis. *IEEE Journal of Oceanic Engineering*, 40(1):71–92, January 2015.

[9] D. P. Williams and A. J. Hunter. Multi-look processing of high-resolution sas data for improved target detection performance. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 153–157, September 2015.