ACOUSTIC-COLOR-BASED CONVOLUTIONAL NEURAL NETWORKS FOR UXO CLASSIFICATION WITH LOW-FREQUENCY SONAR

David P. Williams

NATO STO Centre for Maritime Research and Experimentation (CMRE)
La Spezia, Italy

ABSTRACT

In this work, we contribute a new target classification approach for low-frequency sonar data. More specifically, we illustrate the feasibility of using convolutional neural networks (CNNs) trained on acoustic-color data, a representation that expresses target strength as a function of object aspect and frequency. We show that it is possible, using only limited amounts of this sonar data, to design and train efficient networks with low capacity that avoid overfitting and generalize robustly, even to new objects not seen during training. We demonstrate this in the context of an unexploded ordnance (UXO) classification task using real, measured sonar data collected at sea.

Index Terms— Convolutional neural networks (CNNs), acoustic color, sonar, unexploded ordnance (UXO)

1. INTRODUCTION

When an object is ensonified with low-frequency sonar, the response contains contributions resulting from geometric scattering as well as elastic effects due to the excitation of structural waves that are intimately linked to the material properties of the object [1]. Conventional wisdom maintains that these special phenomena should be valuable for underwater object classification tasks seeking to discriminate particular man-made targets of interest, such as unexploded ordnance (UXO), from other benign objects.

A popular display format that reveals these characteristic signatures is an acoustic-color plot, which is a two-dimensional representation showing target strength as a function of object aspect and frequency [2, 3]. Although angular and frequency dependent scattering clues are observable in this domain, the complicated wave phenomena involved in the process, and especially the complex interactions with the environment [4, 5], make data interpretation challenging.

Interesting investigations have begun isolating individual waves and effects for certain canonical objects [5, 6], but a comprehensive understanding and explanation of such data has not yet been attained. As a result, it has proven exceedingly difficult to define and extract robust acoustic-color-based features that can reliably discriminate objects. That is, manual feature-engineering for acoustic-color data remains an open problem. But it is for precisely this same reason that convolutional neural networks (CNNs) are a perfect match for acoustic-color "imagery." Because CNNs automatically learn the most useful bases in which to represent the data, the extraction of predefined features is obviated.

A standard CNN [7] is a sequence of convolutional layers, nonlinear activation functions, and pooling operations that collectively transform input data (*i.e.,* imagery) into a new representation space in which the classes are easily separable. This work demonstrates the feasibility of using CNNs, trained with only modest amounts of low-frequency acoustic-color data, for UXO classification. This is achieved by recognizing that a crucial quantity in determining the success of CNNs for classification tasks is not the amount of training data *per se*, but rather the relationship between training data and network capacity. As the capacity of a CNN – loosely speaking, the number of free trainable parameters in the model – grows, so too do the training data requirements. Therefore, when faced with extremely limited training data, it is imperative to constrain the CNN's capacity by designing small, efficient networks. In this work, we show that it is possible to design acoustic-color-based CNNs with low capacity that avoid overfitting and generalize robustly, even to new objects not seen during training.

Other studies have explored the use of high-frequency synthetic aperture sonar (SAS) image-based CNNs [8–11], but we are the first to successfully employ low-frequency *acoustic-color* data as input to a CNN. Prior work [12–14] that has attempted to use acoustic-color data for classification has relied on using features derived from template-matching and correlation-based methods. But these approaches have a fundamental shortcoming that make them scale impractically and generalize poorly. To wit, one will never possess data for the entire universe of clutter (*i.e.,* non-UXO) objects, and it is also not feasible to collect sufficient data that encompasses the full variability of the target class, given that the acoustic-color signatures will exhibit a strong dependence on object range, seafloor sediment properties, burial conditions, and UXO decay state. As a result, these correlation-reliant approaches can succeed only when data for a given test object – and collected in very similar conditions – is also present

in the training set. For real-world UXO remediation, this assumption is unrealistic.

The remainder of this paper is organized as follows. In Sec. 2, we present our acoustic-color-based CNN framework. Experimental results of the proposed approach on an object classification task using measured sonar data are shown in Sec. 3. Concluding remarks are made in Sec. 4.

2. ACOUSTIC-COLOR-BASED CNN

2.1. CNN Data-Preparation

Low-frequency sonar data were collected on various UXO and non-UXO objects during the Target and Reverberation Experiment 2013 (TREX13) in the Gulf of Mexico using a rail system [15]. More specifically, data were collected over the frequency band 3-30 kHz for 27 unique objects at ranges 10-40 m on a sandy seafloor.

Each individual data-collection run provided along-track-versus-time scattering data over only a certain aspect (*i.e.,* orientation) span, so multiple runs were executed. For a given object, the raw sonar data was transformed and the results then stitched together [12] to form a full 360° acoustic-color data product expressing target strength as a function of object aspect and frequency. This data has a discretization of 0.5° in aspect and 100 Hz in frequency, meaning a full acoustic-color plot is an "image" – in the general sense of values organized as a two-dimensional array – of size 720 pixels $\times$ 271 pixels. A total of 90 such acoustic-color plots were produced, 54 corresponding to UXO objects (11 unique objects at various ranges) and 36 corresponding to non-UXO objects (16 unique objects at various ranges).

We divide this data into disjoint training and test sets *by object type*, meaning data for a given object exists either only in the training set (13 objects) or only in the test set (14 objects). *This separation is vital because it will allow for true generalization ability of the CNNs to be assessed properly,* a shortcoming of earlier acoustic-color-based classification experiments.

The set of UXO objects in the training set (using the TREX13 nomenclature [12]) include: 155 mm howitzer without collar, 15 mm TP-T, aluminum UXO replica, bullet #1, finned shell #1. The set of non-UXO objects in the training set include: DEU trainer; rock; 55-gallon drum, water-filled with fixture; scuba tank, water-filled, without stem; 2:1 aspect telephone pole section; $2'$ aluminum pipe; solid aluminum cylinder with notch; hollow aluminum cylinder with notch. The objects in the test set are listed later in Table 4. Based on the data division imposed, 47 acoustic-color plots (29 UXO, 18 non-UXO) comprise the training set, and 43 acoustic-color plots (25 UXO, 18 non-UXO) comprise the test set. It is worth reiterating that the objects available for training are distinct from those reserved for testing.

To maximize the amount of data for CNN training, each acoustic-color plot is converted into 720 unique (but highly correlated) images in which each image spans the full frequency band. The aspect span of each image is $[\theta_c - \theta/2, \theta_c + \theta/2]$, so each image is of size $(2\theta+1) \times 271$, and the images are distinguished by the center aspect θ_c. We construct distinct data sets for $\theta = \{10^\circ, 30^\circ, 40^\circ, 90^\circ\}$ so that the impact of data aspect span on CNN performance can be analyzed. Each data set is then further augmented by a factor of 2 by reversing the order of the data aspects (*i.e.,* vertically flipping the acoustic-color data), which is physically justified given the data-collection geometry. The effect of all of the above data augmentation was that the original 90 acoustic-color plots were transformed into a training set of 67680 images and a test set of 61920 images; the images are unique but many contain highly redundant information. (This is another reason it is imperative to create the training/test division by *object type*.) Example input data to the CNNs for different θ are shown in Fig. 1.

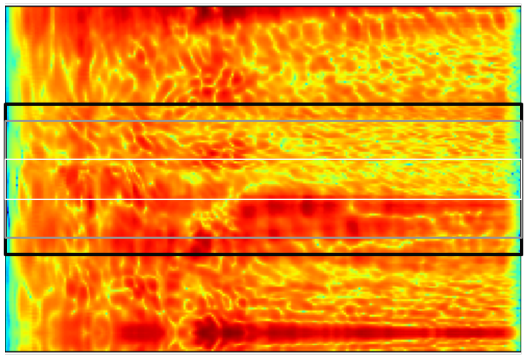


Fig. 1. Example input data to the CNNs of a $3'$ aluminum cylinder (object 7), centered about an aspect of 229° at a range of 30 m, when the data spans 90° (entire image), 40° (delineated by the black box), 30° (gray box), and 10° (white box). The x-axis corresponds to frequency, from 3 kHz to 30 kHz; the y-axis indicates aspect.

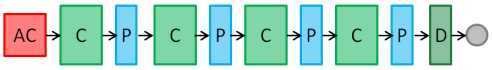


Fig. 2. General CNN architecture for acoustic-color data inputs. *C*, *P*, and *D* denote convolution *blocks* (comprising one or more convolutional layers), pooling layers, and dense layers, respectively. (The specific architecture shown, with 4 convolution blocks, corresponds to CNNs G-I.)

2.2. CNN Design

We carefully design 9 CNN architectures, whose common high-level schematic is shown in Fig. 2. Each CNN contains between 2 and 4 convolution blocks; each block contains an identical number of convolutional layers (between 1 and 3, depending on CNN). Only 4 filters are used in each convolutional layer. ReLU activations are used after each convolutional layer, while a sigmoid activation is used at the output. All pooling layers use average pooling; with an eye toward

generalizability and limiting the number of free parameters, aggressive (*i.e.,* large) pooling factors – as high as 8 – are employed. The filter sizes vary for each CNN, but are typically on the order of a few pixels in each dimension. The number of nodes contained in the dense layer is either 12 (CNNs A-C) or 4 (CNNs D-I). Space constraints prevent us from giving more specific details (*e.g.,* filter sizes and pooling factors) about each CNN.

A CNN is designed for specific input-data dimensions (*i.e.,* number of rows and columns of pixels). When dealing with the acoustic-color data, however, the size of the input data will vary as a function of the aspect span considered. Specifically, the size of the input image will be $(2\theta+1)\times 271$ for data spanning θ aspect degrees. Rather than re-sizing the imagery (*e.g.,* via interpolation) to the largest aspect span considered, which could introduce artifacts and would increase computational effort during training, we instead design a unique CNN architecture for each input-data aspect span considered.

Nevertheless, the CNN architectures are designed to be as similar as possible for different θ. For a given CNN architecture, the same column-wise filter sizes and pooling factors are used regardless of θ (since this dimension is not affected by θ). The row-wise pooling factors for a given CNN are generally kept identical across different θ, while the row-wise filter sizes scale roughly linearly with θ.

For input data spanning $\theta = [10^\circ, 30^\circ, 40^\circ, 90^\circ]$, the mean number of free parameters of the 9 CNNs are $n = [665, 1406, 1833, 3795]$. This should be contrasted with popular (optical-image) CNNs [16–19] commonly used "off-the-shelf" that each contain more than 10^6 parameters. (The sensing modality and image properties of acoustic-color data are so fundamentally different from optical photographs that transfer is not a feasible approach.)

The capacities of our CNNs are intentionally kept so low in order to scrutinize the feasibility of using CNNs when faced with modest amounts of training data. That is, we purposely constrain the networks in an attempt to force them to learn robust filters that will generalize across object class and operating conditions.

2.3. CNN Training

Each raw acoustic-color "image," $\mathbf{A}$, was normalized as $\mathbf{A}' = (\mathbf{A} - 80)/50$ to scale the pixel values approximately into the range $[-1, 1]$. CNN training was performed using the RMSprop optimizer with a learning rate of 0.001, in conjunction with a binary-cross-entropy loss function, until the loss on the training set converged. A batch size of 128 was used, with equal numbers selected from each class. No attempt was made to optimize the learning rate or batch size. The final prediction for a test image was taken as the ensemble (mean) of the predictions on the standard and aspect-reversed input data.

3. EXPERIMENTAL RESULTS

Experiments were conducted to assess the feasibility of employing CNNs with acoustic-color sonar data for discriminating UXO from non-UXO when possessing extremely limited training data. Classification performance is measured in terms of the area under the curve (AUC). Because an auxiliary goal of this work is to enable better understanding of the factors impacting classification success, we also examine performance as a function of various conditions.

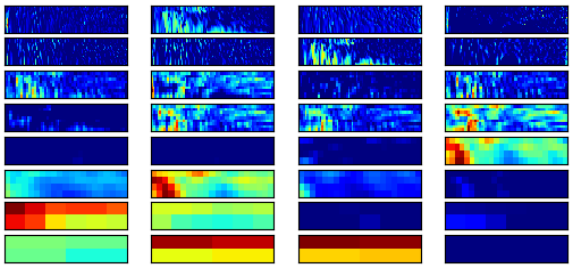
Table 1 presents the performance of the 9 CNN architectures for input data sizes of $\theta = \{10^\circ, 30^\circ, 40^\circ, 90^\circ\}$. Performance as a function of object range, object center-aspect, and object type is shown in Tables 2, 3, and 4, respectively. For the latter, the AUC values correspond to the binary-classification task of discriminating one given object from all objects in the opposite class (the first eight objects in the table belong to the non-UXO class). There are numerous interesting findings from these results.

One of the most remarkable results from Table 1 is that CNN G when using only 10° input data was actually the single most powerful classifier. (It may be germane to note that the beamwidth of dolphin biosonar has been measured to be 10° [20].) The architectures corresponding to CNNs B, D, and G appeared to be the most robust, reliably providing decent performance across different input data aspects. But importantly, employing the ensemble of 9 CNNs consistently improved performance. This valuable result implies that one need not select a single best CNN architecture, but rather that drawing from the unique representations of the individual networks collectively provides superior classification performance. The improvement with larger input data aspect spans is also evident when using the ensemble.

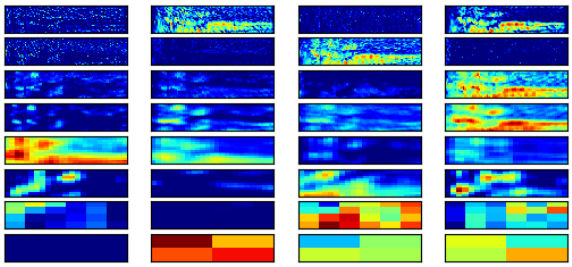
Table 2 shows that classification performance of objects at short range (*e.g.,* 10-20 m) improves dramatically as wider aspect span data is available. This finding has useful implications for data-collection strategies.

In Table 3, it can be seen that objects interrogated at certain sets of aspects, namely $[-60^\circ, 0^\circ)$ and $[180^\circ, 240^\circ)$, which are just off-broadside, become classified much more accurately as the aspect span of data increases. It is possible that more prominent elastic effects are present at those aspects. Slightly off-endfire aspects, namely $[-90^\circ, -60^\circ)$ and $[240^\circ, 270^\circ)$, appear to contain the most discriminatory information even over short aspect spans. Interestingly, these aspects comport with the sensing strategy of marine mammals [21, 22].

Table 4 reveals that the $3'$ aluminum cylinder (object 7) was easily discriminated from the UXO targets with only 10° aspect data, whereas classifying the $2'$ aluminum cylinder (object 17) was challenging. This particular finding suggests that object length information is present even in small spans of aspect. By examining the intermediate responses of the 10° aspect CNNs for the $3'$ aluminum cylinder, one can hopefully discover and isolate interpretable, physics-based clues



(a) 10° input data



(b) 40° input data

Fig. 3. For the data from Fig. 1, the intermediate responses after each convolutional layer (layer-wise by row) of CNN H.

that will lead to increased understanding of the low-frequency phenomenology. For example, for the cylinder data of Fig. 1, it can be seen via the intermediate-layer CNN responses in Fig. 3 (a) that a certain (lower) frequency band appears to contain the discriminatory information that is keyed on to enable the correct classification of the non-UXO object. Comparing Fig. 3 (a) and (b), it is also evident that different features are being leveraged depending on the aspect data provided as input to the CNN.

4. CONCLUSION

The feasibility of using CNNs trained on low-frequency sonar acoustic-color data was demonstrated. It was shown that powerful classifiers could be constructed even from limited training data if the network capacity is intentionally constrained. To perform an honest assessment of algorithm generalizability, we invoked a proper, realistic split of data in which the objects in the training set and test set were disjoint. Future work will attempt to link the discriminatory features isolated in intermediate representations of the CNNs with physics-based phenomena. A preliminary analysis of performance revealed potential clues to pursue.

5. ACKNOWLEDGMENTS

The author thanks Kevin Williams at the University of Washington Applied Physics Laboratory for providing the TREX13 data. This work was supported by the Strategic Environmental Research and Development Program (SERDP).

Table 1. AUC for CNNs of the given input data size

	Input Data			
	10°	30°	40°	90°
CNN A	0.7302	0.7503	0.7975	0.7063
CNN B	0.8491	0.8485	0.8007	0.8091
CNN C	0.7826	0.7510	0.7987	0.6273
CNN D	0.8189	0.8615	0.8321	0.8172
CNN E	0.7775	0.8159	0.8005	0.8002
CNN F	0.8496	0.7813	0.7353	0.8227
CNN G	0.8886	0.8128	0.8037	0.8633
CNN H	0.8085	0.7931	0.8395	0.6561
CNN I	0.7620	0.8287	0.7354	0.8313
$\mathcal{E}$(A-I)	0.8534	0.8754	0.8985	0.9198

Table 2. AUC for ensemble of 9 CNNs of the given input data size, as a function of object range

	Input Data			
Range	10°	30°	40°	90°
[10 m, 15 m)	0.5739	0.5949	0.5573	0.7595
[15 m, 20 m)	0.6859	0.8058	0.7809	0.8290
[20 m, 25 m)	0.9187	0.8874	0.8914	0.9500
[25 m, 30 m)	0.7914	0.8388	0.8871	0.9292
[30 m, 35 m)	0.8863	0.9000	0.9678	0.9877
[35 m, 40 m)	0.9984	0.9988	1.0000	1.0000

Table 3. AUC for ensemble of 9 CNNs of the given input data size, as a function of object center-aspect (0° is broadside)

Center	Input Data			
Aspect	10°	30°	40°	90°
$[-90°, -60°)$	0.9237	0.9263	0.9428	0.9197
$[-60°, -30°)$	0.8155	0.8634	0.8941	0.9271
$[-30°, 0°)$	0.8431	0.9010	0.9007	0.9680
$[0°, 30°)$	0.8835	0.8950	0.9518	0.9400
$[30°, 60°)$	0.8013	0.8180	0.8515	0.8637
$[60°, 90°)$	0.8585	0.8674	0.8502	0.8964
$[90°, 120°)$	0.8589	0.8691	0.8517	0.8962
$[120°, 150°)$	0.8015	0.8183	0.8500	0.8639
$[150°, 180°)$	0.8828	0.8925	0.9511	0.9383
$[180°, 210°)$	0.8448	0.9030	0.9024	0.9682
$[210°, 240°)$	0.8145	0.8622	0.8935	0.9273
$[240°, 270°)$	0.9222	0.9255	0.9427	0.9201

Table 4. AUC (object vs. opposite class) for ensemble of 9 CNNs of the given input data size, as a function of object type

	Input Data			
Object Index – Description	10°	30°	40°	90°
5 – 5:1 aspect telephone pole section	0.865	0.895	0.920	0.940
6 – 55-gallon drum, water-filled	0.993	0.996	0.999	0.995
7 – 3′ aluminum cylinder	0.990	0.997	0.998	0.986
10 – panel target	0.779	0.791	0.896	0.925
14 – scuba tank, water-filled, w/ stem	0.741	0.842	0.881	0.939
17 – 2′ aluminum cylinder	0.550	0.535	0.534	0.686
18 – cement block	0.787	0.882	0.795	0.839
19 – tire	0.920	0.778	0.766	0.678
8 – 155 mm howitzer w/o collar	0.781	0.826	0.865	0.924
12 – 81 mm mortar	0.851	0.815	0.825	0.888
21 – steel UXO replica	0.881	0.900	0.910	0.916
22 – original material UXO	0.873	0.876	0.888	0.905
28 – 155 mm howitzer w/ collar	0.834	0.878	0.920	0.901
29 – bullet #2	0.924	0.942	0.967	0.979

6. REFERENCES

[1] D. Plotnick and T. Marston, "Utilization of aspect angle information in synthetic aperture images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 9, pp. 5424–5432, 2018.

[2] S. Kargl, K. Williams, T. Marston, J. Kennedy, and J. Lopes, "Acoustic response of unexploded ordnance (UXO) and cylindrical targets," in *Proceedings of IEEE OCEANS*, 2010, pp. 1–5.

[3] D. Plotnick, P. Marston, K. Williams, and A. España, "High frequency backscattering by a solid cylinder with axis tilted relative to a nearby horizontal surface," *The Journal of the Acoustical Society of America*, vol. 137, no. 1, pp. 470–480, 2015.

[4] K. Williams, S. Kargl, and A. España, "TREX13 target experiments and case study: Comparison of aluminum cylinder data to combined finite element/physical acoustics modeling," *The Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. 2111–2111, 2014.

[5] S. Kargl, A. España, K. Williams, J. Kennedy, and J. Lopes, "Scattering from objects at a water-sediment interface: Experiment, high-speed and high-fidelity models, and physical insight," *IEEE Journal of Oceanic Engineering*, vol. 40, no. 3, pp. 632–642, 2015.

[6] A. España, K. Williams, D. Plotnick, and P. Marston, "Acoustic scattering from a water-filled cylindrical shell: Measurements, modeling, and interpretation," *The Journal of the Acoustical Society of America*, vol. 136, no. 1, pp. 109–121, 2014.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436, 2015.

[8] D. Williams, "Underwater target classification in synthetic aperture sonar imagery using deep convolutional neural networks," in *Proceedings of the 23rd International Conference on Pattern Recognition (ICPR)*, 2016.

[9] D. Williams, "Demystifying deep convolutional neural networks for sonar image classification," in *Proceedings of the Underwater Acoustics Conference*, 2017.

[10] M. Emigh, B. Marchand, M. Cook, and J. Prater, "Supervised deep learning classification for multi-band synthetic aperture sonar," in *Proceedings of the 4th International Conference on Synthetic Aperture Sonar and Synthetic Aperture Radar*, 2018, vol. 40, pp. 140–147.

[11] N. Warakagoda and Ø. Midtgaard, "Transfer-learning with deep neural networks for mine recognition in sonar images," in *Proceedings of the 4th International Conference on Synthetic Aperture Sonar and Synthetic Aperture Radar*, 2018, vol. 40.

[12] S. Kargl, "Acoustic response of underwater munitions near a sediment interface: Measurement model comparisons and classification schemes," Tech. Rep., SERDP Project MR-2231 Final Report, April 2015.

[13] M. Azimi-Sadjadi, "Multichannel detection and acoustic color-based classification of underwater UXO in sonar," Tech. Rep., SERDP Project MR-2416 Final Report, September 2015.

[14] J. Hall, M. Azimi-Sadjadi, and S. Kargl, "Underwater uxo classification using matched subspace classifier with synthetic sparse dictionaries," *Proceedings of IEEE OCEANS*, pp. 1–9, 2016.

[15] K. Williams, S. Kargl, E. Thorsos, D. Burnett, J. Lopes, M. Zampolli, and P. Marston, "Acoustic scattering from a solid aluminum cylinder in contact with a sand sediment: Measurements, modeling, and interpretation," *The Journal of the Acoustical Society of America*, vol. 127, no. 6, pp. 3356–3371, 2010.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[20] W. Au, *The sonar of dolphins*, Springer Science & Business Media, 2012.

[21] W. Evans and B. Powell, "Discrimination of different metallic plates by an echolocating delphinid," in *Animal Sonar Systems: Biology and Bionics*, 1967, vol. 78, pp. 363–382.

[22] P. Moore, S. Martin, and L. Dankiewicz, "Investigation of off-axis detection and classification in bottlenosed dolphins," in *Proceedings of IEEE OCEANS*, 2003, vol. 1, pp. 316–319.