Evaluation Testbed for ATD Performance Prediction (ETAPP)

Scott K. Ralph[a] , Ross Eaton[a], Magnús Snorrason[a], and John Irvine[b], and Steve Vanstone[c]
[a]Charles River Analytics, Cambridge, MA, USA
[b]SAIC, Burlington, MA, USA
[c]AMRDEC, Redstone Arsenal, AL, USA
email: {sralph|reaton|mss}@cra.com, John.M.Irvine@saic.com, steve.vanstone@us.army.mil

ABSTRACT

Automatic target detection (ATD) systems process imagery to detect and locate targets in imagery in support of a variety of military missions. Accurate prediction of ATD performance would assist in system design and trade stud-ies, collection management, and mission planning. A need exists for ATD performance prediction based exclusively on information available from the imagery and its associated metadata. We present a predictor based on image measures quantifying the intrinsic ATD difficulty on an image. The modeling effort consists of two phases: a learning phase, where image measures are computed for a set of test images, the ATD performance is measured, and a prediction model is developed; and a second phase to test and validate performance prediction. The learning phase produces a mapping, valid across various ATR algorithms, which is even applicable when no image truth is avail-able (e.g., when evaluating denied area imagery). The testbed has plug-in capability to allow rapid evaluation of new ATR algorithms. The image measures employed in the model include: statistics derived from a constant false alarm rate (CFAR) processor, the Power Spectrum Signature, and others. We present performance predictors for two trained ATD classifiers, one constructed using using GENIE Pro™, a tool developed at Los Alamos National Laboratory, and the other eCognition™, developed by Definiens (http://www.definiens.com/products). We present analyses of the two performance predictions, and compare the underlying prediction models. The paper concludes with a discussion of future research.

1. INTRODUCTION

Accurately predicting ATR performance based on image measures has several benefits. The very large total number of parameters, such as sensor parameters (resolution, wavelength), operating conditions (e.g. time of day, humidity, temperature), viewing geometries (e.g. range, angle), and scene content (e.g. contrast, amount of clutter), make it hard to quantify the performance over all relevant parameters. By quantifying performance with respect to a relatively few number of image measures, we can greatly simplify performance prediction. Once the prediction based on the image measures has been shown to be sufficiently robust, quantifying performance in a novel scenario requires only the determination of how the new scenario affects the image measures.

Previous work in ATR development and evaluation has shown that performance is strongly influenced by the operating conditions, which characterize the target, backing clutter, and the sensor. Rather than attempting to enumerate and quantify the full set of operating conditions, the image measures behave as a compact description of the operating condition [1].

Aggressive deployment timelines, as well as the need to assess the ATR on imagery with no ground-truth (e.g. denied area imagery) necessitate the ability to predict performance based purely on measures on the imagery. A robust performance prediction capability that rests entirely on information derived directly from the imagery can support improved system development and tuning, system and sensor design and trade studies, tasking and collection management, and mission planning.

To meet the Army AMRDEC's directorate need for a performance prediction testbed, we have developed a two-phase testbed, depicted in Figure 1. The first phase involves computing a set of image measures over an image training set, and computing the ATR performance on theses same images. The by-product of the training is a mapping from image measures to a predicted ATR performance. The second phase computes performance predictions purely from measures computed on the images.

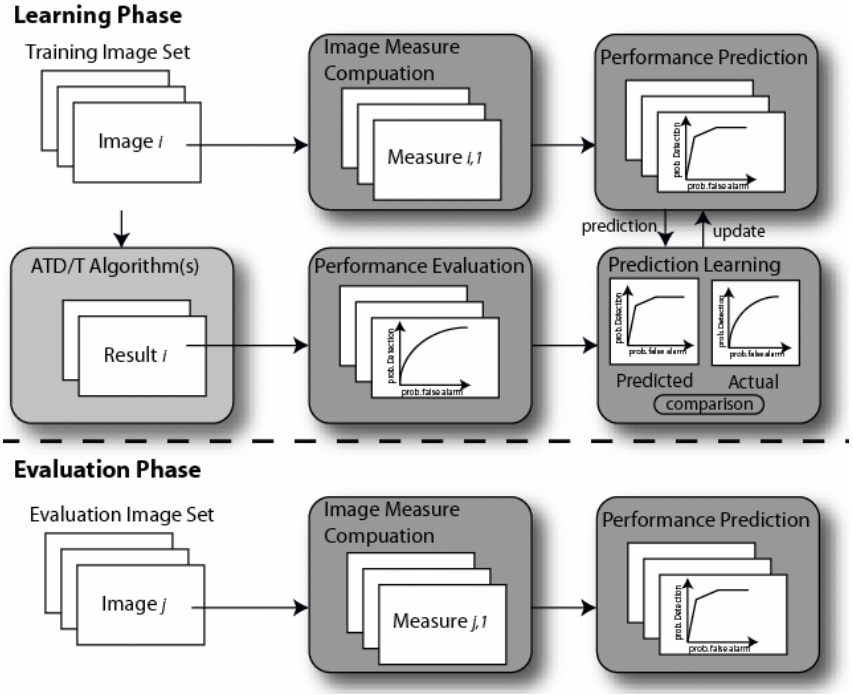


Figure 1: ETAPP performance prediction testbed

2. ARCHITECTURE

The system architecture consists of several modular components, depicted in Figure 2. A suite of test images are read by the Image Stream Adapter that interfaces to the Algorithm Plug-in. The Plug-ins are based on a standardized interface where images will be provided to the various ATR algorithms being tested and results are accepted back in a standard format. The candidate algorithms are scored using START [2], our software framework for ATR scoring and data truthing. This measure of observed algorithm performance is provided to the Performance Learning component that is responsible for iteratively refining its model of predicted performance. The Performance Prediction Engine consists of the various image measures used to form the performance predication using the Performance Estimator.

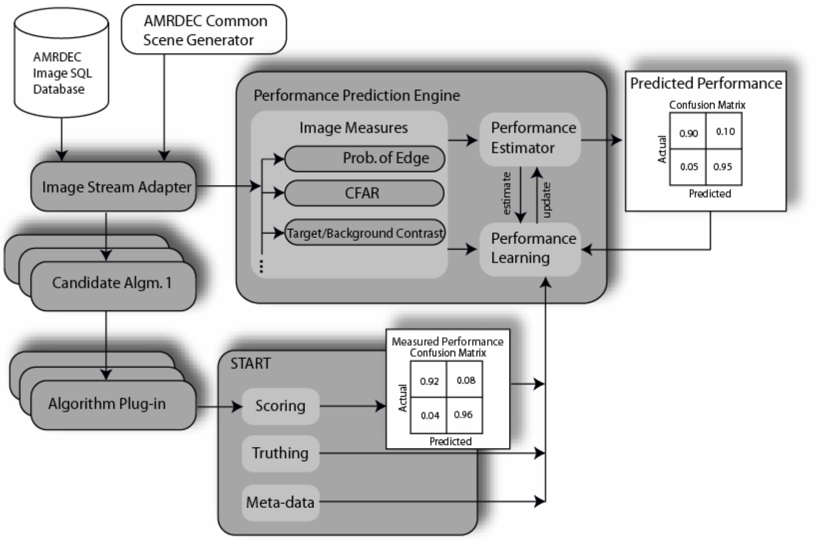


Figure 2: System architecture

3. TEST IMAGERY

In order to obtain a large variance in the set of operating conditions covered by the test imagery, we compiled a set of infrared video images taken from the VIVID data collect at Eglin AFB. A total of 9 video sequences, typically 1000 frames each, were concatenated into a single video, and sampled at regular intervals to produce a set of 100 representative images. These images, depicted in , varied in the size and type of the target under consideration, the viewing geometry, the degree of clutter, the contrast of the target against the background, etc. Additionally we made use of FMTI closing sequence data of an Infrared seeker from our AMRDEC contract.

4. IMAGE MEASURES

In this section we will describe the set of image measures that were used in the construction of the performance prediction.

4.1. MODIFIED CONSTANT FALSE ALARM RATE

The original definition of the Constant False Alarm Rate (CFAR) Detector, depicted in the left portion of Figure 4, determines whether a single pixel value is significantly different from the surrounding region [3;4]. We use a traditional t-test value to determine when the center region significantly differs from the surrounding area.

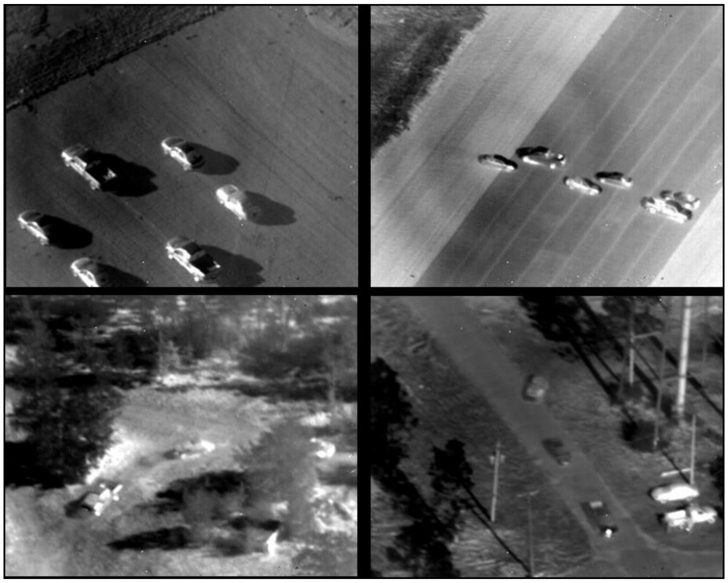


Figure 3: VIVID infrared image samples

The t-test value is given by:

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{((c_1 - 1)\cdot\sigma_1^2) + ((c_2 - 1) * \sigma_{2\cdot}^2)}{c_1 + c_2 - 2}\left(\frac{1}{c_1} + \frac{1}{c_2}\right)}} \qquad (1)$$

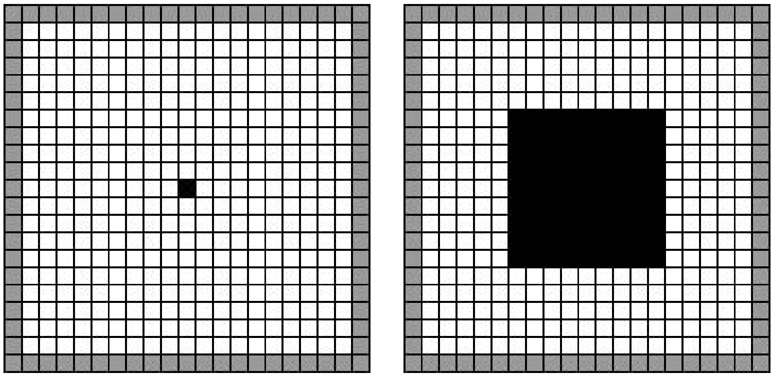


Figure 4: Left: Original CFAR detector. Right: Modified CFAR detector. Black squares are target pixels, gray squares are clutter and white are the buffer area

The CFAR image measure can be computed quickly by using an Integral Image Representation $I(i, j)$, that encodes the sum of the pixel values lying above and to-the-left of the index (i, j).

4.2. NORMALIZED CROSS CORRELATION

Given accurate truth information, we can compute the normalized cross correlation of the image chip pixels at various clutter positions of the image. This gives an indication of how similar the clutter and the target appear in the image. Small cross correlation scores at clutter positions indicate a target that is relatively easily to detect with low false alarm rates, while large cross correlation scores indicate situations where it is difficult to detect with low false alarm rates.

4.3. POWER SPECTRUM SIGNATURE

The Power Spectrum Signature (PSS) is defined in terms of the pixel values $t_{i,j}$ of the image containing the target, and the pixel values, $b_{i,j}$, of the expected image, that is, the image with the target removed. Given these values, the PSS is defined as [5]:

$$\mathrm{PSS} = \left[\frac{1}{\mathrm{POT}} \sum_{i,j} (t_{i,j} - b_{i,j})^2 \right]^{1/2} \qquad (2)$$

where POT is the number of pixels on target.

While there is no way of determining what the pixel values would be if the target was not present, the guiding principal that is often used is that the expected background should not draw the attention of the observer. One simple approach is to compute the Line Expected Background of the image [5], where the background is assumed to vary linearly in the horizontal direction and to remain constant in the vertical direction. To compute the gradient, we sample the background in regions to the left and right of the target (see Figure 5), and compute the corresponding means μ_L and μ_R respectively. The Line Expected Background is then taken to be:

$$b_{i,j} = \left(1 - d/n\right)\mu_L + d/n\,\mu_R \qquad (3)$$

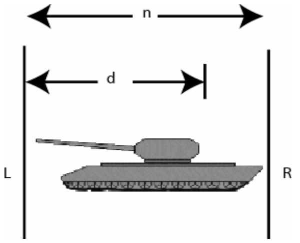


Figure 5: Line Expected Background computation

4.4. CLUTTER METRIC

The Clutter Metric can be used to estimate the amount of clutter in an image [5]. It is computed by sliding a square box over an image and subtracting the mean of the pixels in the box from the value of the pixel that the box is centered on. The box is kept to the interior of the image, so that its borders never leave the image. Ideally, the Clutter Metric should be performed on an image once the targets have already been removed (using the Line Expected Background and the PSS), so they will not be judged as clutter. The equation for computing the Clutter Metric is:

$$CM = {}^{1}\!/_{N} \left(\sum_{i} \sum_{j} \left(b_{i,j} - \overline{B}_{i,j} \right)^{2} \right)^{1/2} \tag{4}$$

where *b* is the value of the *i,j* pixel, B is the mean of the box centered at pixel *i,j*, and N is the number of boxes convolved over the whole image. The size of the box is chosen to be the same approximate size as the target.

4.5. PROBABILITY OF EDGE

To compute the Probability of Edge (POE) we first convolve the image with a Difference of Gaussians (D.O.G.) filter to enhance edges.. The advantage of the D.O.G. is that, since it is implemented as two Separable Gaussian Filters, it can be computed quickly.

If we let τ represent the threshold, and subdivide the image into a set of N regions with twice the apparent size of the target, then the Probability of Edge is defined to be:

$$\mathrm{POE} = \left[\sum_{i=1}^{N} \mathrm{POE}_{i,\tau}{}^{2} \right]^{\frac{1}{2}} \tag{5}$$

where $\mathrm{POE}_{i,\tau}$ is the total number of pixels in the region that exceed the threshold τ .

5. ATD ALGORITHM EVALUATION

To provide automated target detection results, two methods were applied to a set of MWIR data. The imagery was a set of still frames extracted from the VIVID Public Release Set. The two ATD tools were GENIE Pro™ and eCognition™.

5.1 GENIE PRO

Los Alamos National Laboratory’s GENIE (GENetic Imagery Exploitation) system is a machine learning software package using techniques from genetic genetic programming, [6], to construct information extraction algorithms. Both the structure of the information extraction algorithm and the parameters of the individual image processing are learned. GENIE has been applied to a variety of target detection tasks [7;8]. We used GENIE Pro to perform target detection on two sets of infrared imagery: the VIVID data set and the RACER data set.

GENIE Pro has a simple interface for loading imagery, selecting training data, training a classifier, and then apply it to other images. The system is exemplar based with both positive and negative examples used, see (Figure 6, left). Using a simple GUI interface, the operator “paints” the examples and counter-examples in the image. These pixels then define the training set for the classifier. The process begins with a set of primitive image operators, which are concatenated in various ways to produce the classifier. The criteria for evolving the classifier depends on the percent of the training data correctly classified, with a penalty for complexity. If performance is not acceptable, the user can mark up additional training data and iteratively refine the classifier

5.2 eCognition

Definiens developed eCognition to overcome the limitations and weaknesses of traditional pixel-based image classification methods for satellite and aerial imagery. eCognition's unique approach is based on simple concept: important semantic information necessary to interpret an image is not represented in single pixels but in meaningful image objects and their mutual relations. Hence, Definiens Imaging has developed a software-based procedure for object oriented and multi scale image analysis which can be applied to very different tasks. Employing a number of sophisticated capabilities resident in eCognition, the user develops a rule set through an iterative process of testing and refining with the training imagery. The process flow begins with image segmentation, followed by extraction of features from the image segments, and concluding with the classification step. Complex rule sets can include multiple iterations through this cycle to handle challenging problems.

5.3 SCORING PROCEDURES

Since GENIE produces a pixel-level classification, a set of rules are needed to score target detection performance. Although eCognition generates a classification of the image segments, rather than individual pixels, the same scoring rules were applied to the eCognition results. The specific rules were:

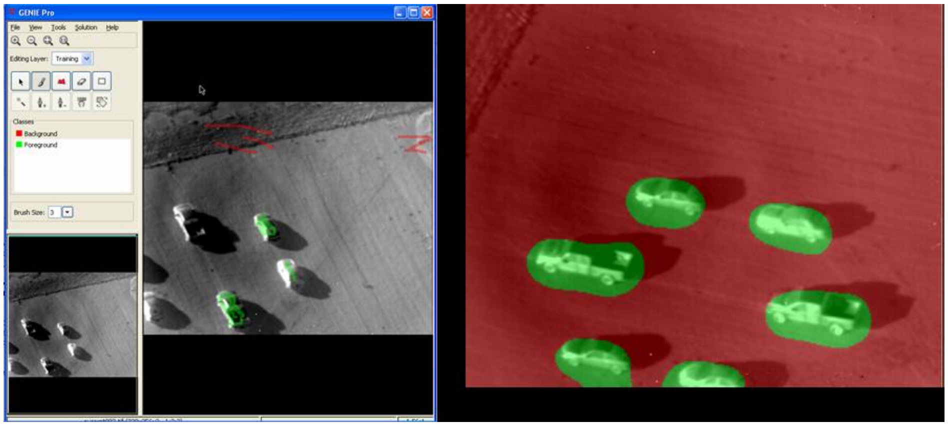
- A contiguous set of target pixels (GENIE) or a single image segment (eCognition) which intersect the true target are considered a detection.
- Two or more target regions intersecting the true target are considered a single detection
- Contiguous regions or image segments that are spatially separated from the targets are considered false alarms.
- A contiguous false alarm region that is the union of several compact, target-size regions are counted as multiple false alarms.

Both classifiers were applied to 80 VIVID images to yield a set of performance results for subsequent analysis. Figure 7 shows sample GENIE classifier results. These measures of ATD performance quantify target detection, false alarms, and a combined measure of effectiveness (MOE):

PD = No. targets detected / No. targets in the image

FAR = number of False Alarms per image

MOE = No. targets detected / (No. targets in image + No. False Alarms)



Editing Layer: Training

Brush Size: 3

Figure 6: (left) User interface for GENIE Pro specifying a classifier for VIVID imagery, and (right) results of applying the classifier

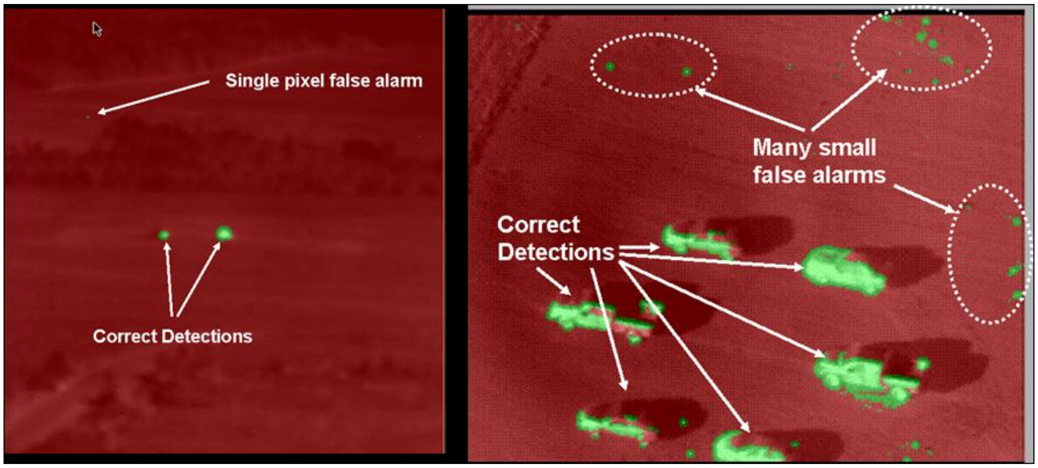


Figure 7: (left) Classifier trained on and applied to RACER imagery, (right) same classifier applied to VIVID imagery

The analysis proceeds in three phases. First we model the relationship between the image metrics and the ATD results from GENIE Pro. A similar analysis is performed using the ATD results from eCognition. Finally, we compare the two sets of models and the relationships between the two ATD methods. The initial analysis included scatter plots, summary statistics, and correlation analysis to identify image metrics which showed some meaningful relationship to the measures of ATR performance.

Earlier research, based only on ATD results generated with GENIE Pro, reported on the stepwise regression analysis that was performed to assess the relationships between each measure of ATR performance and the various metrics [9]. The stepwise approach assists in identifying strong relationships, while excluding redundant or collinear explanatory variables. This initial analysis provided limited predictive power in the sense that the regression models based on the image metrics defined above only account for 30-40 percent of the variance in the three measures of ATR performance. Further investigation extended the earlier work by exploring the relationship between the training set of imagery used to develop the ATD algorithm and the test set for which ATD performance has been scored [10]. To quantify the "distance" from a test image to the training data, the image metrics defined above were computed for the training set. The absolute value of the difference between the image metric for training imagery and for the current test image provides a new image metric. These metrics are denoted with the "D_" prepended to the variable name. In other words, let X be an image metric. Then, for the i^{th} test image, the value of the new image metric based on the relationship to the training data is defined by:

$$D\_M = \| M_i - M_{train} \| \qquad (6)$$

The analysis used all of the original metrics defined above, along with all of the new metrics constructed from the relationship to the training imagery. The full set of metrics was included in a stepwise regression analysis to identify the factors that provided good explanatory power. This approach yielded substantially better explanatory power, with R^2 ranging from 0.72 to 0.80 for the various models Table 1.

Table 1. Improvement in R^2 for GENIE Pro Using Distance From Training Set

Dependent Variable	Old R-square	New R-square
PD	0.54	0.8
FAR	0.29	0.72
MOE	0.21	0.72

The results using GENIE Pro for each measure of ATD performance appear in Table 2 through Table 4. When the image metrics are analyzed jointly in the regression analysis, the final models differ from the specific metrics identified by the simple correlation analysis. This arises from the relationships among the various image metrics and the fact that the stepwise procedure will exclude redundant variables. Similarly the procedure will include variables that offer additional explanatory power, given the variables already entered in the model.

Table 2: Regression model for predicting P(Det) for GENIE Pro

Variable	Coefficient	Std. Error	t-statistic	P value
(Constant)	0.822	0.08	10.3	> 0.0005
D_CM	-0.018	0.002	-7.9	> 0.0005
Mean_POT	0.0001	0.00011	4.9	> 0.0005
NCC_Mean	0.816	0.323	2.5	0.014

R-square = 0.80

Table 3: Regression model for predicting FAR for GENIE Pro

Variable	Coefficient	Std. Error	t-statistic	P value
(Constant)	-5.92	2.39	-2.5	0.016
clut_cfs	1.84	0.285	6.4	> 0.0005

D_clutcfs	1.542	0.397	3.9	> 0.0005
D_POE	-10.13	2.681	-3.8	> 0.0005
POE	-3.515	1.349	-2.6	0.011
PSS	0.002	0.001	2	0.048

R-square = 0.72

Table 4: Regression model for predicting MOE for GENIE Pro

Variable	Coefficient	Std. Error	t-statistic	P value
(Constant)	0.99	0.081	12.3	> 0.0005
D_Target CFAR Mean	-0.019	0.003	-7.1	> 0.0005
D_Clutter CFM	-0.167	0.054	3.1	0.003
D_NCC_SD	1.42	0.633	2.2	0.028

R-square = 0.72

When a similar analysis was conducted with the ATD results generated with eCognition, a similar set of models were produced (Tables 5, 6, 7). A comparison of the two sets of models reveals some interesting findings. First, the image metrics differ for the two sets of models. For example, pixels on target (POT) is an important explanatory variable for the GENIE model for Probability of Detection (PD), but not for the eCognition model. Several factors could account for the selection of different explanatory variables in the two models. The image metrics exhibit a substantial level of multicollinearity. Thus, inclusion of a specific independent variable in the stepwise selection can be sensitive to small differences in the data. A second reason is that the two ATDs rest on fundamentally different approaches. GENIE Pro is a machine learning approach which relies on empirical methods for developing the classifier. In contrast, eCognition is a rule-based approach. Based on expert knowledge, the user develops the classifier by positing relationships and verifying with the training data.

Finally, the models for eCognition have substantially lower explanatory power (as measured by R-square) than the models for GENIE Pro. This is not in anyway a deficiency of the ATD. Rather, it indicates that the image metrics investigated to date are not sufficient to characterize eCognition performance. Further, it suggests that no single approach for performance prediction can be expected, i.e., an “ATD-agnostic” performance prediction model appears unlikely.

Table 5: Regression model for predicting P(Det) for eCogntion

Variable	Coefficient	Std. Error	t-statistic	P value
(Constant)	0.79	0.11	7.03	> 0.0005
PSS	-0.00027	0.0001	-3.05	0.0032
D_PSS	0.00038	0.0001	2.85	0.0057
POE	0.21	0.10	2.12	0.0378

R-square = 0.23

Table 6: Regression model for predicting P(Det) for FAR

Variable	Coefficient	Std. Error	t-statistic	P value
(Constant)	-3.59	1.98	-1.81	0.0740
Clut_cfs	1.81	0.30	6.02	> 0.0005
D_clutcfs	1.31	0.45	2.92	0.0047
D_poe	-10.03	2.56	-3.92	0.0002
POE	-3.99	1.28	-3.12	0.0026

R-square = 0.50

Table 7: Regression model for predicting P(Det) for MOE

Variable	Coefficient	Std. Error	t-statistic	P value

(Constant)	1.0138	0.2440	4.1544	0.0001
PSS	-0.0003	0.0001	-3.7515	0.0004
clut_cfs	-0.0903	0.0306	-2.9466	0.0044
D_PSS	0.0003	0.0001	2.5020	0.0147
POE	0.4661	0.1309	3.5598	0.0007
D_clutcfs	-0.1247	0.0449	-2.7755	0.0071
D_POE	0.5878	0.2615	2.2475	0.0278

R-square = 0.42

In the aggregate, the two ATDs yielded similar performance. Overall detection rates and false alarm rates were very similar. However, the two methods failed in different ways and on different images. From the standpoint of modeling and predicting ATD performance, this suggests that different image metrics and different models will be needed to predict the performance of various ATD algorithms.

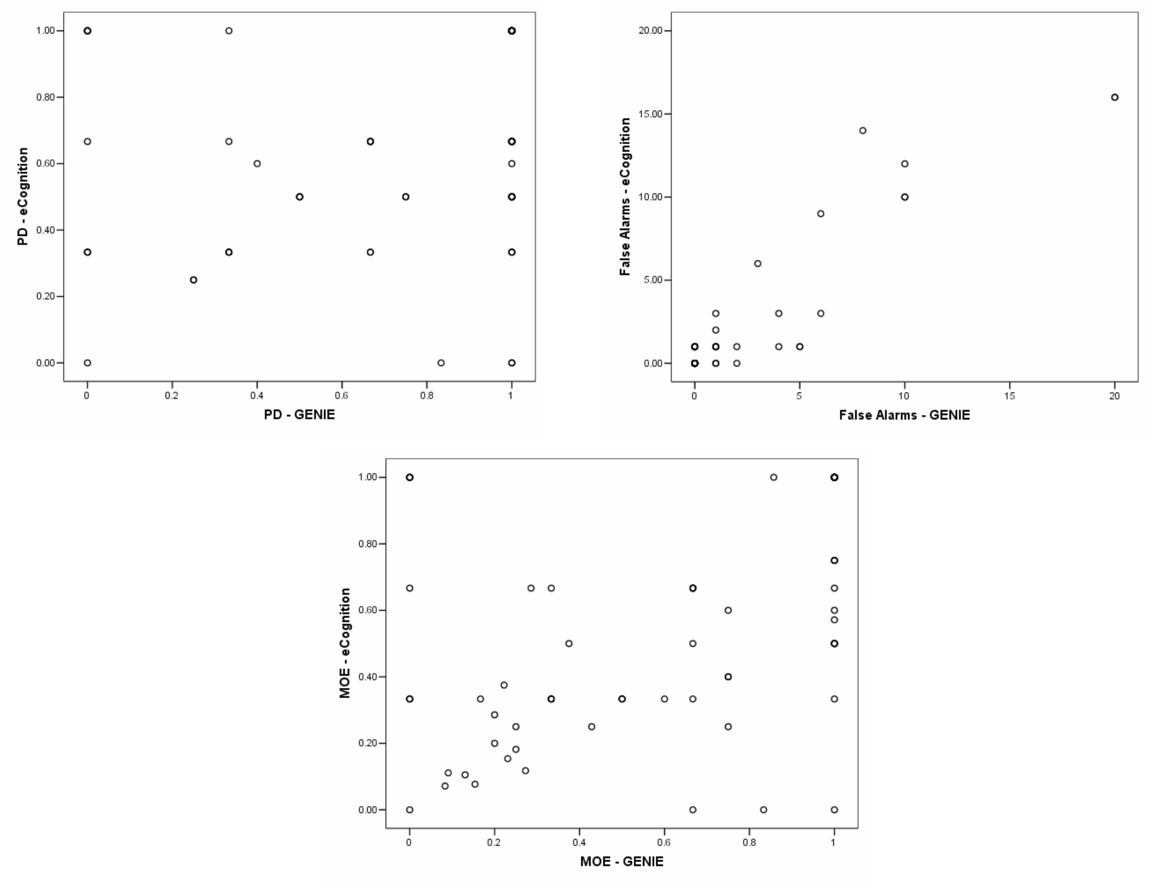


Figure 8. Comparison of the Performance Measures (PD, FAR, MOE) for the Two ATDs

6. CONCLUSIONS

The analysis presented here offers a method for predicting ATD performance based on information extracted directly from the imagery. While the image metrics offer some explanatory power, especially for predicting detection rates, there is room for improvement. The relationship of the training data to the test data is important in predicting ATD performance and inclusion of this information has improved performance of these models over the results reported earlier. Examination of the individual images indicates that performance degrades for images that are substantially different from the training set in terms of the operating conditions. We quantified this effect using the set of image metrics. Thus, the explanatory variables considered for this modeling effort included the actual image metrics for each image and the absolute differences of these metrics from the mean value for the training set, the later representing “distances” from the training set. Stepwise regression analysis was performed to determine the effect these measures play in predicting performance. The result was a substantial increase in the explanatory power of the models.

A comparison of the performance modeling for two different ATD methods reveals the challenges of develop a general ATD performance prediction capability. The two ATD tools examined here exhibit similar overall performance, but differ with respect to individual images and targets. The prediction models developed for one ATD are different, in terms of both specific image metrics and overall explanatory power, than the model for the other ATD. Further investigation is needed to full characterize the two ATDs, but it appears that a general solution remains elusive.

7. REFERENCES

[1] Clark, L. and Velten, V. J., "Image Characterization for Automatic Target Recognition Algorithm Evaluations," *Optical Engineering*, vol. 30, No. 2, no. February 1991, pp. 147-153, 1991.

[2] Ralph, S. K., Irvine, J. M., Stevens, M., Snorrason, M., and Gwilt, D. Assessing the Performance of an Automated Video Ground Truthing Application. Applied Imagery and Pattern Recognition . 2004.

[3] Novak, L. M., Burl, M. C., and Irving, W. W. Optimal Polarimetric Processing for Enhanced Target Detection. IEEE Trans.Aerospace and Electronic Systems Vol. 29[1], 234-244. 1-1-1993.

[4] Novak, L. M., Owirka, G. J., and Netishen, C. M. Performance of a High-Resolution Polarimetric SAR Automatic Target Recognition System. The Lincoln Laboratory Journal Vol. 6[1], 11-24. 1-1-1993.

[5] Wilson, D. L. Image-Based Contrast-to-clutter Modeling of Detection. Optical Engineering Vol. 40[9], 1852-1857. 2001.

[6] Goldberg, D. E., *Genetic Algorithms: In Search, Optimization, and Machine Learning* Reading, MA: Addison Wesley Publishing Company, 1989.

[7] Harvey, N. R. and Theiler, J., "Focus of Attention Strategies for Finding Discrete Objects in Multispectral Imagery," *Proceedings of the SPIE*, vol. 5546 2004.

[8] Theiler, J., Harvey, N., David, N. A., and Brumby, S. B. New Approaches to Target Detection and New Methods for Scoring Performance. 33rd Applied Imagery and Pattern Recognition Workshop: Image and Data Fusion. 2004.