Deep Learning based Object Detection via Style-transferred Underwater Sonar Images

Sejin Lee* Byungjae Park** Ayoung Kim***

* Kongju National University, 1223-24 Cheonan-daero, Cheonan 31080, S. Korea (e-mail: sejiny3@kongju.ac.kr).
** ETRI, 218 Gajeong-ro, Yuseong-gu, Daejeon, 34129, S. Korea (e-mail: bjlp@etri.re.kr).
*** Department of Civil and Environmental Engineering, KAIST, Daejeon, S. Korea, (e-mail: ayoungk@kaist.ac.kr).

Abstract:
Compared to the flourishing researches on terrestrial optical images, deep learning in underwater imaging has not been highlighted. Although some approaches applied deep learning in their underwater imaging still no major application has been found in underwater sonar imaging. Notably, the fundamental limitation in underwater image data would be the main cause of the bottleneck. To alleviate this issue, this paper introduces a simulation-generated dataset for object detection in underwater sonar images. Specifically, this paper focuses on generating real sonarlike style-transferred synthetic sonar images for network training.

1. INTRODUCTION
Object detection is one of the key components in underwater operation. Many underwater robotic applications Cho et al. (2015); Purcell et al. (2011); Reed et al. (2003); Belcher and Lynn (2000); Lee et al. (2013); Williams and Groen (2011); Galceran et al. (2012a) include inspection and mapping where object detection plays a major role. Obvious perceptual preference may be at using optical images, however, the optical image in underwater often is limited by water turbidity and light condition. Imaging sonar, in the literature, has been highlighted by many underwater researchers mainly due to their robustness to the water turbidity and light condition. On the other hand, the imaging sonar faces difficulties in low resolution and noisy images, together with the elevation loss during the image capture. As also reported Inc. (2018), image analysis in sonar image addresses issue regarding these limitations.
Classification from sonar images was tackled in Automatic Target Recognition (ATR) where shadow effect in sonar image has been an issue Galceran et al. Especially when describing a feature point for image matching and classification, the shadow effect changes the appearance of the neighboring pixel around a feature hindering viewpoint invariance feature description. For example in the work by Galceran et al. (2012b), authors targeted a man-made structure using forward-looking sonar. Since the appearance may vary depending on the viewpoint, the authors generated several inferences when detecting an object. Another work by Zhou and Chen (2004) introduced power spectral analysis to classify the seafloor sediment types. In this line of study devised a metric lacunarity to evaluate salient sonar image Williams (2015).
Recent advances in deep-learning-based methods presented promising results. For example, the author Zhu et al. (2017) utilized feature learning capability of deep learning and coupled with Support vector machine (SVM) for target recognition. The target recognition further developed in Williams (2016), who exploited the deep learning on the captured synthetic aperture sonar (SAS) images. In their work, the recognition ability was evaluated over two target objects, revealing promising results against other typical feature-based classifiers. Using real-world data was validated in Williams (2016). The author trained a network using real marine data collected over the years and applied to a classification problem. Another target focused by researchers was a vehicle. As introduced in Kim et al. (2016), authors leveraged deep learning forward-looking sonar and detected vehicle from obtained images. Despite these approaches presented meaningful stepping stone for the deep learning researches in underwater, the data collection and preparation are the main challenges. When preparing the data, most authors relied on real sonar images and training and test were performed using divided real data into training and test sets.
This type of approach for training data preparation would be ideal if the data is available. However, we would like to address this data scanty issue with the training dataset for underwater application. Efforts are found in the literature. For example in McKay et al. (2017) imported pre-trained network using in-air images and applied to in-water sonar images. Another strategy is to generate realist fake images from simulation or CAD. Denos et al. generated the synthetic images by modifying background over the synthesized images. Similarly, another synthetic
dataset was presented in Chen and Summers (2016) using Generative Adversarial Network (GAN).

Unlike the aforementioned studies, our focus is at developing a more complete solution to prepare sonar images to apply deep learning for underwater object detection. The proposed method will be validated using sonar images captured from the real marine environment, showing the following contributions.

- The proposed solution exploits style transfer given a depth image captured from a simulator.
- We validate that the proposed training method can replace the training from real data.

2. TRAINING SET GENERATION

2.1 Base Depth Images

Prior to the style transfer, we prepare a base depth image from a simulator using a depth camera in the UWSim. Dhurandher et al. (2008). When capturing data, camera pose and capturing altitude were diversified. One strategy when using the simulator for dataset generation is to create photo-realistic images. However, we found that it is easier to generate a photo-realistic image in a post-processing phase, not in the simulator level. Thus, we exploit simulator only when generating this basic depth image. We then convert a colormap of this depth map using Gwon et al. (2017). The color-converted image is then added with white noise and normalized.

2.2 Style Transfer

In this paper, two representative styles were considered, water tank and sea. The network learns these two styles through encoder (E), decoder (D), and StyleBank (K).

The StyleBankNet consists of two branches, auto-encoder (E → D) and stylizing branch (E → K → D). Firstly, the reconstruction loss is used to train the auto-encoder branch, targeting to create an output image O that is similar to an input image C.

\[ \mathcal{L}_R(C, O) = \|O - C\|^2, \]  

(1)

Secondly, the perceptual loss was adopted in stylizing branch similarly in Johnson et al. (2016).

\[ \mathcal{L}_P(C, S_i, O_i) = \alpha \cdot \mathcal{L}_c(O_i, C_i) + \beta \cdot \mathcal{L}_s(O_i, S_i) + \gamma \cdot \mathcal{L}_{reg}(O_i) + \delta \cdot \mathcal{L}_{atk}(O_i, S_i), \]  

(2)

\(S_i\) is the \(i^{th}\) stylized image. We used the typical losses as in Johnson et al. (2016), feature reconstruction loss (\(\mathcal{L}_c(O_i, C_i)\)), style reconstruction loss (\(\mathcal{L}_s(O_i, S_i)\)), and regularization loss (\(\mathcal{L}_{reg}(O_i)\)). Additionally in this paper, we introduce a new average top-k intensity (ATKI) loss for sonar image (\(\mathcal{L}_{atk}(O_i, S_i)\)).

\[ \mathcal{L}_s(O_i, S_i) = \sum_{l \in I_i} ||G(F^l(O_i)) - G(F^l(S_i))||^2, \]  

(3)

where \(F^l\) is the feature map and \(G\) is Gram matrix computed from \(l^{th}\) layer of VGG-16 layers \(I_s\).

Modification to the original perceptual loss in Johnson et al. (2016) was made by the ATKI loss. This is to compensate for the characteristic of the sonar image. Motivated by the ATKI Fan et al. (2017), the proposed ATKI loss focuses on the intensity distributions of brighter portion between images.

\[ \mathcal{L}_{atk}(O_i, S_i) = \frac{1}{k} \sum_{j=1}^{k} ||O_G[i] - S_G[i]||^2, \]  

(4)

\(O_G[i]\) and \(S_G[i]\) are the \(j^{th}\) largest intensity values.

2.3 Style Transfer Training

A content set should be prepared for StyleBankNet training. The content set includes base depth images and two major style sets of water tank and sea. Each set consists of 300 images with objects in the center. A single mini-batch uses randomly sampled content images and style images. In each iteration, we did not fix the depth-style images pair for generalization. A \((T + 1)\)-step alternative training strategy is employed to ensure balanced learning Goodfellow et al. (2014).

3. APPLICATION TO OBJECT DETECTION

3.1 Network Architecture

For object detection, the Faster Regions with Convolutional Neural Networks (R-CNN) Ren et al. (2017) was used. Instead of using region proposal algorithms such as EdgeBoxes Zitnick and Dollár (2014) or Selective Search Uijlings et al. (2013), this paper adopted Faster R-CNN. For the Region Proposal Network (RPN) training, we use the following layers; Input layer (32 × 32 × 3), 1st Convolution layer (5 × 5 × 32), ReLu, MaxPooling (3 × 3), 2nd Convolution layer (3 × 3 × 64), ReLu, MaxPooling(3 × 3), 3rd Convolution layer(3 × 3 × 32), ReLu, MaxPooling(3 × 3), Fully-connected Layer(200), ReLu, Fully-connected Layer(2), Softmax Layer, Classification layer.

3.2 Training Image Augmentation

The proposed synthesizing scheme targets various sonar images. In sonar images, the object may look brighter or darker depending on the surrounding environment. To incorporate this characteristic, we include an inverted version of the base depth image. Basically, this is done by inverting the intensity value in the normalized depth image. We also augmented the dataset by including rotation, translation, and scaling.

4. EXPERIMENTAL RESULTS

4.1 Experimental Setup

We train Style transfer and object detection using one NVIDIA GTX 1080 and Adam optimizer. The learning
rates were set to $10^{-3}$ with exponential decay. Weight decay, $\beta_1$ and $\beta_2$ were set to $10^{-5}$, 0.9 and 0.999. We build the StyleBankNet using PyTorch. Each branch of the StyleBankNet uses own Adam optimizer to update the parameters in the StyleBankNet. A total of 100 epochs were performed to train the StyleBankNet. In each iteration, content and style images were randomly resized and cropped for data augmentation. For the evaluation, Intersection Over Union (IOU) larger than 0.25 was considered to be the correct detection.

4.2 Dataset

Real sonar images were prepared by viewing a submerged dummy. For image acquisition, we used a multibeam imaging sonar (Teledyne BlueView M900-90) having a 90° field of view, 20° beam width and 100 m maximum range. Images were acquired from a pool and from the sea. We denote each dataset as follow: depth image from a simulation (SIM), water tank styled image (SIM-POOL), and sea styled image (SIM-SEA2017).

The POOL, was captured in a water tank with a maximum depth of 10 m. During acquisition, the dummy was positioned at about 4 m depth. An Unmanned Surface Vehicle (USV) was equipped with sonar while tilting the sonar at 5° angular interval. The SEA images were obtained in turbid water. During the data collection, the sonar was fixed under the kayak and tilted about 30° downward from the water’s surface. We maintained the distance between the sensor and the target to be around two to four meters.

The POOL dataset was collected twice, and the datasets were named SEA2017 and SEA2018. The dataset of POOL, SEA2017 and SEA2018 has 735, 1045, and 1935 images.

4.3 Simulation Training Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Average precision (IoU 0.25)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEA</td>
<td>0.65</td>
</tr>
<tr>
<td>POOL</td>
<td>0.19</td>
</tr>
<tr>
<td>Styled POOL</td>
<td>0.63</td>
</tr>
<tr>
<td>Styled SIM</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Training from real sonar images would be ideal. To test this idea scenario, we trained the network using images from SEA2017 and tested over images in SEA2018 to establish the baseline. Then other cases will be compared against this baseline. As a result depicted in Table 1, the baseline provides around 0.65 average precision.

Using the baseline, we examined the performance when trained from 1) a water tank (735 images from POOL), 2) stylized images from a water tank (370 images from SIM-POOL) and 3) simulator using style transfer (370 images from SIM-SEA2017). As expected, the performance in terms of average precision is slightly degraded compared to the baseline. However, please note that the effect of style transfer for training was meaningful, revealing a comparable performance to the baseline.

5. CONCLUSION

This paper reported the empirical evaluation of the style transferred training image for underwater object detection. By doing so, we expect the simulation based images could be used to elaborate the training set and overcome the data scanty in underwater deep learning researches. We presented the effectiveness of the proposed image synthesizing by using the real-world training dataset as the baseline.

REFERENCES


