Comparison of Learned Versus Engineered Features for Classification of Mine Like Objects from Raw Sonar Images

Paul Hollesen[1], Warren A. Connors[2] and Thomas Trappenberg[3]

[1] Department of Computer Science, Dalhousie University
hollesen@cs.dal.ca
[2] Defence Research and Development Canada
warren.connors@drdc-rddc.gc.ca
[3] Department of Computer Science, Dalhousie University
tt@cs.dal.ca

Abstract. Advances in high frequency sonar have provided increasing resolution of sea bottom objects, providing higher fidelity sonar data for automated target recognition tools. Here we investigate if advanced techniques in the field of visual object recognition and machine learning can be applied to classify mine-like objects from such sonar data. In particular, we investigate if the recently popular Scale-Invariant Feature Transform (SIFT) can be applied for such high-resolution sonar data. We also follow up our previous approach in applying the unsupervised learning of deep belief networks, and advance our methods by applying a convolutional Restricted Boltzmann Machine (cRBM). Finally, we now use Support Vector Machine (SVM) classifiers on these learned features for final classification. We find that the cRBM-SVM combination slightly outperformed the SIFT features and yielded encouraging performance in comparison to state-of-the-art, highly engineered template matching methods.

1 Introduction

Naval mine detection and classification is a difficult, resource intensive task. Mine detection and classification is dependent on the training and skill level of the human operator, the resolution and design of the sonar, and the environmental conditions that the mines are detected in. Research has occured over the last 25 years into both sensor development and processing of sonar data. Although the sensors and capability of mine countermeasures platforms have improved in this time, the issue of operator overload and fatigue have caused the duty cycles of mine detection and classification to be short, therefore diminishing the effectiveness of Mine Counter Measures (MCM) platforms.

Recent research focuses on development of computer aided tools for detection and classification of bottom objects [1,2,3]. This typically takes the form of a detection phase where mine like objects are selected from the seabed image, and a classification phase where the objects are fitted to a multi-class set of potential mines. This detection and classification process has typically been implemented using a set of image processing tools (Z-test, matched filter), feature extraction, and template-based classification

[1,2,3]. These techniques are effective at finding mines, but are sensitive to tuning the parameters for the processing method, and the sea bottom environment under test [2,3].

Learning algorithms, such as Artificial Neural Networks, have been examined for the mine problem, however success has been limited, and these methods have required the training sets to closely reflect the sea bottom environment of the area where the system will be tested. Earlier work includes using a deep belief network (DBN)[4] which is a stack of multiple Restricted Boltzmann Machines (RBM)[5], to learn to extract features from side scan sonar data. This technique was sucessful in detecting mines with comparable performace to the traditional methods [4].

The RBM learning method is effective, however the Scale-invariant feature transform (SIFT) [6] has been very influential recently in vision and image processing, and has been applied to numerous image processing and feature extraction tasks sucessfully. At the same time, while the original work on RBM/DBN structures for feature learning and classification has shown the power of the DBN for feature extraction, no consideration was given to recent developments with the DBN model including sparseness constraints and a convolutional variation [4]. Imposing sparsity on the RBM regularizes the leaned model by decreasing the weights of nodes whose activity exceeds a prescribed sparsity, therefore simplifying the model it learns and providing a more compact representation of the input. The convolutional approach allows the model to scale to high-resolution imagery and further regularizes the model by reducing the parameter space.

This paper compares feature extraction using SIFT versus a convolutional RBM (cRBM), for the mine classification problem. This also serves to examine how well SIFT generalizes to application domains analogous to visual wavelength imagery. A central argument for using learned rather than carefully selected, contrived features is the ability to apply the model to diverse application domains. This is an interesting domain to explore in this context, as there is a natural 2D, grayscale representation for sonar data, and the mine classification task contains most of the same challenges as generic object recognition: invariance to translation, rotation, luminance, clutter, and noise.

Both techniques were applied to a series of sonar images to extract features, with the output fed to a Support Vector Machine (SVM) for training and classification. As the goal of this effort is to develop a classification system for sonar images of sea bottom objects with comparable performance to highly contrived methods, each technique is treated as a feature extraction method, with the features being passed to an SVM for training and classification. The results were compared to state-of-the-art template matching methods with encouraging results for correct classification of targets.

2 Synthetic Aperature Sonar Imagery

Traditional side scan sonar imagery (e.g. Figure 1a) depicts objects by a strong bright region (highlight) where it is insonified by sound waves followed by a dark region (shadow) cast behind it. The size, shape and disposition of such features are important for both automated and manual methods in mine classification. This imagery may be littered with background noise coming from natural and artificial sources. In imaging

sonars, range resolution is mostly determined by the bandwidth of the transmit pulse, while azimuthal resolution is determined by the length of the receiver array. While bandwidth in modern sonars are sufficient to achieve a high range resolution, azimuthal resolution is difficult to improve due to the engineering limitations of constructing long arrays.

Synthetic Aperture Sonar (SAS) is a recent side scan sonar technique being applied to detecting and classifying mine like objects. This technology is inspired by synthetic aperture radar, which is commonly used on terrestrial and space based radar sensors. Synthetic Aperture Sonar is a technique whereby a longer array length is synthesized by integrating a number of sonar pings in the direction of travel of the sonar, resulting in improved resolution which is also independent of range (e.g. Figure 1b). This provides a powerful tool for the mine detection/classification problem, as the higher fidelity images allow for a richer set of features available for the detection and classification of a sea bottom object.

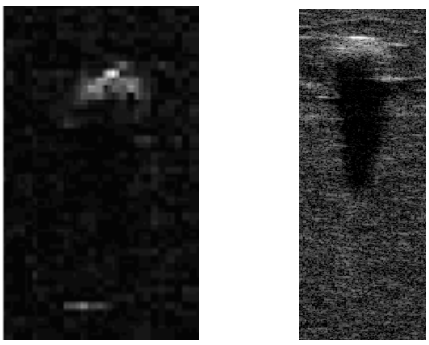

Fig. 1. Sonar images of mine-like objects, showing (a) a Side Scan Sonar image from and (b) an image from the MUSCLE SAS [7] of the same type of object

The data used in this paper was collected by the NATO Undersea Research Center [7] on the MUSCLE Autonomous Underwater Vehicle (AUV). This vehicle is equipped with a 300KHz SAS. The SAS gives an 2.5 cm $\times$ 2.5 cm resolution, at up to 200 meters in range.

2.1 Data Preparation

The SAS dataset was collected in the summer of 2008 off Latvia in the Baltic Sea. The MUSCLE vehicle was used to survey multiple mine-like targets that were deployed as part of the trial, including multiple sonar passes over each target in the field from different angles. The targets were three mine-like shapes, including a cylinder (2.0m $\times$ 0.5m), a truncated cone (1.0m base, 0.5m height), and a wedge shape (1.0m $\times$ 0.6m $\times$ 0.3m). Clutter included numerous rocks and boulders, geographic features of the sea bottom, and a specific rock which was chosen due to its similarity in shape with the truncated cone. The dataset was composed of 65 cylinders, 69 truncated cones, 37

wedges, and 2218 non-mine clutter objects, including 47 rock images that are highly correlated to a target shape.

The raw SAS data can contain as much as ten times the data of a side scan sonar, and the maximum and minimum values for these samples describe a very large dynamic range for the sensor. The data is organized in complex values which describe the amplitude and phase data from the sonar. Although it is appealing to examine the phase component of the SAS data, it is beyond the scope of this work, and is considered in the Outlook section as future work. The data was prepared by removing the phase component, then re-mapping the amplitude component to a decibel (dB) scale.

3 Feature Extraction

Feature extraction is a difficult and error prone task that typically is performed manually. This process is done through an analysis of the sonar data, and careful selection of characteristics which help to describe the class of the object. Modern methods have looked to reduce this complexity through automatically selecting features from the data through decomposition (e.g. PCA, SIFT, Wavelet) in order to have a set of features that uniquely describe the object, and can be used directly for training. Methods such as SIFT have been effective as an automated method for feature extraction, and has been applied to both visual and acoustic images to select a set of features for the object.

Learning methods are appealing for feature extraction, specifically generative models, as they learn the dominant features of the image they are being trained on, and build an internal representation of what elements the object should be composed of. This allows for an unsupervised approach where many images are shown to the learning method, and the learning method determines the features to be selected and modelled. Furthermore, the generative models have the added advantage that it is possible to see the feature filters which have been learned, which gives the researcher a measure of the progress of the learning.

With either method, the goal is the same. We wish to select the most descriptive features for the class to provide the least ambiguous training set to the SVM, allowing it to find easily separable classes, and perform effectively as a target classifier compared to existing manual methods.

3.1 Scale Invariant Feature Transform (SIFT)

SIFT [6] is a method for feature extraction that is invariant to scale, orientation and distortions. We employ dense SIFT feature extraction and explore window sizes ranging from 12 to 24 pixels wide (i.e. spatial bins from 3 to 6 pixels), spaced from 4 to 12 pixels apart.

Similar to our cRBM experiments described later, best results were obtained with the maximum possible window size for this data (24x24), spanning the full width of the image. SIFT extracts a 128-dimensional feature vector from each window, which result in a representation of roughly equal size to that of the cRBM using a spacing of 6 pixels.

3.2 Restricted Boltzmann Machines (RBMs)

The RBM [8] is an energy-based, generative model that can learn to represent the distribution of implicit features of the training data and generate examples thereof. An RBM consists of two layers of nodes, forming the visible and hidden layers. Each layer is fully connected to the others, but is restricted in that there are no connections between nodes within a layer. The energy of the joint configuration of visible and hidden units given the connections between them (ignoring biases for simplicity) is given by

$$E(\mathbf{v},\mathbf{h}) = -\sum_{i=1}^{V}\sum_{j=1}^{H} v_i h_j w_{ij} \tag{1}$$

where v and h are the states of the visible (input) and hidden units, respectively, and w is the connection strengths between each visible and each hidden unit.

Stacks of RBMs can be learned in a greedy, layer-wise fashion, with the output of the previous layer providing the input to the next, forming a Deep Belief Network (DBN) [8]. This enables higher-layer nodes to learn progressively more abstract regularities in the input.

RBM training is accomplished with the Contrastive Divergence (CD) algorithm [9] which lowers the energy (i.e., raises the probability) of the data observed on the visible units and raises the energy of reconstructions of the data produced by the model:

$$\Delta w_{ij} \approx \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model} \tag{2}$$

Using CD, the RBM learns a generative model of the input in a purely unsupervised fashion by measuring the discrepency between the data and the model's reconstructions then 'correcting' the system by slightly altering the weights to minimize reconstruction errors.

We can also regularize the learned model with a sparse representation by decreasing the weights of nodes whose activity exceeds a prescribed sparsity level, s [10]:

$$\Delta \mathbf{w}_j \approx s - \langle h_j \rangle \tag{3}$$

where $\langle h_j \rangle$ is the expected probability of activation which is computed as a decaying average of the activity of that unit over training examples. This has the added benefit of increasing the weights of nodes whose activity is below the target threshold, thus reintegrating nodes whose random initial conditions lead to them being suppressed by the network ('dead nodes'). While this regularization may lead to greater reconstruction error by forcing the network to represent the input with a smaller proportion of nodes, the resulting hidden representation is likely to be more interpretable by subsequent layers or classifiers.

3.3 Convolutional Restricted Boltzmann Machines (cRBMs)

In the cRBM model [11] each hidden node, rather than being fully connected to every input element as in a standard RBM, is connected to only a small, localized region of the image which is defined by the researcher. Furthermore, these connections are shared

by a group of hidden nodes which are collectively connected to every input region. This architecture enables the computationally efficient convolution operation to be used to generate each groups' activation.

If the region of the input image that each node of the cRBM is connected to is significantly smaller than the total input image, as we expect when the input is high resolution imagery, then the cRBM requires orders of magnitude fewer parameters for a similar representation size, since weights are shared by all nodes in a group. This is especially useful when patterns recur in different regions of the input, since any knowledge learned about this pattern is automatically transfered to all input regions.

By pooling adjacent hidden activation within groups, either with the commonly used maximum pooling or the probabilistic maximum pooling method [11], we can attain a degree of translational invariance while also keeping the size of the hidden representation within reasonable bounds. If maximum pooling is used, then we calculate the probability of activation of each node in a pooling window by applying the logistic function to the feedforward activation. In the probablistic maximum pooling method, each pooling window is sampled multi-nomially, so that only one hidden node in a window can be on, and the pooling node is off only if all hidden nodes in its window are off, according to Eq. (4) and (5):

$$P(h^k_{i,j} = 1|\mathbf{v}) = \frac{\exp(I(h^k_{i,j}))}{1 + \sum_{i',j' \in B_\alpha} \exp(I(h^k_{i',j'}))} \tag{4}$$

$$P(p^k_\alpha = 0|\mathbf{v}) = \frac{1}{1 + \sum_{i',j' \in B_\alpha} \exp(I(h^k_{i',j'}))} \tag{5}$$

where $h^k_{i,j}$ is a hidden node in pooling window B_α receiving feedforward input $I(h^k_{i,j})$, and p^k_α is the pooling node for that window.

The representation of each group of hidden nodes is then convolved with its filter to get that group's reconstruction of the input. Summing over all groups' reconstructions yields the networks reconstruction of the input used for CD learning.

For the present experiments we restricted ourselves to a single-layer cRBM. The parameters having the largest impact on classification performance are the filter size and number of filters. Through experimentation, best representations were obtained using 50 filters with width one less than the image width (i.e.22×22). After probabilistic maximum pooling with a 2×2 window size, this filter width results in 50 filters by 36 height representation, with the width collapsed to 1 (1800 dimensional). That is, the width of the image is collapsed in the cRBM's representation by the convolution operation and pooling. This can be seen as a compromise between the conventional and convolutional approach, with minimal transfer of knowledge horizontally. The dataset was amenable to this severe reduction in representation width because targets were centered in the image, with the pooling layer providing sufficient invariance to the small differences in position.

Based on research by Nair and Hinton using the NORB dataset[8], real-valued images can be used at the visible layer of the RBM if training speed is decreased. Therefore a low learning rate of 0.01 for weights and biases was found to be stable for the real valued images and was sufficiently large that learning peaked after 50 epochs through

the training set of 228 images. The learning rate for sparsity regularization was initialized at the learning rate for weights, and then increased to 10 times this rate linearly over epochs. This enables the network to explore representations early in learning since many nodes are active (and thus learning), and then gradually get driven to the desired sparsity level. The target sparsity giving best results was dependent on the representation size and thus the number and size of filters employed. For the 50, 22x22 filter network for which results are reported, a target sparsity of 0.01 in the hidden layer (0.04 in the final pooled representation) yielded best results through experimentation.

4 Results

The original images were 466×119 pixels, though some images had missing rows toward the bottom of the image which were detected and filled with the image mean intensity value. Each image was downsampled by a factor of 5 to remove noise, provide a more computationally tractable representation size, and decrease in-class variation. Image intensities were then normalized to have zero mean and unit standard deviation. Normalization was done per image because the dynamic range varied substantially from one image to the next. Normalizing per pixel across the training set, as is more common, rendered a significant proportion of images undistinguishable from the background due to their low dynamic range.

Classification performance was determined via ten-fold cross validation. This method partitions the training set of data into 10 subsets, where one is retained for the validation of the classification model, and nine subsets are used for training. This process is repeated ten times, where each of the subsets is used once as the validation set. 50 clutter, 10 mine-like rock and 10 of each type of mine was reserved for testing and the model was trained on the remaining data. The small proportion of available clutter examples used for training was chosen so the total clutter in the training set approximated the mean total of the mines.

Classification was performed with an SVM using the libSVM [12] software library. Grid searches were performed for optimal parameters for both a linear and radial basis function (RBF) kernels. For both the SIFT and cRBM feature vectors, the linear kernel gave superior results and was robust over a wide range of the SVM kernel cost parameter.

4.1 Convolutional RBM

We examined the representation learned by RBMs by the reconstruction of the input and its learned filters. Figure 2 provides a sample of sonar images and their reconstructions by one of the cRBMs trained in the course of cross-validation. The reconstructions are significantly smoothed, but with the smoothing generally respecting object boundaries as in nonlinear diffusion.

After learning filters from the training set, the activation probabilities of the convolutional RBM's hidden units were generated for both the training and test sets and passed to the SVM for training and classification. To show not only the correct classification performance but also the missed classifications, a confusion matrix was com-

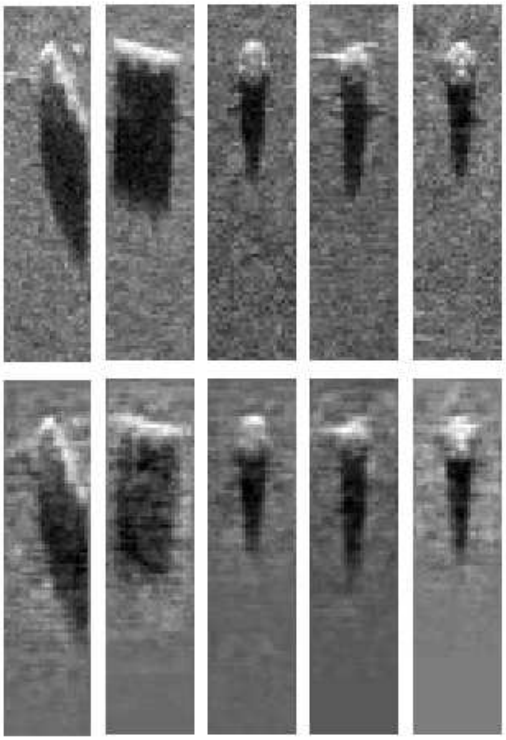

Fig. 2. Sample sonar images [7] (top) and reconstructions produced by the convolution RBM model which resulted in the best classification performance (bottom). The reconstructions are significantly smoothed and the highlight of the object somewhat filled in.

puted for comparison with SIFT and template matching [13]. Table 1 illustrates the results.

Table 1. Confusion matrix for SVM trained on cRBM features

CONFUSION	clutter	cylinder	trunc. cone	wedge	mine-like rock
clutter	0.949±0.013	0.031±0.011	0.006±0.003	0.011±0.003	0.003±0.003
cylinder	0.030±0.048	0.970±0.048	0±0	0±0	0±0
trunc. cone	0.01±0.032	0±0	0.980±0.042	0.010±0.032	0±0
wedge	0±0	0.020±0.042	0.230±0.125	0.740±0.127	0.010±0.032
mine-like rock	0±0	0±0	0.080±0.140	0±0	0.920±0.140

Out of 300 target views (3 types of targets, 10 of each type of target, 10 cross-validations), there were 5 false negatives, and out of 21310 views of non-targets, 1035 false positives. This yields a sensitivity to mines of $\mathbf{.983 \pm .024}$ showing a high rate of correct target classification, and a specificity of $\mathbf{.954 \pm .012}$. While most categories had this high level of classification accuracy, it is interesting to note that a large proportion of wedges (23%) were mis-labelled as truncated cone, due to both the similarity of their appearance in some of the sonar data and the poor representation of wedges in the dataset (37 wedges vs 69 truncated cones). This led to a poor sensitivity for wedges specifically but did not impact the sensitivity to mines in general.

4.2 SIFT

As the SIFT method does not require training, the algorithm was applied to each training image, generating a 128-dimensional feature vector for each 24×24 (overlapping) window. This served as the input for training and testing the SVM. The following confusion matrix in Table 2 illustrates the correct and incorrect classifications using the SIFT features.

Table 2. Confusion matrix for SVM trained on dense SIFT features

CONFUSION	clutter	cylinder	trunc. cone	wedge	mine-like rock
clutter	0.932±0.010	0.011±0.004	0.019±0.006	0.031±0.009	0.008±0.002
cylinder	0.010±0.032	0.980±0.063	0±0	0.010±0.032	0±0
trunc. cone	0±0	0±0	0.950±0.071	0.030±0.068	0.020±0.042
wedge	0.020±0.042	0.030±0.048	0.460±0.158	0.450±0.135	0.040±0.052
mine-like rock	0±0	0.020±0.063	0.040±0.070	0.010±0.032	0.930±0.082

The SIFT features resulted in similar performance to those of the cRBM but with slightly more false positives. The sensitivity to mines was $\mathbf{.970 \pm .025}$ which shows strong classification performance and the specificity was $\mathbf{.944 \pm .008}$. The biggest difference in the performance with respect to the cRBM was that there was significantly

more confusion between truncated cones and wedges. This is an interesting result as it shows that learned features are in particular outperforming in the cases that are difficult to classify.

5 Discussion

Overall the results of this work are encouraging and merit further research into the application of learning methods to sonar imagery and mine classification in particular. Both the SIFT method and the cRBM methods were comparable in performance, with the cRBM performing slightly better than the SIFT feature extraction method. As a basis of comparison, we include below in Table 3 the results from a normalized shadow and echo template-based cross-correlation method [13] which has proven highly effective at classifying targets. These templates are designed for a specific sensor and specific templates are generated for different ranges and therefore are an excellent baseline for learning methods to be compared against.

Table 3. Confusion matrix for template-matching method [13]

NSEM	non-mine	cylinder	trunc. cone	wedge
non-mine	0.94	0.01	0.02	0.03
cylinder	0.03	0.97	0	0
trunc. cone	0.04	0	0.96	0
wedge	0.08	0	0	0.92

As shown in earlier work [4], the RBM/DBN model can effectively extract features of mine-like targets and classify them using traditional side scan sonar data, however this method showed poor performance using the higher resolution data from the SAS sensor (results not shown). We believe that this is caused by the very large dynamic range in the sensor leading the DBN to learn features of the background (noise) distribution at the expense of modelling the object highlight. Although the increase in resolution in the sensor provides a richer set of detailed features of the object being learned, it also has the downside that the learning machines have a tendency to try to model and classify this noise rather than just the object. The cRBM model with enforced sparsity was beneficial in this regard, as the smaller parameter space and sparsity regularized the model and thereby limited the modelling of the background features. To illustrate this, the reconstructions in Figure 2 show a form of smoothing in areas of the image where no target features were present.

The filter sizes providing best results spanned the full width of the image (minus 1 in the case of the cRBM due to 2×2 pooling). This results in an architecture more similar to a conventional network in the horizontal direction but convolutional in the vertical direction. While there is significant error in the reconstructions (Figure 2), the hidden representation from which they are produced has a relatively small number of filters (50) given the large filter size, as well as sparse activation, which proves more interpretable for the classifier. Using smaller filters has the benefit of being able to model

finer features and transfer this learning horizontally, however it results in a significantly larger representation which the classifier had greater difficulty interpreting. In general, cRBM parameterizations which allow more accurate models of the data in the sense of reconstruction error, either by having smaller filters, more filters, or less sparsity, decrease classification performance since they naturally resulted in more complex representations which are more difficult to model with the SVM.

In comparison to the template matching method, the two methods examined in this paper showed comparable performance, with the exception of the wedge shapes, where both methods suffered in comparison to the template based method. Since the highlight of many wedges and truncated cones was little more than a strip of light in many instances, these two classes were confused, with the SVM opting to classify most as truncated cones due to their greater prevalence in the dataset. However, the cRBM distinguished them significantly better than SIFT. Examining the raw data, it was observed that a subset of wedges had some of the brightest highlights in the dataset. This feature, which may have been an artifact of this particular dataset, was likely captured by the the RBM representation but removed by SIFT in its attempt to create an illumination invariant representation. This would explain the cRBM's better performance in distinguishing the wedges and truncated cones whose shape representation was very similar in the sonar image. This effect highlights the benefit of using learned filters rather than engineered features from neighbouring application domains, since features which may be uninformative in one domain (in this case, the illumination of a particular feature) may be informative in the other.

6 Outlook

The results from both the cRBM and SIFT models were encouraging, but also highlight the need for further research. Distinguishing wedges from truncated cones, in particular, proved challenging for our models and demands further attention. In general, the detection and classification of objects from sonar imagery could potentially benefit from additional pre-processing or extensions to the two models, as described below.

6.1 Extraction of features from phase data

As noted in the data preparation section, the raw SAS data is organized as a set of complex numbers that describe both the amplitude and phase of the reflected sound intensities. For the purposes of this paper, the phase element was removed, and just the raw amplitudes considered. Although this phase element is stripped, it is likely that there is some coherent features in the phase information, which could help distinguish non-image related features such as material. If this feature could be extracted, it could be supplied as another feature for classification, or as a method to limit the false alarms during detection and classification phases.

6.2 Spatial Pyramid Matching

In the Spatial Pyramid Matching (SPM) method [14], dense SIFT features are extracted as in the method we employed. SIFT feature vectors are subsequently vector quantized,

and then histograms at multiple levels of resolution (whole image, quater image, ...). A histogram intersection (χ^2) kernel is then employed to classify the histogram representation. This has been very successful in object recognition tasks such as Caltech 101 used in [14]. Preliminary experiments with this method offered poor performance on this dataset, but more work needs to be done in exploring the many parameters of this model to determine if it can be successfully applied to classification of bottom objects in SAS imagery.

6.3 Convolutional DBNs

Our experiments with stacked layers of cRBMs yielded poor performance on the classification task. However, experimentation was hindered by the large computational burden imposed by convolving many filters with many layers of hidden representation. As we and other groups develop software to transfer the computation of these expensive operations to graphics processing units (GPU), this architecture will become much easier to explore and we expect that higher layers will achieve greater invariance to noise and small intra-class variations, as well as uncover more complex regularities in the training data.

Acknowledgements

The authors would like to acknowledge the NATO Undersea Research Center (NURC) for the use of the SAS data for this paper.

References

1. Chapple, P.: Automated detection and classification in high-resolution sonar imagery for autonomous underwater vehicle operations. Technical report, Defence Science and Technology Organization (2008)
2. Fawcett, J., Crawford, A., Hopkin, D., Myers, V., Zerr, B.: Computer-aided detection of targets from the CITADEL trial Klein sonar data. Defence Research and Development Canada Atlantic TM 2006-115. (November 2006) [available at pubs.drdc.gc.ca].
3. Fawcett, J., Crawford, A., Hopkin, D., Couillard, M., Myers, V., Zerr, B.: Computer-aided classification of the Citadel Trial sidescan sonar images. Defence Research and Development Canada Atlantic TM 2007-162. (2007) [available at pubs.drdc.gc.ca].
4. Connors, W., Connor, P., Trapperberg, T.: Detection of mine like objects using restricted boltzmann machines. In: Proceedings of the 23rd Canadian Conference on Artificial Intelligence. (2007)
5. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Comput. **18**(7) (2006) 1527–1554
6. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60** (2004) 91–110
7. Bellettini, A., Pinto, M.: Design and experimental results of a 300 kHz synthetic aperture sonar optmized for shallow-water operations. IEEE Journal of Oceanic Engineering **34** (2008) 285–293
8. Nair, V., Hinton, G.E.: Implicit mixtures of restricted boltzmann machines. In: NIPS. (2008) 1145–1152

9. Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neural Comput. **14**(8) (2002) 1771–1800
10. Hinton, G.E.: A practical guide to training restricted boltzmann machines. Technical Report UTML TR 2010-003, University of Toronto (2010)
11. Lee, H., Grosse, R., Ranganath, R., Ng, A.: Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. (2009)
12. Chang, C.C., Lin, C.J.: LIBSVM: A Library for Support Vector Machines. (2001) Software available at http://www.csie.ntu.edu.tw/ cjlin.libsvm.
13. Myers, V., Fawcett, J.: A template matching procedure for automatic target recognition in synthetic aperture sonar imagery. IEEE Signal Processing Letters **17**(7) (2010) 683–686
14. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. (2006) 2169–2178