# Marine Objects Detection and Segmentation

**By Kushagra Mahajan, Kunal Saini**

## Introduction

We develop a new approach for detecting and segmenting multiple marine objects Ocean Liner, fireboat, container ship, seashore (or only water), speedboat from images based on convolutional neural networks (CNNs). We finetune Alexnet pre-trained on ImageNet on our newly created dataset for the detection task and Segnet for the segmentation task.
We use the DWT filter primarily to determine the resolution and the level of blur of images that can be accurately detected by our model and determine limitations for further improvement. The project is in collaboration with the Indian Naval Research Board.

## Dataset

The dataset consisted of five marine videos captured by the Indian Naval Research Board. These are videos captured from a moving ship of the surroundings which may include other marine objects like ocean liners, speedboats etc. We frames from the five videos using VLC Media Player at 1 frame every 2 seconds of the video. We then randomly created a dataset of 2000 images with roughly a similar distribution of the classes. The size of the JPEG images extracted is 256 * 256. OpenCV was used for resizing the images to the size required by AlexNet (227 * 227 * 3). These images were pixel wise annotated for segmentation by us along with help from some other students in the Computer Vision and Machine Learning Lab at IIIT Delhi.
The training set consisted of 1600 images and the test set consisted of 400 images. The classes that were used for fine-tuning on the dataset: Ocean Liner, fireboat, container ship, seashore (or only water), speedboat. These images were manually annotated by us for the above mentioned classes. We augmented the dataset using techniques like flipping, rotation, translation, adding salt and pepper noise to make the training of the CNN more robust.

## Background Information

### Convolutional Neural Networks (CNNs)

CNNs have become the method of choice for processing visual and textual data. A CNN is composed of one or more convolutional layers with fully connected layers (matching those in typical artificial neural networks) on top. It also uses tied weights and pooling layers.

In comparison with other deep architectures, convolutional neural networks have shown superior results in both image and speech applications. They can also be trained with standard backpropagation. CNNs are easier to train than other regular, deep, feed-forward neural networks and have many fewer parameters to estimate, making them a highly attractive architecture to use.

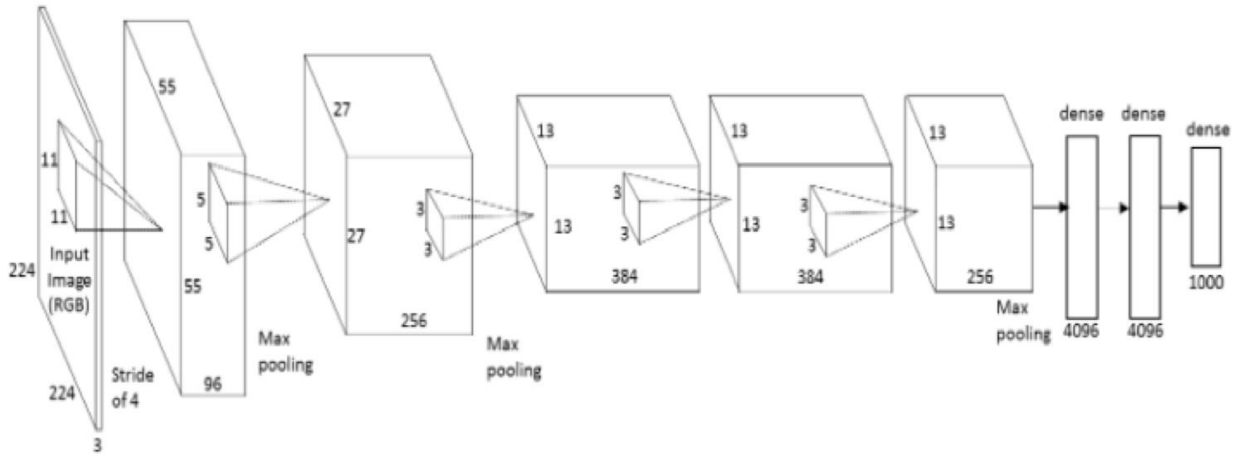


**Fig. 1: AlexNet Architecture**

Fig. 1 shows the architecture of one of the most popular Convolutional Neural Networks ’AlexNet’ named after Alex Krizhevsky. The network achieved a top-5 error of 15.3%, more than 10.8 percentage points ahead of the runner up in the ImageNet Large Scale Visual Recognition Challenge in 2012. The net contains eight layers with weights; the first five are convolutional and the remaining three are fully-connected. The output of the last fully-connected layer is fed to a 1000-way softmax which produces a distribution over the 1000 class labels. We have used AlexNet for classification in this work.

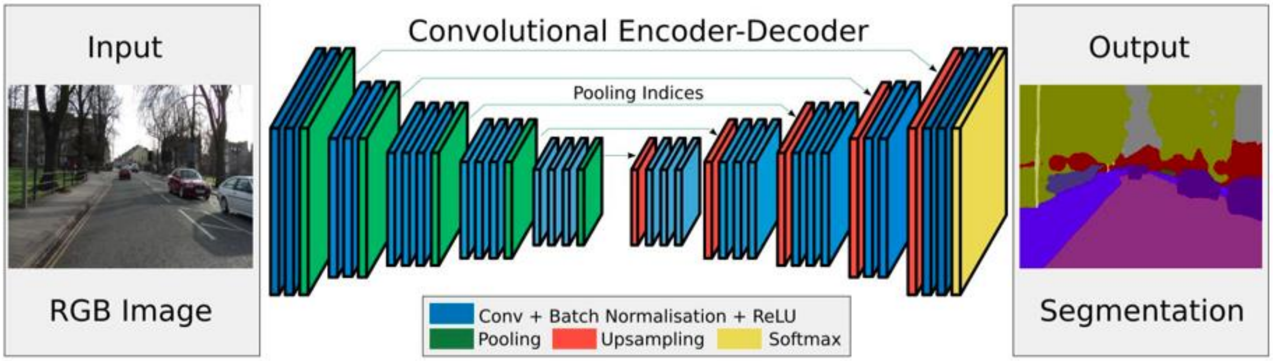


**Fig. 2: Segnet Architecture**

**SegNet** is a deep fully convolutional neural network architecture for semantic pixel-wise segmentation. This core trainable segmentation engine consists of an encoder network, a corresponding decoder network followed by a pixel-wise classification layer. The architecture of the encoder network is topologically identical to the 13 convolutional layers in the VGG16 network as shown in Fig. 2. The role of the decoder network is to map the low resolution encoder feature maps to full input resolution feature maps for pixel-wise classification. The novelty of SegNet lies is in the manner in which the decoder upsamples its lower resolution input feature map(s). Specifically, the decoder uses pooling indices computed in the max-pooling step of the corresponding encoder to perform non-linear upsampling. This eliminates the need for learning to upsample. The upsampled maps are sparse and are then convolved with trainable filters to produce dense feature maps.

## Discrete Wavelet Transform (DWT)

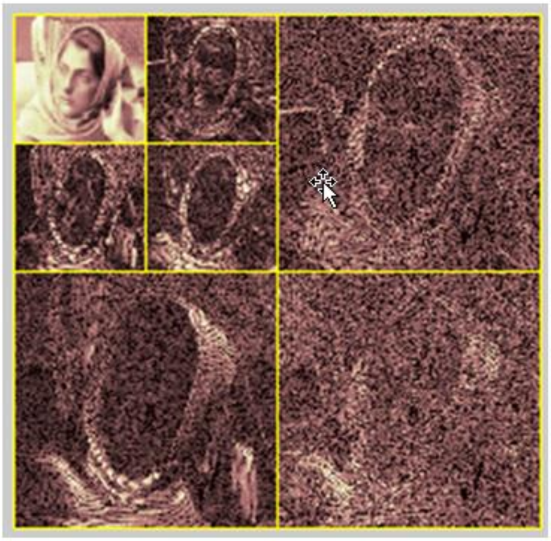


**Fig. 3: DWT Filter**

A **discrete wavelet transform (DWT)** is any wavelet transform for which the wavelets are discretely sampled. As with other wavelet transforms, a key advantage it has over Fourier transforms is temporal resolution: it captures both frequency and location information (location in time). It returns the lowpass (scaling) and highpass (wavelet) filters, Lo and Hi respectively, for the DWT filter bank fb. Lo and Hi are both L-by-2 matrices. L is an even positive integer. The first column of Lo and the first column of Hi are the analysis filters. The second column of Lo and the second column of Hiare the synthesis filters.
The Matlab dwt2 command performs a single level two-dimensional discrete wavelet transform. This kind of two-dimensional DWT leads to a decomposition of approximation coefficients at level j in four components: the approximation at level j + 1(top-left image is the lowpass approximation of the image obtained), and details in the three directions (top-right: horizontal, bottom-left: vertical, and bottom-right: diagonal) as shown in Fig. 3.

## Approach

### For Marine Object Classification

We took Alexnet model pre-trained on ImageNet and finetuned all the layers on 1600 training images of the dataset. The final layer in the model is learnt from scratch and has 5 nodes in the softmax layer (4 classes + 1 background class).

### For Marine Object Segmentation

We took SegNet model pre-trained on the MS-Coco dataset and finetuned the entire model on 1600 training images. We obtain 5 pixel-wise classification maps (4 classes + 1 background class) as output of the model.

## Implementation Details

The proposed method was implemented in Python and C++ using Caffe library. The experiments were carried out on PCs with mid-level graphics card NVIDIA TITAN X and a multi-core 2.1 GHz CPU.

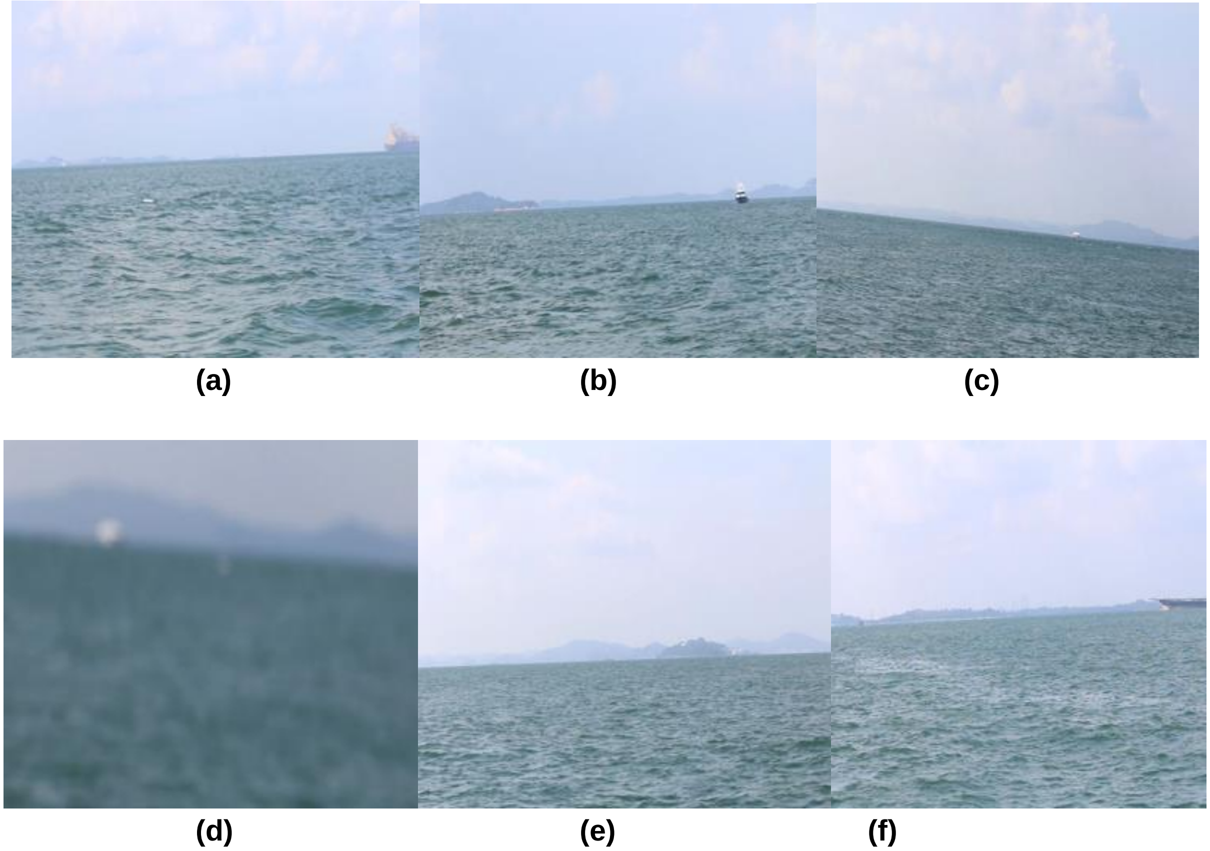
## Experiments

The size of the marine objects in our images varies across a wide range of sizes: from occupying about 40% of the image to occupying only around 2-3% of the image. So, even though the pre-training dataset ImageNet has some classes of marine objects like ocean liner, speedboat etc., transfer learning in this scenario helps to adapt to the specific domain with reduced sizes of the object of interest and of images captured from a moving ship of the surroundings.

The overall classification accuracy achieved on the test set is 63%. The maximum accuracy was achieved by containership (74%) while the minimum was for fireboat (41%). There is a lot of confusion between fireboat a speedboat as distinguishing small sizes of the two is very difficult even for human annotators.

On analysis of the misclassified images, we found that most of the images were either blurred due to the motion blur of the camera resulting in a non-clear object of interest or the object of interest was too small in size for accurate classification.

Some of those misclassified results are:-


(a) (b) (c)

(d) (e) (f)

**Fig. 4: Misclassifications due to blur and small objects**

- In (a), (b) it can be seen that the objects of interest are very small in size and the image is quite foggy.
- In (c), the object is present as it is visible to the human eye, but even the human eye cannot be surely say what is the object.
- In (d), the image is blurred and the object of interest is not at all visible clearly.
- In (e), there is no marine object at all, but it is very difficult to find the difference between (c) (that has some object) and this one.
- In (f), a small part of a marine object is visible but it is aligned with the background hills, so not clear.

The segmentation experiments using SegNet did not give very encouraging results primarily since the size of the objects of interest are extremely small in our original dataset. Hence, for segmentation, we prepared a separate dataset consisting of images in which the object of interest occupied more than 20% of the image. The pixel-wise classification accuracy was as follows:-

- Ocean Liner: 59%
- Fireboat: 42%
- Container ship: 64%
- Speedboat: 58%
- Background: 96%

We applied DWT filter on the images to determine the sizes upto which the images are being detected by our model.

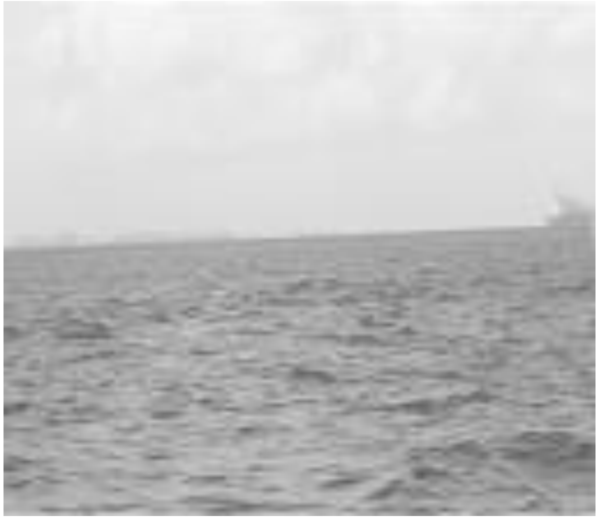


**Fig. 5: Very Small Object of Interest**

In Fig. 5, since the size of the object of interest in the image is too small, none of the image versions (128*128, 64*64, 32*32) are able to detect the presence of the object. In the 256*256 size image, we are able to detect the presence of a marine object.

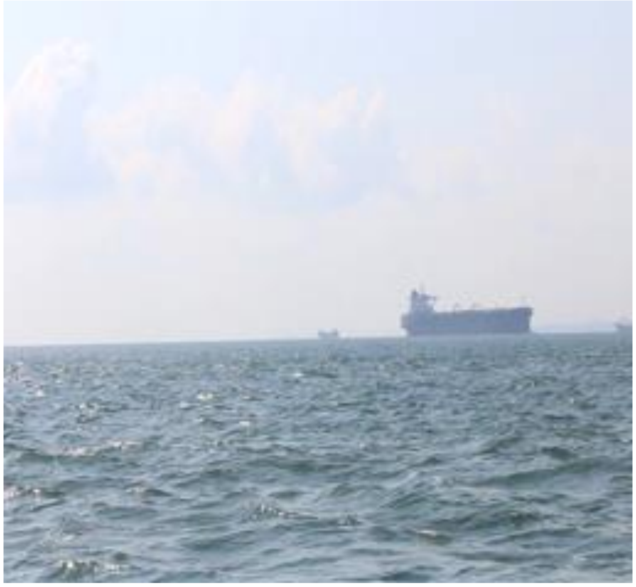


**Fig. 6: Medium-Sized Object of Interest**

In Fig. 6, for sizes 256 * 256, 128 * 128, 64 * 64, the marine object is being detected but for size 32 * 32, we are unable detect the object.

The DWT analysis provides us useful insight about the relative sizes of the image and the object of interest that are accurately classified by our model. They, therefore, allow us to determine the limitations of our model for further improvement.

## RESULT

Through our work, we are able to classify 4 different types of marine objects, segment them out using a state-of-the-art segmentation model, and also quantitatively determine the limitations of our model for further improvement. Future work includes exploring some state-of-the-art work on low resolution object detection and classification like '*Low Resolution Face Recognition Across Variations in Pose and Illumination*' and '*Studying Very Low Resolution Recognition Using Deep Networks*'.

## Acknowledgement

This project is in collaboration with the Indian Naval Research Board. We would like to thank our instructor for the Image Analysis Course, Dr. AV Subramanyam for providing us with this opportunity of working on a deployable system for assisting the Indian Defence forces.

# References

1. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.
2. Jia, Yangqing, et al. "Caffe: Convolutional architecture for fast feature embedding." *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014.
3. Daubechies, Ingrid. "The wavelet transform, time-frequency localization and signal analysis." *IEEE transactions on information theory* 36.5 (1990): 961-1005.
4. Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE transactions on pattern analysis and machine intelligence* 39, no. 12 (2017): 2481-2495.
5. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).
6. Mudunuri, Sivaram Prasad, and Soma Biswas. "Low resolution face recognition across variations in pose and illumination." *IEEE transactions on pattern analysis and machine intelligence* 38, no. 5 (2016): 1034-1040.
7. Wang, Zhangyang, Shiyu Chang, Yingzhen Yang, Ding Liu, and Thomas S. Huang. "Studying very low resolution recognition using deep networks." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4792-4800. 2016.