ATR Performance for Target Clustering

José Melo[a], Samantha Dugelay[a]

[a]NATO STO Centre for Maritime Research and Experimentation (CMRE)
Viale San Bartolomeo 400, 19121 La Spezia
email:jose.melo@cmre.nato.int

Abstract: *The goal of Target Clustering algorithms is to cluster multiple detections of the same target, in a way that can minimize the uncertainty in their location. In order to do so, clustering algorithms usually build on data provided by an Automatic Target Recognition (ATR) software. In particular, they use information on the location of a given detection in a tile, together with an associated weight highlighting how likely a given ATR detection is to be originated from a target.*

In this article we will present an analysis and characterization of the performance of the ATR under use by the CMRE Autonomous Naval Mine Countermeasures (ANMCM) group, for constructive input to the target clustering algorithm. Motivations for this are two-fold. First, it has been observed that in some situations the ATR can fail to completely detect the target. On the other hand, the ATR can sometimes detect a target but assign it a low score which then, depending on the acceptance threshold, can cause the target to remain undetected. Therefore, the output of the ATR needs to be characterized. The analysis will build on experimental data, collected during past trials which cover different operational scenarios and environmental conditions, and curated by an expert.

Results from this analysis will enable to derive an empirically driven probability of detection, but also to statistically characterize the distribution of the ATR scores under different scenarios. It is expected that the obtained results will enable an accurate simulation of an (MCM) mission. Moreover, such results will provide a better link between ATR and clustering algorithm.

Keywords: *automatic target recognition, classification, target clustering*

1. Introduction

Target clustering algorithms use the output of an ATR in order to merge multiple detections of the same target into a single entity, or cluster. Ideally, the formed cluster will benefit from the multiple detections of the same target, by having a better estimate of the target's position. However, in some situations the ATR can fail completely to detect the target. On the other hand, the ATR can sometimes detect a target but assign it a low score which then, depending on the acceptance threshold, can cause a target to remain disregarded. In this article, the goal is to characterize the outputs of an ATR classification algorithm, that will then be used by a subsequent target clustering algorithm.

Figure 1 provides a schematic diagram of an ATR and its interface with target clustering algorithms. In the figure, the ATR obtains 13 detections of the 4 deployed targets, plus two false alarms. Each of the detections consists of an estimated location of the target (x_i, y_i), and an associated classification score, s_i. As illustrated, the target clustering algorithm combines these detections into an appropriate number of clusters. Due to the existence of navigation errors, multiple detections of the same target might present a different location for the same target. Therefore, the target clustering algorithm should also be able to cope with this aspect, providing a single estimate of the position of the target, but also an estimate of its uncertainty. The ATR software here considered consists of a detector [1] and a deep convolutional neural network (CNN) binary classifier [2]. The detector analyses each Synthetic Aperture Sonar (SAS) tile, and identifies possible existing targets. Subsequently, the classifier uses mugshots of those detections to discriminates targets (i.e., mines) from all types of clutter. The output of the ATR software then consists of all the detections and their associated scores.

The remainder of this paper is organized as follows. The next section focus on the experimental data of the different trials, and a review of the ground truth data. Then, Section 3 is devoted to analysing the ATR classification scores. Finally, Section 4, will present some conclusions and future developments.

2. Ground Truth analysis

It is known that some of the detections given by the ATR correspond to actual mine-like targets that have been purposely deployed in the seabed. However, there are also numerous detections which do not correspond to actual deployed targets, and are just an environmental feature, as for example existing rocks. Therefore, the first step on this analysis is to obtain the location of each of the deployed targets. For example, given a SAS tile, which detections given by the ATR correspond to actual targets, and which ones do not?

The data set here in analysis, comprising data collected in five different trials namely MANEX'14, TJMEX'15, NXMEX'15, ONMEX'16 and GAMEX'17, has been curated. This means that an expert in the field has reviewed the tiles and pinpointed the existing mine-like targets, which have been deployed purposely for the execution of the trials. These annotations consist of the pixel location of the centre of each these targets displayed on a given tile, and will be considered "ground truth". Conversely, "non ground truth" detections will be the term used for all the remaining detections.

All sonar data used in this work was collected at sea by MUSCLE, one of CMRE's

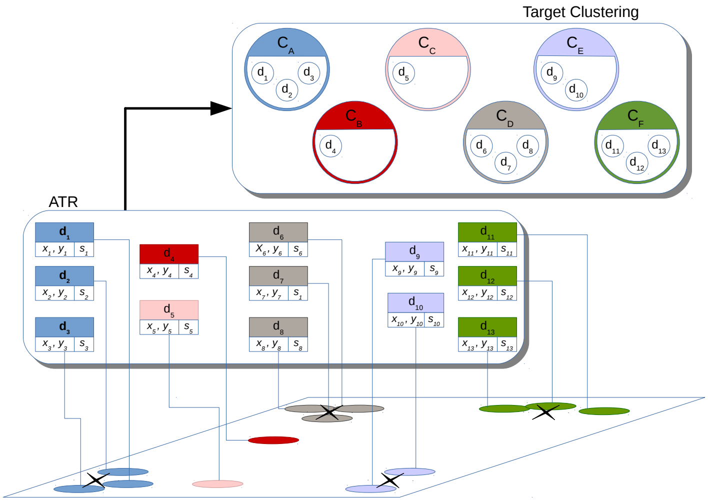


Figure 1: ATR and Target Clustering algorithms. The ATR algorithm provides 11 detections of existing targets, and false alarms. The Target Clustering algorithm merges the 13 detections into 6 different clusters, with the clusters C_A, C_D, C_E and C_F corresponding to the 4 existing targets, and C_B and C_C to the 2 false alarms.

Autonomous Underwater Vehicles (AUVs), which is equipped with a SAS system able to generate high-resolution centimetre level sonar imagery from the sea bottom. The entire data set is composed of a total of 162593 ATR detections, from which 1909 correspond to tiles that contain real targets, as given by ground truth information. Table 1 provides and overview of all the ATR detections data of each trial.

Trial	Detections	Target Observations
MANEX'14	23715	759
TJMEX'15	50420	534
NSMEX'15	65624	323
ONMEX'16	20974	138
GAMEX'17	2220	155

Table 1: Summary of the experimental data sets under analysis

2.1. Missed Detections

Having established the ground-truth, the next step is to assess how many of the targets have been detected or were missed by the ATR. For this, each tile is considered independently from the others. An empirically derived probability of detection, P_d, can then be calculated. Table 2 details the obtained results, categorized according to the respective trial. The column *"Observ."* refers to the number of independent observations of ground truth targets, while the column *"Missed"* refers to the number of times the ATR failed to detect them. The *"P_d"* columns refer to the empirically calculated probability of detection.

By reading Table 2, it can be seen that the probabilities of detection for NSMEX'15, TJMEX'15, and GAMEX'17 are very similar, and above 80%. On the other hand, the probabilities of detection for both MANEX'14 and ONMEX'16 are significantly lower. In

Trial	Day	Observ.	Missed	P_d	P_d
MANEX'14	1	52	52	0.00	0.43
	2	76	76	0.00	
	3	89	87	0.02	
	4	67	38	0.43	
	5	78	30	0.62	
	6	28	10	0.64	
	7	39	11	0.72	
	8	61	23	0.62	
	9	42	23	0.45	
	10	30	15	0.50	
	11	59	16	0.73	
	12	33	12	0.64	
	13	13	10	0.23	
	14	81	24	0.70	
	15	11	7	0.36	
NSMEX'15	1	110	19	0.83	0.82
	2	38	18	0.53	
	3	98	6	0.94	
	4	44	9	0.80	
	5	6	1	0.83	
	6	27	6	0.78	

Trial	Day	Observ.	Missed	P_d	P_d
TJMEX'15	1	73	7	0.90	0.85
	2	30	2	0.93	
	3	58	8	0.86	
	4	177	32	0.82	
	5	31	6	0.81	
	6	165	23	0.86	
ONMEX'16	1	9	5	0.44	0.59
	2	29	6	0.79	
	3	32	15	0.53	
	4	29	11	0.62	
	5	39	20	0.49	
GAMEX'17	1	2	0	1.00	0.88
	2	65	8	0.88	
	3	3	1	0.67	
	4	2	0	1.00	
	5	11	2	0.82	
	6	72	8	0.89	

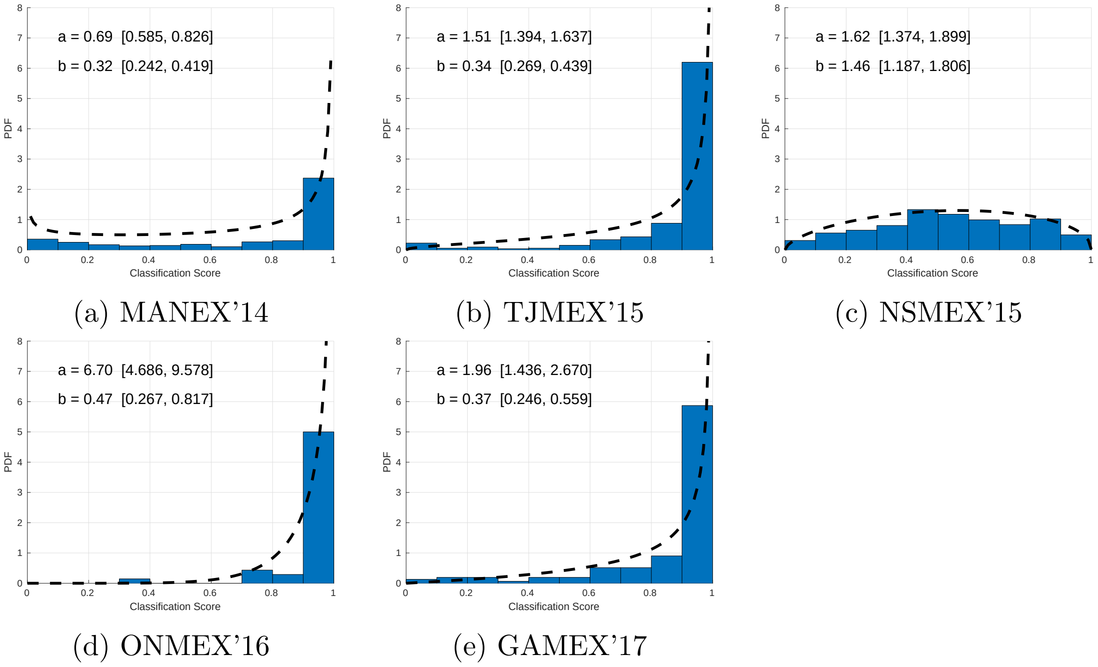
Table 2: Summary of the number of ground truth target observations and missed detections for the different trials

fact, for the MANEX'14 trial, the global P_d is of only 43%. However, it should be noted that the total P_d for MANEX'14 is affected by the extremely low values of P_d for the first 3 days of the trial, when almost none of the observed targets were detected. While it is difficult, at this time, to find a justification for the extraordinarily low probabilities of the detection for these first three days, it is safe to say that such performances are not typical. Therefore, disregarding the first three days, the overall probability of detection achieved during the entire MANEX'14 trial would be of around 60%, which is very similar with what has been achieved during the ONMEX'16 trial.

3. Classification Scores

The classification score provided by the ATR is in fact the output of aforementioned CNN classifier. This score then indicates the probability of a given detection belonging to the targets class. Figures 2 and 3 show the normalized histograms of the classifications scores for all the trials under analysis, where the number of bins was selected according to Sturge's rule. They display, respectively, scores histograms for the ground truth detections, and for all the remaining detections. Additionally, a Beta Distribution has also been fitted to the classification scores of each of the trials, with each of the plots also displaying the value for the a and b parameters of each fit, together with the respective confidence intervals.

When comparing Figures 2 and 3, the main difference between them is the shape of histograms. While in Figure 2 the histograms corresponding to the different trials have a peak in the right most bin, for classification score values around 1, in Figure 3 the opposite happens, with peaks on the left most bin, in the vicinity of zero. The histograms also have an indication of the mode of the classification score, with a value of around 0.95 for the histograms in Figure 2, the ground truth detections, and a value of 0.05 for Figure 3. This is in fact the expected behaviour, which indicates that the ATR is performing



(a) MANEX'14

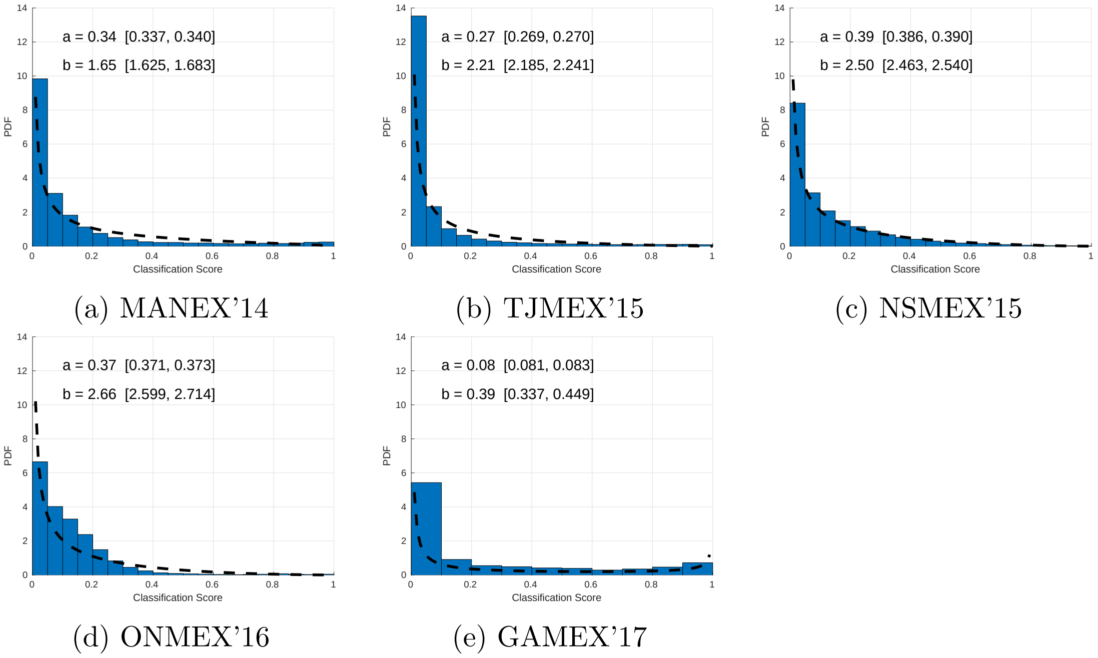
(b) TJMEX'15

(c) NSMEX'15

(d) ONMEX'16

(e) GAMEX'17

Figure 2: Normalized histogram of classification scores of ATR detections for ground truth targets



(a) MANEX'14

(b) TJMEX'15

(c) NSMEX'15

(d) ONMEX'16

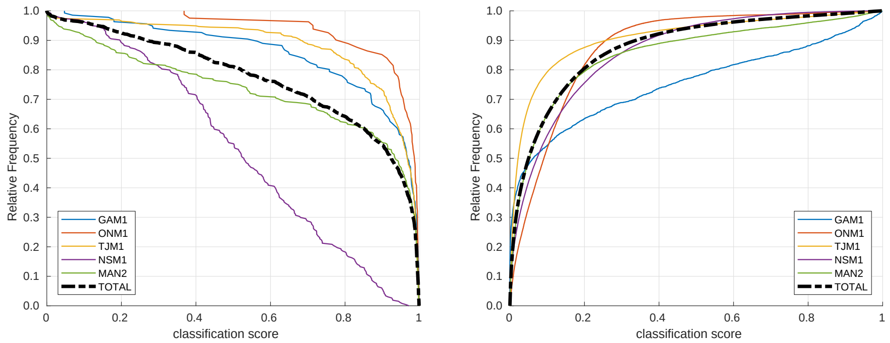
(e) GAMEX'17

Figure 3: Normalized histogram of classification scores of ATR detections for non ground truth targets

well its classification phase. The only exception being Figure 2c, relative to the ground truth detections in the NSMEX'15 trial. In this particular case, the histogram of the ground truth targets detection scores is more evenly distributed among all the possible values for the classification score, with a value for the mode of 0.45. This derives from the fact that, for this particular trial, the ATR was attributing low classification scores to actual deployed mine-like targets. The explanation for this fact comes from the actual nature of the trial, performed in the North Sea, and where the existence of significant water currents was constantly affecting the missions performed by MUSCLE. As a result, the ability of the ATR in correctly classifying mine like targets was affected.

A similar analysis can be made for Figure 3, but in this case the distributions are more similar between each other. Even though the confidence intervals in this case are very narrow, all the plots are very identical, with the main difference being on the height of the peak at the left-hand side. The only exception to that is the plot relative to GAMEX'17, in Figure 3e, where the relative frequency of scores with higher scores is comparably larger than all the others.

Figure 4 presents a more aggregated view of the classification scores for both ground truth and non ground truth targets. Figure 4a presents the complementary cumulative distribution of the classification score for ground truth targets. This plot shows the percentage of the detections that have been assigned a score which is greater or equal than a certain value. For example, for the NSMEX'15 trial, it is possible to see that 100% of the detections have score above 0, 70% have a score above 0.4, and only around 20% of the detections have a score above 0.8. In that sense, it is very clear to see that ONMEX'16 is the trial on which the ATR performed better, with 80% of the detections with a score of roughly 0.8 or above. On the other end, there is NNSMEX'15, where only 5% of the ground truth targets with a classification score above 0.9. A curve for the ensemble of the four trials is also presented. Conversely, Figure 4b presents the cumulative distribution of the ATR classification scores for non ground truth targets. There it is possible to evaluate the percentage of such detections with a classification score below a certain value. For the case of GAMEX'17 trial, one can verify that only 65% of these detections have scores below 0.2, while 20% of them have classification scores above 0.55. It is also possible to see that in both TJMEX'15 and ONMEX'16 the ATR performed particularly well.



(a)

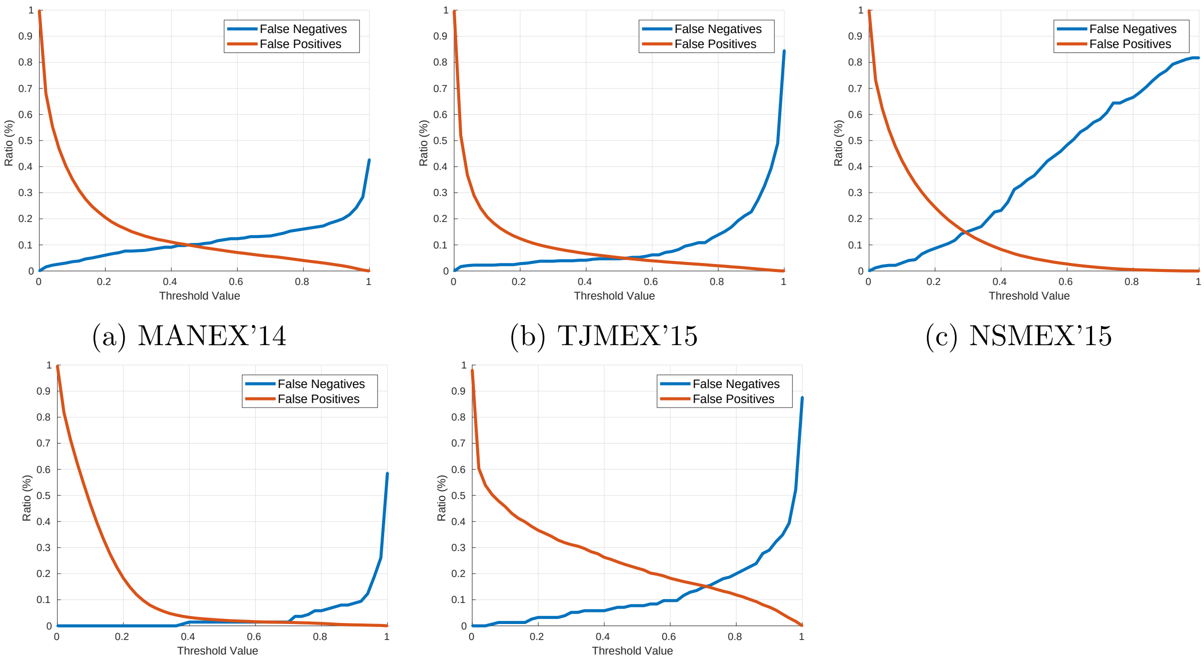
(b)

Figure 4: Cumulative distributions of the classification scores: the complementary cumulative distribution of the classification scores for ground truth targets in 4a, and the cumulative distribution of the scores of non ground truth detections in 4b

3.1. False Positives and False Negatives

From the analysis above, it is possible to conclude that while the classification score gives a fair indication of how likely a given detection is originated from a mine like target, it does not provide an unambiguous indicator. The question that arises is then how to assess if a given detection corresponds to a target? The usual approach is to use a certain threshold level for the classification score to make that decision. Detections with a classification score below that threshold will be discarded, while detections with a score above the threshold will be considered the actual targets. That, however, has certain implications that should not be disregarded. For instance, ground truth target detections with low scores will be disregarded, constituting a false negative situation. Conversely, false positives, also commonly described as a false alarm situations, will arise for ATR detections which do not correspond to real targets. This situations can become particularly adverse when such detections have an associated score with a relatively high value. While it is clear that false positives and false negatives have inherently different costs, addressing this is outside of the scope of the work here presented.

It is clear that there exist two conflicting objectives regarding the false positives and false negatives. On one hand, the possibility of neglecting false negatives should be avoided, as it can result in devastating effects on an operational scenarios. Therefore, an eventual threshold level for the classification score should be set to an arbitrarily low level. However, setting such threshold level very low, will increase the number of false alarms dramatically. As a result, the burden required to process all the detections, in a timely fashion, can be unattainable. Moreover, this would negatively affect subsequent phases of unmanned MCM operations, such as target reacquisition and identification. In



(a) MANEX'14

(b) TJMEX'15

(c) NSMEX'15

(d) ONMEX'16

(e) GAMEX'17

Figure 5: Variation of the rate of False Positives and False Negatives with the classification score threshold

order to avoid this, then an eventual threshold acceptance level should be set arbitrarily high, so that the ratio of false positives is kept as low as possible. Figure 5 presents the plot for the rate of false positives and rate of false negatives for each one of the trials, with a varying threshold acceptance level. The mentioned conflicting objectives are evident.

A rather intuitive approach for that would be to choose a threshold value that would minimize both false positives and false negatives rate. Visually, this threshold level can be identified as the intersection point between both lines. This, of course, assumes that false positives and false negatives have a similar cost. Arguably, this might not be the case in operational scenarios as the cost of missing a real mine can be too high. Figure 5 also highlights another aspect of choosing a threshold level. By comparing the plots in Figures 5a to 5e, it is unequivocal that for the different trials, the false positives and false negatives lines intersect with different classification scores. While for NSNEX'15 this intersection occurs for a value of around 0.3, for GAMEX'17 this occurs for a value of around 0.7. These are the extreme cases, and for other trials this happens for intermediate values. This suggests that and adaptive threshold level should be used, that is dependent on operational and environmental conditions.

4. Conclusion and Outlook

This report presented an analysis performed to the output of the ATR. The motivation for this analysis is to obtain suitable characterization of the ATR output, under different operational environments, so that it can be used by a target clustering algorithm, which is an ongoing effort.

This analysis started with examining the missed detections and empirical probabilities of detection of the different trials. Taking into consideration the specifics of each trial, particularly the MANEX'14, results show that, roughly, the trials can be divided within two categories according to the empirical probabilities of detection: one with P_d of around 60%, and one with P_d of around 85%. Furthermore, the classification scores have been considered. It has been shown how the rate of false positives and the false negatives can be strongly influenced by the value of the chosen threshold level for the scores. A naive approach for choosing such threshold, on which both the rate of false positives and false negatives is minimized, highlighted the need for a more robust and adaptive way to select such threshold. Finally, the distribution of the classification scores has been approximated with a Beta distribution, that allowed to compare the similarities between data from all the trials.

REFERENCES

[1] David P. Williams. The mondrian detection algorithm for sonar imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2):1091–1102, Feb. 2018.

[2] David P. Williams. Demystifying deep convolutional neural networks for sonar image classification. In *Proceedings of the Underwater Acoustics Conference*, Skiathos, Greece, Sep. 2017.