DETECTION OF NORTH ATLANTIC RIGHT WHALES WITH A HYBRID SYSTEM OF CNN AND DICTIONARY LEARNING

Ali K Ibrahim , Hanqi Zhuang , Nurgun Erdol , and Ali Muhmed Ali
Electrical Engineering and Computer Science
Florida Atlantic University
Boca Raton Fl, USA
aibrahim2014, zhuang, erdol, amuhamedali2014@fau.edu

***Abstract*—In this paper, a hybrid approach of using Convolutional Neural Network (CNN) and dictionary learning is proposed for the detection of North Atlantic Right Whale (NARW) up-calls. The CNN was applied to extract features from NARW up-calls after these sound signals are converted into spectrograms. The features were then used to train a dictionary with a dictionary learning algorithm. The hybrid system of CNN and dictionary learning was compared with other methods using empirical data. It is shown in the paper that the proposed hybrid system produces superior results for detecting NARW up-calls.**

***Index Terms*—North Atlantic Right Whale, Convolutional Neural Network, Dictionary Learning**

I. INTRODUCTION

The North Atlantic Right Whale (NARW) is one of the critically endangered whales whose birth rate is too low to compensate its death rate [1,2]. The number of NARWs recently recorded in the east coast of North America is only about 300-500 [3]. Tracking their numbers and migration paths in order to prevent accidental deaths and promote their reproduction is vital to sustaining their existence. Acoustic monitoring of NARWs can be accomplished by detecting their up-calls, their signature vocalizations, which are narrow band signals with frequency swings in the range of 50-250 Hz [4]. Early attempts at detection consisted of single-stage algorithms that used edge detection [5] and time-frequency domain convolution [6]. They had relatively high levels of false positive errors [7]. Later methods with feature extraction and classification capabilities [7, 8] were able to drastically reduce the probability of false alarm. Various multiple-stage methods were also proposed in [9]. Recently, researchers devised detection schemes based on convolutional neural networks [10].

In our earlier work [11-13], we applied a number of feature extraction algorithms to process NARW up-calls. A successful attempt is to combine Discrete Wavelet Transform with Mel-frequency Cepstral Coefficients. The up-calls were then classified with a number of popular classifiers. We reported a detection accuracy of 92% using these hybrid feature sets with the SVM classifier. We argued also the use of DWT with MFCC to extract features from NARW calls is intuitive and inspired by the practice of marine biologists.

A promising method for detection and classification is sparse representation and dictionary learning (DL)[14]. In sparse coding, a dictionary is the core containing major characteristics learned from input data. Sparse coding reduces representation complexity while improving accuracy [15-17]. Dictionary learning further decreases the size of the dictionary, and hence improves the efficiency of the classification algorithm. The effectiveness of dictionary learning algorithms is, however, highly correlated with the data used to train the dictionary.

Convolutional Neural Networks are a type of deep neural network designed specially to explore the geometry of images in order to accommodate prior knowledge to the model [18,19]. CNNs are built upon a hierarchical architecture in order to be able to learn deeper features presented by the image data set. Its scalability is one of the main advantages. Nevertheless, CNNs have some drawbacks. Since the loss surface is non-convex, a global minimal is not assured. Moreover, the last layer of the traditional CNN is the fully-connected softmax layer, which is a classifier less efficient, in general, than a sparse classifier.

In this paper, we propose a hybrid system to overcome the drawbacks of each technique while taking advantage of the best characteristics of both techniques. The algorithm begins with a convolutional neural network whose sole purpose is to obtain features of NARW up-calls. The extracted features are then used to train a dictionary using a dictionary learning algorithm. Extracting features with CNNs overcomes the limitations of hand-crafted features, and applying dictionary learning classifiers for detection improves significantly the detection accuracy. It is shown with empirical studies that the hybrid system produces superior accuracy performance in comparison with other algorithms, including the CNN with the softmax layer for NARW up-call detection.

II. DATASETS

Cornell University researchers deployed an auto-detection buoy network near the Cape Cod Bay area. Cornell University also distributed a dataset of NARW up-calls and non-calls for

ISBN 978-90-827970-1-5

researchers around the world. The dataset consists of 30,000 training samples and 54,503 testing samples. Each sample is a 2-second .aiff sound clip with a sample rate of 2 kHz. Dataset contains mixture of right whale calls, non-biological noise and other sounds. The task is to create an algorithm for detecting right whale calls. The spectrogram of an up-call sound is shown in Fig. 1. The original labels have some errors. In our research, we checked every audio clip to make sure that the labels were consistent with the audio signals.

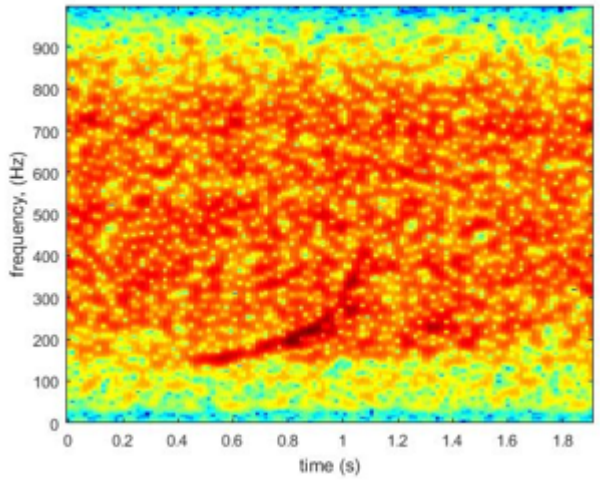

Fig. 1. A typical NARW up-call sound spectrogram.

III. SYSTEM DESCRIPTION

A. Convolutional Neural Networks

A convolutional neural network is a type of deep neural network specialized mainly in treating image data. Taking into consideration that images are more or less homogenous in their local patches, one uses weight sharing which describes the idea of different units within the same layer to use identical weights in order to create filter banks that extract geometrically related features from an image dataset [19]. The process is progressed hierarchically over many layers in order to obtain higher level features at later layers. Fully connected layers are added at the top of the architecture in order to function similar to perceptron classifiers. The network is normally trained using backpropagation techniques. The filter banks used in CNNs have an attribute kernel size, which is related to the number of weights used to extract the features of a patch of the input image. In the proposed hybrid model, the kernel size is chosen to be 5x5. The first convolutional layer was defined to have 32 kernels and the second one to have 64 filters. Each convolutional layer is followed by a max pooling layer. The final layer is the softmax layer with 64 fully-connected nodes.

B. Dictionary Learning

Sparse learning is a way which aims at finding a sparse representation of a given dataset in the form of linear combination of its basic atoms which is stored column-wise in a dictionary. Researchers have devised a number of optimization techniques to find a better dictionary. One of the early attempts is the K-SVD method introduced in [14]. The K-SVD algorithm is based on solving the following problem:

$$< D, X >= \arg\min_{D,X} \|Y - DX\|_2^2 s.t \|x_i\|_0 \leqslant T_0 \quad (1)$$

where T_0 is the sparsity constraint, making sure that each sparse representation x_i contains not more than T_0 non-zero entries. The Dictionary D is obtained by K-SVD algorithm on the training samples Y. The main goal of the discriminative K-SVD, an extension to the K-SVD algorithm, is to use the given label matrix, H, to learn a linear classifier W$\in R^{\text{mxM}}$, while obtaining a sparse representation of the signal, x_i, and returning the most probable class this signal belongs to. A straightforward approach, described in [20], is to solve the following problem :

$$W = \arg\min_{D,X} \|H - WX^*\|_F^2 + \lambda \|W\|_F^2 \quad (2)$$

where λ is the regularization parameter. This problem has the following closed form solution:

$$W = H(X^*)^T (X^*(X^*)^T + \lambda I)^{-1} \quad (3)$$

The drawback of this solution is that learning the classifier W is done independently from learning the dictionary D and the sparse codes X and is thus suboptimal due to the fact that the dictionary learning procedure does not take into account the fact that its output will be used to train a classifier. To overcome the sub-optimality of the K-SVD algorithm for classification discussed above, [17] proposes to incorporate the classification error term directly into the dictionary learning formulation in (1), forcing the algorithm to simultaneously learn the dictionary and the classifier. The authors formulate the joint dictionary-classifier learning problem in a couple of ways, one of which is as follows:

$$\begin{aligned} < D, W, X >= \arg\min_{D,X} \|Y - DX\|_F^2 + \|H - WX\|_2^2 \\ s.t \|x_i\|_0 \leqslant T_0 \\ = \arg\min_{D,X} \| \begin{bmatrix} Y \\ \sqrt{\alpha} H \end{bmatrix} - \begin{bmatrix} D \\ \sqrt{\alpha} W \end{bmatrix} X \|_F^2 \end{aligned} \quad (4)$$

where α is a regularization parameter. The authors showed experimentally that this indeed increases the discriminative power of the resulting classifier. The D-KSVD algorithm is summarized below:

C. hybrid system

In order to build a model for NARW detection based on the hybrid CNN-DL Model, we train the CNN using the backpropagation algorithm with a softmax layer. The output of the remaining layers (before the softmax layer) is used to extract features from the spectrograms related to the NARW up-calls. The proposed system is shown in Fig.2.

ISBN 978-90-827970-1-5 © EURASIP 2018

Discriminative K-SVD

Input: $Y \in R^{n\times N}, H \in R^{m\times M}, \gamma, T_0$

Output: $D \in R^{n\times N}, W \in R^{m\times M}, X \in R^{m\times M}$

1. Initialize:
1.1. Compute D^0 using an initialization scheme of choice, e.g., by concatenating class-specific dictionaries found with K-SVD.
1.2 Compute X^0 for Y and D^0 using sparse coding.
1.3 Compute W^0 using Eq. 3 for $\lambda = 1$:

$$W^0 = Hz^0$$

where $z^0 = X^{0^T}\left(X^0 X^{0^T} + I\right)^{-1}$

2. K-SVD:

Solve PD_KSVD; use $\left(D^{0^T}, \sqrt{\gamma} W^{0^T}\right)^T$ to initialize the dictionary.

3. Normalize $D \leftarrow \left\{\frac{d_1}{\|d_1\|_2}, \frac{d_2}{\|d_2\|_2}, \dots\dots, \frac{d_K}{\|d_K\|_2}\right\}$

$$W \leftarrow \left\{\frac{W_1}{\|d_1\|_2}, \frac{W_2}{\|d_2\|_2}, \dots\dots, \frac{W_K}{\|d_K\|_2}\right\}$$

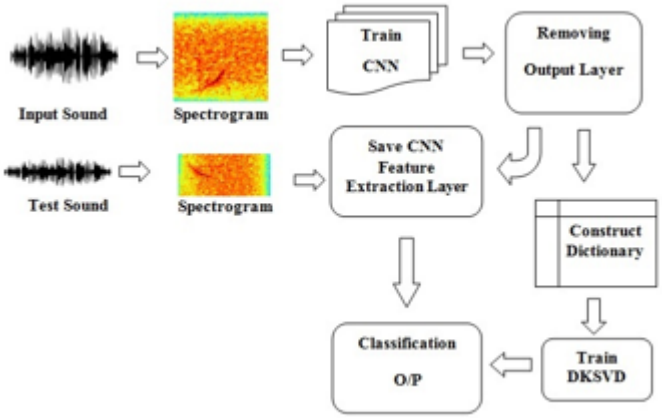

Fig. 2. The proposed hybrid system.

IV. EXPERIMENTAL RESULTS

As was mentioned in Section II, the dataset used in this research consists of the sound files of the NARW up-calls. In our experiments, we used 30 thousand sounds for training and 5 thousand for testing. The label of each sound was validated by a human observer before it was used for either training or testing. Each sound was converted to the spectrogram image by using 0.08 sec frame length, 50% overlapping, and hamming windowing. The output of the CNN was called CNN features, and was used to train the dictionary (5880x25000) with the D-KSVD algorithm. Three different detection rates were used to analyze the detection results:

$$Up\,call\,detection\,rate(UCDR) = \frac{T_p}{P} \quad (5)$$

$$Non\,Up\,call\,detection\,rate(NUCDR) = \frac{T_N}{N} \quad (6)$$

$$False\,alarm\,rate(FAR) = \frac{F_P}{N + F_P} \quad (7)$$

In the above equations, Tp is True Positive (correct detection); Tn is True Negative (correct detection); Fp is False Positive (wrong detection); and Fn is False Negative (wrong detection). Furthermore, P is number of correctly detected cases and N is number of wrongly detected cases. Table I shows the average test accuracy of ten experiments from the CNN and CNN-DL algorithms. Note that when the CNN algorithm is used for both feature extraction and classification, a softmax classifier was adopted. To comparison, we also included results in which features were extracted using Mel-Frequency Cepstral Coefficients (MFCC). It is observed that the highest rate of correct detection is achieved by using the CNN as a feature extractor from spectrogram with dictionary learning as a classifier. The detection rate 92.37 % and a false alarm rate at 1.42%. The lowest detection rate is by using MFCC features with DL. The second row of Table 1 also reveals that MFCC together with CNN achieved 89.72% up-call detections with a false alarm rate at 5.24%.

TABLE I
DETECTION RESULTS BY USING SLANTLET AND LRSDL

Method	UCDR	N-UCDR	FAR
Spectrogram+ CNN	86.32%	97.87%	7.48%
MFCC+CNN	89.72%	96.78%	5.24%
MFCC+DL	76.82%	97.16%	3.64%
Spectrogram+CNN+DL	92.37%	97.37%	1.42%

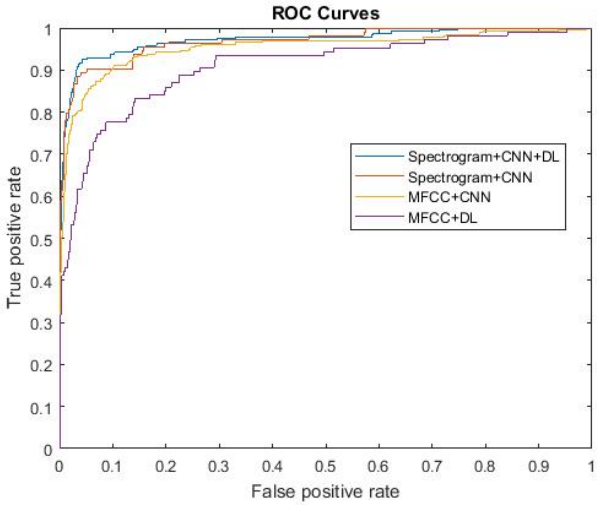

Fig. 3. ROC plot of different classifiers.

To visualize the overall classification results, the Receiver Operating Characteristics (ROC) curves, which are the plot of

ISBN 978-90-827970-1-5 © EURASIP 2018

true positive rate (correctly classified up-calls) against false positive rate (non-up-calls classified as up-calls), are shown for all scenarios in Fig. 3. The closer the ROC curve follows the vertical axis and then the top border of the figure, the more accurate the classifier is. It can be seen that the CNN-DL algorithm consistently outperforms other algorithms.

V. CONCLUSIONS

Our results show that a hybrid system of CNN and DL is capable of achieving good performance for the detection of NARW up-calls. Using CNNs for feature extraction removes the need of handcraft feature extraction algorithms, thereby avoiding any limitations associated with any of such procedures. Replacing softmax classifiers with more powerful sparse coding classifiers based on dictionary learning improves further the detection accuracy. With sparse coding classifiers, one obtains a sparse (in general more compact) representation of the input signal. Intuitively, this is a good thing for detection and classification, since a simpler model with fewer parameters may have a smaller risk of overfitting the training data. We will study this in the future research.

REFERENCES

[1] D. K. Mellinger, K. M. Stafford, S. E. Moore, R. P. Dziak, and H. Matsumoto, "An overview of fixed passive acoustic observation methods for cetaceans," Oceanography, vol. 20, pp. 36-45, 2007.

[2] S. M. Van Parijs, C. W. Clark, R. S. Sousa-Lima, S. E. Parks, S. Rankin, D. Risch, and I. C. Van Opzeeland, "Management and research applications of real-time and archival passive acoustic sensors over varying temporal and spatial scales," Mar. Ecol. Prog. Ser., vol. 395, pp. 21-36, 2009.

[3] R.R. Reeves, B.D. Smith, E.A. Crespo, and G. Notarbartolo di Sciari, "Dolphins, Whales and Porpoises: 20022010 Conservation Action Plan for the Worlds Cetaceans," IUCN/SSC Cetacean Specialist Group, Chapter 4, 2003.

[4] C.W. Clark, "The Acoustic Repertoire of the Southern Right Whale, a Quantitative Analysis," Anim. Behav. 30, 1060-1071,1982.

[5] D. Gillespie, "Detection and classification of right whale calls using an edge detector operating on a smoothed spectrogram," Can. Acoust. Vol 32, pp. 39-47,2004 .

[6] D. K. Mellinger and C. W. Clark, "A method for filtering bioacoustic transients by spectrogram image convolution," Proceedings of the IEEE: OCEANS, vol. 93 no. 3, pp. 122-127 1993.

[7] I. R. Urazghildiiev, C. W. Clark, T. P. Krein, and S. E. Parks, "Detection and recognition of North Atlantic right whale contact calls in the presence of ambient noise," ," IEEE J. Ocean. Eng. vol. 34, pp. 358-368,2009.

[8] D. K. Mellinger, "A comparison of methods for detecting right whale calls," Can. Acoust. 32, 55-65,2004.

[9] I. R. Urazghildiiev, and C. W. Clark, "Acoustic Detection of North Atlantic Right Whale Contact Calls Using the Generalized Likelihood Ratio Test," J. Acoust. Soc. Am.120, 1956-1963,2006.

[10] S. Evgeny, "North atlantic right whale call detection with convolutional neural networks," Proceedings of the 1st Workshop on Machine Learning for Bioacoustics, ICML, 2013.

[11] M. Esfahanian, H. Zhuang, and N. Erdol, "parse Representation for Classification of Dolphin Whistles by Type," J. Acoustical Society of America EL, Vol. 136 (1), July, 2014.

[12] M. Esfahanian, H. Zhuang, and N. Erdol, "On Contour-based Classification of Dolphin Whistles by Type,"J. Applied Acoustics, Vol. 76, pp. 274-279, Feb 2014.

[13] A K Ibrahim, H. Zhuang, N. Erdol, and A. Muhamed Ali, "A new approach for North Atlantic Right Whale up-call detection," IEEE International Symposium on Computer, Consumer and Control , Xian China,2016.

[14] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," IEEE Trans. Signal Proc. 54, no. 11,pp.4311-4323,2006.

[15] A K Ibrahim, H. Zhuang, N. Erdol, and A. Muhamed Ali, "An Approach for Facial Expression Classification," International Journal of Biometrics (IJBM). 9, 2017.

[16] A K Ibrahim, H. Zhuang, N. Erdol, and A. Muhamed Ali, "EEG seizure detection by integrating slantlet transform with sparse coding," Signal Processing Conference (EUSIPCO), 2017 25th European, 459-462,2017.

[17] Ali K Ibrahim, Laurent M. Chrubin, Hanqi Zhuang, Michelle T Schrer Umpierre, Fraser Dal-gleish, Nurgun Erdol, B. Ouyang, and A. Dal-gleish, "An approach for automatic classification of grouper vocalizations with passive acoustic monitoring," J. Acous. Soc. Ame., 143:2, pp. 666-676,2018.

[18] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 22782323, 1998.

[19] O. Abdel-Hamid, A. r. Mohamed, H. Jiang and G. Penn, "Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition," in Proceedings IEEE Int. Conf. Acoustics, Speech and Signal Processing, Kyoto, Japan, pp. 4277-4280,2013.

[20] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in Proc. IEEE Conf. Comp. Vis. Pattern Rec., pp. 2691 2698, 2010.

ISBN 978-90-827970-1-5 © EURASIP 2018