Sonar Image Translation Using End to End Network for Underwater Recognition

Jin Hua, Mengzhao Liu, Daxin Xin and Yixin Liu
College of Electronic Information Engineering, Xi’an Technological University, Xi’an 710021, China
Email: huahua_dz@aliyun.com

Abstract. With the rapid development of sonar technology, the classification of underwater sonar images is of great significance. In order to recognize underwater sonar images more effectively, side scan sonar is used for data acquisition. After pre-processing, six types of sonar images are selected. Combining with the characteristics of underwater sonar images, relevant theories and methods of convolutional neural network are studied to establish an underwater sonar CNN recognition model, which achieves 98.5% recognition success rate. Experiments show that the recognition model can improve the accuracy, speed and robustness of underwater sonar image classification, and can meet the practical application requirements.

1. Introduction
At present, the classification methods of underwater sonar images are mainly to extract the texture features of underwater sonar images by different methods and to complete the classification by matching the features. However, due to different feature extraction methods and different underwater sonar images used for classification, underwater sonar image classification methods cannot be widely used. By obtaining the common features of the same type of images, an excellent model which is widely used in underwater sonar image classification is trained and used to identify, classify or predict underwater sonar images in practice.

2. Data Set Preparation and Network Structure

2.1. Data Set Preparation
Underwater sonar images belong to a small sample data set. The dataset is divided into six categories, including rocks, wrecks, rocks, sandlines, terraces, and cracks. After obtaining the original data, it cannot be directly used as the training data of the model, and some preprocessing operations are needed [1]. After the standard naming, the data set was grayed into 1200 training sets and 600 test sets. According to the central limit theorem, environmental noise can be approximated by Gaussian distribution. Gaussian white noise with mean of 0 and standard deviation of 25, 35 and 45 was added to the image to simulate environmental noise. Pretreatment mainly studies the denoising method of environmental noise.

2.2. Pretreatment
Some indicators are needed to evaluate the denoising effect of the algorithm. The commonly used evaluation criteria are peak signal-to-noise ratio (PSNR), mean square error (MSE) and average absolute error (MAE) [2]. The denoising effect of terrace image is shown in figure 1.
The PSNR of BM3D is still the highest among the four algorithms, and both MSE and MAE are lower than the other three algorithms. As shown in Table 1, the median filter is better than the gaussian filter in the rock denoising effect, but the gaussian filter in the terrace image is slightly better than the median filter.

![Noisy image](sigma=25) ![Average neighborhood](sigma=25) ![Median filtering](sigma=25) ![Gaussian filter](sigma=25) ![BM3D](sigma=25)

![Noisy image](sigma=35) ![Average neighborhood](sigma=35) ![Median filtering](sigma=35) ![Gaussian filter](sigma=35) ![BM3D](sigma=35)

![Noisy image](sigma=45) ![Average neighborhood](sigma=45) ![Median filtering](sigma=45) ![Gaussian filter](sigma=45) ![BM3D](sigma=45)

**Figure 1.** The denoising effect of terrace image.

**Table 1.** Comparison of denoising effect of terraced fields.

| σ  | Denoising algorithm     | PSNR  | MAE    | MSE    |
|----|-------------------------|-------|--------|--------|
| 25 | Average neighborhood    | 21.19 | 0.0446 | 0.0037 |
|    | Median filtering        | 24.42 | 0.0364 | 0.0026 |
|    | Gaussian filter         | 24.97 | 0.0362 | 0.0021 |
|    | BM3D                    | 27.36 | 0.0181 | 0.0012 |
| 35 | Average neighborhood    | 20.78 | 0.0487 | 0.0042 |
|    | Median filtering        | 24.06 | 0.0382 | 0.0031 |
|    | Gaussian filter         | 24.58 | 0.0371 | 0.0027 |
|    | BM3D                    | 26.96 | 0.0199 | 0.0016 |
| 45 | Average neighborhood    | 20.15 | 0.0496 | 0.0028 |
|    | Median filtering        | 23.82 | 0.0398 | 0.0039 |
|    | Gaussian filter         | 24.17 | 0.0393 | 0.0032 |
|    | BM3D                    | 26.51 | 0.0214 | 0.0019 |

### 3. Network Structure

In order to find the area of interest of the sonar image and align it in the coordinate system, ROI{net} takes the original sonar image as input and outputs the area of interest image of the sonar image. As can be seen from the following figure 2. The first half of ROI{net} is a pre-trained VGG16 network, splice after Pool3, and the top is LRN, which is used as the feature extractor of the network. In this way, L2 normalized feature maps can be generated to retain the spatial information in sonar images. The second part of the feature map connected to ROI{net} is a fully connected homing network. The two hidden layers...
of FC1 and FC2 have 1024 and 512 neurons. After these two layers, relu activation function and Dropout are used to avoid the overfitting phenomenon of neural network. The local network is composed of feature extraction network and regression network, and then output normalized coordinates. The normalized coordinates $\theta$ are forwarded to the grid generator which transforms a regular, square grid $G$ to a deformed grid $T_\theta(G)$ based on $\theta$ and the normalization range is between -1 and 1. A bilinear sampler takes the deformed grid $T_\theta(G)$ as an input and samples the original hand image $I$ to form a regular grid of $h_{ROI} \times w_{ROI}$ pixels, which is the ROI image [3]. ROInet is connected to FERnet, which is responsible for the extraction and recognition of sonar image features. To extract sonar image features, the ROI image was fed into another independent CNN, which was also trimmed after Pool3, with VGG-16 of LRN at the top. Although the feature extraction network is the same between them, there is no weight sharing. LRN is an output layer, $f$sROI is space said sonar ROI. Spatial dimensions are defined in the grid generator in ROInet. So LRN connects to a Dropout layer and a full connection layer FC4, and the final output is a 512-dimensional vector $f_{PD}$. The vector $f_{PD}$ is passed to the other Dropout layer and the full connection layer FC5, which returns the sonar image tag.

![Figure 2. Model structure.](image-url)
4. Training Strategy

In this experiment, at the beginning, in order not to destroy the convolutional layer of pre-training, the best way is to train only the top layer of random initialization. At the same time, you can train the bottom layer. In the first stage, the localised network is trained to initialize ROI net, which is used to extract ROI of sonar images. In the second phase, FERnet is trained based on ROI. In the third stage, the model was trained end-to-end, and the ROI net initialized in stage I was connected with FER net for training. In short, stage III is end-to-end training, and Stage II is not. In the second stage, sonar ROI is input, and in the third stage, complete sonar images are input. While training the final model, the ROI input to FERnet depends on the input sonar image and ROI net [4].

As can be seen from the following table 2. In S1-S5 strategy, after ROI net is connected to FERnet, FERnet conducts training according to the sonar image training network. In the S0 strategy, only FERnet is trained separately. The top levels of module C, FC4 and FC5, participate in the training of all strategies. After 20 epochs, the local network was microadjusted and training was stopped at 40 Epochs. ROI net extracts ROI images and inputs them into FERnet. Of course, strategy S0 is not an end-to-end exercise. In the S1 strategy, all layers of FERnet and ROI net were off, parameters in the C module only updated N epochs and closed Drop1 and Drop2. Strategy S2 is the same as strategy S1, but Drop1 and Drop2 are open. Strategy S3 was the same as S2 and S4, but after 20 epochs, the learning rate was reduced to 0.0001, and the B module in ROI net was fine-tuned. S3 closed Drop1 and Drop2, and S4 opened Drop1 and Drop2. Finally, strategy S5 was the same as strategy S4, but after 35 epochs, Drop1 and Drop2 were turned off [5].

| Table 2. Details of the different training strategies. |
|-----------------------------------------------|
| FERnet | ROI net |
| fc4, fc5 | fc1, fc2, fc3 | Drop1, Drop2 |
| S0 | II | C | [1, N] | / | / | / | / |
| S1 | III | C | [1, N] | / | [1, N] | off | [1, N] |
| S2 | III | C | [1, N] | / | [1, N] | on | [1, N] |
| S3 | III | C | [1, N] | B | (20, N) | off | [1, N] |
| S4 | III | C | [1, N] | B | (20, N) | on | [1, N] |
| S5 | III | C | [1, N] | B | (20, N) | on/off | [1, 35] / (35, N) |

5. Experiment and Analysis

ADAM optimizer is used in this network, with momentum of 0.8, batch size of 64 and learning speed of 0.001. After network training, it will converge. The ROI image is 224×224 resolution. In ROI net, adjust the size of the input image I for 56×56. ROI image size is set to 112×112 pixels, so $f s ROI$ size is 14×14×256. Drop1, Drop2, Drop3 and Drop4 value is set to 0.5, 0.1, 0.2 and 0.5. Leakyrelu was used as the activation function, and the parameter value in Leakyrelu was 0.2 [6]. In layer FC1-FC5, the weights and biases are randomly sampled from a Gaussian distribution of 0 mean and 0.001 variance [7].

In order to verify the robustness of the model, different noise levels are added to the data in the normal test data set to simulate different underwater environments [8]. As can be seen from the following table 3, with the increase of noise intensity, the PSNR value keeps going down and the image quality becomes worse and worse. When the number of test samples remains unchanged, the effective recognition number and the recognition rate decrease [9].

The results show that the test results of this method in the existing data set are superior to other recognition and classification methods. Meanwhile, the change of recognition rate of various methods with the change of noise intensity is recorded in table 3. Without adding noise, the recognition rate reached 98.5%. With the increase of noise intensity, the accuracy of all recognition methods decreases. When the standard deviation is over 55, the recognition rate of the model is lower than 60%; when the standard deviation is 95, the recognition rate is 27.1%; the recognition rate of other methods is basically...
lower than 10%, and the robustness of the model is stronger than that of other methods. The BP model has the lowest recognition rate and low robustness, and the classification accuracy of SVM and AdaBoost is lower than that of other convolutional neural network models, which proves the advantage of applying the convolutional neural network and the feasibility and practicability of applying the model in this paper to sonar images [10].

| σ  | LeNet5 | GoogleNet | VGG16 | SVM  | BP   | AdaBoost | Ours |
|----|--------|-----------|-------|------|------|----------|------|
| 0  | 87.5%  | 90.8%     | 93.1% | 82.9%| 80.3%| 84.2%    | 98.5%|
| 15 | 85.1%  | 89.6%     | 90.3% | 81.2%| 77.8%| 82.7%    | 96.0%|
| 25 | 78.4%  | 80.4%     | 82.7% | 73.4%| 63.1%| 77.3%    | 87.3%|
| 35 | 70.6%  | 72.1%     | 73.1% | 61.3%| 52.7%| 72.6%    | 81.2%|
| 45 | 59.0%  | 60.8%     | 61.8% | 50.7%| 49.3%| 66.1%    | 68.8%|
| 55 | 47.5%  | 48.7%     | 51.7% | 38.9%| 44.7%| 50.9%    | 60.1%|
| 65 | 34.6%  | 42.2%     | 39.2% | 29.6%| 38.5%| 38.7%    | 54.5%|
| 75 | 21.7%  | 23.6%     | 30.8% | 18.7%| 31.4%| 23.8%    | 38.0%|
| 85 | 12.1%  | 13.3%     | 15.2% | 6.8% | 24.8%| 8.9%     | 32.8%|
| 95 | 5.4%   | 7.2%      | 7.8%  | 4.7% | 11.3%| 3.5%     | 27.1%|

6. Conclusion
In this paper, six kinds of sonar images are selected and combined with the characteristics of underwater sonar images, relevant theories and methods of convolutional neural network are studied, and an underwater sonar CNN recognition model is established to classify these six kinds of targets. In the pretreatment, BM3D algorithm can remove the environmental noise better than the other three algorithms. Five training strategies were compared and tested to choose the best one. The experimental results show that the robustness of the model is tested by adding noise, and the recognition rate and robustness of the model are higher than those of other methods after comparative analysis. General research of image data from the same waters, the experimental conclusion to some extent, the lack of universality, want the future to the sonar image can collect more data, or experimental study results are integrated to more waters, establish the database of different waters and classification criterion, let different bottom sediment classification of underwater environment problem can get differentiation treatment, also let the unknown underwater environment according to the forecast.

Acknowledgement
This paper was supported by National sonar Key Laboratory Open Fund (6142109KF2018); National Fund Youth Project (11804263).

References
[1] Agostinelli F, Anderson M R and Lee H 2013 Adaptive multi-column deep neural networks with application to robust image denoising Advances in Neural Information Processing Systems 1493-1501.
[2] Schmidhuber J 2014 Deep learning in neural networks: An overview Neural Networks 61 85.
[3] Ioffe S and Szegedy C 2015 Batch normalization: Accelerating deep network training by reducing internal covariate shift arXiv preprint arXiv:1502.03167.
[4] Radford A, Metz L and Chintala S 2015 Unsupervised representation learning with deep convolutional generative adversarial networks arXiv preprint arXiv:1511.06434.
[5] Gu S, Zhang L, Zuo W, et al. 2014 Weighted nuclear norm minimization with application to image denoising IEEE Computer Vision and Pattern Recognition 2862-2869.
[6] Chuang M C, Hwang J N and Williams K 2016 A feature learning and object recognition framework for underwater fish images IEEE Transactions on Image Processing 25 (4) 1862-1872.
[7] Chen Q 2012 Research on Underwater Target Recognition Based on Underwater Acoustic Image (Harbin Engineering University).

[8] Gaurav K and Pradeep K B 2014 A detailed review of feature extraction in image processing systems ACCT2014 pp 5-12.

[9] Buades A, Coll B and Morel J M 2005 A non-local algorithm for image denosing IEEE Computer Vision Pattern Recognition Conference Record 2 60-65.

[10] Li Y, Xie W and Li H 2016 Hyperspectral image reconstruction by deep convolutional neural network for classification Pattern Recognition 63 371-383.