SCIENCE AND TECHNOLOGY ORGANIZATION
CENTRE FOR MARITIME RESEARCH AND EXPERIMENTATION

Reprint Series **CMRE-PR-2017-005**

Multi-view SAS image classification using deep learning

David P. Willilams, Samantha Dugelay

November 2017

Originally presented at:

OCEANS'16 MTS/IEEE Monterey

About CMRE

The Centre for Maritime Research and Experimentation (CMRE) is a world-class NATO scientific research and experimentation facility located in La Spezia, Italy.

The CMRE was established by the North Atlantic Council on 1 July 2012 as part of the NATO Science & Technology Organization. The CMRE and its predecessors have served NATO for over 50 years as the SACLANT Anti-Submarine Warfare Centre, SACLANT Undersea Research Centre, NATO Undersea Research Centre (NURC) and now as part of the Science & Technology Organization.

CMRE conducts state-of-the-art scientific research and experimentation ranging from concept development to prototype demonstration in an operational environment and has produced leaders in ocean science, modelling and simulation, acoustics and other disciplines, as well as producing critical results and understanding that have been built into the operational concepts of NATO and the nations.

CMRE conducts hands-on scientific and engineering research for the direct benefit of its NATO Customers. It operates two research vessels that enable science and technology solutions to be explored and exploited at sea. The largest of these vessels, the NRV Alliance, is a global class vessel that is acoustically extremely quiet.

CMRE is a leading example of enabling nations to work more effectively and efficiently together by prioritizing national needs, focusing on research and technology challenges, both in and out of the maritime environment, through the collective Power of its world-class scientists, engineers, and specialized laboratories in collaboration with the many partners in and out of the scientific domain.

NOTE: The CMRE Reprint series reprints papers and articles published by CMRE authors in the open literature as an effort to widely disseminate CMRE products. Users are encouraged to cite the original article where possible.

Multi-view SAS Image Classification Using Deep Learning

David P. Williams and Samantha Dugelay
NATO STO CMRE
La Spezia, Italy

Abstract

A new approach is proposed for multi-view classification when sonar data is in the form of imagery and each object has been viewed an arbitrary number of times. An image-fusion technique is employed in conjunction with a deep learning algorithm (based on Boltzmann machines) so that the sonar data from multiple views can be combined and exploited at the (earliest) image level. The method utilizes single-view imagery and, whenever available, multi-view fused imagery, in the same unified classification framework. The promise of the proposed approach is demonstrated in the context of an object classification task with real synthetic aperture sonar (SAS) imagery collected at sea.

1. Introduction

The aspect at which an object is imaged by a sonar can have a profound effect on its ability to be detected and classified. For example, cylindrical objects viewed at endfire (where there are fewer pixels on target) are significantly more challenging to detect than cylindrical objects viewed at broadside. Similarly, the sonar image of a rock may look nearly indistinguishable from that of some man-made objects when imaged at certain aspects, but not others. For reasons like these, it can be valuable to collect multiple views, at different aspects, of an object. Doing so can provide a more complete picture of the (unknown) object and also reduce the likelihood of observing the object at only one *unfavorable* aspect. In turn, this richer information should allow more accurate and more confident predictions regarding an object's identity.

Translating the additional information provided by the multiple views into better predictions requires a classification approach that can adeptly handle the extra data. In this work, we address this challenge by proposing a new approach for multi-view classification when data is in the form of imagery and each object has been viewed an arbitrary number of times. The topic is addressed in the context of mine countermeasures (MCM) where the objective is to classify mine-like objects detected in sonar imagery as either targets (*i.e.,* mines) or clutter (*e.g.,* rocks). The promise of the approach is demonstrated on real sonar imagery collected at sea by CMRE's autonomous underwater vehicle (AUV) called MUSCLE.

The remainder of this paper is organized as follows. Sec. 2 discusses the multi-view classification problem and outlines the proposed approach. Sec. 3 describes the sonar data used in the experiments. Experimental results and a discussion are presented in Sec. 4, before concluding remarks are made in Sec. 5.

2. Multi-view Classification

In binary classification, the objective is to learn a classifier that will properly classify each data point (*e.g.,* object) as belonging to one of two classes (*e.g.,* target or clutter). The standard procedure is to extract a set of features for each object, and then learn appropriate classifier weights by making use of labeled training data for which the true class of each object is known. The features quantify characteristics of the objects that are hopefully useful for discriminating between the two classes. In our sonar classification task, each object is initially represented by an image, from which features can subsequently be extracted.

Suppose a set of sonar data was collected in which each object in the data set was viewed a certain (possibly different) number of times, at different aspects. For example, some objects may have been imaged only once, others may have been imaged twice, and still others may have been imaged some arbitrarily large number of times. When data from multiple aspects of (a subset of) objects are available, a necessary task is to determine how that multi-view information should be combined. When the eventual goal is to perform binary classification – declaring a new unknown object to be

a target or clutter – there are three general possibilities, the distinguishing element being the stage at which the data is fused: (i) at the prediction level, (ii) at the feature level, and (iii) at the image level.

The simplest approach is to combine the multi-view data for an object at the prediction level. Under this scenario, a standard *single-view* classifier can be employed. More specifically, normal feature extraction would be performed on each image (*i.e.,* view) separately, and each resulting feature vector (*i.e.,* data point) would be treated as if no relationship existed. Single-view classifier training would ensue, and a prediction (*e.g.,* the probability of belonging to each class) would be effected for each view. Only at this final stage would the multiple predictions corresponding to the views of the same object be combined, for example by simple averaging. But this perspective fundamentally ignores, for a significant portion of the classification stage, the obvious dependence shared by the object's views. Moreover, since each view is treated equally – both favorable views that potentially contain valuable discriminating information, and unfavorable views for which information is lacking – the performance gains that can be expected from having multiple views are necessarily limited.

The second approach for combining multi-view information would fuse data at the feature level. In this scenario, the same set of features would be extracted from each image independently. Therefore, if an object had been imaged three times (at different aspects), there would be three unique feature vectors associated with the object. An object imaged only once would be represented by only a single feature vector. Fusing data at the feature level would mean either combining the multiple feature vectors of an object in some way, or preserving the multiple feature vectors and expanding the feature space of the eventual classifier. The former is not sound because it does not make sense to average the values of a given feature – say, the length of the object highlight – obtained from different aspects, since the geometry and physics can be fundamentally different in each view. The latter is untenable because not every object is guaranteed to have the same number of views. In this case, an expanded feature space to accommodate the object with the highest number of views would result in substantial amounts of "missing data" [1], thereby complicating classifier learning. (In particular, it would be likely that insufficient training data was available for the feature dimensions associated with the highest numbers of views.)

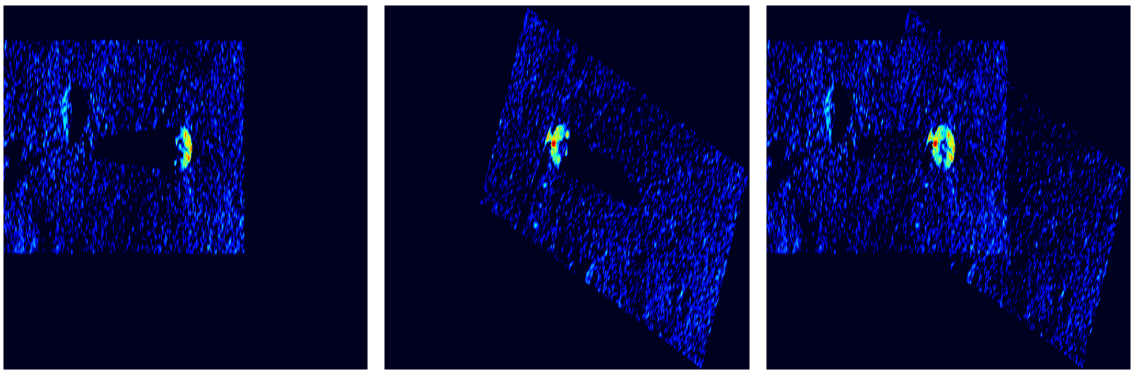
The third possible approach for combining multi-view information is unique to tasks for which "raw" data[1] is in the form of imagery. Combining data at the image level would mean fusing the multiple views of a given object into a single composite image. By combining data at this highest, earliest level – the image level – the dependence shared among an object's multiple views can be exploited fully. The idea behind multi-view image fusion is that the joint, composite image will contain more information than the individual views, in the sense that classification will be easier. Results obtained from circular synthetic aperture sonar processing [2], which attempts to reconstruct a composite image from full 360° surveys around an object, support such a hypothesis.

Once multi-view fused imagery is obtained, the matter of classifier learning must still be resolved. A standard approach in which a set of features are extracted (from the multi-view imagery) is plausible, but would require the development of new features, since the set of features developed for single-view imagery is likely to be inappropriate for the new data format. For example, a feature based on shadow length is meaningful for single-view sonar imagery, but not multi-view imagery where shadow information may be destroyed in the fusion process. (Additionally, missing data corresponding to the "multi-view features" can result if not every object is viewed multiple times.) So rather than attempting to develop a new set of features to be extracted from fused multi-view images, we adopt a fundamentally different approach that relies on image-based classification without resorting to intermediate feature extraction. That is, the pixels of the (fused) images are themselves used as the data on which classifier learning is conducted.

In this work, we adopt a "deep learning" [3] approach to classification that operates directly on the imagery. In our case, the imagery is the output of a multi-view fusion algorithm [4] when multiple views of an object are available. However, when only a single view of an object is available, the single-view image can be used in the same model with no modifications. This flexibility to use a single, unified framework for both multi-view fusion imagery and single-view imagery makes the method particularly attractive.

The specific classification approach employed here is based on a deep Boltzmann machine, developed in [5], featuring an architecture with two hidden layers of weights. One key to this particular approach is the greedy pre-training of each layer, with this effectively initializing the classifier weights to reasonable

[1] Significant processing is required to transform the receiver-element-level sonar ping-return data into imagery, but from the perspective of classifier learning the sonar imagery can be viewed as raw data.



(a) (b) (c)

Figure 1. Example SAS images of a truncated cone from the AMiCa data set, displayed in a common reference frame, viewed at aspects of (a) 359° (with the port sonar) and (b) 23° (with the starboard sonar), and (c) the resulting fused image. The bounding box of each image contains an area of approximately 10 m $\times$ 10 m.

regions of the very high-dimensional parameter search space. Learning is performed by maximizing the log-likelihood of the model via gradient descent. The exploration of more sophisticated deep classifiers will be a topic of future work, but this relatively simple algorithm is sufficient for our multi-view proof-of-concept purposes.

3. Multi-view Data

To assess the feasibility of the deep learning approach for object classification with multi-view sonar imagery, we utilize data collected by CMRE's MUSCLE AUV. This experimental, state-of-the-art AUV is a 21-inch diameter vehicle from Bluefin that is equipped with a synthetic aperture sonar (SAS) system developed by Thales. The center frequency of the SAS is 300 kHz, and the bandwidth is 60 kHz. The system enables the formation of high-resolution sonar imagery with a theoretical along-track resolution of 2.5 cm, and a theoretical across-track resolution of 1.25 cm, usually out to a range of 150 m.

To compile a data set for classification, suspicious mine-like objects of interest were automatically detected in scene-level imagery by applying an integral-image-based detection algorithm [6]. The resulting "mugshots" of each detected object were then passed on to the classification stage. For this work, the mugshot imagery of objects that have been viewed multiple times were combined using an active contour-based image-fusion algorithm [4] that fuses the multiple views into a single composite image. It is this multi-view fusion imagery – as well as the single-view imagery – that was used in the subsequent classification experiments.

Multi-view fusion imagery was available from two previous sea trials, AMiCa and CATHARSIS 2, but only for deployed objects (target shapes and calibrated rocks), not clutter. Therefore, the limited-scope binary classification task considered here was to discriminate truncated cones from calibrated rocks of a similar size.

The AMiCa trial was conducted in May-June 2010 near Tellaro, Italy; the CATHARSIS 2 trial was conducted in October 2009 near Elba, Italy. Example (mugshot-level) imagery from the AMiCa trial of a truncated cone, along with the fusion result from [4], is shown in Fig. 1. (In this work, the term "aspect" is taken to mean the direction of AUV travel when the object is viewed, which is orthogonal to the sonar's imaging direction. The convention used here is that an aspect of 0° points up the page and increases clockwise.)

It should be noted that the fusion process destroys the well-defined coordinate system of the single-view images (*i.e.,* axes corresponding to along-track and across-track dimensions), so it is important that the pixel dimensions are made equal in each direction prior to the fusion. For the data considered, each pixel of the multi-view fusion images spans an area of 1.5 cm $\times$ 1.5 cm. The single-view images are upsampled via interpolation so that their pixel dimensions match those of the multi-view fusion images, making all data uniform in this sense.

For the deep classifier, the size of the imagery is reduced to an area of approximately 1.25 m $\times$ 1.25 m (corresponding to 83 pixels by 83 pixels), centered around the object highlight. (The peak correlation with a simple binary template was used to automatically lo-

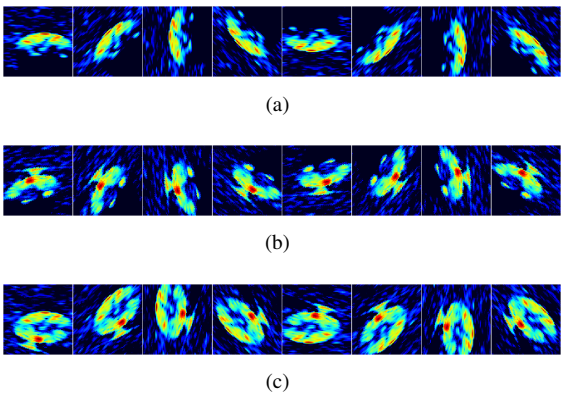


Figure 2. Example (cropped) images, of a truncated cone, at different rotations from the AMiCa data set used for the deep-classifier training for the approach employing (a)-(b) rotated single-view images, and (c) multi-view fusion images. Each series is derived from rotating an original unrotated image (the initial image in each series). Each image spans an area of approximately 1.25 **m** $\times$ 1.25 **m.**

cate the highlight center of each image.) This choice is made so that the resulting classifier focuses only on the highlight information – essentially, the object shape – and not the seabed background. However, the useful shadow information has still been exploited in the previous detection stage.

Because the multi-view fusion images are not in a standard viewing or reference frame, each cropped image – both single-view and multi-view – is also rotated at a series of angles (in increments of $\Delta = 45^\circ$ here). This procedure, which is performed to support classifier robustness vis-à-vis aspect invariance, serves to also augment the amount of training data available.

Examples of the unrotated images used in the classification phase, as well as the derived series of rotated single-view images and rotated multi-view images, for a truncated cone are shown in Fig. 2. The object shape information provided by the multi-view fusion result is particularly striking.

4. Experimental Results

4.1. Experimental Set-Up

To assess the utility of the multi-view fusion, a series of classification experiments was conducted. Four approaches were evaluated, two that employed only single-view data and two that exploited multi-view data in different ways. The first approach used the *unrotated* single-view imagery to learn a deep classifier. The second approach used the *rotated* single-view imagery to learn a deep classifier. The third approach, the proposed method, used the (rotated) multi-view fusion imagery to learn a deep classifier. The fourth approach exploited the dependence between an object's multiple views only at the prediction stage. Specifically, this approach used the deep classifier learned from the rotated single-view imagery, but then averaged the classifier's predictions of the multiple views that were associated with the same object. This last case corresponds to performing multi-view fusion at the prediction level, whereas the third case performs multi-view fusion at the earlier image level.

Table 1. Number of images in the AMiCa data set.

Method	AMiCa	
	Class 1	Class 2
Single-View Unrotated Images	2526	4424
Single-View Rotated Images	20208	35392
Multi-View Fusion Images	10104	17694

Table 2. Number of images in the CATHARSIS 2 data set.

Method	CATHARSIS 2	
	Class 1	Class 2
Single-View Unrotated Images	34	170
Single-View Rotated Images	272	1360
Multi-View Fusion Images	136	679

The number of images available in each data set for each approach, broken down by class, is summarized in Tables 1-2. Class 1 corresponds to the calibrated rocks, while class 2 corresponds to the truncated cones. Because more data was available from the AMiCa trial, that imagery was treated as labeled training data, while the CATHARSIS 2 imagery was treated as test data upon which classification was to be performed.

The deep learning method noted in Sec. 2 was used as the classification approach in all experiments. All deep-classifier parameter settings were kept fixed and identical for all approaches. A two-hidden-layer architecture was employed, with 100 units in layer 1 and 196 units in layer 2. The number of epochs used to greedily pre-train each of the two layers was set to 1000, the number of epochs used to subsequently train the two-layer Boltzmann machine was 500, and the number of epochs used to fine-tune the learned machine via back-propagation was 20. The training data was augmented

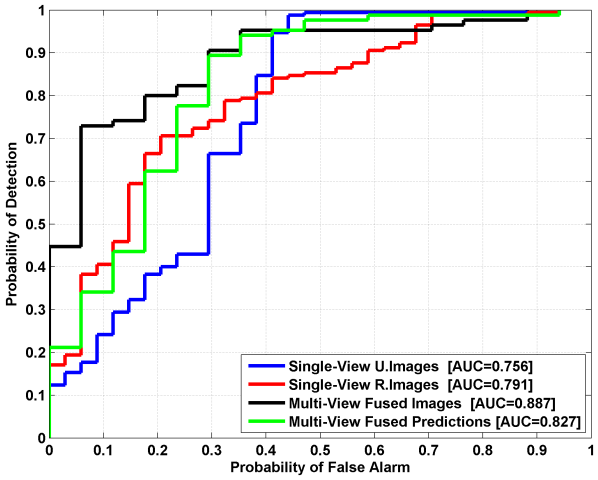


Figure 3. Classification performance on CATHARSIS 2 data after learning the deep classifiers by training on AMiCa data. For the single-view cases, the legend indicates whether the images were unrotated ("U") or rotated ("R").

by "mirroring" each image about the vertical axis, effectively doubling the size of the data sets. To ensure that the training set was balanced (in terms of the number of data points from each class), a subset of data points from the class with more examples was randomly selected and used during each training epoch. To accelerate the learning phase, training was conducted in batches, with 100 data points per batch (50 from each class).

4.2. Results

For the approaches that rotated a given image at a series of 8 angles (*i.e.,* in $\Delta = 45^\circ$ increments), the final prediction for a test image was taken to be the mean of the 8 predictions. The classification performance for the four approaches considered is presented in terms of receiver operating characteristic (ROC) curves in Fig. 3. The area under the ROC curve (AUC) [7], which is a scalar summary measure of an ROC curve, is shown for each approach in Table 3. It can be observed that the proposed multi-view fusion approach achieves the best performance on this limited experiment.

For the deep classifier trained using *unrotated* single-view images, the learned weights for layers 1 and 2, as well as the corresponding "basis images" for the entire deep classifier – constructed as a linear combination of the two layers of weights [8] – are shown in Fig. 4. The analogous results for when the deep classifier was trained using *rotated* single-view images are shown in Fig. 5. The analogous results for when the deep classifier was trained using multi-view fusion images are shown in Fig. 6.

Table 3. Classification performance on CATHARSIS 2 data after training on AMiCa data.

Method	AUC
Single-View Unrotated Images	0.756
Single-View Rotated Images	0.791
Multi-View Fused Images	**0.887**
Multi-View Fused Predictions	0.827

4.3. Discussion

Based on the limited study conducted, the use of a deep classifier in conjunction with multi-view fusion imagery is promising. Because explicit feature extraction is avoided, the various negative issues that arise with multi-view data are elided. However, the image fusion process can introduce other challenges. The classification approach implicitly assumes that the image fusion is robust and accurate. But if an image-fusion result is poor – say, due to one view being characterized by inferior data quality – the classification stage will be adversely affected. Poor fusion imagery in the training phase will contaminate the learning process, while poor fusion imagery in the test phase can undermine the multi-view prediction.

The single-view SAS imagery is well-structured in the sense that there is a consistent, physical target-sensor relationship that results in along-track and across-track image axes. The image fusion process destroys this structure. Potentially valuable shadow information is also usually lost in the fusion process. For the multi-view fusion to be worthwhile, the performance gains due to more complete (highlight) imagery must outweigh these drawbacks.

The experiments here addressed a binary classification task of limited scope, attempting to distinguish truncated cones from calibrated rocks. The true problem of interest for MCM is discriminating targets (*i.e.,* mines) from clutter of all types. Unfortunately, imagery of only purposely *deployed* objects – target shapes and calibrated rocks – has been associated (via manual ground truth) so far. An automatic clustering algorithm to properly link multiple views of a given clutter object is still needed, but AUV navigation errors complicate this task. Nevertheless, a complete study of the benefits of multi-view fusion imagery will eventually need to address the general classification goal.

Figure 4. For the deep classifier trained using *unrotated single-view* images from the AMiCa data set, (a) the learned weights for layer 1, (b) the learned weights for layer 2, and (c) the learned basis images for the entire classifier.

Figure 5. For the deep classifier trained using *rotated single-view* images from the AMiCa data set, (a) the learned weights for layer 1, (b) the learned weights for layer 2, and (c) the learned basis images for the entire classifier.

Figure 6. For the deep classifier trained using *multi-view fusion* images from the AMiCa data set, (a) the learned weights for layer 1, (b) the learned weights for layer 2, and (c) the learned basis images for the entire classifier.

Attempting this more complex classification task would require more data to be available for training. With more data, more sophisticated deep classifiers can be employed, such as deep convolutional neural networks [9] with many more layers of hidden units; we have already begun exploring this approach for single-view classification, for which promising initial results have been obtained [10]. However, even with more advanced classification methods, it is likely that environmentally adaptive methods [11] developed for traditional "shallow" architectures can still be useful.

Future research will investigate the use of meta-features extracted from the image background [12] – to characterize the seabed or environment – as a way to encourage the deep classifiers to favor certain subsets of data during training. This would be necessary because it is unlikely that the diverse clutter class would be fully represented in all environments of interest.

Finally, it is also of interest to generate multi-view image fusion data sets from additional sea trials so that more extensive experiments can be conducted. For example, it would make sense to train the deep classifiers using multi-view imagery collected during sea trials at different geographical locations to increase the variability within the data set and to improve the robustness of the subsequent learned classifier. The great diversity of the clutter class is again the driving factor for needing to augment the available training data.

5. Conclusion

A preliminary investigation into the use of multi-view fusion imagery for object classification tasks was undertaken. The promise of the proposed framework, which exploits multiple views of a given object at the image level, rather than later feature or prediction levels, was demonstrated on real SAS data collected by the MUSCLE AUV. Importantly, the approach offers a way to utilize both single-view and multi-view imagery in a single consistent framework. The fusion imagery used in the study was generated by an active contour-based method, while a deep classifier based on Boltzmann machines was employed for learning. The study provides one potential starting point from which the value of adaptive surveys [13] – wherein multiple views of an object are obtained – can be assessed. Future research will focus on addressing classification tasks with wider scope, and doing so with more sophisticated (deep) classifier architectures.

References

[1] P. García-Laencina, J. Sancho-Gómez, and A. Figueiras-Vidal, "Pattern classification with missing data: a review," *Neural Computing and Applications*, vol. 19, no. 2, pp. 263–282, 2010.

[2] T. Marston, J. Kennedy, and P. Marston, "Coherent and semi-coherent processing of limited-aperture circular synthetic aperture (CSAS) data," in *Proc. IEEE OCEANS*, 2011, pp. 1–6.

[3] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," 2016, book in preparation for MIT Press. [Online]. Available: http://www.deeplearningbook.org

[4] S. Dugelay and W. Fox, "Active contours for synthetic aperture sonar snippet registration," in *Proc. IEEE OCEANS*, 2015, pp. 1–6.

[5] R. Salakhutdinov and G. Hinton, "Deep Boltzmann machines," in *International Conference on Artificial Intelligence and Statistics*, 2009, pp. 448–455.

[6] D. Williams, "Fast target detection in synthetic aperture sonar imagery: A new algorithm and large-scale performance analysis," *IEEE Journal of Oceanic Engineering*, vol. 40, no. 1, pp. 71–92, 2015.

[7] J. Hanley and B. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, pp. 29–36, 1982.

[8] H. Lee, R. Grosse, R. Ranganath, and A. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proc. 26th International Conference on Machine Learning (ICML)*, 2009, pp. 609–616.

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

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

[11] D. Williams and E. Fakiris, "Exploiting environmental information for improved underwater target classification in sonar imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 10, pp. 6284–6297, 2014.

[12] D. Williams, "Fast unsupervised seafloor characterization in sonar imagery using lacunarity," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 11, pp. 6022–6034, 2015.

[13] D. Williams, F. Baralli, M. Micheli, and S. Vasoli, "Adaptive underwater sonar surveys in the presence of strong currents," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 2604–2611.

m

Document Data Sheet

Security Classification		Project No.
Document Serial No.	**Date of Issue**	**Total Pages**
CMRE-PR-2017-005	November 2017	9 pp.

Author(s)

Williams, D.P., Dugelay, S.

Title

Multi-view SAS image classification using deep learning.

Abstract

A new approach is proposed for multi-view classification when sonar data is in the form of imagery and each object has been viewed an arbitrary number of times. An image-fusion technique is employed in conjunction with a deep learning algorithm (based on Boltzmann machines) so that the sonar data from multiple views can be combined and exploited at the (earliest) image level. The method utilizes single-view imagery and, whenever available, multi-view fused imagery, in the same unified classification framework. The promise of the proposed approach is demonstrated in the context of an object classification task with real synthetic aperture sonar (SAS) imagery collected at sea.

Keywords

Multi-view Classification, deep learning, Synthetic Aperture Sonar (SAS).

Issuing Organization

Science and Technology Organization
Centre for Maritime Research and Experimentation

Viale San Bartolomeo 400, 19126 La Spezia, Italy

[From N. America:
STO CMRE
Unit 31318, Box 19, APO AE 09613-1318]

Tel: +39 0187 527 361
Fax:+39 0187 527 700

E-mail: library@cmre.nato.int