The Classification of Underwater Acoustic Targets Based on Deep Learning Methods

Hao Yue, Lilun Zhang, Dezhi Wang*, Yongxian Wang and Zengquan Lu

Academy of Marine Science and Engineering, National University of Defense Technology, Changsha, China

*Corresponding author

***Abstract*—The underwater target classification is a challenging task due to the complexity of marine environment and the diversity of underwater target features. Most of the-state-of-the-art target recognition systems depend on feature extraction schemes based on expert knowledge in order to effectively represent the target signatures. In contrast, 16 different kinds of underwater acoustic targets are categorized in this paper by using Convolution Neural Network (CNN) and Deep Brief Network (DBN), which can achieve the accuracy up to 94.75% and 96.96% respectively in both supervised and unsupervised fashions. To compare with the results of traditional machine learning methods, we also use Support Vector Machine (SVM) and Wndchrm to do the same job and the latter is originally a tool applied for the biological image analysis. The results show that deep learning methods can achieve higher recognition accuracy when classifying the underwater targets from their radiation noises.**

Keywords-underwater target; classification; recognition; deep learning; DBN, CNN

I. Introduction

The key to underwater target recognition is the feature extraction. The current feature extraction methods mainly include time domain feature extraction, spectral estimation techniques, time-frequency analysis and so on. With the expansion of the underwater acoustic datasets, the original feature extraction methods are gradually ineffective. Therefore, it is of great significance to carry out new underwater target recognition methods.

Deep learning is a new field in machine learning research and it was proposed by Hinton in 2006 [1]. In recent years, it has made a breakthrough in the fields of speech analysis, image recognition and so on. One of the outstanding properties of deep learning is that it can capture the deep features hidden in the target signals through multi-level network architecture without structure features designed artificially.

DBN and CNN are the two famous deep learning methods, which respectively use unsupervised and supervised learning models. The first real multi-layer learning algorithm CNN proposed by Y. Lecun has been successful in handwriting recognition. In 2012, the deep convolution neural network was applied to ImageNet and achieved astonishing results [2]. DBN is an unsupervised model raised by Hinton in 2006 and it was designed to solve the deep structure-related optimization problems [3]. In 2012, it was the first time to apply this unsupervised learning method for the construction of acoustic models and achieved a great success [4]. Our work aims to use these two networks to identify the underwater acoustic targets. In order to show the effectiveness of deep learning in underwater target recognition, we adopt the popular Mel Frequency Cepstral Coefficient (MFCC) to preprocess the dataset. Usually, MFCC is used to extract the characteristics of speech signals, in our experiment it is used as a feature extraction technique of underwater targets. As one of the most popular classifiers, SVM is then used to do the subsequent classification task.

Meanwhile, in this paper, we still use another traditional method to compare with the deep learning algorithms. This method identify the target through extracting the features of LOFAR (Low Frequency Analysis Recording) spectrum converted by raw audio data. In our implementation, we used a tool called Wndchrm to do this job. It was developed by Lior Shamir and used in the classification of whale calls in 2014 [5] [7]. Now, we apply it to the recognition of underwater targets.

The rest of this paper is organized as follows. Section 2 briefly describes the related work on the recognition of underwater targets. Section 3 briefly introduces the approach we used. Section 4 introduces our experiment dataset. Section 5 presents the details of implementation. In Section 6 we discuss the experimental results and errors. Finally, we conclude our work in section 7.

II. Related Work

In [6], a convenient open toolkit LIBSVM was developed by CC Chang, which is easy for users to implement the SVM algorithm efficiently.

In [7], Lior Shamir et al. applied Wndchrm to the classification and recognition of whale calls, which greatly improved the accuracy and efficiency compared with the hand engineered feature extraction schemes.

In [8], Suraj Kamal et al. applied DBN to passive target recognition tasks. And the results showed generative DBN can extract more stable and more expressive features of the target than the schemes based on expert knowledge of underwater acoustic signal processing. And it has achieved 90.23% in accuracy on a dataset with 40 categories of underwater targets.

III. Approach

A. Overview of Our Approach

In this paper, we use DBN and CNN to classify the 16 classes of underwater targets respectively, at the same time, we also used traditional SVM and Wndchrm to do the same job and analyzed their accuracy and properties.

B. DBN

DBN consists of a stack of Restricted Boltzmann Machines (RBM) trained in greedy manner one layer at a time, and the DBN structure used in this paper shown in Figure I. The training process of DBN can be divided into pre-training and fine-tuning. In the process of pre-training, each RBM is trained without supervision separately to ensure that the feature vectors are mapped to different feature spaces as much as possible. The learned weights of RBM stack is used to prime the input layer of a traditional back propagating neural network(BPNN) classifier which attributes the class labels to corresponding classes and the entire network is fine-turned by back propagating the classification error.

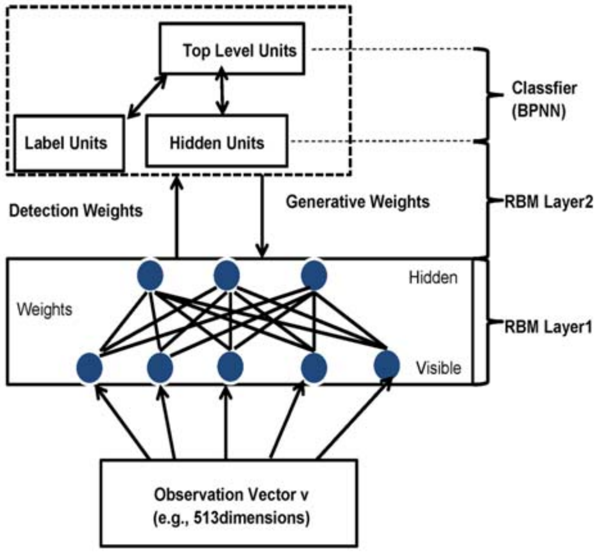

FIGURE I. ARCHITECTURE OF DBN

C. CNN

CNN mainly consists of two types of layers which are convolution and pooling (sampling) layers. The role of the convolution layer is to extract the various features of the image and the pooling layer is used to abstract the original characteristic signal. CNN processing is also divided into forward training and back propagating. The forward training outputs predictive probability vector through extracting and pooling target features and the back propagating is used to feedback through evaluating the error between prediction and ground truth. The network structure used in this paper is shown in Figure II.

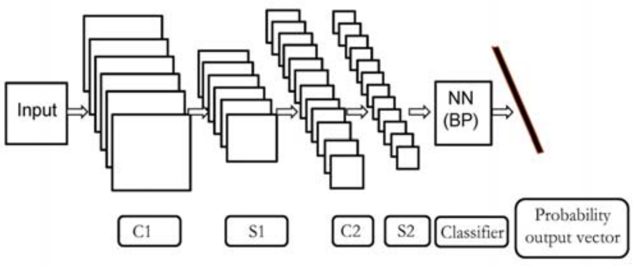

FIGURE II. ARCHITECTURE OF CNN

D. WNDCHRM

It’s a traditional method by extracting the features of LOFAR spectrum of the underwater targets because it can reflect the features of the signal in time-frequency domain [9]. In this paper, we use the short-time Fourier transform (STFT) to obtain the LOFAR spectrum of the training samples, and then use the Wndchrm tool to train different class of LOFAR spectrum. The software works by first extracting image content descriptors from the raw image, image transforms, and compound image transforms just as for Fourier transform(FFT), Gabor transform etc. In the extraction process, 11 different algorithms are used. Then, the most informative features are selected by Fisher Score algorithms, and the feature vector of each image is used for classification and recognition. The process is showed in Figure III.

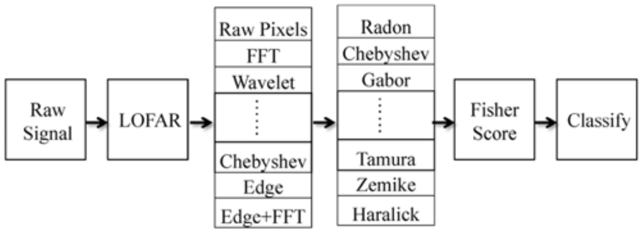

FIGURE III. ARCHITECTURE OF WNDCHRM

IV. DATASET

The dataset used in this study is obtained from the Historic Naval Sound and Video database website. It consists of 16 different underwater target recordings in World War II, such as cruiser, torpedo, submarine and acoustic signals of other underwater targets. The audio files are all in WAV format. Through the preliminary analysis of these underwater target noises, we found that the low frequency features are dominant. So, we just only care about the low frequency part of the noise, for that we resampled the signal to 2kHz from 22kHz. For preparing the training set, we sliced each audio file into multiple overlapping frames and each frame is 512ms with a 5ms step. Then we take a Fourier transform (FFT=1024) for each frame. Finally, we obtained 5000 samples for each class, so the amount of total samples is 80000(16 * 5000). We randomly selected 3/4 of each type for training, 1/4 for testing.

V. IMPLEMENTATION

In this section, we detail the experimental settings and the model implementation.

A. Experimental Settings of FFT+DBN

For training a DBN, initially the RBM should be trained in an unsupervised manner with the training set. We designed the first RBM with an input of 513 dimension since the FFT (1024) spectrum is symmetric with the y axis. Then, we found a 513-100-100-16 nodes construction of entire DBN. The training numepoch of both RBM layers is set to 10, and the activation function is Sigmoid. In fine-tuning step, we select Softmax function to calculate the output probability vector and set the numepoch to 20. Other parameters are shown in Table I.

TABLE I. THE PARAMETERS SETTING OF DBN

Parameters	***Momentum***	***Learning rate***	***Batch size***
Value	0.5	0.1	10

B. Experimental Settings of FFT+CNN

Usually, CNN is used for extracting two-dimensional features, so we reshaped the feature vector which consists of 512 obvious neurons into a two-dimensional feature matrix with the size of 16*32. In our experiment, we designed a CNN net with a 6c-2s-12c-2s structure. In detail, it consists of 2 convolution layers and 2 pooling layers. The first convolution layer consists of 6 feature maps and the second convolution layer consists of 12 maps, but the size of convolution kernels are all 5*5. After convolution, a 2*2 mean pooling is used for abstracting the extracted features by convolution layer. Finally, a full-connection is designed as the input of a BP neural network for obtaining all the features extracted by different feature maps. In addition, the loss function is set to Mean Square Error, learning rate set to 1, batchsize set to 10 and the numepoch set to 15.

C. Experimental Settings of STFT+WNDCHRM

First, we sliced each type of audio recording into 200 frames, then we take a short-term Fourier transform (FFT = 1024, Hann Window) to obtain the LOFAR spectrum, as shown in Figure IV. So, we obtain 200 samples for each class and select 3/4 of them for training and the rest 1/4 for testing randomly. After training 20 times, we can obtain the mean accuracy.

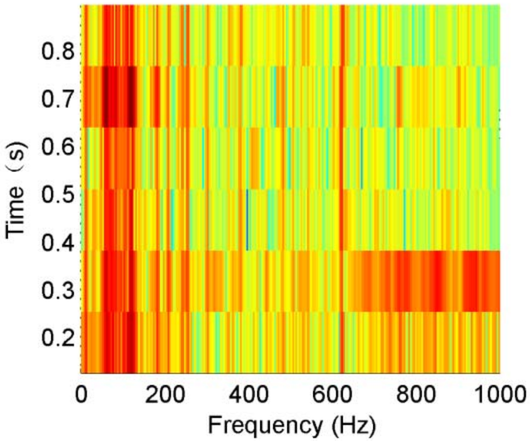

FIGURE IV. AN EXAMPLE OF LOFAR SPECTROGRAM

D. Experimental Settings of MFCC+SVM

Similarly, each audio recording was sliced into multiple overlapping frames and each frame is 100ms with a 25ms step, then the MFCC of each frame was calculated and used as the input feature of a single sample. Figure V is an example of MFCCs feature spectrum of an underwater target. In our experiment, we had 1000 samples of each class and selected 3/4 of them for training and the rest for testing randomly.

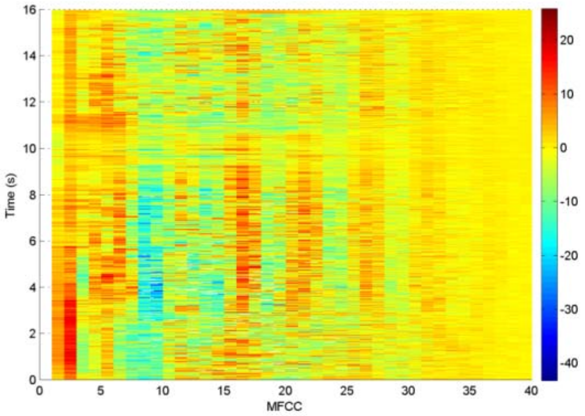

FIGURE V. AN EXAMPLE OF MFCCS SPECTROGRAM

VI. RESULTS ANALYSIS

Experimental results of the four approaches mentioned above are compared in Table II and Table III.

TABLE II. THE ACCURACY OF DIFFERENT APPROACH

Approach	***MFCC+SVM***	***WNDCHRM***	***FFT+CNN***	***FFT+DBN***
Accuracy	86.6%	92.15%	94.75%	96.99%

Table II shows the accuracy for each approach from the results, we can see that MFCC+SVM and Wndchrm are also feasible for recognizing the underwater acoustic targets and could obtain an acceptable accuracy, although the former is usually used in speech recognition and the latter often used in biological image.

The deep learning methods can achieve significantly better results than the traditional SVM and Wndchrm. Furthermore, by comparing these two deep learning methods, FFT+DBN can obtain a higher accuracy with 96.99% when recognizing the underwater targets by using acoustical recordings.

At the same time, we compared the pros and cons of these four approaches through combining the characteristics of the model itself and the details during our implementation. We present the brief information in Table III.

TABLE III. THE COMPARISON OF DIFFERENT APPROACHS

	Pros	***Cons***
MFCC+SVM	Suitable for small dataset, low computational complexity	Not suitable for big data and complex selection of the optimal parameter
WNDCHRM	Suitable for 2D analysis, a collection of multiple algorithms	Not suitable for big data, Computing redundancy
FFT+CNN	Suitable for 2D analysis and large dataset, high accuracy	Supervised learning, requiring labelled samples, relatively high computational cost
FFT+DBN	Unsupervised method without labelling samples, high accuracy	Complex model parameters and structures, high computational cost

VII. CONCLUSIONS AND FUTURE WORK

In this paper, the deep learning methods are applied to the underwater target recognition problem. We use CNN and DBN to classify the underwater acoustic targets and make a comparison with the analysis tool Wndchrm and the method combing MFCC and SVM. The results show that it can achieve higher accuracy when using deep learning models to classify the underwater targets from their acoustic noises. However, we all know that the deep learning model is effective just when it is driven by a large amount of data while the audio dataset of underwater targets is often difficult to obtain. So, it’s necessary to apply large underwater acoustic datasets in the applications of deep learning methods in the classification of underwater targets in further studies.

ACKNOWLEDGMENT

This study was funded by the National Key Research and Development Project of China (No. 2016YFC1401800) and the Scientific Research Project of NUDT (No.ZK16-03-46, No.ZK16-03-31).

REFERENCES

[1] G. E. Hinton, R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” Science 2006, vol. 313, pp. 504-507.

[2] A. Krizhevsky, I. Sutskever I, G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012, pp. 1097-1105.

[3] G. E. Hinton, S. Osindero, Y. W. Teh. “A fast learning algorithm for deep belief nets,” Neural Computation 2006, vol. 18, pp. 1527-1554.

[4] A. Mohamed, G. E. Dahl, G. E. Hinton. “Acoustic Modeling Using Deep Belief Networks,” IEEE Transactions on Audio Speech & Language Processing 2012, vol. 20, pp. 14-22.

[5] L. Shamir, N. Orlov, D. M. Eckley, T. Macura, J. Johnston, I. G. Goldberg, “Wndchrm – an open source utility for biological image analysis,” Source Code for Biology and Medicine, 2008, pp. 3-13.

[6] C. C. Chang, C. J. Lin, “LIBSVM: A library for support vector machines,” Acm Transactions on Intelligent Systems & Technology, 2011, vol. 2, no. 3, pp.27-32.

[7] L. Shamir, C. Yerby, R. Simpson, A. M. von Benda-Beckmann, P. Tyack, F. Samarra, et al, “Classification of large acoustic datasets using machine learning and crowdsourcing: Application to whale calls,” Journal of the Acoustical Society of America, 2014, vol. 135 No. 2, pp. 953-962.

[8] S. Kamal, S. K. Mohammed, P. R. S. Pillai, M. H. Supriya, “Deep learning architectures for underwater target recognition,”.Sympol 2013, pp. 48-54.

[9] Z. Y. Song, Y. P. Ding, X. L. Zhao, L. Weng, “The Method of Underwater Target Recognition Based on LOFAR Spectrum,”, Journal of Naval Aeronautical & Astronautical University, 2011.vol. 3, pp. 47-50