A Template Matching Procedure for Automatic Target Recognition in Synthetic Aperture Sonar Imagery

Vincent Myers and John Fawcett

***Abstract*—A method for classifying objects in sonar imagery is proposed. Motivated by the high-resolution achievable by modern imaging sonars, a novel template matching technique is developed that compares a target signature generated from a simple acoustic model with the actual image of an object being classified. The approach uses both the correlation with target echoes as well as projected acoustic shadow, and is tested on data obtained from a synthetic aperture sonar during experiments at sea. It is compared to two commonly used methods that are based on normalized cross-correlation, and results show that the proposed method outperforms the standard methods in terms of receiver-operating characteristic (ROC) curves as well as confusion matrices.**

***Index Terms*—Automatic target recognition, synthetic aperture sonar, template matching.**

I. Introduction

THE task of automatically classifying objects in high-resolution sonar imagery as one of a number of known targets or as a benign nontarget, is made difficult by the uncertainty associated with interpreting 2-D projections of 3-D objects. An approach to this problem consists of physically modeling the possible targets to create a number of templates; then, the degree to which the templates match the image of the object under consideration is computed and used as a classification feature. The sum of absolute differences (SAD) and normalized cross-correlation (NCC) [1] are examples of template matching methods which have met with some level of success in many fields of computer vision. However, the direct application of these techniques to imagery produced by means of an active acoustic sensing system such as high-frequency sonar has not resulted in satisfactory classification performance, which may partly be due to the relatively poor resolution and noise levels of many available sensors [2].

As a result, a number of approaches have been put forward [3]–[5] which first segment the image pixels into areas of echo, shadow and background in an attempt to diminish the effect of noise, and then perform the template match using some distance metric. For instance, Reed *et al.* [3] created a number of fuzzy membership functions based on the Hausdorff distance between modeled and measured shadows, and Quidu [4] computed the distance between the Fourier descriptors of the modeled object shadows and the measured one. Fawcett [5] used a number of cross-correlations between modeled templates and target images as features for a kernel ridge regression classifier. An alternative approach, proposed by Coiras [6], is to invert the 2-D projection back to a 3-D representation of the object, and to then perform the matching in the 3-D space.

The introduction of robust synthetic aperture sonar (SAS) techniques [7] has resulted in much higher resolution imagery and therefore more detailed object signatures than traditional sonar arrays. Template-matching for SAS imagery has also been considered by Midelfart *et al.* [8]. This letter aims to revisit correlation operations for template matching, in particular with respect to high-resolution SAS images. A new method is proposed which exploits both the target shadow and the target echo while being robust with respect to variations in the absolute levels and the background of the image. This method is shown to outperform normalized cross-correlation for both segmented and unsegmented images.

II. NSEM Template Matching Procedure

The procedure described here, called the normalized shadow-echo matching (NSEM) method, relies on generating idealized target images (templates) with a computer model; these are then matched to the real target images under consideration. This eliminates the need for training data which are generally difficult to obtain in many underwater applications. Here, a ray tracing method, which is a simplification of work by Bell [9], has been implemented as the basic acoustic propagation and scattering models. The method takes into account the sensor beam patterns via the array construction, and was extended for SAS sensors by synthesizing a longer array based on theoretical gains [10]. The target shapes are input as a set of facets, and the ray-tracing model predicts the amplitude of the echo using Lambert's law [11] at the intersections of the rays with the scattering surface. This produces values in the range $[0\ldots1]$ based on the grazing angle of the ray at the target facet. The ray model also predicts where the object will cast an acoustic shadow on the seabed and sets these values to -1. The pixels belonging to the surrounding seabed are set to 0. An object can be viewed from any aspect angle, and the characteristic of high-frequency imaging sonar is such that different aspects can produce a very different signature for the same target, depending on its shape. The result of the modeling step is a template $T^{(\gamma,\theta)}$ for each target type $\gamma\in\Gamma$ and aspect angle $\theta\in\Theta$.

Images of objects are detected and extracted from large swathe images through simple matched filter operations with high detection rates as well as high rates of false alarm. Each object image, I, is then roughly segmented into five classes

Manuscript received March 17, 2010; revised May 03, 2010; accepted May 05, 2010. Date of publication June 01, 2010; date of current version June 16, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Patrizio Campisi.

The authors are with the Defence R&D Canada-Atlantic, Dartmouth, NS B2Y 3Z7 Canada (e-mail: vincent.myers@drdc-rddc.gc.ca).

Digital Object Identifier 10.1109/LSP.2010.2051574

corresponding to bright echo (+1), echo (+0.5), background (0), shadow (−0.5) and dark shadow (−1). These are defined in terms of the image median $\tilde{x}$ and a given image pixel of amplitude x is classified as

$$x = \begin{cases} -1, & \text{if} \quad x < a_1\tilde{x} \\ -0.5, & \text{if} \quad x < a_2\tilde{x} \\ +0.5, & \text{if} \quad x > b_1\tilde{x} \\ +1, & \text{if} \quad x > b_2\tilde{x} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where a_1, a_2, b_1, b_2 are user-defined values which define how much the amplitude must deviate from the median value. The largest four regions of echoes are kept, as is the largest shadow region. The rest of the image is defined as background and those pixels receive the value 0. Segmenting the image using a simple linear operation, such as removing the mean of the image, does not yield an acceptable mapping of the data, since the pixel amplitudes for the echo/shadow/background classes are not distributed linearly. Echoes have a greater dynamic range than the background and shadow regions, while shadows tend to be concentrated near the low end of the distribution of amplitudes. The segmentation method based upon (1) robustly maps image values into intervals useful for classification. This type of segmentation has been used previously for automated detection algorithms [12].

The resulting segmented image, I', can then be compared to each of the modeled target templates. The templates are decomposed into their echo T_E and shadow T_S components, where, for a pixel at location (i,j), $T_E(i,j) = \max(T(i,j), 0)$ and $T_S(i,j) = \min(T(i,j), 0)$ The template matching procedure $f(T,I)$ between template T and the image I is computed as:

$$f(T,I) = \max\left(\frac{I' \bigotimes T_E}{1 + I'_E \bigotimes \overline{T}_E} + \frac{I' \bigotimes T_S}{1 + I'_S \bigotimes \overline{T}_S}\right) \tag{2}$$

where $\bigotimes$ is the cross-correlation operator and $I'_E(i,j) = \max(I'(i,j), 0)$ and $I'_S(i,j) = |\min(I'(i,j), 0)|$ correspond to the echo and shadow parts of the segmented image. The variables $\overline{T}_E$ and $\overline{T}_S$ are called the *complementary* templates and are defined as

$$\overline{T}_S(i,j) = \begin{cases} 1, & \text{if} \quad T_S(i,j) = 0 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

The complementary echo template is similarly defined. In addition, each element $T(i,j)$ is first normalized with respect to the L_1 norm of T in order to make (2) invariant to the total number of pixels in a given template. The complementary templates are used to penalize areas of echoes or shadows that fall outside the ideal templates. Otherwise, a small template matched with a large object could produce very high values of f which is not a desirable result. The complementary filter reduces this effect. Therefore, the value of f is the peak of the sum of the cross-correlation between the shadow and echo templates, normalized by their respective complementary templates. The measure of (2) is symmetric with respect to the highlight and shadow regions, giving as much weight to the fit of the highlight region as it does to the fit of the shadow region.

The final classification rule $g(I)$ returns a pair (γ, θ) that denotes the best target and aspect template match for all the possible targets and aspects:

$$g(I) = \begin{cases} \underset{\gamma\in\Gamma,\theta\in\Theta}{\arg\max}\, f\left(T^{(\gamma,\theta)}, I\right), & \text{if } f\left(T^{(\gamma,\theta)}, I\right) > \tau \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where τ is a threshold which defines the degree to which I must match any of the templates in order to be classified as a target. If none of the templates exceed τ, then (4) returns a value of 0 (meaning "no target"); otherwise, NSEM returns the pair corresponding to the highest target-aspect.

III. Experimental Results

The NSEM method was tested on sonar data gathered during experiments at sea. The system used was the MUSCLE AUV, designed by the NATO Undersea Research Centre (NURC)[1] [13]. It is comprised of an autonomous vehicle equipped with a synthetic aperture sonar operating at a frequency of 300 kHz, with a 60 kHz bandwidth. This sensor was designed to provide a resolution of roughly 2.5 × 2.5 cm at upwards of 200 m in range. The experimental data was gathered during the summer of 2008 in the Baltic Sea, near Latvia. A number of mine-like targets were deployed and systematically surveyed from different aspects in order to gather a data set of high-resolution images. The targets consisted of a cylinder of length 2.0 m by 0.5 m diameter, a truncated cone shape with a 1.0 m diameter base and 0.5 m height, and a wedge shape of length of roughly 1.0 m by 0.6 m by 0.3 m, with some protrusions such as fins. A large number of rocks, boulders and other clutter objects were also surveyed. The database contained 37 views of the wedge, 69 views of the cone, 65 views of the cylinder and a total of 2351 clutter objects.

The target-aspect templates were generated using the following principles. Since the cone is azimuthally symmetric about its vertical axis its signature does not vary with aspect and only requires a single template; the cylinder was modeled at every 10 degrees of aspect, resulting in 36 templates; and due to its considerably more complex shape, the wedge was generated at every 5 degrees of aspect, resulting in 72 templates, for a total of 109 templates. The templates were also generated for a discrete set of ranges, 20 m to 250 m in steps of 5 m, from the sonar since the geometry of the problem is such that the image of an object will vary as a function of this range, particularly the shadow. A nominal sonar altitude of 13 m (the average altitude during the experiments) was used in the model. From the recorded range of the target, the closest range index for the precomputed templates is computed and the corresponding set of 109 templates comprising the set of different simulated targets and aspects selected for classification.

The NSEM method, is compared with two other techniques.

1) Normalized cross-correlation (NCC1) of the template with the raw image pixels. In order to remove the effect of very large echo values, the image is clipped at three times the median value of the image amplitude.

[1] This data was obtained under the Joint Research Project for Detection and Classification. Thanks to J. Groen and D. Williams of NURC.

TABLE I
CLASSIFICATION CONFUSION MATRICES FOR THE THREE EXAMINED METHODS. (CLASSES: CYLINDER (C), WEDGE (W), TRUNCATED CONE (TC), AND NONTARGET (N))

	NSEM				NCC1				NCC2			
	W	TC	C	N	W	TC	C	N	W	TC	C	N
W	**0.92**	0	0	0.08	**0.62**	0.14	0.03	0.22	**0.81**	0.03	0	0.16
TC	0	**0.96**	0	0.04	0.03	**0.91**	0	0.06	0.01	**0.90**	0	0.09
C	0	0	**0.97**	0.03	0.02	0	**0.92**	0.06	0	0	**0.95**	0.05
N	0.03	0.02	0.01	**0.94**	0.07	0.02	0.02	**0.88**	0.03	0.01	0.08	**0.88**

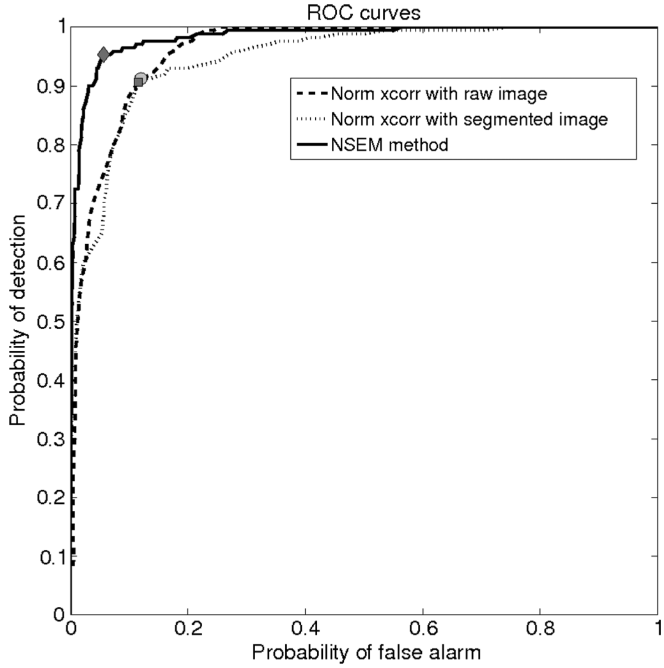

Fig. 1. ROC curves for the three methods considered. The most accurate point, as defined in Eq. (5) is also shown for the three curves. This is the point which is used to compute the confusion matrices shown in Table I.

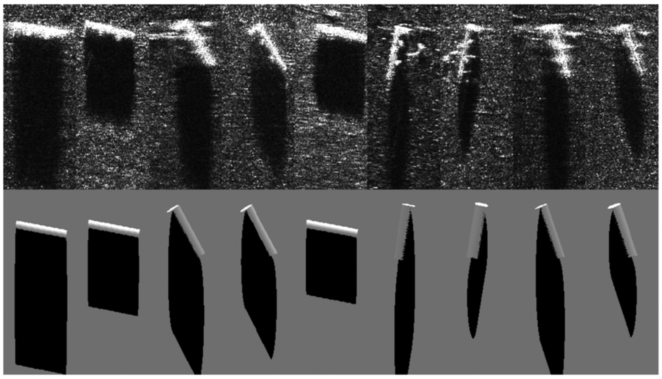

Fig. 2. Large cylinder from several aspects and ranges with the corresponding best matching templates. The first two images correspond to broadside aspects and the last four images correspond to nearly end-on aspects.

2) Normalized cross-correlation (NCC2) of the template with the same segmented image as with the proposed NSEM method.

The computed templates are only slightly smaller (119 × 460) than the size of the target image (119 × 466) (these are dimensions in the vertical and horizontal directions, respectively). For many of the templates, a large fraction of the template is zero. However, these zero areas define the complementary shadow and highlight filters which are used in (2). For the NCC1 and NCC2 methods, these complementary filters are not used, therefore, the most the zeros surrounding the template are removed before applying them. The normalized cross-correlation is implemented within the MATLAB routine (Image Processing toolbox) NORMXCORR2. The parameters for (1) were set to $a_1 = 0.4$, $a_2 = 0.45$, $b_1 = 2.0$, and $b_2 = 5.0$; these were determined empirically and fixed for the entire experiment. Fig. 1 shows the results of the three methods as a receiver-operating characteristic (ROC) curve, obtained by varying the threshold τ. The methods based on normalized cross-correlation achieve adequate performances, both obtaining a probability of correct classification of over 90% with a false alarm rate of less than 12% at their most accurate point. Here, the most accurate point is defined as the point which achieves:

$$\max\left[(p/nt) + (q/nc)\right] \tag{5}$$

where p is the number of targets (cylinders, cones and wedges) correctly classified as targets and q is the number of nontargets correctly classified as such, nt is the total number of targets and nc is the total number of nontargets. The method NCC1 using the raw (clipped) image outperformed NCC2 on the segmented image for all points on the ROC curve. However, both methods are outperformed by the NSEM method, which achieved a probability of detection of over 95%, with a false alarm rate of 6% at its most accurate point on the ROC curve. Table I shows the confusion matrices for the most accurate points of the ROC curve for the three methods.

The confusion matrices allow one to determine the breakdown of the target classification, which shows that the NSEM method obtains greater within-class accuracy than the other methods. This is information that is not captured in the ROC curves, since it only considers the target-non target problem: for instance, a cone classified as a cylinder is still correctly classified as a target, and increases the probability of detection in the ROC analysis. For the NSEM method, not only is the correct target class correctly chosen, the correct target aspect is also obtained. Examples can be seen in Fig. 2 and Fig. 3 for the cylindrical and wedge targets.

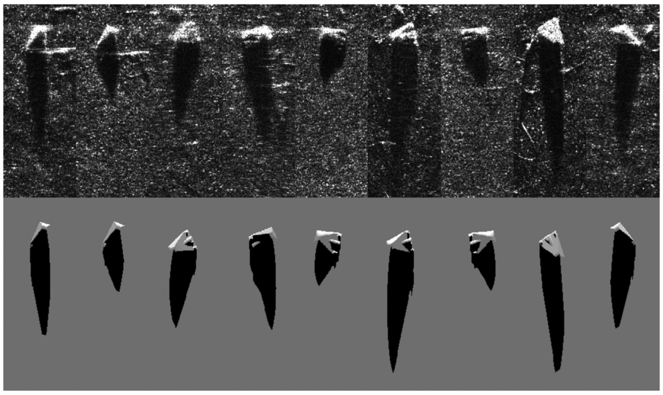

Fig. 3. Wedge target as seen from several aspects and ranges with the corresponding best matching templates.

IV. Conclusions

In this letter, a novel target classification method for high-resolution sonar imagery was presented. This method is based on matching an object's signature image with a number of computed templates generated using a simple ray-tracing method. The matching procedure implements a generalized cross-correlation measure based upon the goodness of fit of the shadow and highlight regions of an object image to a template, normalized by the poorness of the fit of these regions to that of the complementary template. The proposed template matching technique was tested on a large data set of real sonar data and was shown to outperform two other methods based on normalized cross-correlation. The reason for this can be attributed 1) to the the separate consideration of the echo and shadow of a target image and 2) the penalization of the score by the complementary template. For many target images, the shadow can be much larger than the echo, particularly as one goes further out in range. By treating these separately, one is able to control the contribution of each component to the overall similarity measure—in this letter, they are weighted equally in (2)—rather than having the larger shadow dominate the score. Since the background is set to 0, echoes and shadows which lie outside of the template will not be penalized by the cross-correlation methods causing higher scores than one would want for that situation. By penalizing the score by correlation with the complementary template (where the background is now set to 1), this situation is remedied and the template fitting is made more precise.

Analysis by confusion matrix shows that the proposed method is able to precisely classify the target images, while rejecting clutter. In addition, the method is able to correctly determine the target's view aspect. The estimated aspect of a target(as well as the target type) is valuable information which can be used to determine an optimal additional view aspect if multi-aspect classification is desired. For instance, if a given object is classified as a cylinder with an aspect that implies that the base of the cylinder is facing the sensor (an *end-on* aspect), then a second view which obtains the perpendicular aspect along its length (a *broadside* aspect) will yield an image with many pixels on-target, which can be used as corroborating evidence for a more definitive classification. This kind of information will be exploited in the future to develop adaptive control techniques for an autonomous underwater vehicle to automatically and dynamically determine which views to obtain of an object to increase the overall classification accuracy.

Of course, this method requires that the target types are known, and that a 3-D model can be created in order to generate the templates. In the case when the target types are not known, other techniques must be used. Also, in the data set shown in Section III, all of the target objects were proud and lying in a flat, upright position and only the horizontal rotation was variable. In general, however, targets may be found in any vertically rotated position, i.e. upside-down, as well as partially buried. If templates are to be generated and matched for all of these cases, then the problem quickly becomes very large. Although generating templates can be time consuming, this computational burden can be alleviated by generating them into a large library and selecting the appropriate ones when a target is found, as was done in this letter. While a cost of a single calculation of (1) is negligible, the time required to match with the full library of templates increases linearly. Care should be taken to optimize the implementation of (1), as well as limit the number of candidate templates to avoid excessively long calculation times.

References

[1] R. Brunelli, *Template Matching Techniques in Computer Vision: Theory and Practice*. Hoboken, NJ: Wiley, 2009.

[2] V. Myers and M. Pinto, "Bounding the performance of sidescan sonar automatic target recognition algorithms using information theory," *IET Radar Sonar Navig.*, pp. 266–273, 2007.

[3] S. Reed, Y. Petillot, and J. Bell, "Model-based approach to the detection and classification of mines in sidescan sonar," *Appl. Opt.*, vol. 43, no. 2, pp. 237–246, 2004.

[4] I. Quidu, J. Malkass, G. Burel, and P. Vilbe, "Mine classification based on raw sonar data: An approach combining Fourier descriptors, statistical models and genetic algorithms," in *OCEANS 2000 Conf. Proc.*, 2000, vol. 1, pp. 285–290.

[5] J. A. Fawcett and V. Myers, Computer-Aided Classification for a Database of Images of Minelike Objects Defence R&D Canada, Tech. Rep. TM 2004-272, 2005.

[6] E. Coiras, J. Groen, V. Myers, and B. Evans, "Estimation of 3D shape from high resolution sonar imagery for target identification," in *Proc. Int. Conf. Detection and Classification of Underwater Targets*, 2007, pp. 37–44, Inst. Acoust..

[7] M. Hayes and P. Gough, "Synthetic aperture sonar: A review of the current status," *IEEE J. Oceanic Eng.*, vol. 34, pp. 207–224, 2009.

[8] H. Midelfart, J. Groen, and O. Midtgaard, "Template-matching methods for object classification in synthetic aperture sonar images," in *Proc. Third Int. Conf. Underwater Acoustic Measurements*, 2009.

[9] J. Bell, "A Model for the Simulation of Sidescan Sonar," Ph.D. dissertation, Heriot-Watt Univ., Edinburgh, U.K., 1995.

[10] L. Cutrona, "Comparison of sonar system performance achievable using synthetic aperture techniques with the performance achievable by more conventional means," *J. Acoust. Soc. Amer.*, vol. 58, no. 8, pp. 336–346, 1975.

[11] F. B. Jensen, W. A. Kuperman, M. B. Porter, and H. Schmidt, *Computational Ocean Acoustics*. New York: Springer, 2000.

[12] G. Dobeck, J. Hyland, and L. Smedley, "Automated detection/classification of seamines in sonar imagery," in *Proc. SPIE*, 1997, vol. 3079, pp. 90–110.

[13] A. Bellettini and M. Pinto, "Design and experimental results of a 300 kHz synthetic aperture sonar optimized for shallow-water operations," *IEEE J. Oceanic Eng.*, vol. 34, pp. 285–293, 2009.