Assessing Perception Quality in Sonar Images Using Global Context

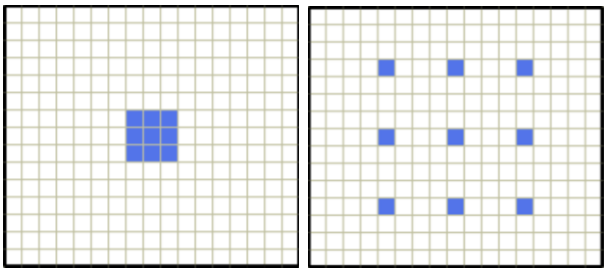
Robert DeBortoli, Austin Nicolai, Fuxin Li, Geoffrey A. Hollinger
Collaborative Robotics and Intelligent Systems (CoRIS) Institute, School of Mechanical, Industrial & Manufacturing Engineering
Oregon State University, Corvallis, Oregon 97331
Email: {debortor, nicolaia, fuxin.li, geoff.hollinger}@oregonstate.edu

***Abstract*— In this work we address the problem of identifying informative images gathered by underwater sonar. We define informative frames as those containing enough clarity and substance (e.g. target of interest) to be useful in higher level tasks such as map building or the inspection of underwater environments. This classification is important, as noise degrades many of the sonar images captured to the point of uselessness, and analyzing them for these higher level tasks is computationally wasteful. To identify high-quality frames we leverage the atrous convolution architecture which is composed of dilated filters that utilize a more global context than standard convolutional filters. This is particularly important in the sonar domain, where compared with standard cameras there is a lack of strong local features and a large amount of noise. We test our model on sonar imagery of underwater objects gathered using a sonar mounted on an underwater vehicle. We show that our atrous model identifies high-quality frames with a higher average precision than previous approaches when completing transfer learning without sacrificing performance in situations where the training data is visually similar to test data.**

I. Introduction

Underwater vehicles are now used in a variety of applications, including ship hull inspection, underwater mine detection, and underwater surveying [1], [2], [3]. During such missions, oftentimes exteroceptive sensory data in the form of camera or sonar images, is captured for mission support. However due to the large amount of data being captured, it is often difficult for a human to extract meaning in realtime [4]. In conjunction with this difficulty, humans uniquely have the ability to complete higher-level tasks such as the annotating of image features in sonar imagery [5]. This motivates the need for introspection on such vehicles: human operators often cannot monitor large streams of image data in realtime; however they are capable of completing higher level tasks for mission success. In this work we enable the automatic processing of underwater sonar data for proposal of high-quality information to a human operator who can use this for tasks such as finding an object of interest or the 3D reconstruction of underwater environments.

Due to the turbidity of water, sonar sensors, as opposed to standard cameras, are oftentimes the preferred sensor in underwater exploration and monitoring missions. Standard underwater imaging sonars can produce images at tens of Hz. However many of the frames produced do not contain useful information (e.g. target of interest, clear imagery of the environment). The poor quality is often the result of noise in the images, the possible sources of which include multipath reflection off the seafloor, non-diffuse reflection of the acoustic wave off of the object, and the interference of projected and reflected acoustic waves [6]. An example of a sonar target in the shape of an X that is corrupted in image space can be seen in Fig 3b. In our experiments we found up to 64% of the incoming images did not contain enough clear features for a human operator to recognize the object being displayed. Thus the perceptual algorithms (or human in charge of the mission) waste valuable time analyzing poor and uninformative imagery. In this work we develop a perceptual quality assessment network which classifies sonar images as either containing useful information or not. This approach would relieve researchers of the need to constantly observe this data as well as allow autonomous systems to spend computational resources more efficiently.



(a) Dilation rate of 1 (b) Dilation rate of 4

Fig. 1: Example of traditional 3x3 filter (left) and a 3x3 dilated filter (right). The dilated filter uses a larger neighborhood but still only uses 9 pixels to compute the output.

Traditional algorithms for image quality assessment that were developed for cameras are insufficient for use in sonar as the amount of noise is much higher and the resolution of sonar imagery is much lower [7]. A more global context than traditional methods allows for noise and larger features to be identified more reliably.

To achieve this increase in global context, in this work we utilize the atrous convolution architecture which has previously not been examined on sonar data. Although initially applied in the signal processing domain, atrous convolution has gained recent popularity in the computer

This work was funded in part by Department of Energy contract DE-EE-0006816.

vision community for tasks such as dense feature extraction as well as multi-scale image analysis [8], [9], [10]. In contrast to a standard convolutional neural network (CNN) this architecture uses dilated filters as seen in Fig. 1. Dilated filters use pixels in a larger neighborhood around the point of interest than standard convolutional filters. This increase in size of context leads to a more generalizable classifier that is less dependent on strong local features.

In this work we develop a network that is extendable: it is able to identify informative frames containing objects different than those trained on. We show that our network is able to achieve this transfer learning capability without compromising performance in situations where the test data is similar to the training data. We also show that our model is able to classify images in realtime on sonar data taken of different objects across different deployments of our vehicle. Both realtime processing and transfer learning capabilities are important characteristics in deploying systems that will be able to reason about the world around them in a safe and efficient manner.

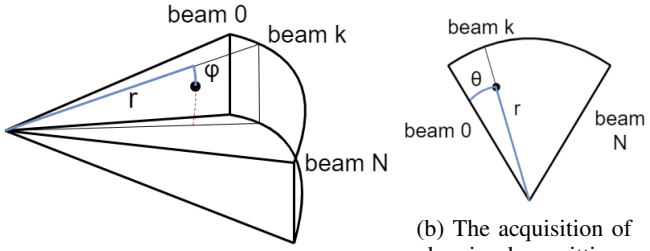
II. Background

A. Sonar

In order to understand the need for a global context in sonar image classification, it is important to detail the sonar imaging process. In this work we use a Gemini Tritech 720i Multibeam Sonar onboard a tethered Seabotix vLBV300 underwater vehicle. The tether provides realtime data transmission of sonar images to an offboard computer which can be used for realtime processing and display. The sonar is a low-cost and portable model designed for mounting on small underwater Remotely Operated Vehicles (ROVs). It is in a class of imaging sonars including the SoundMetrics DIDSON and Aris Explorer 3000 as well as BlueView's M900-90 which have been used in previous work [5], [11], [12]. In this work we develop a method that is appropriate for use with imaging sonars in general, including the models listed above.

The sonar insonifies objects by emitting an acoustic wave and measuring the time-of-flight and strength of return for the wave reflected back from the environment. The time-of-flight measurement gives the range r to the object and sending out multiple waves, or beams, as shown in Fig. 2b gives bearing measurement θ to an object. The strength of return defines the pixel value (0-255) for that return in sonar image space. As can be seen in Fig. 2, during this imaging process, the elevation angle ϕ of the object is lost.

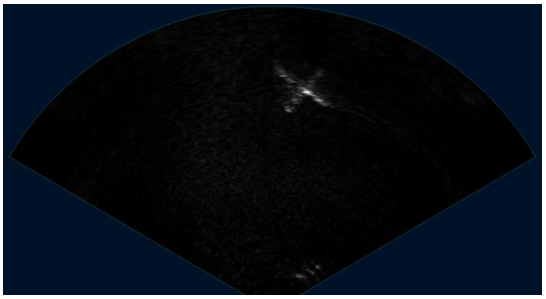
Given the goal of finding informative frames, the image quality assessment problem for sonar can been approached in two general ways: global image analysis or more localized image feature extraction. Kalwa and Madsen developed a global method for predicting image quality by training a neural network using 3 features: the mean pixel value (as a measure of noise), entropy (to measure complexity of the image), and substance (a custom metric which measures the amount of strong returns in the image) [13]. Our preliminary tests of this method suggest that because our images include



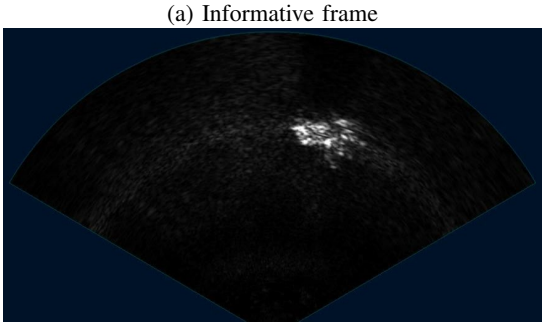
(a) Computation of range and elevation angle

(b) The acquisition of bearing by emitting multiple beams.

Fig. 2: Mapping from Euclidean to polar coordinates.



(a) Informative frame



(b) Non-informative frame

Fig. 3: Examples of informative (an object is easily recognizable) and non-informative frames while inspecting a target in the shape of an X. In our experiments we found on average about 39% of the captured frames to be informative.

nonuniform noise as well as noise that contains many strong returns, that using only these three features is not expressive enough for our applications.

Feature extraction from sonar imagery has received much research attention lately. Ji, et al. use standard computer vision techniques (Canny edge detection and Hough transforms) to extract features from sonar imagery. However these techniques rely on the use of strong gradients in the image to identify points of interest [11]. Given the strong gradients of corrupted imagery, as seen in Fig. 3b, these approaches are not robust for our applications. Johannsson, et al. identify objects by using the strong image gradients between the objects and their shadows [14]. While this approach works well in their data gathering process, the presence of a strong shadow is the result of viewing the object from a low altitude and thus is not a fully general approach. We note that shadows were not prevalent in the sonar imagery we

collected. Aykin and Negahdaripour use the brightest pixels in the image as features which are then clustered to form objects [15]. Given that noise can often appear as strong returns in sonar imagery, this approach has the potential to return many false positives and is thus not appropriate for our purposes.

There also exists work similar to ours in finding representative imagery for use in summarizing underwater missions. Kaeli, et al. create summaries of camera imagery over a long sequence by quantifying, based on a prior, how "surprising" it is to see certain images [4]. They quantify this in part by extracting Quantized Accumulated Histogram of Oriented Gradient features from each image. While this method works well for camera imagery, and allows for the acoustic transmission of summaries to a human operator, local feature extraction is not robust to classifying sonar imagery.

In this work we demonstrate that a balance of local and global feature extraction, rather than each individually, improves the average precision of our classifier. We show that global analysis is not as powerful as CNN methods for extracting local features and that current CNN methods cannot generalize as well as methods with dilated filters.

It is also important to place this work in the context of previous work in the area of identifying useful data. There exists a large amount of work in selecting particular data for active learning (to improve the speed to converge for a learner or to improve the diversity of training data) [16], [17], [18]. This objective has been addressed in a variety of ways, many times by extracting local features from the data. For example, Demir and Bruzzone extract SIFT features from top-down imagery of outdoor environments in order to select informative images that decrease the uncertainty and increase the diversity of the training set, and are representative of the distribution for the images being evaluated [17]. Holub et al. reduce the number of training examples needed for training a classifier by using the Spatial Pyramid Match Kernel of Lazebnik method of matching[18]. This method also relies on extracting SIFT features.As previously discussed, such low-level features often rely on information that is not reliable in sonar imagery.

While these works have a similar objective to ours (the selection of informative frames) our criteria for identification is different than previous work. We evaluate images not on their utility to a learner, but rather on their *usefullness in being presented to a human operator for the completion of higher level tasks*. This is encapsulated in our training process, where the binary ground truth for each training image is only true if an image contains enough clear information to be utilized by an operator for higher-level tasks.

Atrous filters allow us to achieve this objective by using a larger context than normal filters. This is particularly useful in underwater exploration missions, where a human operator does not have the ability to evaluate incoming imagery in realtime for their utility in higher level tasks.

B. Atrous Convolutional Neural Networks

The building block for convolutional neural networks is the convolution operator, in which a filter k of kernel size rxr is slid over an image I and convolutions are computed for some output pixel $[l, m]$ of image Y as:

$$Y[l,m] = \sum_{i=1}^{r}\sum_{j=1}^{r} k[i,j]I[l-i, m-j] \tag{1}$$

In atrous convolution, the output pixel $[l, m]$ of image Y is computed from image I using a dilated filter k of size rxr which uses pixels spaced at some dilation rate d:

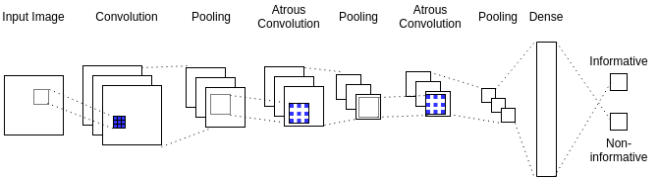
$$Y[l,m] = \sum_{i=1}^{r}\sum_{j=1}^{r} k[i,j]I[l-id, m-jd] \tag{2}$$

An example of a filter with dilation rate 4 can be seen in Fig. 1b. This increase in global context is important for sonar images because large-scale noise is prevalent while strong local features are not.

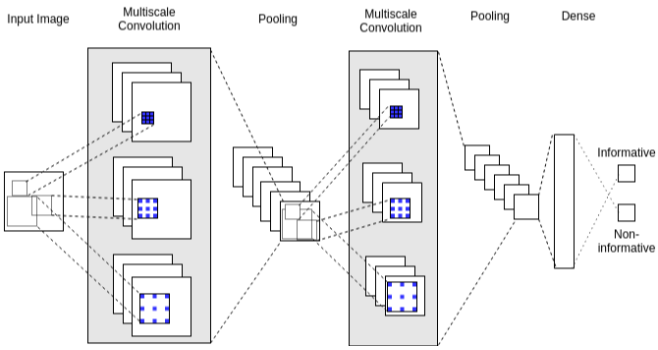
While atrous networks have yet to be used for analysis of sonar imagery, standard convolutional networks have gained recent popularity in this area. Kim, et al. use a CNN to classify the occurrence of another vehicle in sonar images [7]. They use a sliding window approach not only for object detection but also for object localization in image space. We compare to this method and show that atrous convolution provides a more generalized approach that allows for a higher average precision when given images of objects it has not been trained with. Kim, et al. extend this approach to incorporate the You Only Look Once method and thus achieve realtime capability [19]. In our method we maintain realtime performance by increasing the stride length of the sliding window.

Atrous convolution has gained much recent popularity in the computer vision community, and many approaches have been developed for utilizing these dilated filters. Yu, et al. use atrous filters in series (similar to Fig. 4a) to increase the performance of image segmentation by incorporating objects at multiple scales without decreasing image resolution (as would occur in max pooling) [10]. Chen, et al. use atrous filters with different dilation rates in the same layer (similar to Fig. 4b) to increase performance in image segmentation [20]. We examined both the series and single-layer architectures as well as the parallel pretrained architecture and found the series architecture to perform the best on sonar imagery.

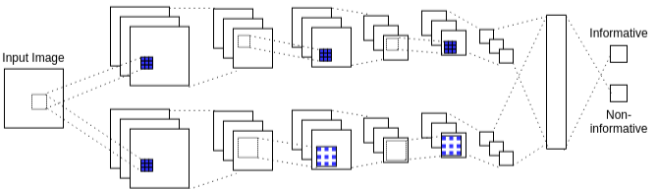
There has also been extensive work using deep learning in sonar image quality analysis and object detection with Synthetic Aperture Sonar (SAS). Williams and Dugelay use a deep neural network to distinguish cone shaped objects from rocks [21]. They address the problem of noise by using multiple views of the objects. In this work we address noise and remove the requirement for multiple images by using dilated filters. McKay, et al. developed a CNN based method using pretrained networks for use with SAS [22]. While they achieve impressive results, we note that their imagery is captured from far above the object and their noise



(a) Series architecture



(b) Single-layer (multiscale) architecture



(c) Pretrained parallel network architecture

Fig. 4: Representations for the atrous convolution architectures that were tested.

is artificially added after data collection. In our imagery the noise is naturally occurring and our viewing distance to the object is very small. These mission trajectories allow for close inspection of features in the environment but result in non-uniform degradation of the object as well as various regions of strong returns that do not necessarily correspond to objects.

III. Architecture

As summarized in Fig. 4, we tested multiple configurations for the atrous architecture including: placing the filters in series, incorporating multiple dilation rates in a single layer, and utilizing two parallel and pretrained networks. The series method (Fig. 4a) allowed dilated filters to interact with the image features once passed through a standard convolutional layer. The single-layer architecture (Fig. 4b) allowed filters with different dilation rates to extract features over the same input image. The filters were then combined, pooled, and sent to the next layer. Finally, the parallel network (Fig. 4c) was designed by separately pretraining two networks (one using standard convolution and one using atrous convolution). Once trained, the dense layers were removed and the two networks were combined into one dense layer. To test each architecture and set of parameters we held out 1000

TABLE I: Performance of different atrous architectures

Architecture	Best average precision (transfer learning)
Series	**0.52±0.04**
Single-layer	0.47±0.02
Pretrained parallel	0.51±0.02

frames of the *Multi* dataset which was used for testing the models (see Section IV-A for a full dataset description). We evaluated the architectures based on their average precision performance while evaluating imagery of objects unseen in training. Within each architecture we tested predefined and discrete values for multiple parameters including: kernel size, number of filters, dilation rate, and number of layers.

As shown in Table I, we found aligning the atrous filters in series, as seen in Fig. 4a, achieved the highest average precision. While the pretrained parallel approach performed almost identically, the series architecture requires no pretraining and is thus preferable.

Our network evaluates entire sonar images by a sliding window approach. From sonar images of size 256x235, subwindows of size 102x94 are extracted from the original frame and resized to 32x32 before being fed into the network. The subwindow size was chosen to be twice the size of the average hand labeled bounding box of objects in our images. This allowed for more contextual information to be leveraged in classification. To determine a threshold on the model output for informative predictions, we generated a precision recall curve from the model performance on the validation data and chose a threshold corresponding to a precision of 0.75. This resulted in a threshold of 0.99. We chose a threshold corresponding to a high precision as due to high data capture rates, we value precision in detecting informative frames over the recall of many informative frames. While evaluating each subwindow, if the network found an object with over 0.99 probability, the entire frame was classified as informative. A stride length of 10 between subwindows was used to allow sonar frame prediction in realtime without significant loss in performance.

IV. Experiments and Results

A. Datasets

The data used for this work came from 3 separate deployments of our Seabotix vLBV300 Remotely Operated Vehicle. In each of these deployments the data was captured in a passive manner, that is a human operator drove the vehicle to inspect the object of interest while the sonar image data was recorded. Thus, each dataset is a video or contiguous set of images containing both informative and non-informative frames at naturally occurring intervals.

Dataset X_1 contains data collected of the X target (see Fig 7a) in the Oregon State University pool during our first deployment. Dataset X_2 contains data collected of the same X target in the pool during a different deployment with different environmental noise characteristics. Finally, dataset $Multi_1$ contains data from all four of the objects in Fig. 7. Each dataset contains a different number of frames

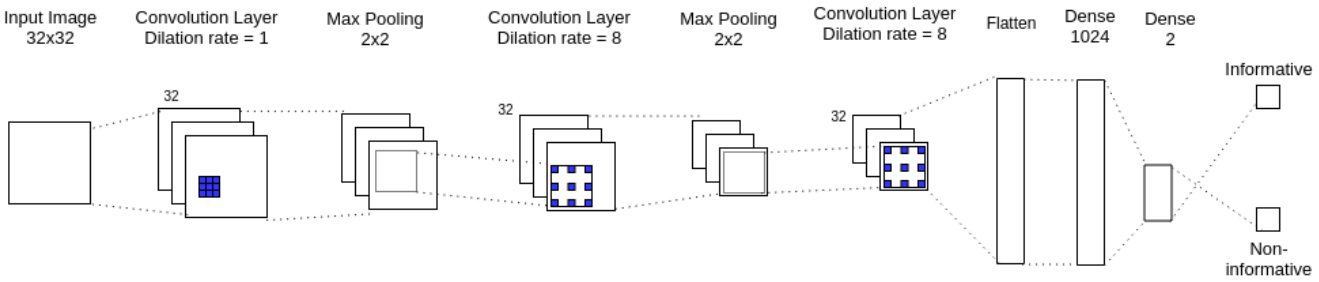


Fig. 5: The atrous convolutional network used. The parameters (including the use of the atrous filters in series) were tuned by testing discrete options.

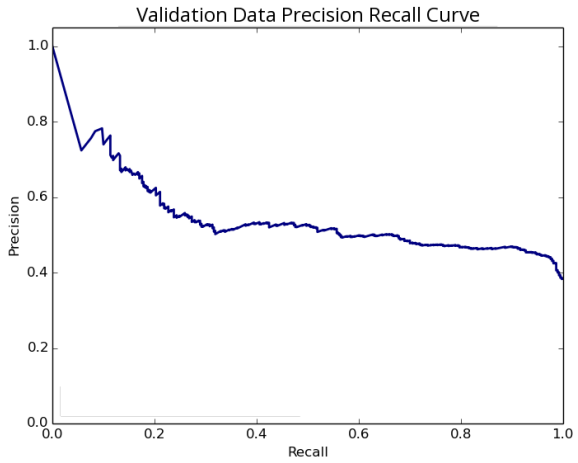


Fig. 6: The precision recall curve generated from validation data used to choose a threshold for informative vs. non-informative frames. We chose a threshold corresponding to a precision of 0.75, allowing us to propose informative frames at a high rate.

with objects that could be confidently identified by a human operator. These distributions are summarized in Table II. The low number of high quality frames underpins the need for introspection in such vehicles, as much of the data captured is non-informative in completing higher level tasks.

The ground truth labels for these frames are from the hand labeling of frames where we identify a frame as informative if it consists of an object that can be clearly identified. This identification often can only be achieved with the existence of multiple distinguishing features. For example, identifying the four corners of an object would not be enough to identify the frame as informative because it could be the X or the Square target. However, noticing lines intersecting in the middle of the target would be enough to confidently identify the X shaped target. Clear identification of an object, as opposed to simply detecting a feature in an image, is important as this imagery is used for higher-level tasks such as monitoring or inspection. In such tasks, images containing information the human can accurately interpret are much more useful than those containing features that must be tracked and grouped together over time. It should be noted that while we make the condition for an informative frame the occurrence of a set of distinguishing features, our method is still flexible. By evaluating our method on the transfer learning case (where the training and test datasets contain imagery of different objects) we ensure that our method is not overfitting to certain shapes or features.

TABLE II: Summary of the datasets.

Dataset	Total number of frames	Percent informative frames
X_1	5000	36%
X_2	1107	56%
$Multi_1$	5000	39%

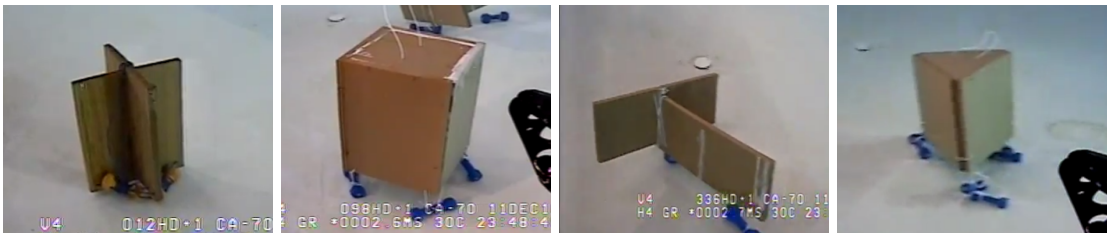
Notes: The total number of frames contains both informative and non-informative frames. The ground truth label for a frame is determined by a human operator determining if they can confidently identify an object in the frame.

B. Experimental Design

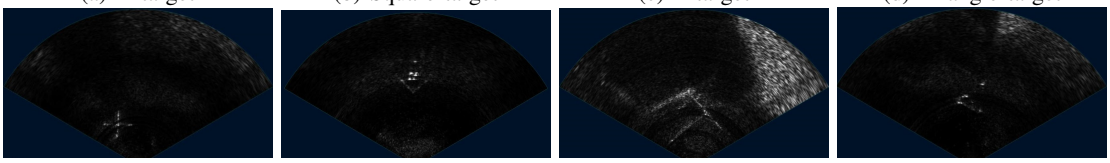
In developing models for use in the field it is important to not only be able to identify objects that have been trained on, but also be able to reason about objects not seen in training. We thus tested the capabilities of our model in both situations. In each of the experiments we primarily trained on data containing the X object (datasets X_1 and X_2). The difference between experiments was the varying of 2 binary parameters, namely seeding the model with multi-object data and the presence of the X object in test data. Seeding the model with multi-object data gave the model 600 frames of multi-object data to train on, about 33% of which were informative frames. Because the multi-object data contained imagery of all four objects (and thus the X) explicitly removing the X from the test set evaluated the transfer learning capabilities of the model.

Experiment 1 seeded the training of the model with multi-object data and allowed the test data to contain imagery of the X. Experiment 2 seeded the training of the model however imagery of the X was no longer present in the test data. Experiment 3 tested the transfer learning capabilities of the model: multi-object data was not used during training and test images did not contain the X.

Training was done by manually cropping subwindows that either contained a full object (informative) or contained parts of an object, environmental noise, or simply background imagery (non-informative). For training without seeding 31301



(a) X target (b) Square target (c) T target (d) Triangle target



(e) X in sonar (f) Square in sonar (g) T in sonar (h) Triangle in sonar

Fig. 7: Each target and its representation in sonar image space. (a)-(d) are images taken from the vehicle underwater.

subwindows were used and with seeding 31901 subwindows were used.

We compare our results to three other architectures, namely the Discrete Cosine Transform (DCT) method of global image analysis, the CNN method developed by Kim, et al. (called "two-layer CNN") and a deeper CNN we developed inspired by Kim, et al. (called "three-layer CNN") [23], [7]. The DCT method extracts the frequency components of the image as a whole and uses these components as features for use in a neural network classifier. We use a 2D DCT, defined for pixels m, n of an image A of size MxN as:

$$B_{pq} = \alpha_p \alpha_q \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_{mn} cos \frac{\pi(2m+1)p}{2M} cos \frac{\pi(2n+1)q}{2N}, \tag{3}$$

where

$$\alpha_p = \begin{cases} \frac{1}{\sqrt{M}}, p = 0 \\ \sqrt{\frac{2}{M}}, 1 \leq p \leq M-1 \end{cases} \tag{4}$$

$$\alpha_q = \begin{cases} \frac{1}{\sqrt{N}}, q = 0 \\ \sqrt{\frac{2}{N}}, 1 \leq q \leq N-1 \end{cases} \tag{5}$$

and B_{pq} is the coefficient on the basis cosine function defined by the constant p and q [23]. The intuition for this method is that by converting the image to the frequency domain, noise and strong features are easily separable and thus noise does not impair the classification performance. The two-layer CNN is a standard CNN architecture consisting of 2 convolutional layers (with 64 and 128 filters respectively) max pooling, and a dense layer of size 1024. Our three-layer CNN improves upon this by adding an additional layer and increasing the size of the filters from 3x3 to 5x5, both features we found to be beneficial in preliminary testing.

C. Results

To evaluate each approach we generated precision recall curves and compared the average precision of each method. The results presented in Table III are averaged over 20 runs. With the goal of deploying our network on a real system, we run our method on real sonar data played back in realtime. Representative curves for each experiment (best viewed in color) are shown in Figs. 8-10. We note that all of the experiments are on highly unbalanced datasets and thus the resulting curves are lower than expected for a classifier.

In Experiment 1, where multi-object data is used to seed training, we find that our method achieves a higher average precision than the DCT method and the two-layer CNN, and performs similarly to our three-layer convolution method. As shown in Fig. 8, it is important to note that the DCT based method of global analysis lacks the power of deep learning approaches in capturing the low-level features that define similar looking objects. For Experiment 2, where the test data does not contain imagery of the X, but the training is seeded, we find that our model significantly outperforms the others. This is intuitive as the use of dilated filters allows for a more generalized approach to feature extraction. The results of this experiment are shown in Fig. 9. In Experiment 3 we train on imagery of the X object and test on imagery of other objects and find that our method outperforms all other methods. This is expected, as the atrous filters allow for a larger neighborhood of influence when computing the output for features that are generally on a larger scale than that of normal camera images. As shown in Fig. 10 it is interesting to note that the DCT based method outperforms the purely CNN methods, demonstrating that transfer learning capability benefits greatly from a global perspective on the images.

TABLE III: Average precision for each method

	Avg. Precision		
Method	Exp. 1	Exp. 2	Exp. 3
DCT	0.57±0.01	0.55±0.02	0.48±0.02
Two-layer CNN [7]	0.69±0.04	0.58±0.05	0.42±0.04
Three-layer CNN	0.70±0.05	0.64±0.04	0.42±0.03
Atrous Convolution (ours)	**0.72±0.02**	**0.68±0.03**	**0.54±0.04**

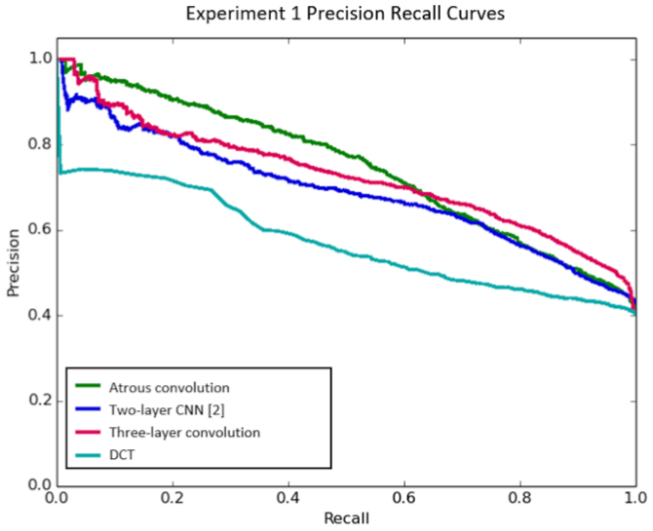


Fig. 8: Representative precision recall curves of each approach tested on an unbalanced dataset with the seeding of multi-object data and the inclusion of the X object in the test set.

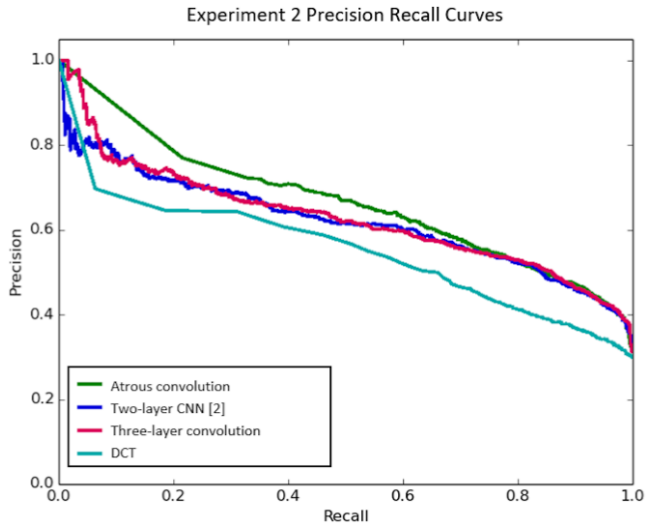


Fig. 9: Representative precision recall curves of each approach tested on an unbalanced dataset with the seeding of multi-object data but without the X object in the test set.

V. Conclusion

In this work we developed a method for perceptual quality assessment in underwater sonar images. We showed that using dilated filters is particularly appropriate in sonar imagery since the features are often on a larger scale than those of analogous camera features. We showed that this method achieves a higher average precision than state-of-the-art methods in the transfer learning case, a situation important in deploying vehicles into the world. This ability to assess sensory data increases the autonomy of underwater vehicles, which are oftentimes monitored by one or more dedicated human operators.

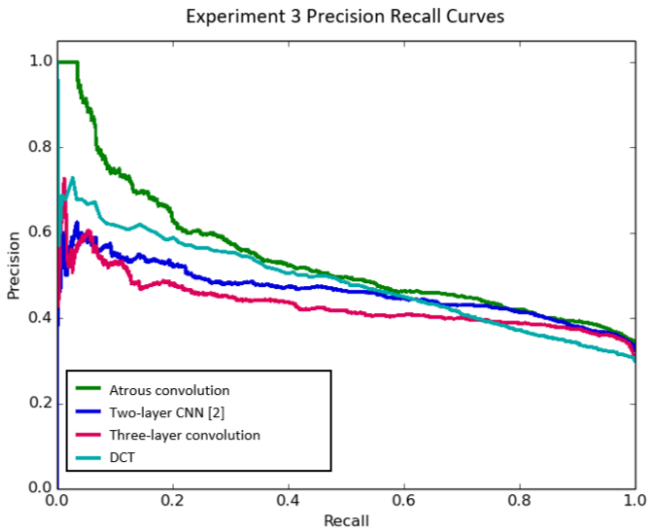


Fig. 10: Representative precision recall curves of each approach tested during the transfer learning case on a highly unbalanced dataset.

One such example of increased autonomy involves the underwater environment reconstruction process. Given the absence of a generalized approach to automatic feature extraction from sonar imagery, oftentimes a human operator must hand-label the features [5]. These hand labeled features can then be used later for tasks such as environment reconstruction. In hand labeling the frames, much time is spent by the human analyzing imagery that does not contain useful features, thus creating a large time gap between data collection and reconstruction. Our work would allow for a proposal system where only frames with a clear object are given to the human for hand labeling, greatly reducing the amount of time the human spends analyzing imagery and decreasing the gap between data collection and reconstruction.

There also still exists work in utilizing the location of the subwindow in the image space to complete missions more intelligently. Previous work used this information to track another vehicle's trajectory [7]. Of particular interest for us is utilizing this information for further automating the reconstruction process. Our Seabotix vLBV300, like many underwater systems, maintains its local position using a Doppler Velocity Log (DVL). This data can be used to project objects from the image space to the real world, allowing for more intelligent mission planning.

References

[1] F. S. Hover, R. M. Eustice, A. Kim, B. Englot, H. Johannsson, M. Kaess, and J. J. Leonard, "Advanced perception, navigation and planning for autonomous in-water ship hull inspection," *The International Journal of Robotics Research*, vol. 31, no. 12, pp. 1445–1464, 2012.

[2] D. P. Williams, "On optimal auv track-spacing for underwater mine detection," in *Proc. IEEE International Conference on Robotics and Automation*, 2010, pp. 4755–4762.

[3] M. Johnson-Roberson, O. Pizarro, S. B. Williams, and I. Mahon, "Generation and visualization of large-scale three-dimensional reconstructions from underwater robotic surveys," *Journal of Field Robotics*, vol. 27, no. 1, pp. 21–51, 2010.

[4] J. W. Kaeli, J. J. Leonard, and H. Singh, "Visual summaries for low-bandwidth semantic mapping with autonomous underwater vehicles," in *Proc. IEEE/OES Conference on Autonomous Underwater Vehicles*, 2014, pp. 1–7.

[5] T. A. Huang and M. Kaess, "Towards acoustic structure from motion for imaging sonar," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015, pp. 758–765.

[6] M. D. Aykin and S. S. Negahdaripour, "Modeling 2D lens-based forward-scan sonar imagery for targets with diffuse reflectance," *IEEE Journal of Oceanic Engineering*, vol. 41, no. 3, pp. 569–582, 2016.

[7] J. Kim, H. Cho, J. Pyo, B. Kim, and S.-C. Yu, "The convolution neural network based agent vehicle detection using forward-looking sonar image," in *Proc. IEEE/MTS OCEANS Monterey*, 2016, pp. 1–5.

[8] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform," in *Wavelets*. Springer, 1990, pp. 286–297.

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *arXiv preprint arXiv:1606.00915*, 2016.

[10] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, November 2015.

[11] Y. Ji, S. Kwak, A. Yamashita, and H. Asama, "Acoustic camera-based 3d measurement of underwater objects through automated extraction and association of feature points," in *Proc. IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, Sept 2016, pp. 224–230.

[12] E.-h. Lee and S. Lee, "Development of underwater terrain's depth map representation method based on occupancy grids with 3D point cloud from polar sonar sensor system," in *Proc. IEEE International Conference on Ubiquitous Robots and Ambient Intelligence*, 2016, pp. 497–500.

[13] J. Kalwa and A. Madsen, "Sonar image quality assessment for an autonomous underwater vehicle," in *Proc. IEEE World Automation Congress*, vol. 15, 2004, pp. 33–38.

[14] H. Johannsson, M. Kaess, B. Englot, F. Hover, and J. Leonard, "Imaging sonar-aided navigation for autonomous underwater harbor surveillance," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 4396–4403.

[15] M. Aykin and S. Negahdaripour, "On feature extraction and region matching for forward scan sonar imaging," in *Proc. IEEE/MTS OCEANS Hampton Roads*, 2012, pp. 1–9.

[16] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial intelligence*, vol. 97, no. 1, pp. 245–271, 1997.

[17] B. Demir and L. Bruzzone, "A novel active learning method in relevance feedback for content-based remote sensing image retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2323–2334, 2015.

[18] A. Holub, P. Perona, and M. C. Burl, "Entropy-based active learning for object recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2008, pp. 1–8.

[19] J. Kim and S.-C. Yu, "Convolutional neural network-based real-time rov detection using forward-looking sonar image," in *Proc. IEEE/OES Conference on Autonomous Underwater Vehicles*, 2016, pp. 396–400.

[20] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, June 2017.

[21] D. P. Williams and S. Dugelay, "Multi-view sas image classification using deep learning," in *Proc. IEEE/MTS OCEANS Monterey*, 2016, pp. 1–9.

[22] J. McKay, I. Gerg, V. Monga, and R. Raj, "What's mine is yours: Pretrained cnns for limited training sonar atr," *arXiv preprint arXiv:1706.09858*, June 2017.

[23] J. Makhoul, "A fast cosine transform in one and two dimensions," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 1, pp. 27–34, 1980.