Standard SAR ATR Evaluation Experiments using the MSTAR Public Release Data Set

Timothy Ross*, Stephen Worrell+, Vincent Velten*, John Mossing#, Michael Bryant*

*Sensor ATR Technology Division
Sensors Directorate
Air Force Research Laboratory
AFRL/SNA 2010 Fifth St.,
WPAFB, OH 45433-7001

Sverdrup Technology Inc
4200 Colonel Glenn Highway
Suite 500
Beavercreek, OH 45431

+Wright State University
Dept. of Electrical Engineering
3640 Colonel Glenn Highway
Dayton, OH 45435-0001

ABSTRACT

The recent public release of high resolution Synthetic Aperture Radar (SAR) data collected by the DARPA/AFRL Moving and Stationary Target Acquisition and Recognition (MSTAR) program has provided a unique opportunity to promote and assess progress in SAR ATR algorithm development. This paper will suggest general principles to follow and report on a specific ATR performance experiment using these principles and this data. The principles and experiments are motivated by AFRL experience with the evaluation of the MSTAR ATR.

Keywords: ATR, Automatic Target Recognition, Evaluation, SAR, Data Sets, Performance, Standards

1. INTRODUCTION

Standard evaluation procedures are proposed for use with the MSTAR data to promote generally sound evaluation practices and to make results comparable across different evaluations. We are interested in encouraging the various ATR research and development activities to focus on the truly challenging technical problems and to minimize data-specific local optimizations. The developers are interested in producing convincing evidence of genuine advances in ATR science. Having broadly accepted, sound, and comparable evaluation practices will facilitate both of these aims. This paper consists of two main sections, suggested general principles for experiments using the MSTAR Public Released data and then a specific experiment using these principles with this data.

MSTAR is developing the next generation SAR ATR, and has conducted three data collections to support the development and testing of this algorithm. The collections were made in September '95, November '96, and May '97. The MSTAR collected data is perhaps unique in its importance to the ATR community. There is a significant quantity of images for statistical significance, it is well truthed, it was collected with a state-of-the-art radar, and it has a variety of military vehicles under a variety of controlled conditions.

The MSTAR Public Release data is a sub-set of the MSTAR data from the 9/95 collection. It is available upon request from Veda Inc. (MSTAR data managers for AFRL and DARPA). The data consists of X-band SAR images with 1 foot by 1-foot resolution. The target images contain one of three T72 Main Battle Tanks (MBTs), one of three BMP2 Armored Personnel Carriers (APCs), or a BTR70 APC. There are images of a test object (Slicy) available also. The target images are 128x128 pixels and were collected near Huntsville, Alabama by Sandia National Laboratory (SNL) using the STARLOS sensor. Contact Veda on the web at http://www.mbvlab.wpafb.af.mil/public/MBVDATA for more information about the data.

The data used in the specific experiment below includes the Public Release set and some additional data from the September '95 and November '96 collections.

Part of the SPIE Conference on Algorithms for Synthetic Aperture Radar Imagery V
Orlando, Florida • April 1998
SPIE Vol. 3370 • 0277-786X/98/$10.00

2. EVALUATION PRINCIPLES

2.1 Introduction

Specific measures are suggested below, but whether these measures or other measures are used, they should be fully defined. It is further recommended that performance measures be chosen to match previous studies whenever possible. A principal goal of this paper is to suggest a set of measures (that are being used within the MSTAR program) that might form a core set to be used in follow-on evaluations.

Performance results should include some indication of the confidence interval around any reported value. A specific confidence interval calculation is suggested below, but other expressions of the basis for a result, such as the number of images tested, would provide the same kind of information.

The evaluation report should specify exactly which images were used for training or tuning the ATR system and which images were used in testing the ATR system. There should be little or no intersection between these two sets.

We also encourage "extended operating condition (EOC)" conscious partitioning of the data between training and testing sets. That is, the MSTAR Public Data Set is NOT a random sample of the set of operationally interesting SAR images. Therefore, selecting training and/or tests sets at random from the Public Data Set does not provide meaningful insight into performance beyond that specific data set. Instead, evaluators are encouraged to train at a specific EOC point and to test at various controlled deltas from the training conditions. Such sensitivity results are more meaningfully extended to operational scenarios. See References 3 and 4 for more background on EOCs.

2.2 Measures of Performance

We are generally interested in the ability of an ATR to discriminate targets from non-targets and to distinguish the different classes of targets. The analysis tools used for this include classical Receiver Operating Characteristic (ROC) curves, confusion matrices, and probabilities of correct decisions. Average ATR discrimination scores (e.g., mean-square-error) also provide insight into confusability and performance limits. Standards from the ATR Working Group (ATRWG) are recommend and followed here where applicable, see Reference 1.

In the definitions of performance measures, we use the following terms in a formal sense.

Class - a vehicle category such as BTR60, M109, M35,..., or a higher level category such as MBT, APC, Truck,... In MSTAR both of these particular categories are of interest and the former is called a "type" and the latter a "class," an MSTAR "class" will generally include several MSTAR "types." In this paper we are only concerned with the classes: BMP2, T72, BTR70, and Other.

Target - any vehicle that is a member of a class that the ATR is intended to detect and classify, for the experiment below the "targets" are the BMP2, BTR70, and T72, so in that case the M109 is not a "target."

Confuser - a vehicle that is not a target, e.g., the M109 in the experiment below.

Classify - to assign a class-label to an image.

Detection - the declaration of any target class-label.

Although there are many other EOCs, we are primarily concerned here with serial-number variants, version variants, configuration, damage, and SAR depression angle. These EOCs are defined as:

Serial Number Variant - vehicles that are of the same class and version, but are different serial numbers.

Version Variant - differences from the manufacturer, targets of the same class but were built to different blueprints,

Configuration - adding or removing something by design, and

Damage - something added, removed, or moved when it was designed to be.

"Controlled" variants are those noted in the data headers and "Uncontrolled" variants are those where vehicles differ from the nominal condition, but are not specifically truthed. Reference 5 lists variations in the targets of interest here. Note that the word "variant" does not indicate any specific kind of variations, it should always be used in conjunction with an EOC. In MSTAR it has been useful to distinguish "major" and "minor" variants. Major variants are defined as those significantly affecting a baseline ATR system's performance. Minor variants are defined as differences noticeable in the photographs or other truth data that do not significantly affect the baseline ATR. Of course, the major-minor variant distinction is a function of the baseline ATR and a subjective judgment about whether an effect is significant, but it has still proven to be useful.

Specific figures-of-merit include:

Probability of Detection (Pd): number of targets detected / number of targets tested.

Probability of False Alarm (Pfa): the probability that a specific set of confuser images will be detected as targets, number of confusers detected / number of confusers tested.

False Alarm Rate (FAR): the number of false target declarations per square kilometer of test imagery. FAR applies to systems that search an area, while Pfa applies to systems that work on a given region-of-interest (ROI). In MSTAR the focus-of-attention (FOA) module and the system as a whole (including the FOA module) have a FAR. The down-stream modules that work on ROIs from FOA have a Pfa.

Probability of Correct Class (Pcc): there are three kinds of Pcc,

Unconditional Pcc: number of targets correctly classified / number of targets tested.

Conditional Pcc: number of targets correctly classified / number of targets detected (i.e., conditioned on detection).

Conditional and Unconditional Pcc depend on the ROC operating conditions, so it is necessary to specify the Pd associated with these numbers. A Pd of 0.9 is typically used in MSTAR and recommended as a standard operating point.

Forced-decision Pcc: number of targets correctly classified / number of targets tested when the ATR is forced to classify all ROIs, i.e., the Pcc when the operating point is for a Pd of 1.0.

ROC Curves: There are two kinds of ROC curves of interest:

Pd vs FAR: for systems where FAR is meaningful. FAR is typically measured on "clutter" imagery (images where there are no targets). A CD of clutter imagery is included in the Public Release set.

Pd vs Pfa: for systems where Pfa is meaningful. Pfa is typically measured using non-target vehicles ("confusers") from the target scenes. False alarms from clutter scenes using the front-end of an ATR system can more appropriately serve as Pfa test images for the back-end of the system. In the experiment below, non-target vehicles are used. In a sense, these confusers are especially hard cases of what would actually leak into a classifier, giving an upper-bound on Pfa.

Confusion Matrices show the Pcc for each test vehicle against each trained class. There is a row in the matrix for each test vehicle and a column for each decision category. The decision categories should include a non-target decision (a.k.a. a rejection).

Cost Measures: there are three kinds of cost that MSTAR tracks, all three are recommended when reporting an ATR's performance (see Reference 4 for details).

Data Processing Costs: the time and memory required to run the ATR (e.g., CPU seconds and maximum RAM usage).

Data Storage Costs: the size of the ATR algorithm (e.g., in number of templates stored or disk space required).

Data Collection Costs: the number of measured (or synthetic) images required to train and tune the ATR.

Confidence Intervals: standard procedures are recommended for characterizing uncertainty in numerical results (Reference 1). A confidence level of 0.95 (Zc = 1.96) was used for all confidence intervals reported here and are recommended as a standard. The confidence interval width for probability estimate $\hat{p}$ with n samples was computed as $\pm Zc\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, when valid. This expression is generally valid for $n > 30$, $np > 5$, and $n(p-1) > 5$. If any of these conditions are not met, we use the Law of Large Numbers' confidence interval of $\pm\frac{1}{\delta}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ for $\delta = 1 - ConfidenceLevel$, or 0.05 in our case. This confidence interval is approximately $\frac{\sqrt{c}}{\delta n}$ for $\hat{p} = {}^{c}\!/_{n}$ and $\hat{p}$ very small. If $\hat{p} \approx 1$ then the confidence interval is approximated for $1-\hat{p}$. For ATR scores with a sample variance of $\hat{S}$, the confidence interval for the estimate of score means is computed by $\pm Zc \hat{S}/\sqrt{n}$. When there are a large number of samples involved in the estimates, the computed confidence intervals may be tighter than is actually meaningful (i.e., variations not accounted for in the statistical model may be significant compared to those included in the model). Experience in MSTAR suggests that 100 samples begin to provide meaningful results and that more than 300 samples have little payoff. Of course, the above statistics are based on an assumption that there has been a random draw from a fixed population. So, for example, the confidence intervals tell you whether an estimated Pcc is close to the Pcc of the entire test set. It does not tell us anything about how the Pcc might vary with other test sets. The authors know of no test sets that are sufficiently comprehensive that ATR performance measures from them should be taken as an estimate of how the ATR would perform in a fully operational setting.

2.3 EOC Testing

To the extent that evaluators are interested in training their system at one EOC state and testing at a different EOC state (which we strongly encourage) results will be more comparable if the same nominal conditions are used for training. Rather than trying to anticipate specific experiments that will be conducted with the Public Data Set and specifying specific training images, we propose a set of nominal EOC conditions. Although this will not always be the case, there are many instances where an evaluator must arbitrarily select a nominal EOC state, especially in such cases we encourage evaluators to use the nominal set defined here. Since various researchers will have different interests in a given experimental result, having this commonly accepted EOC nominal will allow an expanding basis of comparable results at nominal and off-nominal conditions.

There are two vehicle classes (the T72 and BMP2) that have three different serial numbered vehicles each in the Public Data Set. The BMP2 serial numbers are 9563, 9566, and C21. All three BMP2s are of the same version variant. C21 is recommended as the nominal BMP2 serial number. Reference 5 lists the BMP2s' configuration, articulation, and damage states. Note the C21 is the more common configuration and has the least damage.

The T72 serial numbers are 132, 812, and S7. There are two versions of the T72, one represented by S7 and the other represented by 132 and 812. 132 is recommended as the nominal T72 serial number. Reference 5 lists the T72s' version, configuration, articulation, and damage states. Note the 132 is without fuel drums (as is S7) and is the same version as 812.

There are two depression angles available in the Public Data Set, 15 degrees (target and clutter scenes) and 17 degrees (target scenes only). The 17 degree data is recommended as the nominal depression angle.

There is an 360 degree coverage in aspect for each target in the Public Data Set. The left aspects (0 - 180 degrees) are recommended as nominal aspect angles. If a sub-set of aspect angles are used within a continuous range, we recommend using the aspect angles closest to the whole numbers that are multiples of the separation (e.g., for a 2 degree separation, recommend using the images closest to 0, 2, 4, 6, ... degrees). Further we recommend the separations be whole numbers with 5 degree separation being of special interest.

The performance measures and test conditions recommended above are oriented towards full ATC/ATR systems - that detect or classify targets. The Public Data Set is also well suited for use in a variety of pre/sub-ATR processing tests. If a test concentrates on a single image (or small set of images) then we recommend that image Hb03787.015 from the T72 17 degree data for SN 132 and image Hb06158 from the first clutter disk be included. While various tests may require additional images, evaluators are encouraged to include results on these images to promote the comparability of results.

3. AN EXPERIMENT

3.1 Introduction

The objective of this experiment is to illustrate the use of proposed standardized evaluation measures, suggest report formats and provide ATR developers some simple baseline results for comparison.

3.2 Experiment Description

Figure 1. shows the distribution of targets considered in the baseline experiment. All training occurred at a 17^0 depression angle while testing was performed at a 15^0 depression angle. Training was limited to the three targets shown on the left while testing was conducted over all targets, including the confuser vehicles. Confuser vehicles considered included the M-109 and M-110. All data considered in this paper consists of x-band, one foot SAR imagery data collected in support of the Moving and Stationary Target Acquisition and Recognition (MSTAR) program.

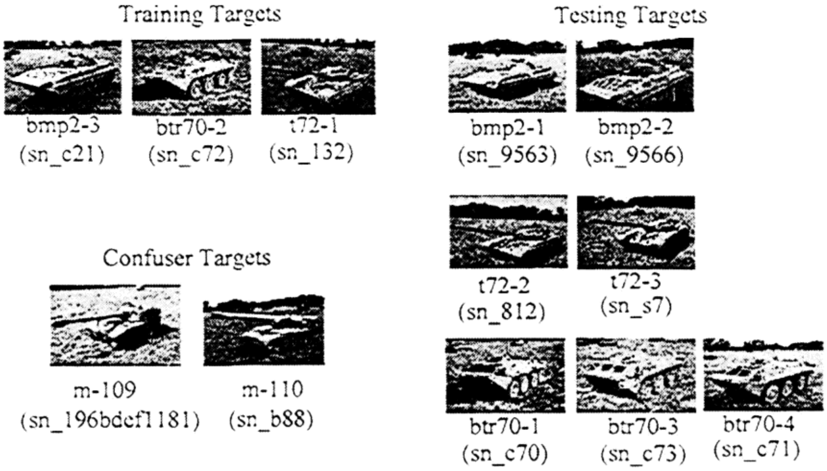

Figure 1. Baseline Experiment Target Types

3.3 Algorithm Description

3.3.1 Template Formation Process

The target classification algorithms considered in this paper are template based. The training/template formation process consists of registering and estimating the mean target signature over small aspect windows. The baseline templates were formed over 10^o aspect windows using the following registration process:

$$(x_{opt}, y_{opt}) = \underset{x_s\, y_s}{\operatorname{argmin}}\left(\sum_{x=1}^{N}\sum_{y=1}^{N} (|M_k(x,y) - S(x - x_s, y - y_s))|\right)$$

where;

(x_s,y_s) = Translation Variable
$w(x,y)$ = Binary Mask
$S(x,y)$ = Chip Magnitude

and

$$M_k(x,y) = [(k-1)*M_{k-1}(k-1)/k+k*S(x-x_{opt},y-y_{opt})]$$

$$(k = 1 \ldots \text{Number Chips/Template})$$

3.3.2 Match Metric

As in the case for template formation, a registration process also occurs when matching an observation to a template. The observation is optimally registered across all stored templates using the following error criteria:

$$(\mathrm{x}_{\mathrm{opt}}, \mathrm{y}_{\mathrm{opt}}) = \underset{\mathrm{x}_s\,\mathrm{y}_s}{\operatorname{argmin}}\Big(\sum_{x=1}^{N}\sum_{y=1}^{N} |\mathrm{w}(\mathrm{x},\mathrm{y})(\mathrm{M}_k(\mathrm{x},\mathrm{y}) - \mathrm{S}(\mathrm{x}-\mathrm{x}_s, \mathrm{y}-\mathrm{y}_s))|\Big)$$

where;

$(\mathrm{x_s},\mathrm{y_s})$ = Translation variable
w(x,y) = Binary mask defined by thresholding the template
S(x,y) = Chip magnitude

Following, registration of the observation to all templates, the best match is selected using the following criteria:

$$(\mathrm{C},\Theta) = \underset{i,j}{\operatorname{argmin}}\Big(\sum_{x=1}^{N}\sum_{y=1}^{N}|\mathrm{w}(\mathrm{x},\mathrm{y})(\overline{\mathrm{M}}_{\mathrm{ij}}(x,y) - \overline{S}(x_{opt},y_{opt}))\Big) \quad (\mathrm{i}=1\ldots\mathrm{L}, \mathrm{j}=1\ldots\mathrm{M})$$

where;

C,Θ = Class and Pose that minimizes error measure

$$\overline{\mathrm{M}}_{\mathrm{ij}}(x,y) = \frac{\mathrm{w}(\mathrm{x},\mathrm{y})\mathrm{M}_{\mathrm{ij}}(x,y)}{\sum_{x}\sum_{y}\mathrm{w}(\mathrm{x},\mathrm{y})\mathrm{M}_{\mathrm{ij}}(x,y)}$$

= Power Normalized Template for ith pose and jth class

$$\overline{S}(x_{opt},y_{opt}) = \frac{\mathrm{w}(\mathrm{x},\mathrm{y})\mathrm{S}(x_{opt},y_{opt})}{\sum_{x}\sum_{y}\mathrm{w}(\mathrm{x},\mathrm{y})\mathrm{S}(x_{opt},y_{opt})}$$

L = Number of templates per class (36)
M = Number of classes (3)
w(x,y) = Binary mask

3.4 Results

Figure 2. shows the results using the above formulation when considering the proposed standardized reporting metrics. The upper left plot shows the probability of detection ($\mathrm{P_D}$) as a function of the probability of false alarm ($\mathrm{P_{FA}}$), where $\mathrm{P_{FA}}$ is the probability that a confuser vehicle is less than the detection threshold. The upper right and lower plots in the figure show the unconditional probability of correct classification ($\mathrm{P_{CC}}$) v.s. ($\mathrm{P_{FA}}$), where $\mathrm{P_{CC}}$ is $\mathrm{P_D}$*$\mathrm{P_{CC/D}}$. As expected, in all cases $\mathrm{P_{CC}}$ is considerably higher for the training vehicle, clearly illustrating the importance of testing across serial number variants.

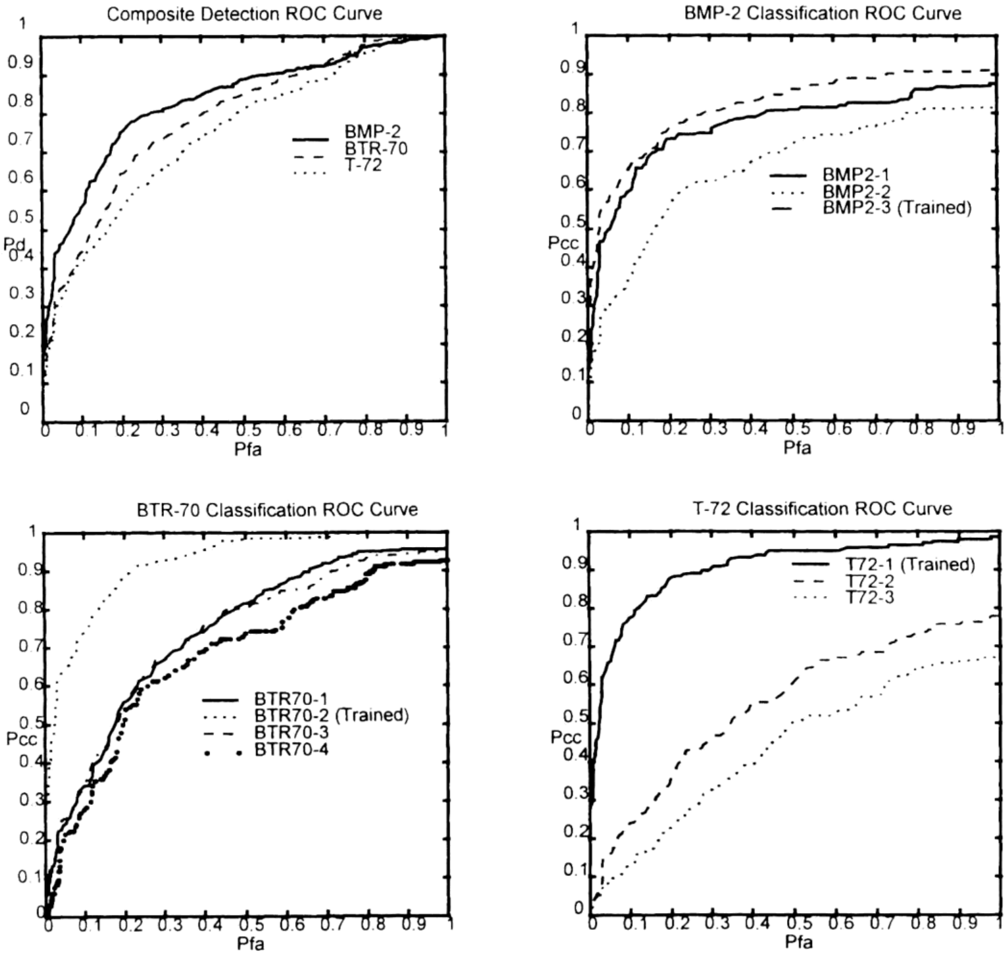

Figure 2. Standardized Reporting Formats and Metrics

Tables 1. and 2. show the classification and in-class rejection rates for each vehicle. The * marks the trained vehicle. Results in both number of samples and fractional scores are presented. All reported results are based on a threshold corresponding to $P_D = 0.9$. The confidence intervals are computed as suggested above with a confidence level of 0.95.

Table 1. Sample Classification and Rejection Rates ($P_D = 0.9$)

	BMP2	BTR70	T72	Rejections
BMP2-1	**161**	18	0	16
BMP2-2	**150**	31	0	15
BMP2-3*	**175**	8	0	13
BTR70-1	2	**252**	2	18
BTR70-2*	0	**270**	0	3
BTR70-3	5	**243**	1	25
BTR70-4	4	**165**	0	27
T72-1*	1	2	**188**	5
T72-2	13	35	**112**	35
T72-3	8	28	**131**	24

Table 2. Fractional Classification and Rejection Rates (P_D = 0.9)

	BMP2	BTR70	T72	Rejections	Confidence
BMP2-1	**0.8256**	0.0923	0	0.0821	+/- 0.020
BMP2-2	**0.7653**	0.1582	0	0.0765	+/- 0.020
BMP2-3*	**0.8929**	0.0408	0	0.0663	+/- 0.020
BTR70-1	0.0073	**0.9197**	0.0073	0.0657	+/- 0.017
BTR70-2*	0	**0.9890**	0	0.0110	+/- 0.017
BTR70-3	0.0182	**0.8869**	0.0036	0.0912	+/- 0.017
BTR70-4	0.0204	**0.8418**	0	0.1378	+/- 0.020
T72-1*	0.0051	0.0102	**0.9592**	0.0255	+/- 0.020
T72-2	0.0667	0.1795	**0.5744**	0.1795	+/- 0.020
T72-3	0.0419	0.1466	**0.6859**	0.1257	+/- 0.020

Based on these results and previous studies have there may be several major EOC variants at work here. The BTR70-3 and T72-3 are both major version variants. T72-2 is a major configuration variant (having two large fuel drums attached to the back of the vehicle). BTR70-4 is a major configuration variant (possibly because of the three gas cans stored externally and not having the triangular shelves that the training vehicle has). BMP2-2 has configuration, articulation and damage variants (especially related to its spotlights) that may account for its poor performance. The trained BTR70 is from the November '96 collection while the test BTR70s are from the September '95 collection, therefore the background is slightly different. That may explain part of the difference between trained and untrained performance there. The EOCs covered here include a modest amount of depression angle, where the training and test images differ by about 2 degrees. The high recognition rates for the trained vehicles suggest that this depression angle difference is not a major variant.

4. SUMMARY

The recent public release of high resolution Synthetic Aperture Radar (SAR) data collected by the DARPA/AFRL Moving and Stationary Target Acquisition and Recognition (MSTAR) program has provided a unique opportunity to promote and assess progress in SAR ATR algorithm development. This paper suggests general principles to follow and reports on a specific ATR performance experiment which used these principles and this data. A key principle is that testing should be done with conscious isolation of Extended Operating Conditions (EOCs). A key result of the experiment is that the effects of various EOCs are noticeable, but even with the MSTAR data it can be difficult to isolate the effects of specific EOCs.

5. REFERENCES

1. ATRWG, "Target Recognizer Definitions and Performance Measures," ATRWG No. 86-001, February 1986.
2. ATRWG, "Applications of Confidence Intervals to ATR Performance Evaluation," ATRWG No. 88-006, October 1988.
3. Keydel, Eric R., Shung Wu Lee, John T. Moore, "MSTAR Extended Operating Conditions, A Tutorial," SPIE Vol. 2757, pp.228-242, March 1996.
4. Ross, Timothy D., Lori A. Westerkamp, Edward G. Zelino and Thomas J. Burns, "Extensibility and Other Model-Based ATR Evaluation Concepts", pp.213-222, SPIE'97 Algorithms for Synthetic Aperture Radar Imagery IV, April 1997.
5. Ross, Timothy D., Jeff Bradley, Micheal O'Connor, "MSTAR Data Handbook for Experiment Planning", AFRL/AAC with Sverdrup Technology Inc., October 1997.