START for Evaluation of Target Detection and Tracking

Scott K. Ralph[a], Mark R. Stevens[a,] John Irvine[c], Jeremy Marvel[a] ,

Magnús Snorrason[a], and Andrew Rice[b]

[a] Charles River Analytics, Cambridge, MA

[b] AFRL/SNAA Wright Patterson AFB, Dayton, OH

[c] SAIC, Burlington, MA

ABSTRACT

A major challenge for ATR evaluation is developing an accurate image truth that can be compared to an ATR algorithm's decisions to assess performance. We have developed a semi-automated video truthing application, called START, that greatly improves the productivity of an operator truthing video sequences. The user, after previewing the video selects a set of salient frames (called "keyframes"), each corresponding to significant events in the video. These keyframes are then manually truthed. We provide a spectrum of truthing tools that generates truth for additional frames from the keyframes. These tools include: fully-automatic feature tracking, interpolation, and completely manual methods. The application uses a set of diagnostic measures to manage the user's attention, flagging portions in the video for which the computed truth needs review. This changes the role of the operator from raw data entry, to that of expert appraiser supervising the quality of the image truth.

We have implemented a number of graphical displays summarizing the video truthing at various timescales. Addition-ally, we can view the track information, showing only the lifespan information of the entities involved. A combination of these displays allows the user to manage their resources more effectively.

Two studies have been conducted that have shown the utility of START: one focusing on the accuracy of the automated truthing process, and the other focusing on usability issues of the application by a set of expert users.

Keywords: Scoring, truthing, registration, evaluation, automatic target recognition, motion imagery

1. Introduction

Computer vision applications that Involve detecting and tracking entities throughout a video sequence require large sets of image truth data to evaluate their performance. Traditionally, this has been performed solely by a human operator who, on a frame-by-frame basis, defines the position and extent of entities in the given frame (say by drawing a bounding-box around each of the objects). This is an extremely error-prone and labor-intensive process, and in many instances it is this associated cost that limits the use captured image sequences. In automatic target recognition (ATR) problems, for example, a significant percentage of the overall evaluation effort consists of developing accurate image truth, and then comparing the image truth to the ATR decisions to assess the correctness of these decisions.

Charles River Analytics has developed an application, called the *Scoring, Truthing and Registration Toolkit (START)* that automates the truthing of video data for ATR algorithm evaluation by the Air Force Research Lab [1-6]. Additionally, it has also been used in other problem domains, such as the detection and tracking of humans and cars in a parking lot scenario [7], automated sign-detection [8], and license plate detection on an autonomous robot [9]. The START user interface is depicted in Figure 1. We presently define an object

boundary by an oriented rectangle, however this could be easily extended to include general polygons and other boundary descriptions. Examples of the domains in which it has been applied is depicted in Figure 2.

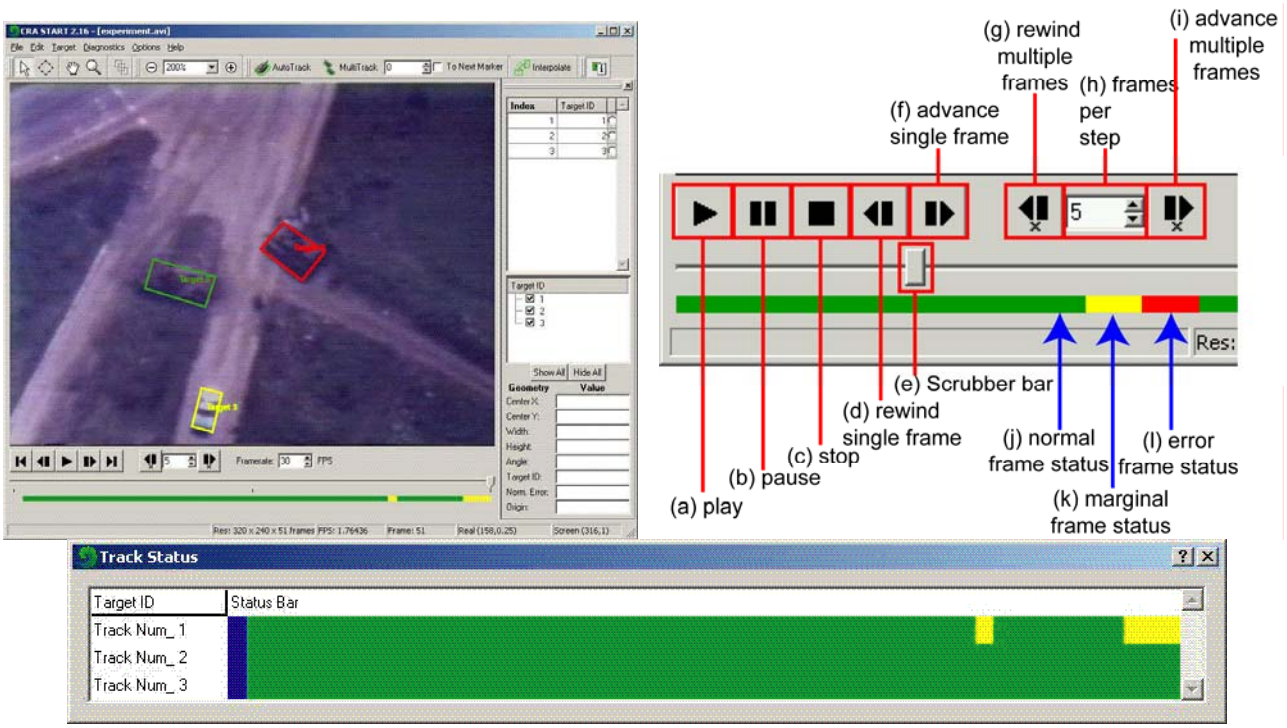


Figure 1: START Application user interface illustrating the truting of a military vechicle (top), and track status visualization (bottom)

The goal of a semi-autonomous video truthing application is to provide object boundary information for each of the entities in the video, while only requiring the user to provide the image truth for a relatively sparse number of frames in the video. Borrowing from the computer animation domain the user specifies the image truth at various frames, called keyframes, throughout the video. Features taken from the image, are tracked from one frame to the next to produce image truth for the intervening frames. This START workflow shifts the responsibility of the user from a video-truther to a more supervisory quality-assurance role. The overriding philosophy of the START truthing tool is that the most critical resource that must be managed effectively is the user's attention.

1.1. START Features

The following features of START are provided to maximize the effectiveness of the truthing:

Speed of Truthing: The user only supplies the exact location of the targets in a very sparse set of image frames.

User's role shifted to QA: The user now takes on the role of quality assurance. This involves playing back the video sequence and verifying that the boundaries of the targets correspond sufficiently closely to their assessment. This assessment can also involve the examination of the diagnostic measures computed during the tracking process.

Diagnostics: START makes use of a set of diagnostics whose role is to identify regions of the video in which the reliability of the automated tracking is likely to perform poorly. The user can then review the identified regions and ensure that the computed regions are correct by modifying the regions manually, or by changing the search parameters. These diagnostics measure 1. the quality/certainty of the target positions produced, 2. the apparent motion of the target, and 3. its placement relative to other targets to determine when the automated process could be unreliable.

Figure 2: START application appled to SUSAN, VANESSA, and SESAME

Video Preview and Annotation: In addition to the standard video playback controls (play, pause, fast-forward *etc*.), there is a scrubber-bar that allows random access to frames in the video. The user can also place bookmarks at points along the video that correspond to significant events. These events are things not determined quickly and accurately using computer vision: The introduction of new, or exiting of old targets from the field of view; the onset of significant obscuration (potentially causing tracking problems); rapid changes to the behavior of entities in the frame (such as a rapid acceleration); and others.

Even in cases where these conditions are detectable autonomously, the detection must be very fast (*i.e*. at least frame-rate) to be of use; if the user can detect these conditions more effectively then they should do so.

A color-coded status display below the scrubber bar indicates the status of frames throughout the video. Red and yellow colors indicate places in the video where the tracking algorithm(s) were not confident in their predictions, and user intervention is needed.

Track Visualization: We provide a separate track visualization, depicted in Figure 1, showing each track in the video, as well as the status of each entity for every frame of the video sequence.

Suite of Tracking Tools: We provide a number ways of computing image truth, ranging from fully autonomous methods, to completely manual methods. The applicability of each of the methods is dependent on numerous factors affecting both the apparent motion in the frame, as well as the difficulty in determining a match. These include: the motion of the object (linear, non-linear, erratic, *etc*.), the motion of the sensor (fixed, steady motion, vibration/erratic, *etc*.), level of obscuration, amount of deformation, and others. The fully-autonomous tracking (called *AutoTracking*) uses cross-correlation of a template image extracted from the keyframe. The matching considers four degrees of freedom (D.O.F.) in the matching: *x*- and *y*-position in

the frame, the angle Θ, and a uniform scale term. We provide two interpolation methods, linear and a cubic-spline, that can be used effectively when there is smooth or no camera motion, and occlusion does not allow the cross correlation to perform well. We also provide two methods of manually specifying the image truth: a streamlined method of specifying the center of the object, when scaling and rotation effects are small; and a fully manual method where the boundary is exactly specified on each frame.

2. Truthing Strategies

START provides a suite of tools that enable the operator to rapidly truth video data, and include algorithms for automatically computing target boundary information (see Figure 3), as well as user interfaces that integrate with the operator's truthing workflow.

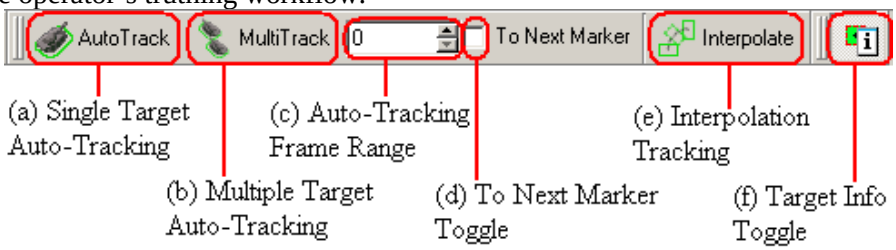


Figure 3: The controls of the variety of truthing strategies

2.1. Video Markers

The first step in the truthing process is to analyze the key events in the video, such as the entry/exit of targets in the field of view, rapid changes in the direction and speed of vehicles, significant lighting changes, and other factors. The user can review the video using the VCR-style playback controls, and by sliding the scrubber bar back-and-forth to rapidly locate locations in the video of interest. Once located, the user can annotate the video timeline by placing "marker" positions along the video timeline, as depicted in Figure 4.

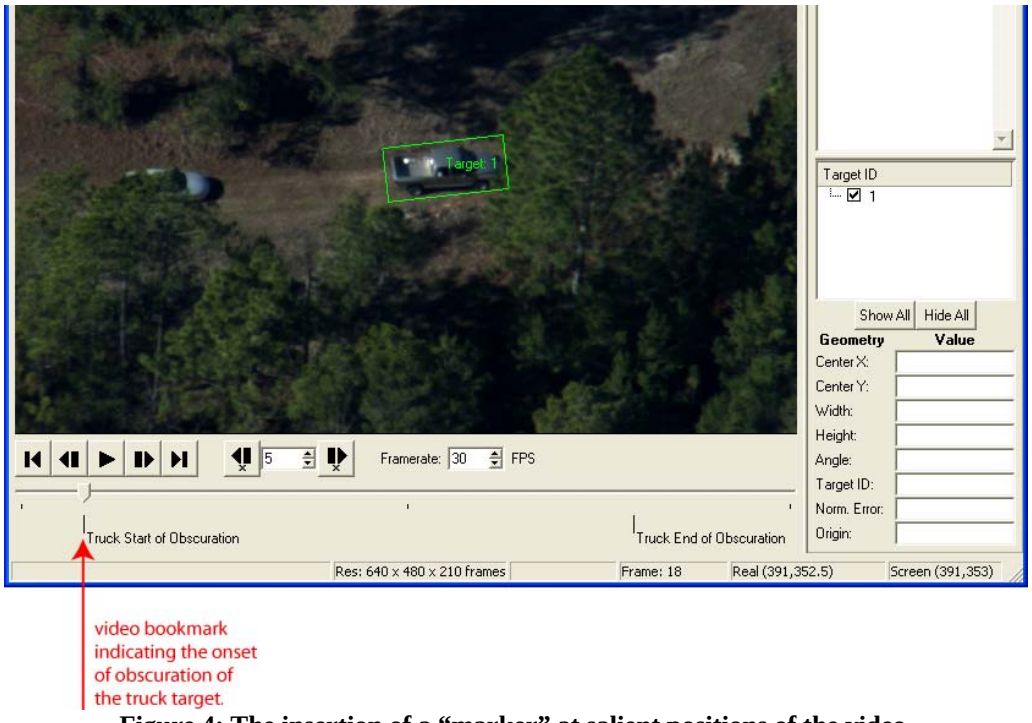


Figure 4: The insertion of a "marker" at salient positions of the video.

2.2. Autonomous Truthing via Normalized Cross Correlation Search

The fastest truthing algorithm START provides, called *AutoTrack*, uses normalized cross correlation computed from an image chip of the target, matching the chip in the keyframe image to images in the subsequent frames. This truthing strategy is applicable when the appearance of the target is relatively stable, for example, when large amount of obscuration is not present, and rapid changes from light-to-shadow illumination does not occur. The matching algorithm matches the template undergoing changes in position, orientation, and scale. The user can tailor the range of possible translations, orientations, and scales to be considered using the dialog depicted in Figure 5. The larger the range, and the finer the increment of the search parameters, the better the quality of the match, however this is at the cost of computational speed.

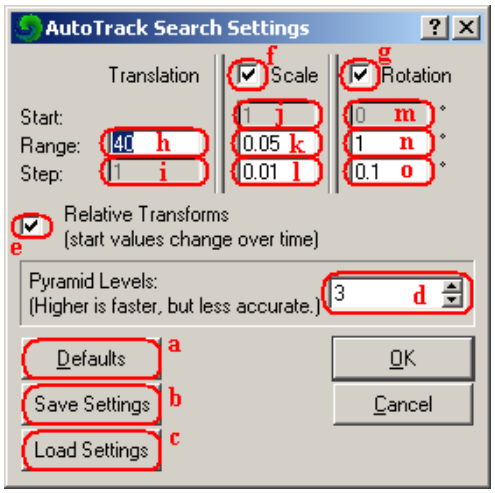


Figure 5: AutoTrack parameter specification

The speed of the matching is greatly improved by using a multi-resolution matching algorithm, first matching the target chip at a very coarse scale, and iteratively improving this by considering finer-and-finer details in the higher resolution images. This matching process is depicted in Figure 6.

START allows the automatic computation of target truth for all targets in a given image, as is applicable when all of the targets are undergoing a similar motion and viewing geometries, or only a single target in the image.

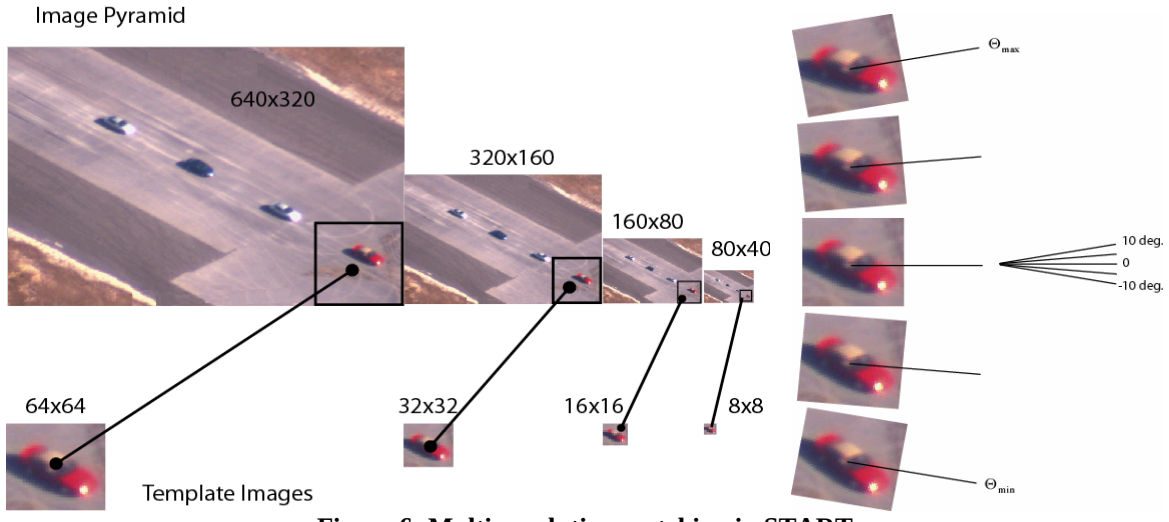


Figure 6: Multi-resolution matching in START

2.3. Truthing using Interpolation

In instances where targets undergo a large degree of occlusion, such as a tank going under a bridge (see Figure 7), the user can use the interpolation truthing strategy. In this mode, the user defines a set of keyframes along its path, both prior to, and after the occlusion (in the example, say at fames 10, 20, & 30). Using linear, or cubic interpolation, the position, orientation, and size of the target is computed for all subsequent frames.

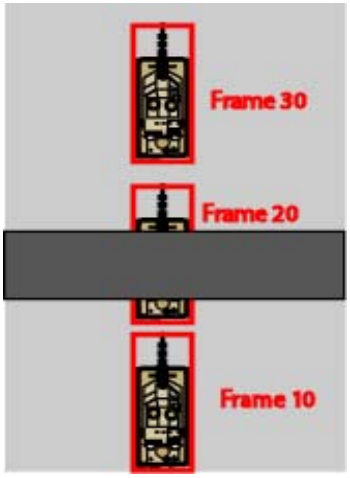


Figure 7: Using the interpolation truthing strategy in the presence of occlusion

2.4. Manual Truthing Strategies

In cases where neither tracking-based truthing strategies can be used, and interpolation gives insufficient accuracy (e.g. the sensor is undergoing turbulent motion), we offer a number of manual methods of defining the image truth:

- **Copy-and-paste:** In this strategy the user can copy the truth of a given target (alternatively all targets), advance to the next frame of interest, and paste the target (alternatively all targets) into the new frame. The position, size, and orientation of each target can then be tweaked individually, directly with the mouse, or using the keyboard (described below).
- **Single Click:** Using this option, the user can copy an instance of a truthed target, and then using a single click, specify the position of the center of the target in the subsequent frame. The frame is automatically advanced to the next frame.
- **Keyboard Nudging:** This method of specifying image truth allows the user to rapid cycle through the set of targets and modifying the orientation and position by small fixed increments. This mode is particularly useful when the user is reviewing a segment of the truth data.

3. Diagnostic Measures and Target Track Lifetime Display

The overriding design principle of START was to efficiently manage the user's attention and increase the throughput of the video truth operator. Two key properties of the current state of the video truthing process are displayed to the user: a confidence level of the computed image truth as estimated by the internal tracker, and a temporal display of the target tracks with associated status.

The confidence level of a computed truth region is computed using a set of user-configurable thresholds of the following:

1. The normalized cross-correlation score value of the best match position, orientation, and scale
2. The percentage of the target region hypothesized to be occluded. A pixel is hypothesized to be in occlusion when the gray values differ by more than a specified threshold (see Figure 9).
3. The distance between the search regions of the target and its nearest neighbor.

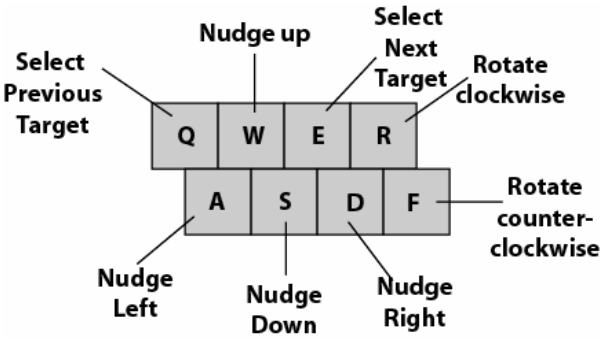


Figure 8: Keyboard commands for modifying truth regions in a frame

For ease of interpretation, the confidence score is displayed as a color code: green representing a high confidence score, yellow indicating a marginal confidence score, and red indicating a low confidence score. The lowest scoring target is displayed directly above the scrubber bar (see Figure 1, top). The user can then click on yellow or red regions to further investigate which targets in the frame need further attention. The status of each target region is also displayed in the track-status display (see Figure 1, bottom).

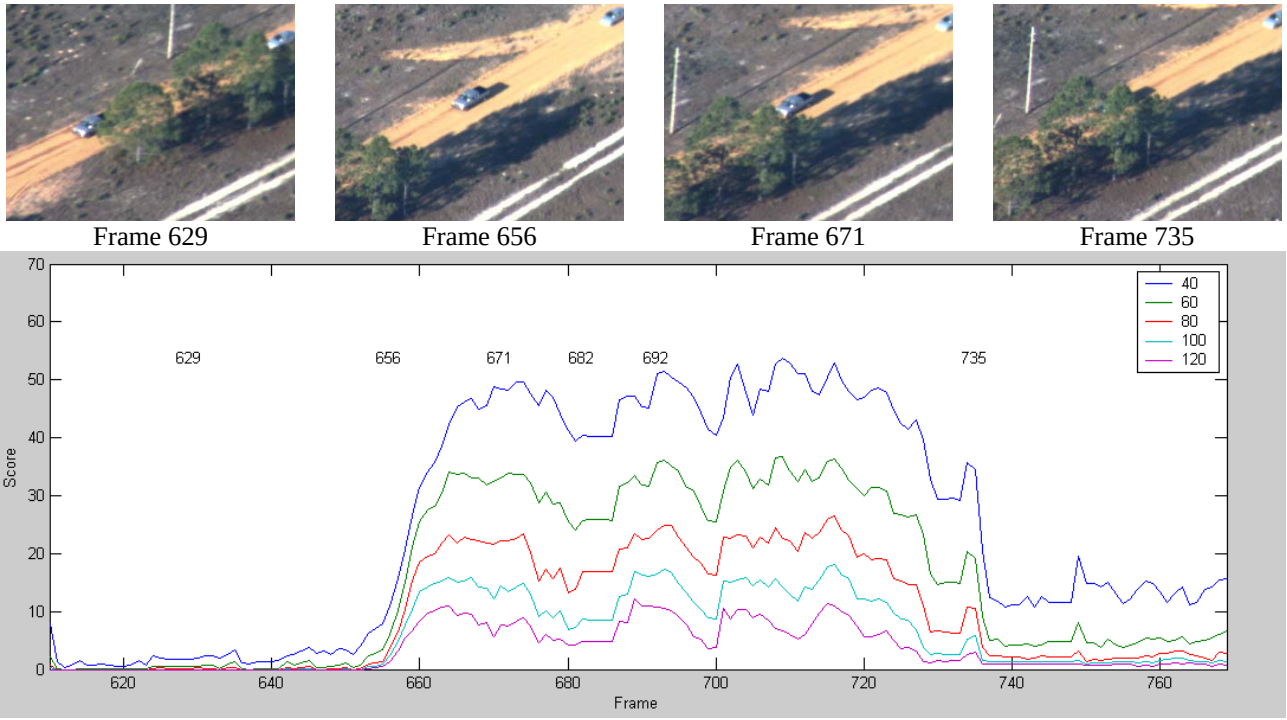


Figure 9: Diagnostic measure used to predict the presence of an occlusion

4. Usability Studies

We have performed two studies assessing the performance of START truthing application. The first study, presented in [5], assessed the accuracy of the truth as compared to expert truthers. The second study assessed the usability and speed of the tool, comparing it to existing truting tools at AFRL.

The accuracy study involved eight test subjects that truthed every frame in a short sequence [5]. Each target was given a minimally-bounding oriented rectangle. Figure 10 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.

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 10. 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. 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 vindicates 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. Additionally the bounding box dimensions and orientation envelopes are also depicted in Figure 10 and Figure 11.

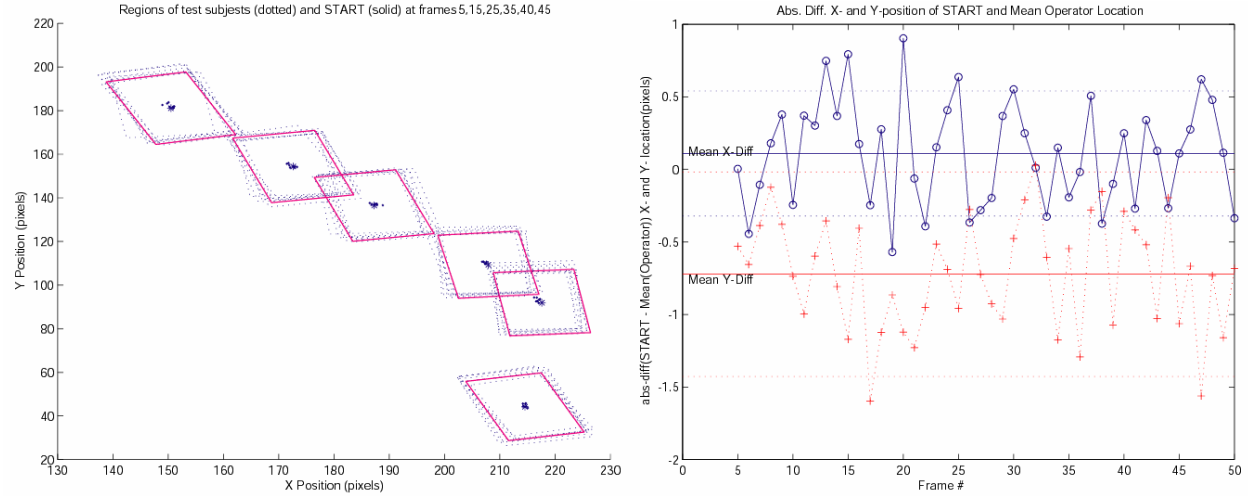


Figure 10: Bounding boxes of START and test subjects (left), and positional errors of START versus mean user (right)

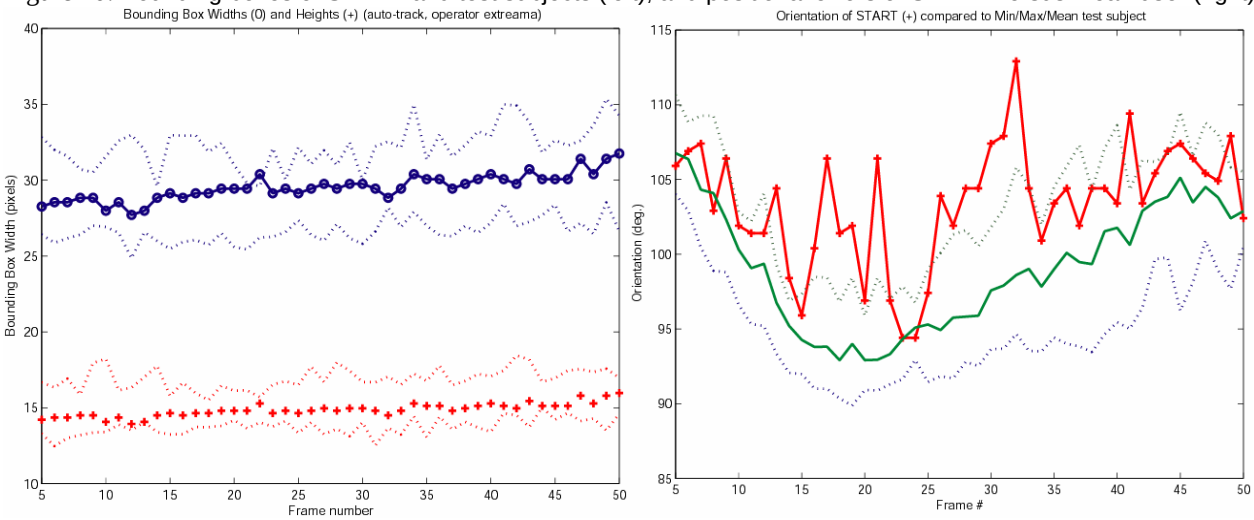
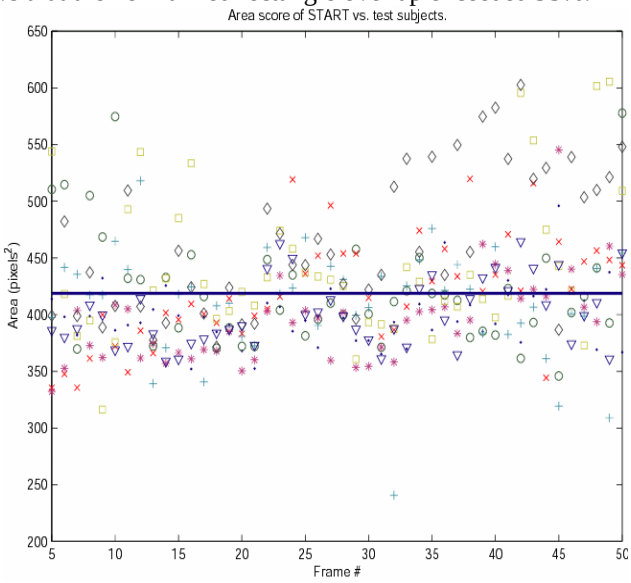


Figure 11: START Orientation versus test subjects (left), and bounding box dimensions of START compared to test subjects (right)

We then computed the area overlap of each of the test subjects bounding rectangles with the rectangle produced with START, the results of which are depicted in Figure 12. The average area overlap is shown in the table of Figure 12, and shows that the normalized rectangle overlap exceeded 93%.



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

Figure 12: Area score results of START as compared to test subjects

For the second evaluation, the approach was a user-oriented evaluation consisting of three elements:

1. Part 1 of the evaluation involved exercising various options and features available in START to truth the same clip several different ways. This phase explored different strategies for truthing, which aided in understanding how best to use START.
2. The second phase of the evaluation was a comparison of START to a tool currently available at AFRL for truthing video data, namely Truth Tool V (TTV). This phase focused on performance and timing, providing a quantitative assessment of the benefits and limitations of START. The primary measure of performance is the time required to truth the imagery. This phase concluded with a questionnaire to elicit subjective feedback on the strengths and weaknesses of START.
3. The final step in the evaluation was a focus group. During the focus group, video clips were projected on a screen and the evaluation participants discussed their strategies for truthing using START. This step provided feedback on how best to approach truthing a clip, given various conditions.

The first step of the evaluation revealed some general principles about how best to exercises the features in START. Typically, it is wise to view a clip at least once before actually collecting any image truth. If the targets are not occluded and the viewing geometry is relatively constant, the auto-track and multi-track capabilities offer rapid effective methods for extracting most of the truth data. When aspect angles change, it may be necessary to update the reference for the auto-tracking on a frequent basis. The diagnostics assist in determining when various interventions are needed to correct the auto-tracking. For vehicles that move in and out of occlusion, including partial occlusion, the interpolation functionality provides a quick and effective method for establishing the vehicle track.

In the second step of the evaluation, truthing with START exhibit approximately a factor f 3 time savings when compared with the legacy truthing tool know as TTV (figure 13). The finding is consistent across the analysts who participated in the evaluation (figure 14). The subjective feedback derived from questionnaires indicated generally favorable feeling about START:

- Training: General the participants felt well prepared for the evaluation, although analysts 3 received only limited training
- Overall START performance was judged to be fairly strong and analysts preferred START over the legacy tool
- Benefits to user were evident in terms of time savings, rapid development of metadata, and reduced operator fatigue

The evaluation concluded with the focus group session. During the focus group it was clear that experienced users were consistent in their use of the wide range of tools and functionality in START.

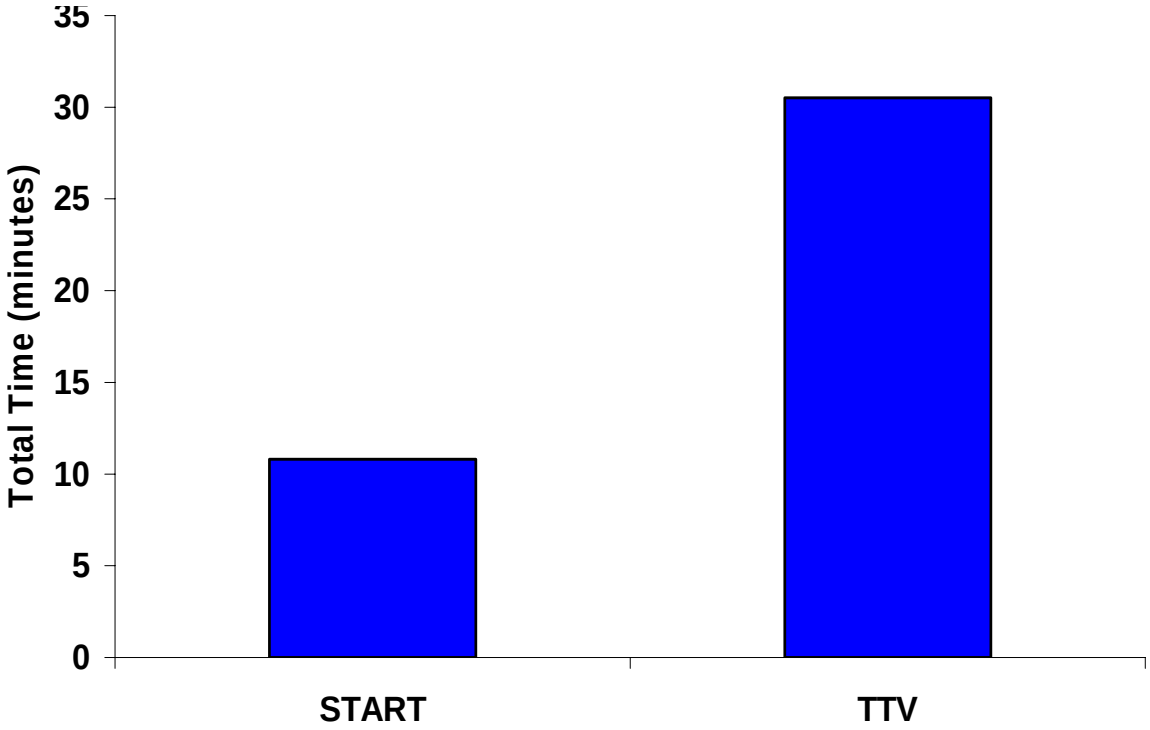


Figure 13: Mean Truthing Time per Clip Using START and TTV

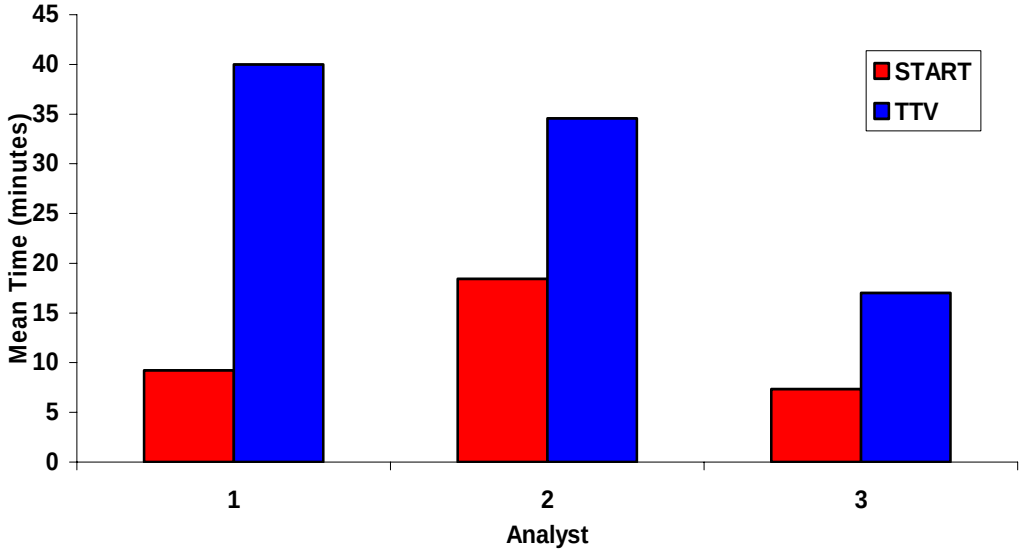


Figure 14: Mean Truthing Time for START and TTV by Analyst

5. Conclusions

START provides a powerful tool for truthing video imagery. Both objective and subjective evaluations show that START consistently and dramatically outperforms the legacy software available for truthing motion imagery. The analysts generally found the START interface was user-friendly and easy to learn. The various capabilities enable the experienced user to recognize certain situations and select the specific strategy for rapidly and accurately truthing the clip.

BIBLIOGRAPHY

[1] Ralph, S., Irvine, J., Stevens, M., and Snorrason, M. START: A Tool for Rapidly Generating Image-Truth and Evaluating ATR Performance. Proceedings of the WACV, Brekenridge, CO. 2005.

[2] Irvine, J., Ralph, S., Stevens, M., Marvel, J., Snorrason, M., and Gwilt, D. START for Evaluation of Target Detection and Tracking. Proceedings of SPIE Defense and Security. Vol. 5807. 2005.

[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.

[4] Irvine, J., Ralph, S., Stevens, M., Kenyon, S., Anderson, D., Snorrason, M., and Gwilt, D. A Scoring Truthing and Registration Toolkit for Evalutaion of Target Detection and Tracking. Proceedings of the SPIE Defense and Security. Vol. 5426. 2004.

[5] 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.

[6] Stevens, M., Ralph, S., and Snorrason, M. Interactive Truthing Tools for Moving Platforms and Moving Targets. Automatic Target Recognition Working Group. 2004. Eglin, FL.

[7] Stevens, M., Pollak, J., Ralph, S., and Snorrason, M. Video Surveilance at Night. Proceedings of SPIE Defense and Security, Orlando, FL. Vol. 5810. 2005.

[8] Goodsell, T., Snorrason, M., Cartwright, D., Stube, B., Stevens, M. R., and Ablavsky, V. Sign detection for autonomous navigation. Unmanned Ground Vehicle Technology V. Vol. 5083. 2003. Orlando, FL, USA, SPIE.

[9] McBride, J., Snorrason, M., Goodsell, T., Eaton, R., and Stevens, M. Single Camera Stereo for Mobile Robot Surveilance. IEEE Conference on Computer Vision and Pattern Recogniton, Workshop on Advanced 3D Imaging for Safety and Security. 2005.