A Comparative Analysis of Human and Algorithm Track Smoothing

Lee B. Spence, Jason Rathje, and Mary Cummings[1]

Abstract

Deciding if and what objects should be engaged in a Ballistic Missile Defense System (BMDS) scenario involves a number of complex issues. The system is very large and the timelines may be extremely short. The magnitude of the task and the short timelines drive designers to highly automate these systems. On the other hand, the critical nature of BMD engagement decisions suggests exploring a human-in-the-loop (HIL) approach to allow for judgment and knowledge-based decisions, which provide for potential automated system override decisions. This BMDS problem is reflective of the role allocation conundrum faced in many supervisory control systems, which is how to determine which functions should be mutually exclusive and which should be collaborative. Clearly there are some tasks that are simply too fast or computationally intensive for humans to make useful contributions, especially in time-pressured environments. On the other hand, there are tasks that are tractable by humans in the available time, and these may be completed without automation or with basic computer assistance. Between these extremes are a number of cases in which degrees of collaboration between the human and computer are possible. This paper motivates and outlines an experiment that is quantitatively investigating human/automation tradeoffs in the specific domain of tracking problems.

Introduction

Deciding if and which objects should be engaged in a Ballistic Missile Defense System (BMDS) scenario involves a number of complex issues. The system is very large, it has many interconnected elements, and is physically spread over an area that is a significant fraction of the Earth. The information for such decisions may be incomplete and/or inconclusive, and, given the enormity of the decision-making task, the timelines may be extremely short, on the order of minutes. The magnitude of the task and the short timelines drive designers to highly automate these systems because of computational speed, repeatability, and high consistency. On the other hand, the grave nature of BMD engagement decisions suggests exploring a human-in-the-loop (HIL) approach to allow

[1]Author's Affiliations: Lee B. Spence, MIT Lincoln Laboratory, Lexington, MA; Jason Rathje and Mary Cummings, MIT Department of Aeronautics and Astronautics, Humans and Automation Laboratory, Cambridge, MA. Corresponding author: Spence@ll.mit.edu

This work was sponsored by the Department of Defense Missile Defense Agency under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and not necessarily endorsed by the Sponsor.

DISTRIBUTION STATEMENT A. Approved for Public Release; distribution is unlimited.
Approved for Public Release. 09-MDA-4286 (24 FEB 09)

for judgment and knowledge-based decisions (Rasmussen, 1983), which provide for potential automated system override decisions.

This BMDS problem is reflective of the role allocation conundrum faced in many supervisory control systems, which is how to determine which functions should be mutually exclusive and which should be collaborative (Sheridan, 2002). Clearly there are some tasks that are simply too fast or computationally intensive for humans to make useful contributions, especially in time-pressured environments. On the other hand, there are tasks that are tractable by humans in the available time, and these may be completed without automation or with basic computer assistance. Between these extremes are a number of cases in which degrees of collaboration between the human and computer are possible.

Because humans can reason inductively and generate conceptual representations based on both abstract and factual information, they also have the ability to make decisions based on qualitative and quantitative information (Fitts, 1951). In addition, allowing operators active participation in decision-making processes provides not only safety benefits, but promotes situation awareness and also allows a human operator, and thus a system, to respond more flexibly to uncertain and unexpected events. Thus, decision support systems that leverage the collaborative strength of humans and automation in supervisory control planning and resource allocation tasks could provide substantial benefits, in terms of both human and computer impacts on system performance.

This paper motivates and outlines an experiment that is quantitatively investigating human/automation role allocation tradeoffs in the specific domain of tracking problems. Computers, in the form of predictive algorithms, are often relied upon for proper connection of object radar track segments that may contain erroneous associations. However, this task is one with clear vision-based pattern recognition elements, so it is possible that human operators could perform as well, or better than, the automation. The possibility that humans can make contributions to these problems has support in a number of studies that have investigated the ability of humans to perceive lines in data that is incomplete or occluded (Fulvio et.al., 2008). The proposed experiment is assessing how well humans and a specific algorithm compare in the task of correlating track segments generated by processing a batch of radar data. The ultimate goal of this research is to determine an empirically-based rationale for a collaborative human-computer track correlation decision support system.

Background

There are a number of functions that radars typically perform (Sullivan, 2004). These include detection, the determination of the presence of targets in the radar's field of regard; tracking, the process of ascertaining where the objects have been in space and where they will be in the future; and object assessment, the process of using information in the measured radar return signal to assess targets' characteristics. Of these functions, we propose that tracking, in some applications, can be improved with a mixed-initiative approach to human-automation role allocation (Horovitz, 1999).

Radar tracking occurs in the coordinates of range, azimuth, elevation and their respective rates. While this leads to a six dimensional characterization of the targets, there are cases where the tracker may become confused about which returns should be matched with which targets, and this erroneous association can result in tracks being mislabeled. Figure 1, which plots a set of re-entry breakup tracks as a function of time and relative range alone, is an illustration of several types of incorrect associations. As can be seen in Figure 1, where different tracks are denoted by different colors, there are cases at crossings where the automated tracker switches object labels. In addition, track identification can switch even in non-crossing situations. For the track data to be most useful, it is important that these misidentifications be identified and corrected.

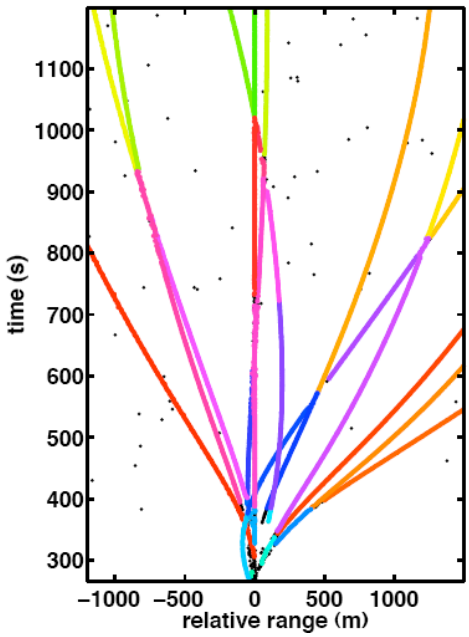

Figure 1: Tracker output with track swaps

While there are a number of possible algorithmic approaches to identifying and removing these misidentifications, one that will be considered here is an unpublished multi-target smoothing (MTS) algorithm developed by under MDA's Project Hercules by Hendrick Lambert of MIT Lincoln Laboratory. Figure 2 shows the result of applying the MTS algorithm to the entire batch of data in Figure 1 (the original data is repeated for ease of comparison). In this example, all of the incorrect identifications have been corrected.

The MTS algorithm operates on batches of raw track data, and first removes any segments judged to be outliers and then breaks the tracks into individual segments. Then the algorithm seeks to reconnect the segments by examining all possible combinations to determine the best fit to the data in a mean-square sense. It is possible to think of the human operators as following a similar procedure – the outliers are removed because they "don't look right" and the remaining data is fit to what "looks best".

Preliminary observations of algorithm performance presented to operators resulted in them seeing associations that the algorithm did not make with the available data. This observation suggests that humans may be superior to the algorithm in associating correct segments when the data spans are shorter as compared to longer time spans like in Figure 2 where the algorithm shows excellent performance for a re-entry breakup tracking.

If humans can be shown to be superior to automation in cases with the shorter time spans, the use of humans in this process could potentially provide additional engagement time that would be operationally significant. In order to gain insight as to whether human operators can indeed identify track swaps better than, or at least as good as, the automation, an experiment is underway to compete the human versus the automation.

Interface Design

In order to conduct the experiment to investigate human operator performance in track resolution, an interface was designed to allow a human the ability to interpret and connect radar track data (Figure 3). The design of this interface was guided by principles that direct that effective displays should allow users to have appropriate control, both in solution creation and editing, provide them with error correction and appropriate feedback, and keep the interface as simple as possible, especially for such a time-constrained task (Nielsen, 1993). While the interface was not designed to be an actual operational interface, it was designed to evaluate concepts for an operational interface insofar as possible.

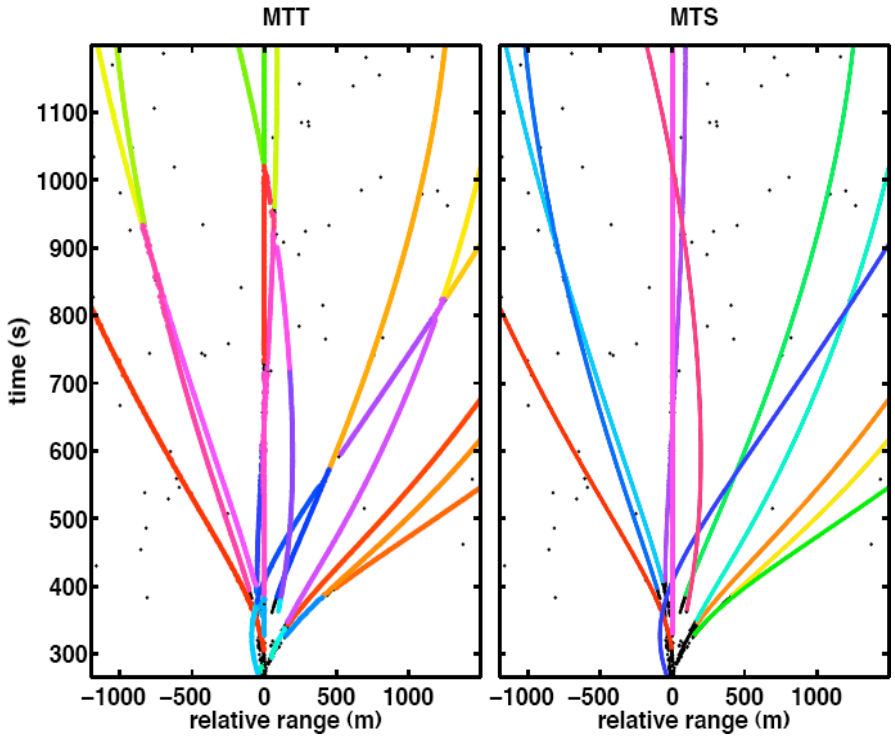

Figure 2: Correction of track swaps by the MTS algorithm (right side)

In order for the user to efficiently produce best-fit tracks, simple graphing ability, including an editing capability, was vital. The subjects also needed some control over the way the data were displayed. From an operational standpoint, to keep the user aware of the time available for fitting the tracks, a counter was placed in the interface. The resulting interface has four major components: Plotting area, Interaction panel for track specification, Plot appearance, and Timer. These are discussed below.

The data plot, which constitutes the majority of the interface in Figure 3, displays the simulated data that is presented to the subjects, and is the working area for the subjects to select their "best fits" to the tracks. As seen in Figure 1, the simulation data contains color-coded track segments. However, because the different colors have no specific meaning to the operator and could bias the operator in terms of track selection, the radar track data is displayed in gray in order to let the subjects connect the data as they think best. The option of utilizing color coded data in a subsequent experiment may be considered for future work.

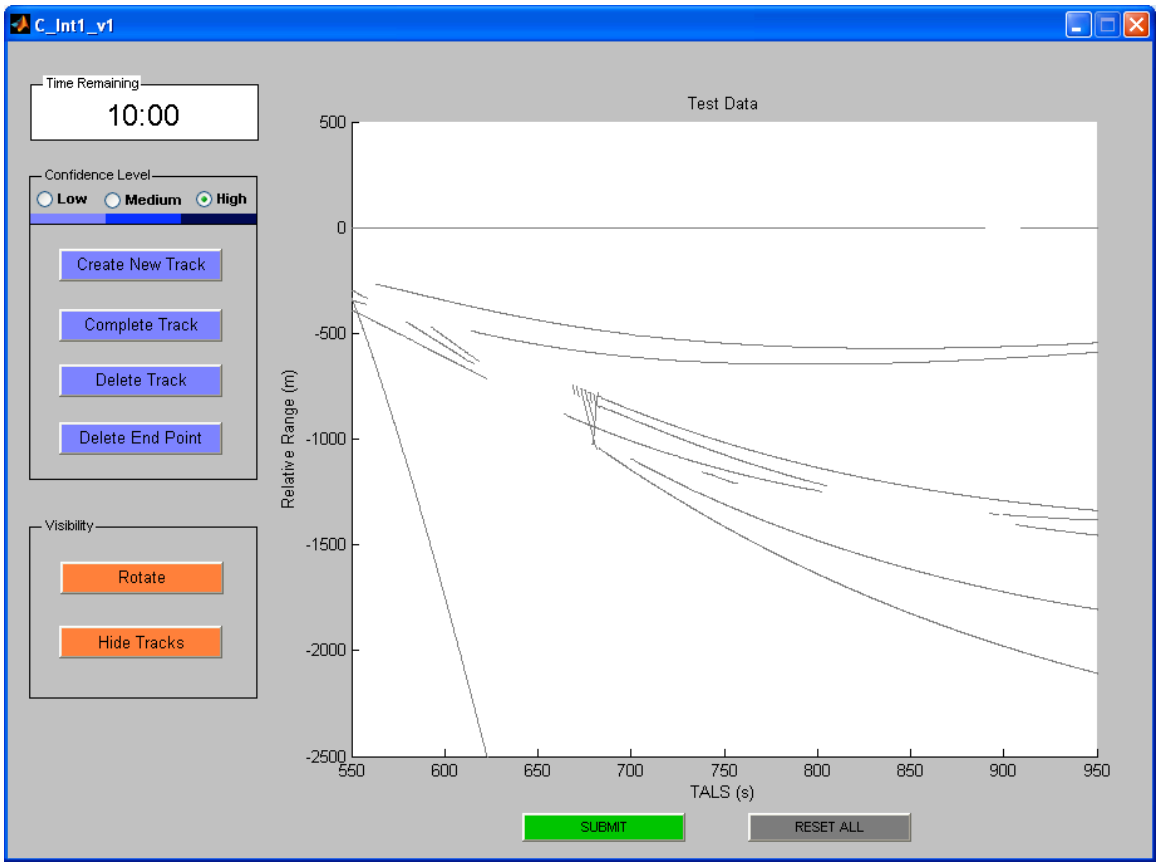

Figure 3: The Experiment Interface

The panel of buttons at the top left of Figure 3 is used for plotting the fitted trajectories and manipulating them. The user can initiate, complete, and edit tracks using this panel. In addition, this panel allows the user to select a confidence that he/she feels appropriate

for each segment of each track. This estimate is also color-coded along the track, such that it is clear when confidence estimates change from point to point along the track.

The panel of buttons in the lower left of Figure 3 allows the user to change the plot appearance during the experiment. One button allows the user to select a "right/left" or "up/down" data display and allows for adjustment for individual preferences as well as allowing operators to observe a different orientation to spot potential patterns. A second button allows the user to inhibit the display of previous tracks that may be interfering with a current selection and darkens the data that is displayed to alert the user to the fact that visibility of previous selections has been turned off.

Lastly, because the tasks must be completed in a specified time, a timer is included in the interface to keep the user aware of the time available for the task. The timer's size and font were chosen to be salient while not taking away from the task at hand. In the present experiment the timer counts down from 10 minutes.

Experimental Concept and Design

In order to investigate the ability of both the human and an algorithm to identify and correct the track swaps of interest, a data set was needed. The data used for this experiment are simulated outputs of a hypothetical radar's real-time multi-target tracker, containing association errors similar to those discussed earlier. Data were selected for the experiment that challenge the algorithm and operator, particularly when sparse data is available to execute a best fit. The data spans were selected to test cases that could result in humans performing better than the algorithm. Over a long time period, the algorithm performs well, however at shorter times like 50% of these times it does not do as well. Thus for the experiment, time periods of 30%, 60% and 100% of the interval over which the algorithm works well were investigated to assess performance over cases in which the algorithm would be expected to work well and cases in which it would not. Since in operational scenarios these might represent time savings on the order of 5 minutes, superior human performance would be very significant from an operational perspective. The data were also sorted by degree of difficulty. Cases with shallow crossing angles and high track density are considered the most difficult cases.

The experiment is being conducted by presenting human subjects with representative data segments on the interface, and asking them to fit lines to the segments that they perceive as actually connected. In parallel with the human subject testing, the same data presented to the subjects were also presented to the MTS algorithm. The results for both cases are scored using the underlying truth data that were generated as part of the simulation. The independent variables for the experiment are summarized in Table 1. Given the three independent variables in Table 1, this is a 2x2x3 fully crossed within-subjects experiment.

Table 1: Independent Variables

Variable	Levels
Decision source	Human and Computer
Degree of difficulty	Hard (narrow crossing angles or confusing situations) and Easy (wide angles and straightforward)
Data span	30, 60 and 100% of the available data interval

Scoring

Trajectory completeness and fit, as defined below, are the primary variables used to score performance.

Trajectory Completeness

The user, or algorithm, should correctly identify each track present in the data. The user, or algorithm, is not given this knowledge prior to the experiment, so they may miss some tracks (Missed Truth Lines) or may plot additional tracks (False Lines). These scores are tallied for each plot.

Trajectory Fit

In addition to completeness, it is important to quantitatively determine how well the user (or algorithm) does. The closer the user matches the corresponding truth trajectory, the better the score. This score is calculated from goodness-of-fit measures used in linear regression modeling. The truth trajectory is assumed to be the model (regression fit), while the line plotted by the participants or the algorithm is assumed to represent the observations. The mean squared error (MSE) and the root mean squared error (RMSE) are calculated as shown below (Levine, 2001). Currently N, the number of samples along the line, of 500 is used to calculate the results. This was selected by observing that, while the results changed significantly from N = 100 to 500, there was little change beyond this.

$$MSE = \frac{\sum_{i=1}^{N} \left(Y_{estimate_i} - Y_{truth_i} \right)^2}{N}$$

$$RMSE = sqrt(MSE)$$

Figure 4 is an outcome of scoring a pilot test, showing the truth lines (solid lines) compared to the user's plotted data (dashed) and the quantitative results. Each color is a different track and truth pair. Table 2 shows the score averaged over all of the lines in the test figure.

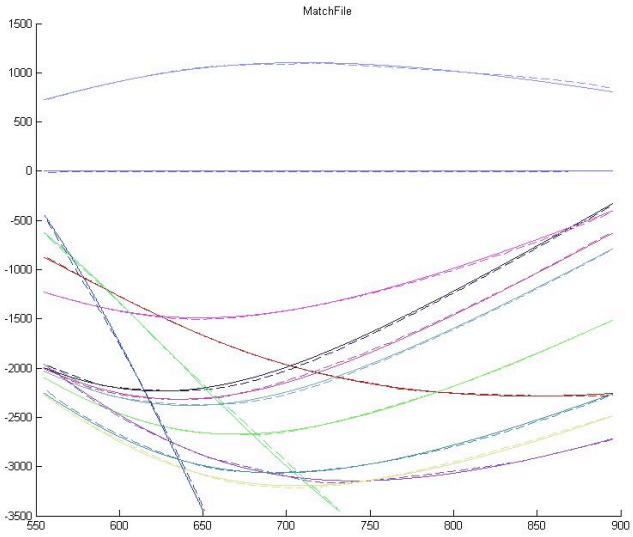

Figure 4: Pilot Test Example Performance.

Table 2: Example Pilot Test Scores

Missed Truth Lines	0
False Lines	0
Average MSE	303 (m^2)
Average RMSE	15.4 (m)
Time to Solution	8 m 57 s

In analyzing the actual experiment results, a hypothesis test is used to compare the human and algorithm performance. If the results as a function of data span show that the operators can perform better with the shorter data spans than the algorithm, this means that the operators should likely be employed in some time critical situations to return accurate tracks that are better than the algorithm's. Before interpreting results, however, there is an important consideration of screen resolution, discussed in the next section, which has to be considered as part of the scoring procedure.

Screen Resolution Limits

The monitor used for the experiment is a Dell 2001FP. The display area horizontal dimension, D_H, is about 41 cm (1280 pixels) and the vertical dimension, D_V, is about 31 cm (1024 pixels). When performance is scored, the difference between results and truth is calculated in meters. However, the difference between human results and truth is ultimately limited by the ability of the human to resolve points on the screen. It has been estimated that the human eye can resolve points to about a minute of arc (Ware, 2004). At a nominal distance from a monitor of 50 cm, this is a point separation of about 0.015 cm

or 0.15 mm. On the other hand, the extent of a pixel on the screen can be approximated as $D_V/1024 = 0.03$ cm or 0.3 mm. Clearly the pixel extent and not the resolution of the eye dominates the observer's power of resolution, and it will be assumed that, if a human performance is within one pixel of the truth, that is the best that can be achieved, and the human should be given credit for a "perfect" performance when this occurs.

Status and Initial Results

The interface for testing has been completed. Simulated radar data and the associated truth data have been generated. Pilot testing has been completed and the algorithm has been run on the simulated data and scored. Testing with human subjects has been completed on twenty nine subjects chosen from MIT Lincoln Laboratory employees with no special requirements, aside from adequate vision to complete the test, placed on subject participation. The subject's participation was part of their regular workday and they were not paid directly for participation. After informed consent was obtained, the subjects were given an initial orientation to the test and they participated in practice sessions of increasing difficulty. The final level of difficulty was chosen to represent the hardest case they would expect to encounter on the test.

Detailed scoring of human and algorithm performance is ongoing. However, detailed scoring of two cases completed to date has indicated one case in which the algorithm performs better and one case in which the human subjects perform better. Scoring of the remaining cases will have to be completed before general conclusions can be drawn.

Summary and Further Work

While many function allocation approaches to human supervisory control take a mutually-exclusive approach, particularly in BMD settings, we propose that there are scenarios where a more collaborative approach to decision making and action could result in superior systems performance. To this end, an experiment is being conducted to address human/automation tradeoffs in the specific case of improving radar tracking data. An experimental interface is complete, simulated experimental data, including truth data has been generated, and testing is underway.

Two areas have been identified for further effort. One area is relative performance of the humans and algorithm as the data quality degrades. The data used for the current experiment is of high signal-to-noise ratio, so the results could change once this ratio diminishes. Another future research area comes from noting that, while the algorithm works with six dimensional data, the operator is expected to resolve the track ambiguities with only two dimensions. It may be possible to provide the operator with another dimension or two, like range-rate, and the impact of this on performance could be investigated.

Acknowledgements

This work benefited significantly from useful discussions on experiment design with Birsen Donmez and from assistance in experiment execution from Allison Loftin.

References

Fitts, P. M. (Ed.). (1951). *Human Engineering for an Effective Air Navigation and Traffic Control system*. Washington, DC: National Research Council.

Fulvio, J. M. et.al. (2008), *Precision and consistency of contour integration, Vision Research, **48**, pp. 831-849.*

Horovitz, E. (1999). *Principles of Mixed-Initiative User Interfaces*. Paper presented at the ACM SIGCHI Conference on Human Factors in Computing Systems.

Levine, D. M., et al. (2001). Applied Statistics for Engineers and Scientists. Upper Saddle River, New Jersey 07458, Prentice Hall.

Nielsen, J. (1993). *Usability engineering*. Cambridge, MA: Academic Press

Rasmussen, J. (1983). *Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models.* Paper presented at the IEEE Transactions on Systems, Man, and Cybernetics SMC-13.

Sheridan, T. B. (2002). Humans and Automation: System Design and Research Issues, John Wiley & Sons, Inc., Santa Monica, CA.

Sullivan, R. J. (2004). Radar Foundations for Imaging and Advanced Concepts. Scitech, Raleigh, NC.

Ware, C. (2004), Information Visualization, Morgan Kaufmann Publishers, San Francisco CA.