Assessing the Performance of an Automated Video Ground Truthing Application

Scott K. Ralph[a], John Irvine[b], Mark R. Stevens[a] , Magnús Snorrason[a], and David Gwilt[c]
[a]Charles River Analytics, Cambridge, MA
[b]SAIC, Burlington, MA
[c]AFRL/SNAA Wright Patterson, AFB, Dayton, OH
email: {sralph|mrs|mss}@cra.com, John.M.Irvine@saic.com, David.Gwilt@wpafb.af.mil

Abstract

Present methods of quantifying the performance of ATR algorithms involves the use of large video datasets that must be truthed by hand, frame-by-frame, requiring vast amounts of time. To reduce this cost, we have developed an application that significantly reduces the cost by only requiring the operator to grade a relatively sparse number of data "keyframes". A correlation-based template matching algorithm computes the best position, orientation and scale when interpolating between keyframes.

We demonstrate the performance of the automated truthing application, and compare the results to those of a series of human operator test subjects. The START-generated truth is shown to be very close to the mean truth data given by the human operators. Additionally the savings in labor producing the results is also demonstrated.

1. INTRODUCTION

Automated Target Recognition (ATR) technology offers the potential to process and exploit large volumes of sensor data and provide meaningful information to military users. Successful development and transition of ATR technology depends on many factors, including rigorous test and evaluation across a range of operating conditions [1-3].

A significant part of the evaluation effort is the process of developing accurate image truth and comparing the image truth to the ATR decisions to assess the correctness of these decisions. In this paper, we present a complete truthing system we call the Scoring, Truthing, And Registration Toolkit (START).

The START image truthing tools consists of a series of image and track editors that allow a user to construct a truth sequence given an input AVI/MPEG video stream. There are several key components to the truthing tool. The first is the ability to define a key-frame. A key-frame is defined as a frame taken from the larger video sequence for which the user will specify truth information. The user is responsible for selecting where in the video sequence the key-frames are defined, and they can add and delete them throughout the truthing process. Selection of the given key-frames should take into consideration:

- The addition or deletion of a target from the field of view of the frame
- Presence of fast-moving objects in which the interpolation becomes ill conditioned
- Occlusion or close proximity of two targets making two targets difficult to distinguish
- The performance of the feature matching (if any) when computing the interpolation

Once the key-frames are defined, the targets in each frame are hand segmented and labeled. We currently use color-coding to distinguish between targets in the frame and use bounding boxes to delineate the target extent. We are adding the ability to produce higher resolution truthing masks using convex polygons. The system architecture for START is depicted in Figure 2. For the discussion, we offer the following definitions of truthing and scoring [3]:

Truthing: The process of developing or constructing information about the location and characterization of targets in the imagery or sensor data. The product of this effort is the "image truth." Image truth is not synonymous with "ground truth." Ground truth locates the targets with respect to a standard datum (e.g., latitude and longitude, UTMs, etc.), whereas image truth relates the target information to the pixels within the image.

Scoring: The process of comparing the algorithm's (or analyst's) decisions to image truth to assess the correctness of each decision. Note that "decisions" include the explicit algorithm output, as well as implicit output. Explicit output includes the declaration of a target, which could be either a correct detection or and false alarm. An implicit decision is the failure to declare a target when one is present, resulting in a missed detection.

The goal of the truthing portion of start is to allow the user to grade a relatively sparces set of frames, called *key-frames*, from which subsequent target locations can be computed automatically. For example, consider the case of grading a sequence of tank locations as depicted in Figure 1. Traditionally each of the four frames, (a)-(d) would have to be truthed manually. Using the automatic truthing application, the user can specify the target location of (a), and automatically compute the target positions at each of (b), (c), and (d). Once computed the truth information is then displayed to the operator for approval or modification, or instances where the tracking algorithm is unable to determine the location, the frames are flagged for further operator intervention. This shifts the operator labor from a data-entry role, to more of a quality-assurance (QA) monitoring role.

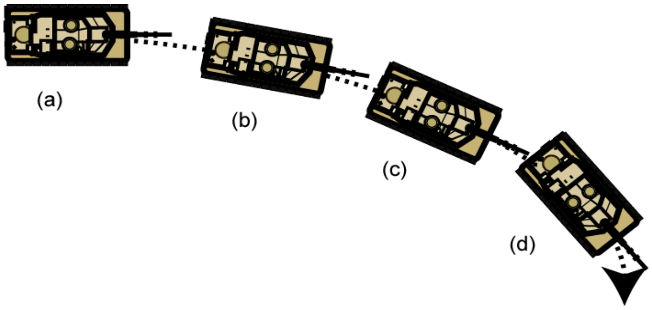


Figure 1 : Example of a tank track

The rest of the paper describes the START Grading Application, followed by a description of the set of scoring metrics provided by START. Next, we describe an experiment with 8 human operators, and characterize the performance of the automatic tracking algorithm to the test subjects.

2. START GRADING APPLICATION

The architecture of the START application, depicted in Figure 2, consists of two components: the truthing component, and the scoring component. The truthing component consists of the operator-GUI as well as a correlation-based tracking algorithm that computes the position, orientation, and scale of a target in new frames from the image data specified in keyframes. The scoring component provides a generic interface to which a candidate ATR algorithm provides data to be scored. A set of performance measures is provided for the given input video.

The START GUI, depicted in Figure 3, allows the user to browse the video data and view the target locations as entered by the operator, or computed by the application. The target locations consist of oriented-rectangles in which a reference edge (pointed to by an arrow) disambiguates the orientation. For better accuracy the user is able to zoom-in and pan to the area of interest and place the target to arbitrary accuracy in the image.

3. START AUTO-TRACKING

The START application uses a correlation-based approach to track the graded target from a given keyframe to subsequent video frames. The correlation search has four total degrees of freedom, two positional, one rotational, and one uniform scale. The user is able to select the range of possible rotations, translations and scales to be considered (see Figure 4). The scale and rotation are sampled at discrete values. To improve efficiency the search employs a multi-resolution approach in which an image pyramid is constructed. At each level of the pyramid a coarser representation is constructed by approximating 4 pixels in a 2x2 neighbourhood of the preceding level by a single pixel.

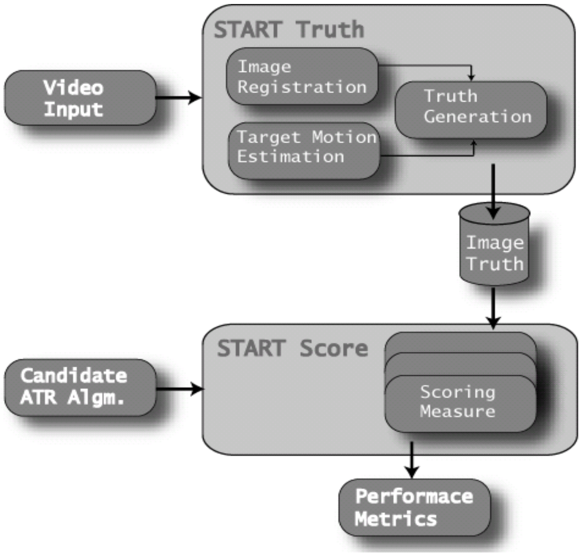
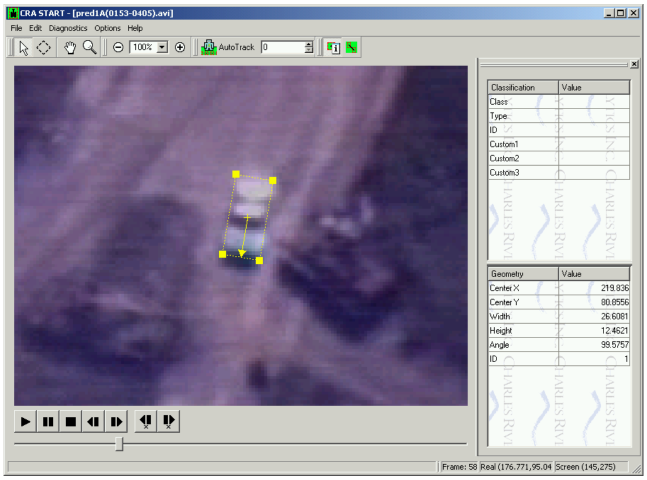


Figure 2: START system diagram

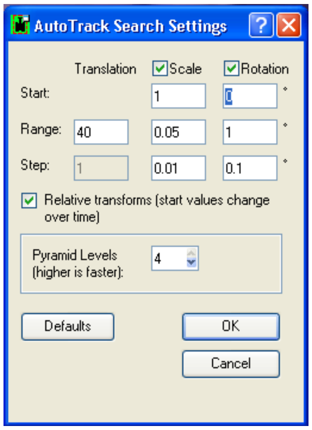


Classification	Value
Class	
Type	
ID	
Custom1	
Custom2	
Custom3	

Geometry	Value
Center X	219.836
Center Y	80.8556
Width	26.6081
Height	12.4621
Angle	99.5757
ID	1

Frame: 58 Real (176.771,95.04 Screen (145,275)

Figure 3 : The START truthing application



	Translation	☑ Scale	☑ Rotation
Start:		1	0 °
Range:	40	0.05	1 °
Step:	1	0.01	0.1 °

☑ Relative transforms (start values change over time)

Pyramid Levels (higher is faster): 4

Figure 4 : START Auto-Track Search parameters

4. START SCORING METRICS

At present START supports four scoring metrics (depicted in Figure 5):

Point Scoring: This is the simplest approach, representing the truth and ATR target locations as a point at the respective center of mass. The Euclidean distance between the center points must be smaller than some small threshold in order to be classified detection.

Centroid Scoring: This approach combines point and area-based properties of the target points, and requires that the truth target region has been segmented. If the centroid of the ATR detection falls within the truth target region, then it is classified detection.

Touch Scoring: This approach requires that both the image truth and the ATR decision define regions in the image. Detection occurs if the intersection of the truth target region and the ATR detection region is non-empty.

Area Scoring: Like touch scoring, this approach requires segmentation of both the truth and ATR decision regions. The approach is to assess the relative area of the intersection of truth and detection normalized by the union of the truth and detection regions. If the relative area exceeds a threshold the target is classified as a detection.

Chamfer Distance: This method is often applied to specific features, such as edges of a segmented object. It can, however, be computed for pixels based on the truth target and the ATR decision. The chamfer distance is defined by the average minimum distance between truth pixels and detection pixels.

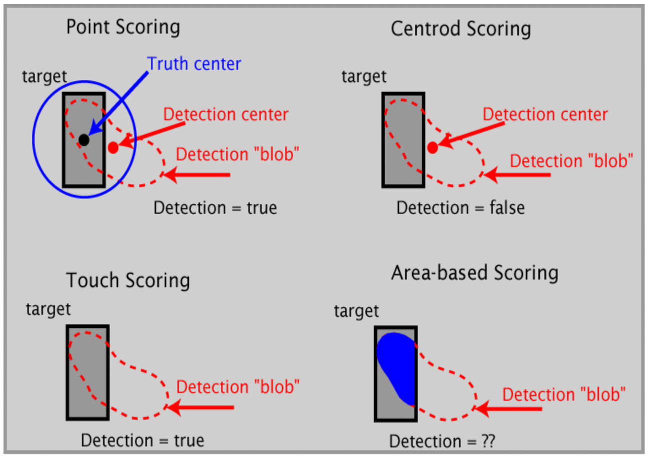


Figure 5: Scoring metrics computed by START

5. Experimental Results

To test the effectiveness of the Auto-tracking algorithm in grading an image we used a short portion (250 frames) of a Predator video (640x480, 8x3-bit color). The video, shown in Figure 3, contains a single truck traveling down and to the left in the image frame, passing an intersection that has a stationary tank. The orientation of the truck varies over a range of approximately 15 degrees.

In total eight human operator test subjects were involved in grading the image sequence. Each subject was given a short demonstration of the GUI controls and allowed to practice for 15 minutes on some practice video sequence. Since the “best” bounding-box is a subjective measure, the test subjects were told to strive for maximum consistency across frames rather than asking them to emulate a particular style of grading. The grading task consisted of positioning an oriented rectangle around the target at 5-frame intervals, for a total of 50 truthed frames. For comparison, an operator defined an initial bounding box for the first frame, and the START Auto-track algorithm was used to truth the remaining 249 frames.

5.1. START Accuracy

Figure 6 shows the START computed frames (given in solid red lines) and the set of test subject oriented bounding boxes (given as blue dotted-lines), for six representative frames along the trajectory. From this graph, it is apparent that the START generated ground-truth consistently lies in the envelope of human-graded frames.

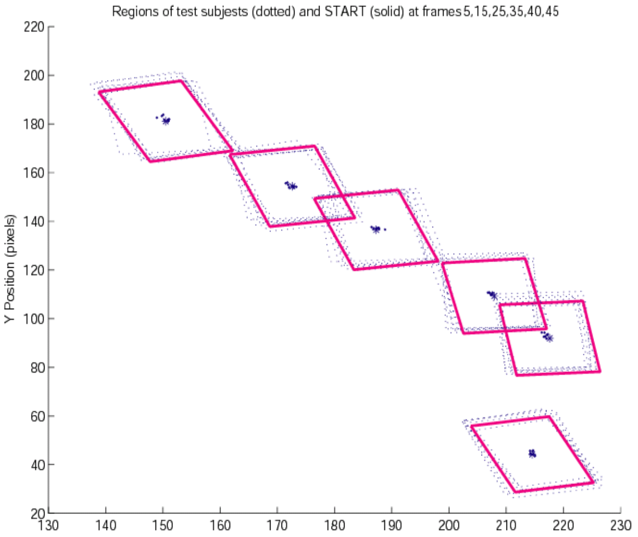


Figure 6 : Bounding boxes of START and test subjects

To further characterize the performance of the test subjects, we computed the variance of the position of the center point over the set of eight subjects at each of the frames. The results are shown in Figure 7. The standard deviation of the *x*- and *y*-components vary from (0.16-0.73) and (0.35-1.3) pixels respectively, with a mean standard deviation of 0.43 and 0.705 pixels. This shows that the human operators are very good at localizing the center of the target bounding boxes.

To compute the variance in the bounding-box center positions produced by START we first computed the mean test subject bounding box center for each frame and computed the differences, as shown in Figure 8. From this plot it is apparent that a large proportion of the START center-point positions lie within one standard deviation from the mean test subject position.

One apparently troubling feature of the positional differences is that neither the *x*- nor *y*-components appear to be zero mean, but rather show a consistent bias of (+0.18, 0.71) pixels as compared to the mean test subject center point. While the bias in the *x*- dimension is quite small, the relative large bias of 0.71 pixels in the *y*-axis can largely be explained by the initial position of the bounding-box at the initial frame, which is 0.53 pixels lower in the frame than the mean test subject's position. While this does vindicate the performance of the START tracking algorithm, it does demonstrate the necessity of accurately defining the bounding boxes at the keyframes. Errors made when defining the keyframe are propagated forward at each step of the tracking algorithm.

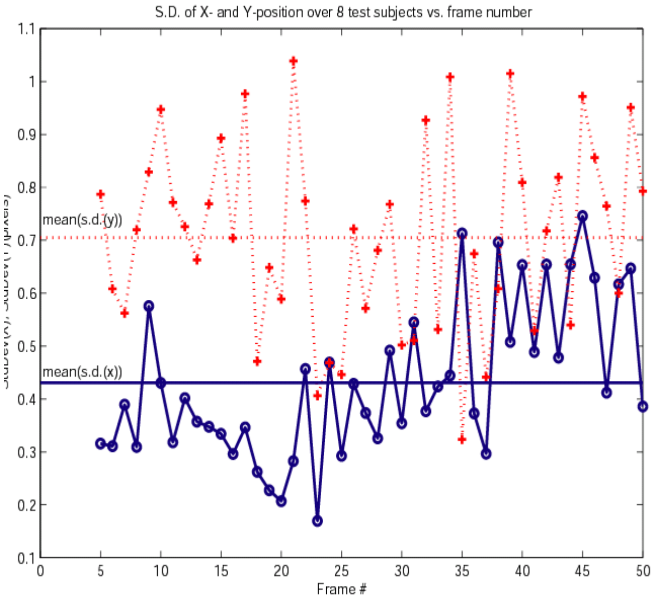


Figure 7 : Positional variance among the 8 test subjects

Assessing the ability of START to correctly determine the orientation of the bounding region is given in Figure 9 which shows that START is consistently able to resolve the apparent orientation of the target to within 10 degrees, a level of performance that is very close to the set of test subjects. This is corroborated by looking at the ensemble of bounding boxes in Figure 6.

The variation in the bounding box width and heights are shown in Figure 10, which shows that there up to a 5 pixel variation in the widths and 8 pixel variation in the heights. While this variation is large, it is to be expected given the lack of common guidelines in the grading process. It is clear that the bounding boxes produced by START lie in the envelope of the test subjects.

A last measure of the performance of the START application is to compute the area overlap of the START regions with each of the test subject regions. The area scores for each of the 8 test subjects is given in Figure 11, and summarized in Table 1. We see that the mean normalized area overlap corresponds to approximately 94%, indicating that the regions are very similar to the test subject's.

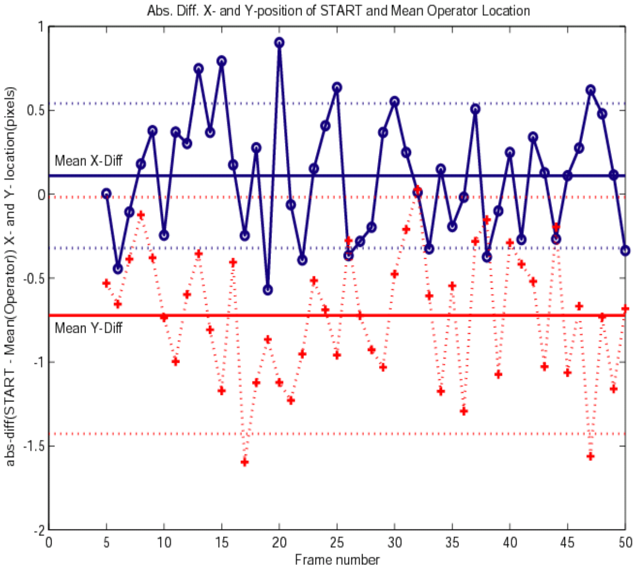


Figure 8 : START positional errors compared to mean test subject

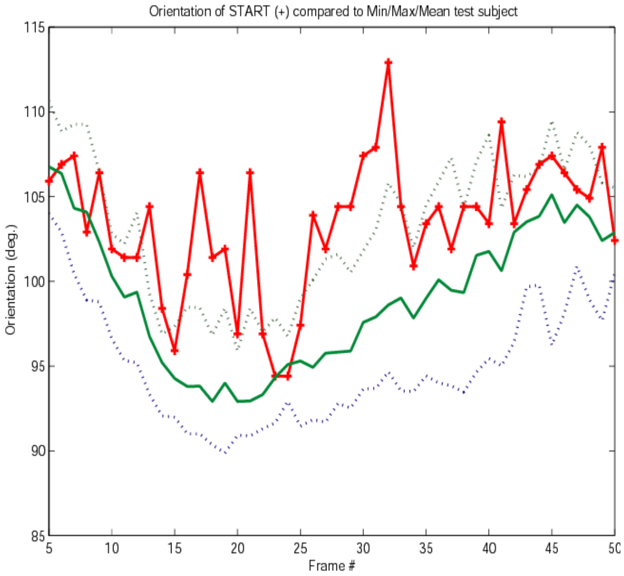


Figure 9 : START Orientation compared to operators

5.2. Speed of Automated Truthing Tool

The time each of the 8 test subjects took to truth each frame was 35±5 seconds per frame, for a total of 25—33 minutes per subject for the entire video. The START application took just 0.6 seconds per frame, for or ½ minute total. This is a factor of more than 50 in performance.

One issue that remains, however, that the speed with which the frames are graded is dependent on the ranges and sampling intervals of each of the degrees-of-freedom on which the search is performed. For example, testing the rotational angles [-10,10] deg. takes twice as long if the precision is ½ degree as compared to 1 degree. At present these ranges of translation, scale and rotation are set by the operator, however, work is currently underway to help automate this process so that the accuracy and efficiency are effectively traded-off.

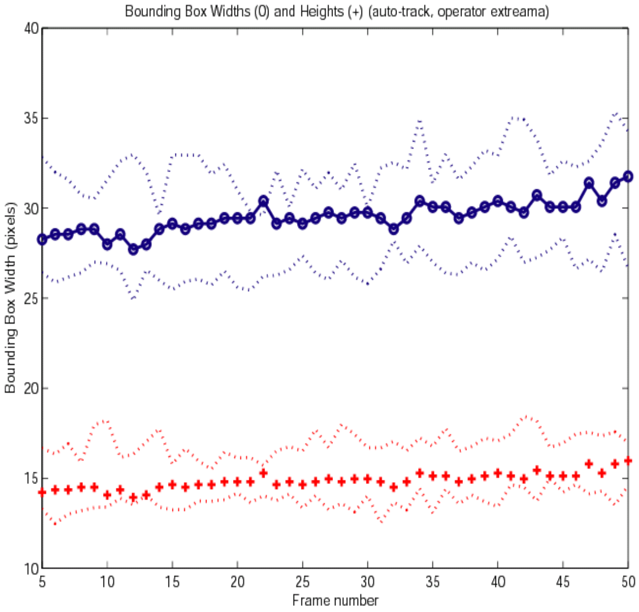


Figure 10 : Bounding box width and height variation in test subjects and START

Table 1: Summary of area scores

Mean Width	29.029
Mean Height	15.391
Mean Area	446.791
Mean Area Score	418.814
Mean Normalized Area Score	0.937

6. Future Work

We are still in the early phases of the automated-tracking portion of start, and there remains many potentially fruitful avenues to investigate in improving the tools abilities. At present only rectangular target regions are supported, however, we plan to extend this to include arbitrary polygonal-shaped target regions. This would allow much better fidelity when modelling complex target shapes such as an aircraft etc.

Currently there is no capability of the auto-tracking mechanism to use prediction to aid in the search of the target in the field of view. If the imaging system is stationary, or alternatively, if we are able to register the frames to a common reference, then we can use a Kalman or Alpha-Beta filter to predict the motion of the target. The predicted position of the target can then be used to reduce the computational complexity of the search. Additionally, when the target image becomes degraded due to obscuration or occlusion, the tracker can provide the best guess and present it to the operator for verification.

Current research has focused on improving the speed and accuracy of the template matching. This includes using gradient-based approaches to solve for the best parameters defining the affine transformation of the best match. Additionally we are investigating the use of edge-based approaches which have the potential to greatly improve the localization of the target.

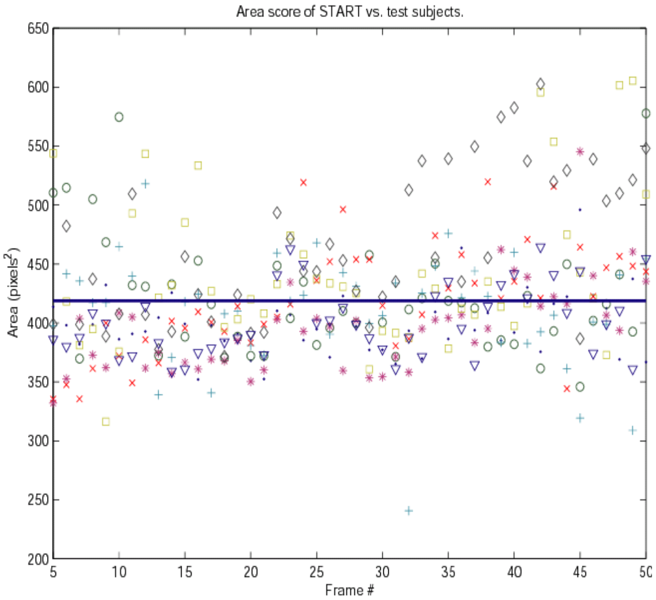


Figure 11 : Area score of START over 8 test subjects

7. Conclusions

In this paper, we presented the framework for a complete truthing system we call the Scoring, Truthing, And Registra-tion Toolkit (START). The toolkit has two components which meet two separate needs: the truthing component that automates the process of generating high-quality truth information, and the scoring component which attempts to standardize the scoring of candidate ATR algorithms according to a set of performance measures.

We have compared the performance of the automated tracking algorithm to a set of 8 test subjects, and have found that the accuracy produced by the automated tool is of a very similar quality as the mean human performance. Additionally, the speed with which the tool is able to compute the ground-truth information is approximately 50 times faster than a human.

8. Acknowledgements

The authors would like to thank Sam Kenyon and Jeremy Marvel for their assistance in the development of the

START Truthing Tool, and in their helpful input during the development process.

9. REFERENCES

[1] Irvine, J. M., "Assessing Target Search Performance: The Free Response Operator Characteristic (FROC) Model," *Optical Engineering*, 2004.

[2] Irving, J. M. Evaluation of ATR Algorithms Employing Motion Imagery. 30th Applied Imagery and Pattern Recognition Workshop, Sponsored by IEEE Computer Society. 2001.

[3] Stevens, M. R., Snorrason, M., and Jarratt, M. A Scoring, Truthing, and Registration Toolkit for ATR Evaluation. Proceedings of SPIE. Vol. 5094. 2003. Orlando, FL.