NASA Contractor Report 177583

USAAVSCOM Technical Report TR-90-A-009

User Acceptance of Intelligent Avionics: A Study of Automatic-Aided Target Recognition

Curtis A. Becker, Brian C. Hayes, and Patrick C. Gorman

CONTRACT NAS2-12849
May 1991

National Aeronautics and Space Administration

US ARMY AVIATION SYSTEMS COMMAND

AVIATION RESEARCH AND TECHNOLOGY ACTIVITY
MOFFETT FIELD, CA 94035-1099

NASA Contractor Report 177583 USAAVSCOM Technical Report TR-90-A-009

User Acceptance of Intelligent Avionics: A Study of Automatic-Aided Target Recognition

Curtis A. Becker, Brian C. Hayes, and Patrick C. Gorman

Bio-Dynamics Research and Development Corp.
1000 Willagillespie Rd., Suite 200
Eugene, Oregon 97401

Prepared for
Ames Research Center
CONTRACT NAS2-12849
MAY 1991

National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California 94035-1000

US ARMY
AVIATION
SYSTEMS COMMAND

AVIATION RESEARCH AND
TECHNOLOGY ACTIVITY
MOFFETT FIELD, CA 94035-1099

Abstract

User acceptance of new support systems typically has been evaluated after the systems are specified, designed, and built. The current study attempts to assess user acceptance of an Automatic-Aided Target Recognition (ATR) system using an emulation of such a proposed system. The detection accuracy and false alarm level of the ATR system were varied systematically, and subjects rated the tactical value of systems exhibiting different performance levels. Both detection accuracy and false alarm level affected the subjects' ratings. The data from two experiments suggest a cut-off point in ATR performance below which the subjects saw little tactical value in the system. An ATR system seems to have obvious tactical value only if it functions at a correct detection rate of .7 or better with a false alarm level of .167 false alarms per square degree or fewer.

INTRODUCTION

A wide variety of both current and proposed avionics systems are intended to meet the needs of Army rotorcraft pilots. These systems range from electronic maps and notepads to autonomous systems that locate a target, navigate to an optimal firing point, select the proper weapon, and engage the target. The trend is for such systems to become increasingly complex and intelligent, and to assume more of what are traditionally the pilot's responsibilities. Given the life or death consequences in the aviation environment, aviators who use complex systems must have both a high level of skill in operating those systems and a high level of confidence in the capabilities of their systems. Pilots must understand fully what the systems can and cannot do, and they must develop a feel for the conditions under which their systems are more likely and less likely to function as needed.

User confidence in any system is some function of the experienced value of the system's information handling. One important component of value appears to be the overall accuracy of the system's output. If the user can clearly see errors in the system's knowledge base, reasoning, or other aspects of system operation, the value as perceived by the user may well suffer. Subtle or undetected errors can also have profound effects. If the user focuses on only the most abstract level of system output, there could be a false sense of certainty leading to a catastrophic outcome such as that encountered by the U.S.S. Vincennes (Friedman, 1989). In this incident which resulted in the shooting down of Iran Air Flight 655, many of the high level system outputs fit the situation-specific expectations of the users, and little if any effort was directed toward understanding what lay behind the high-level outputs. In this case, several small errors masked by users' confidence in their system cascaded into a tragedy.

Below the highest level, other system characteristics may determine the perceived value of a system. These include user access to system-internal decision rules and criteria, and the availability and quality of information against which the user may validate the system's operation and output. Each of these factors represents a way for the user to check on the system's accuracy at run-time, on a case by case basis, and begin to build an experience base about those conditions that result in better system performance and those conditions that lead to worse performance. Another factor is the timeliness of system output. In many cases, the value of information or advice deteriorates as a

function of the temporal distance from the conditions that gave rise to the information or advice.

To begin our examination of the parameters of user acceptance of intelligent systems, we have addressed the factor of overall system accuracy. If a system is to achieve any value in the user's eyes, obviously it must perform at a level better than chance. On the other hand, an intelligent system may not need to be perfect, an often requested level of performance. This leaves us with a broad band of system accuracy levels to examine. Our starting assumption was that somewhere in that broad band there is a system accuracy level that corresponds to a breakpoint for user-perceived value. Performance below that level is unacceptable, while performance above that level begins to accrue significant value in the eyes of the user.

APPLICATION OF INTEREST

The intelligent system under study here is an Automatic-Aided Target Recognition (ATR) system that is being considered for inclusion in future Army helicopters. An ATR system receives its inputs from a bank of electronic sensors, performs a set of pattern matching operations, and presents the results of its analysis to the helicopter pilot. In a combat environment, the general rule is that the first one to detect the foe controls the engagement. The envisioned role of an ATR system is to enhance the pilot's capacity for early detection of his adversary and thereby increase the pilot's chances for survival and mission success.

There are several views on just how an ATR system will function in future helicopters. The view that places the least demands on technological sophistication is one that requires the pilot to position the aircraft, execute a sensor scan, withdraw from the scan position, and examine the ATR results after moving to a relatively safe location. This usage scenario affords the ATR system a substantial amount of time to complete its analysis. At the other end of technological requirements is the view that the ATR system will operate continuously, providing its outputs to the pilot in near real time. This version of the usage scenario requires the ATR system to scan its environment rapidly and to analyze its inputs rapidly. From a timeliness perspective, the second view seems preferable, but as noted above, timing is but one important property of such a system.

The detailed specification of an ATR system includes properties of the sensors used by the system, the pattern matching algorithms, and the format in which the ATR results are to be presented to the pilot. For our purposes, we need only know the general characteristics of the sensors and the basic approach to the algorithms. The sensor array can include devices for visible light, infrared, radar, acoustic, or other components of the spectrum. An adequate model of the pattern matching algorithms can be found in standard Signal Detection Theory (Green & Swets, 1966).

As for the output format for the ATR results, there are numerous implementation from which to choose. The output can range from a simple text list of detected objects and their location, to a symbolic overlay on an electronic map, to an overlay on visible source data, to an abstract three-dimensional rendering that includes spatially located sounds. For our first study, we selected a presentation format that used a symbolic overlay on top of infrared

imagery. All of the pilots in our subject pool had some experience with infrared imagery, and we could gain access to a large sample of appropriate video-taped infrared tactical scenes.

EXPERIMENT 1

Our goal for this first study was to begin mapping the domain of system accuracy and the ways that it relates to user-perceived value. For an ATR system, performance accuracy can be well specified in terms of the system's correct detection rate and its false alarm level. In addition, a "reality check" can be independently established from knowledge about the actual location of targets. By comparing actual target locations with targets as detected by an ATR system, subjects can begin to assess and understand the accuracy of the system.

For Experiment 1, video-taped tactical scenes were obtained from a field test of early ATR systems. The various scenes were populated with a broad range of military vehicles as well as with a variety of target-competitive clutter. From the video tape, static images of infrared views of the scenes were captured and used as a background upon which to indicate the location of ATR detected targets. Subjects examined an ATR-processed image, and then rated the ATR system's apparent performance and tactical value for that image. Following the ratings, actual target locations were marked, and subjects were able to compare ATR-detected targets and false targets with the real targets contained in the image. Across a series of such examine-rate-evaluate trials, we expected subjects to learn the operating characteristics of the ATR system and to adjust their ratings to fit the system.

Method

Subjects: Nineteen current Army helicopter pilots, all males, served as the subjects for this study. The pilots ranged in rank from CW2 through Captain, and they participated in the study during a training session for the Army's Light Helicopter program assessment effort. Three of the pilots served to de-bug and validate the procedures, and the remaining 16 provided the data reported below.

Equipment: Two testing stations were configured with a Sony PVM-2030 20-inch monitor, a Tektronix Tek-Touch touch screen mounted on the monitor, and a Commodore-Amiga 2000 computer with an auxiliary Syquest SQ555 44Mb removable disk cartridge system.

Task Description: Each trial in the study consisted of six component displays. An example of these displays is shown in Figure 1. The first component, shown in Figure 1a, presented an infrared image of a tactical scene with red arrows (shown in Figure 1 as white arrows) superimposed over the scene to indicate the location of objects detected in that scene. Subjects were told that the red arrows showed the results of ATR system processing. They were to examine the display, decide which of the indicated objects were real targets and which were not, and then to search the infrared scene for real targets that the ATR system might have missed. When finished, the subject proceeded to the next component display by operating either the touch panel or the system mouse.

The second component display, shown in Figure 1b, added a rating scale to the initial display. The scale was presented vertically, in red, along the left edge of the display, labeled "HI" at the top, "LO" at the bottom, and

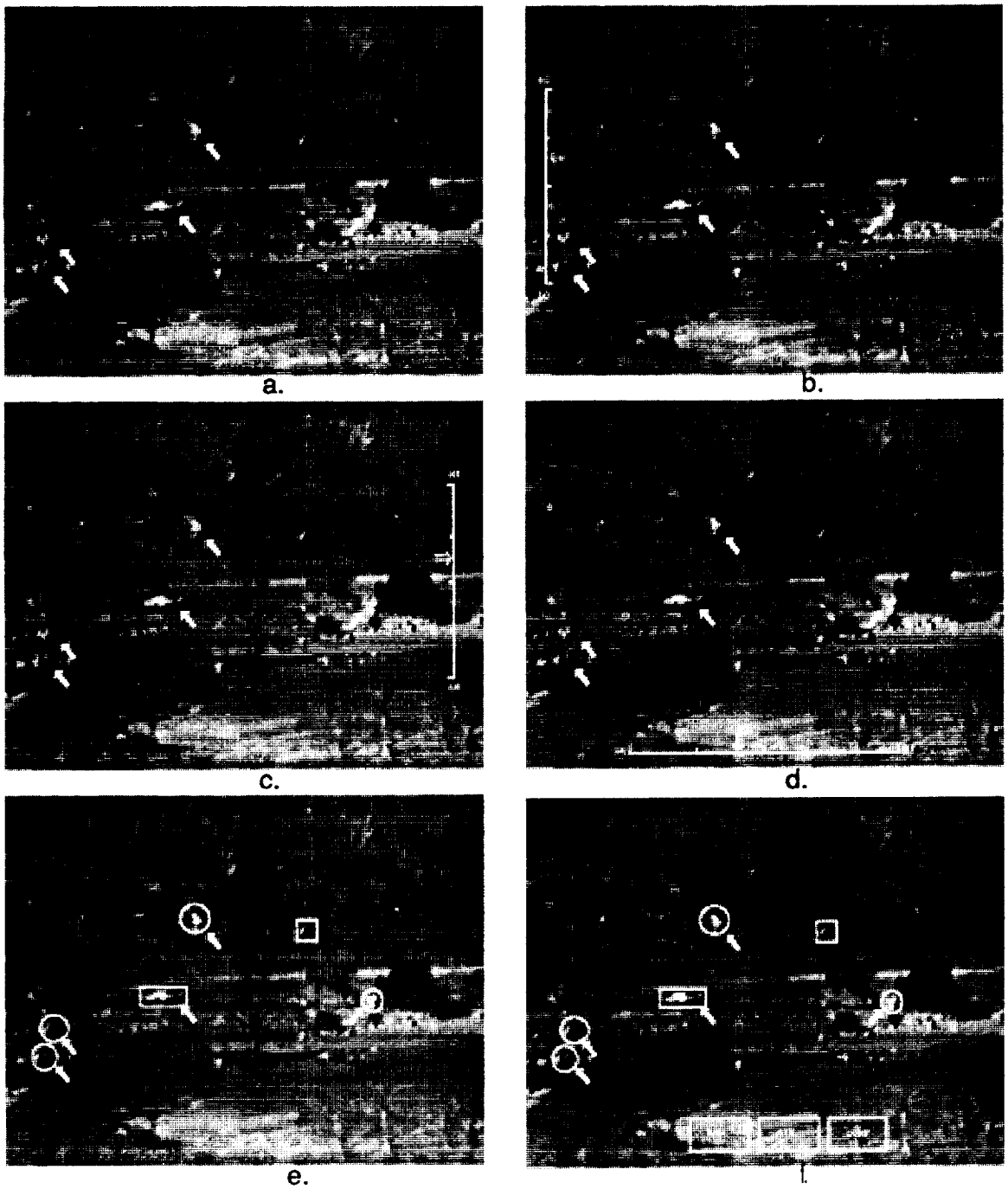

Figure 1 -- The six component displays for one trial showing a .5 correct detection rate with four false alarms. The top four displays show the initial ATR output and the three rating scales with white arrows instead of red ones. The bottom two displays show a variation of the feedback display with false alarms marked by circles instead of yellow rectangles.

with three intermediate unlabeled markers dividing the scale into four equal parts. Again using either the touch panel or the mouse, the subject positioned a pointer along the scale to indicate his judgement of the ATR system's *correct detection rate* for the current scene. The pointer could be

re-positioned until the subject was satisfied, at which time, he proceeded to the next component display.

The third display, in Figure 1c, erased the correct detection scale from the left-hand edge and added a scale, similar in appearance but in yellow, to the right edge of the display. This scale was to be used to indicate the judged level of ATR *system false alarms* in the current scene. For the false alarm judgement, the subject was told to base his rating on the number of apparent false alarms in the scene. Specifically, of the significant objects indicated, are there many false alarms or only a few.

After completing the false alarm judgement, the subject went on to the fourth component display (Figure 1d), one that removed the false alarm scale and added a scale horizontally across the bottom of the display in green. For this scale, the subject was told to rate the *tactical value* of the information in the display. The instructions for this scale told the subject that his goal was to identify and engage a target and that the tactical value rating was an indication of the usefulness of the ATR display in meeting that goal.

Following the tactical value rating, the subject proceeded to the feedback display (Figure 1e). In this display, the infrared scene was overlayed with the original red arrow indicators along with indicators of the "ground truth" for the scene, that is, the location of all real targets. Specifically, the feedback display included red rectangles drawn around all real targets in the scene and yellow rectangles drawn around all false alarms reported by the ATR system. In this display, correctly detected real targets were indicated by the presence of both a red arrow and a red rectangle. Real targets that the ATR system missed were marked with just the red rectangle, and ATR system false alarms were indicated by a red arrow combined with a yellow rectangle encircling the false object. (In Figure 1e, rectangles mark the real targets and circles indicate false alarms.)

After inspecting the feedback display, the subject went on to the final component display (Figure 1f) that allowed him to review the trial by selecting the raw infrared scene for further inspection, the original ATR output display, or the feedback display. This review display was intended to allow the subject a further opportunity to understand the characteristics of the ATR system that was represented by the trial. As described below, the apparent performance of the ATR system was varied widely and systematically.

Materials and Design: Video tapes from an infrared camera provided the source for all base displays used in this study. The tapes were created during the Multi-Sensor Fusion Demonstration conducted in 1988 at Ft. Hunter-Liggett. A total of 25 test sites were included with a variety of military vehicles placed at different locations within the sites across the numerous trials of the field test. Using a FrameGrabber system and a Commodore-Amiga, the base displays were digitized at a resolution of 320 X 200 pixels with a 16-level grey scale. A total of 160 base frames were captured from the 25 test sites included on the video tapes. Each base display was then graphics edited to create a total of three variants of the base yielding 480 stimulus displays. The editing allowed us to enhance objects in the scene, blur objects, and even add and remove objects to meet the requirements of our system accuracy treatment conditions. Each variant of a base display was assigned to a different system accuracy condition.

	Correct Detection Rates			
	.3	.5	.7	.9
No. of Real Targets	78	78	72	69
No. of Correct Detections	24	39	51	63
Max. per Display	2	3	3	4
Min. per Display	0	0	1	1
No. of Missed Targets	54	39	21	6
Max. per Display	3	2	2	1
Min. per Display	1	0	0	0

Table 1 – The total number of real targets per 30 displays and the breakdown into correct detections and missed targets.

The design included four levels of correct detection fully crossed with four levels of false alarms. The correct detection rates were .3, .5, .7, and .9. The false alarm levels were 1.5, 1.0, 0.5, and 0.167 false alarms per square degree of scene. Each of the displays represented about three square degrees.

To represent an ATR system that seemed to perform at a particular level of accuracy, we selected 30 displays and marked correct detections, false targets, and missed targets in each display. A different set of 30 displays was used for each of the sixteen conditions defined by the factorial combination of the four levels of correction detection and the four levels of false alarms. Within each condition, the individual displays were sequenced such that the average system performance level was met for each subset of ten displays. For example, in representing an ATR system with an average correct detection rate of .5 and a false alarm level of 1.0 per square degree, we

	False Alarm Levels			
	1.5	1.0	0.5	0.167
No. of False Alarms	135	90	45	15
Max. per Display	8	6	3	1
Min. per Display	2	1	0	0

Table 2 – The total number of false alarms per 30 displays.

allowed both the detection rate and the false alarm level to vary for individual displays. Averaged across each subset of ten displays, though, the detection rate was exactly .5 and the false alarm level was exactly 1.0 per square degree. By varying the ATR system properties from frame to frame, we attempted to more realistically represent the operation of a functioning ATR system with its probable variations as a function of terrain and target types.

Table 1 presents the mapping of real targets to the four levels of correct detection rates used here. Also included are the maximum and minimum numbers of real targets assigned to individual displays. Table 2 shows the same data for the four levels of false alarms.

In all, sixteen sets of 30 displays were constructed, one set for each combination of correct detection rate with false alarm level. Each subject saw all sixteen sets with the order of the sets determined by a 16 X 16 balanced Latin square. In this design, each treatment condition occurred once in each ordinal position across subjects, and each treatment condition preceded and followed every other condition equally often, thereby balancing both first-order and second-order sequence effects. In creating the Latin square, the only constraint on random assignment was

that no level of detection rate or of false alarms immediately follow itself.

For each of the sixteen treatment conditions, the set of 30 displays was presented in three different orders. As stated above, each subset of ten displays in a set matched the detection rate and the false alarm level for the condition. Across every three subjects, the ordering of displays was rotated such that each of the three subjects saw a different set of ten displays as his final ten. This counterbalancing of displays across subjects means that no single display can have an inordinate influence on the findings of this study. The displays upon which subjects base their ratings can be treated as a statistically random factor.

General Procedure: Each subject in the study was given a 30-45 minute instructional session during which the procedures and ratings were explained. A practice set of sixteen displays was presented with one display representative of each of the sixteen treatment conditions. After the instructions, subjects began working through their sequence of the sixteen treatment conditions according to the order specified by one of the rows of the balanced Latin square. They completed one or two of the conditions during the remainder of the first session. The remaining conditions were completed across three or four subsequent 45-75 minute sessions, depending on the availability of the subjects.

As stated above, each of the displays included in the practice set represented one of the sixteen treatment conditions. This fact was emphasized as the subjects worked through the practice set. The goal here was to expose the subjects to the full range of system detection rates and false alarm levels so that they could anchor their use of the three scales to the range detection rates and false alarm levels used here.

Results and Discussion

The data of interest were taken from the last ten trials in each condition for each subject. The first twenty trials, then, were used to allow the subjects to identify the characteristics of the represented ATR system and to adapt their ratings appropriately. A complete analysis showed no significant interactions between the subsets of ten trials and any of the other factors, although there were numerical trends suggesting that the subjects began a condition using the middle of the scales and modulated their judgements to fit the treatment condition.

For the statistical analyses, the subjects' placements of the markers along the scale displays were translated into numerical scores between 0 and 100 with 0 corresponding to the "LO" labelled end of the scales and 100 corresponding to the "HI" end.

The statistical analyses presented below used an analysis of variance based on the Latin square. Such an analysis uses a pooled, residual error term to assess the significance of all factors. The residual error term in this study includes the usual subject interaction sources as well as variance attributable to the stimulus displays. Since the exact set of displays seen by the subjects was varied across subjects thereby confounding stimulus variance with subject variance, all statistical conclusions can be generalized to the stimulus population as well as to the subject population.

The first two rating scales, those for detection rate and false alarms, were included in this study so that we

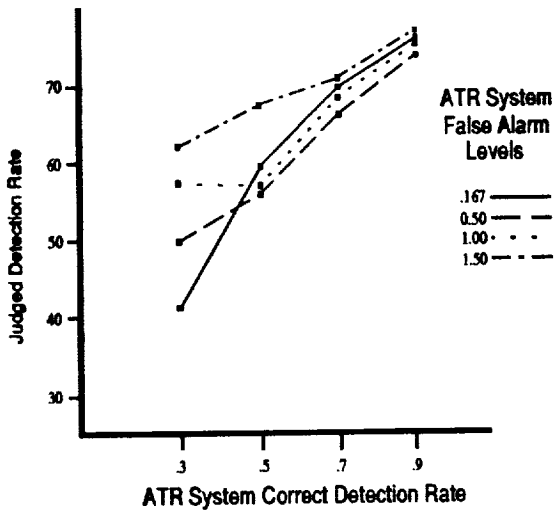

Figure 2 -- Subject judgements of correct detection as a function of ATR system detection rates and false alarm levels.

could determine first whether the subjects were sensitive to the manipulations of system accuracy and second whether the subjects were using the rating scales in a sensible way. The data for judged detection rates are shown in Figure 2. There was a significant main effect of the correct detection rate represented in the ATR system, $F_{3,210} = 42.78$, $p < .01$, and a significant main effect of the ATR system's false alarm level, $F_{3,210} = 7.49$, $p < .01$. The effect of system detection rate is obvious in Figure 2. As the system's detection rate increased, so did the subjects' judgements of those rates. The data also show that the judged detection rates were overestimated at low levels of system detection and underestimated at the highest level. These findings are consistent with a large body of subjective rating literature on magnitude estimation showing a less than one-to-one mapping between actual magnitudes and subjective magnitudes.

The main effect of ATR system false alarm level is less obvious. The highest level of false alarms, 1.5 per square degree, yielded judged detection rates substantially higher than the other three levels of false alarms. Most of this effect seems confined to the lower levels of system correct detection rates as evidenced by a significant interaction between system correct detection rate and system false alarm level, $F_{9,210} = 2.03$, $p < .05$. One interpretation of this interaction is that subjects may have been biased by the "more is better" notion. With a large number of false alarms in a display, there are many indicators of detected objects. The subjects' judgement of detection rate may have been based partly on the erroneous assumption that some proportion of the marked objects were real targets. Alternatively, it could have been that the information in the infrared scene was not good enough to allow the subject to reject all of the false alarms. That would inflate the number of objects initially accepted as real targets and hence inflate the judgement of correct detection rate.

The data from the subjects' judged false alarms are shown in Figure 3. The results here are straightforward. There was a significant main effect of

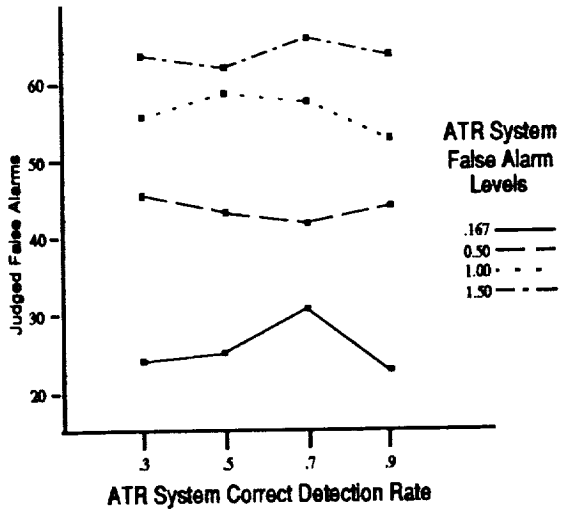

Figure 3 -- Subjects judgements of false alarms as a function of ATR system detection rates and false alarm levels.

ATR system false alarm level, $F_{3,210}$ = 153.88, $p < .01$. Higher system false alarm levels were matched by higher judged false alarms. Again, these data are consistent with other evidence about subjective estimates. Specifically, as the absolute magnitude of the judged quantity increases, the just noticeable difference between two magnitudes increases. Thus, a linear difference in absolute physical magnitude often translates into a decelerating curve for the subjective judgements. In these data there was no main effect of system detection rate, $F_{3,210}$ = 0.81, and no interaction of detection rate with system false alarm level, $F_{9,210}$ = 1.20.

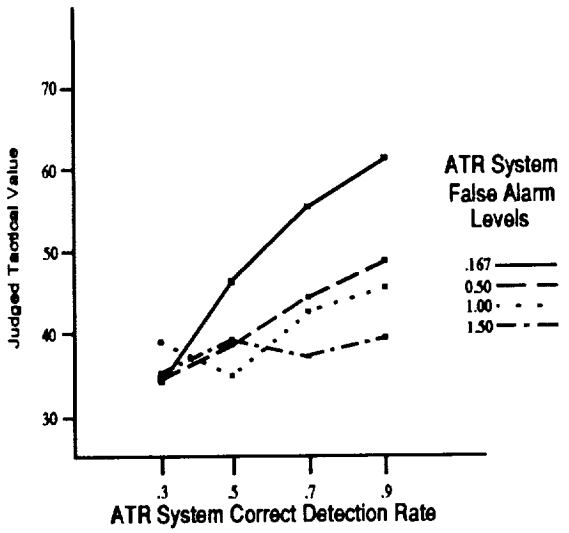

Figure 4 -- Subject judgements of tactical value as a function of ATR system detection rates and false alarm levels.

These first two sets of rating data were collected to determine whether our manipulations of ATR system properties were detectable by our subjects and whether the subjects were performing in accord with findings from a large body of other work on subjective judgements. The results described so far suggest that our manipulations were effective and that our subjects were performing consistently. This provides a sound basis against which to evaluate the data from the subjects' ratings of tactical value.

The initial hope for the tactical value ratings was that they would help to identify some kind of break point, that is, a system accuracy point below which an ATR system is quite unacceptable and above which an ATR system would be considered valuable by the users. The tactical value data are shown in Figure 4. In these data, there was a significant main effect of system correct detection rate, $F_{3,210}$ = 14.35, $p < .01$, a significant main effect of system false alarm level, $F_{3,210}$ = 12.60, $p < .01$, and a significant interaction of the two factors, $F_{9,210}$ = 2.69, $p < .01$. The tactical value ratings increased with increases in system correct detection rates, and the ratings also increased with decreases in system false alarm levels. The interaction lies in the fact that at the lowest detection rate, all four false alarm levels were rated equally bad; while at the higher detection rates, the lowest level of false alarms diverged substantially from the other conditions.

It is the interaction in the tactical value data that best meets our initial hope for this study. From the graph in Figure 4, there appears to be a break point at about 50 on the tactical value scale. The two data points above 50 seem fairly well separated from the cluster of data points below 50. Several *post hoc* comparisons support that conclusion, although weakly. The two highest rated conditions are the only ones for which appropriate, rank-ordered paired-comparisons approached significance. To allow further assessment by the reader, the standard error of the mean derived from the Latin square residual error term is 3.06.

The tactical value ratings support some interesting conclusions that appear to contradict some of the statements made by our subjects. First, it was said that an ATR system needs to be virtually perfect before it can be of any value. That does not seem to be true in our data. Even our low-end systems were rated as having some tactical value. A second common statement was "I don't care how many false alarms you give me, just don't miss anything". If this statement were true, the ATR system false alarm level should not have affected the tactical value ratings, and that was clearly not the case.

The findings from this initial study are encouraging. The results from the correct detection and false alarm ratings clearly suggest that our manipulations of ATR system accuracy were detected by the subjects, and the findings from the tactical value ratings suggest that these manipulations are important determiners of user acceptance of an ATR system. We view this initial study as a baseline against which improvements and enhancements can be evaluated. From the data in Figure 4, it is clear that the best system represented in this study was rated slightly above 60 on the tactical value scale. A rating of about 60 on a 100-point scale cannot be construed as a particularly strong endorsement. Alternatively, the numerical results may be attributed to scaling effects. Subjective judgements often tend to cluster around the mid-point of the scale, so the absolute level of the ratings may not be fully indicative of absolute levels of acceptance.

In designing this first study, we focused on the type of ATR system that places the least demands on technological sophistication. From this baseline, there are a wide range of potential system improvements that can be explored. The next study incorporates one such improvement, and it expands the parameter range for system false alarm level.

EXPERIMENT 2

We have two concerns about the first study that can be addressed relatively easily. First, the stimulus displays from Experiment 1 used a fairly coarse digitizing mode of 320 X 200 pixels, and the displays were presented on a 20" diagonal screen. These two factors in combination yield low quality visual detail. This may have contributed to a dissatisfaction with the information contained in the displays. The subjects may not have been able to discern real targets from false alarms. In Experiment 2, the displays were digitized using the 640 X 400 display mode, and they were presented on a 9" diagonal monitor. Because our source video tapes were third-generation VHS copies of high-resolution (875-line) original tapes, this manipulation cannot be characterized in terms of video resolution. However, the increase in pixel-level resolution along with the decrease on the size of the monitor did result in an obvious and sizable increase in the visible quality of the stimulus displays. Because of the characteristics of the manipulation and some of the procedural problems of inducing the subjects to use the rating scale in a way comparable to that of Experiment 1, we consider this factor as a preliminary look at the potential effects of display quality.

The second concern focuses on the question of whether our sample of system accuracy goes far enough. The best system represented in Experiment 1 had a correct detection rate of .9 and a false alarm level of 0.167 per square degree, and this system received the highest tactical

value rating. It could be that higher detection rates or lower false alarm levels would result in even higher ratings. To further explore the upper bounds of system accuracy, Experiment 2 includes a lower false alarm level of 0.067 false alarms per square degree.

METHOD

Subjects: The same nineteen pilots from Experiment 1 participated in this study. One subject was run to de-bug the procedures, and the remaining 18 contributed the data reported below. The time lag between participating in the first and second studies varied from 3 to 5 months. Therefore, before participating in Experiment 2, all subjects were re-run on a subset of the conditions from Experiment 1 in an attempt to ensure consistent use of the rating scales between the two studies and to remind the subjects of the appearance of the displays from Experiment 1. Specifically, three conditions from Experiment 1 (detection rate/false alarm levels of .3/1.5, .5/1.0, and .9/.167) that spanned the range of correct detection rates and false alarm levels were used prior to the first data collection session. In Experiment 2, subjects were run in three sessions, and, prior to the second and third sessions, one block of trials from Experiment 1 was presented using displays from the .7 correct detection rate .50 false alarm per square degree condition.

	Correct Detection Rates		
	.5	.7	.9
No. of Real Targets	72	72	72
No. of Correct Detections	36	51	63
Max. per Display	3	4	4
Min. per Display	0	0	1
No. of Missed Targets	36	21	9
Max. per Display	2	2	1
Min. per Display	0	0	0

Table 3 – The total number of real targets per 30 displays and the breakdown into correct detections and missed targets for Experiment 2.

	False Alarm Levels		
	1.0	0.167	0.067
No. of False Alarms	90	15	6
Max. per Display	5	1	1
Min. per Display	1	0	0

Table 4 – The total number of false alarms per 30 displays for Experiment 2.

Equipment: The only change in equipment from Experiment 1 was the change from a 20" monitor to a 9" Sony CPD-9000 monitor.

Task Description: The task and the sequence of events during a trial were identical to those used in Experiment 1. Because of the change in pixel-level resolution, several minor changes were required. First, all scales were displayed in white instead of in the colors used for Experiment 1. Second, we were limited to a maximum of seven total correct detections, misses and false alarms in a single display. Third, the format of the feedback display used colored, labeled pointers to indicate correct detections, missed targets, and false alarms instead of the colored rectangles used for Experiment 1. All of these changes were necessitated by the hardware limitations of the 640

X 400 display format.

Materials and Design: As stated above, a new set of base displays was captured using a 640 X 400 pixel resolution with a 16-level gray scale. As in Experiment 1, the base displays were then edited to meet the requirements for the various treatment combinations used here with each variation of a base display assigned to a different treatment condition. The descriptive statistics for the conditions of Experiment 2 are shown in Tables 3 and 4.

The design included three levels of correct detection rate crossed with three levels of false alarms. The correct detection rates were .5, .7, and .9, and the false alarm levels were 1.0, 0.167, and 0.067 false alarms per square degree. All three of the detection rates were used in Experiment 1 as were two of the three false alarm levels. Thus, we have six comparable conditions in the two experiments to assess the effects of the change in visible quality. Experiment 2 also allows us to evaluate the effect of a further decrease in the level of false alarms.

General Procedure: Each subject participated in three sessions. The first session included a repeat of ten-trial blocks from three of the conditions from Experiment 1, as specified above, followed by a nine-trial practice set using the displays for Experiment 2. During the rest of the first session, the subject completed three of the conditions of Experiment 2. The second and third sessions began with a ten-trial block of one of the conditions from Experiment 1 followed by three 30-trial blocks from Experiment 2. This procedure was followed to ensure that the subjects remembered the systems represented in the first experiment and the quality of display used in that study.

The order of treatments in Experiment 2 was determined by a 9 X 9 Latin Square constructed with the same constraints as in Experiment 1. No level of correct detection rate or false alarm level immediately followed itself. The first nine subjects were assigned to a row of the square, and the second nine subjects were run using the same Latin Square in reverse order. This achieved the same level of sequence balancing as in Experiment 1. That is, each treatment condition occurred equally often at each serial position in the sequence, and each treatment condition preceded and followed every other condition equally often.

Results and Discussion

A preliminary analysis examined the data for interactions of the Latin Square factor with other factors and for interactions of the 10-trial subsets with other factors. No interactions were found; so, the following data analyses are based on the last ten trials for each subject for each treatment condition using a residual error term pooled across the two Latin Squares.

The findings for both the judged

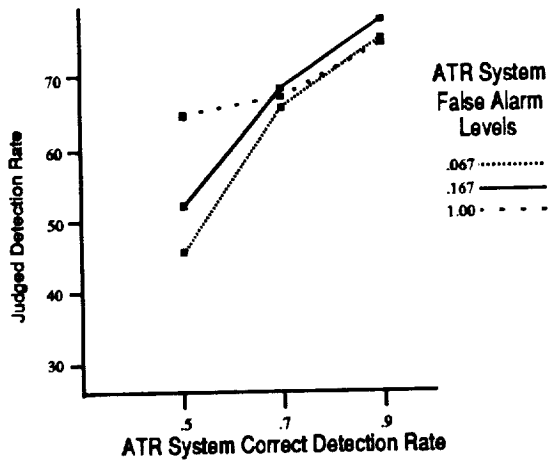

Figure 5 -- Subject judgements of correct detection as a function of ATR system detection rates and false alarm levels in Experiment 2.

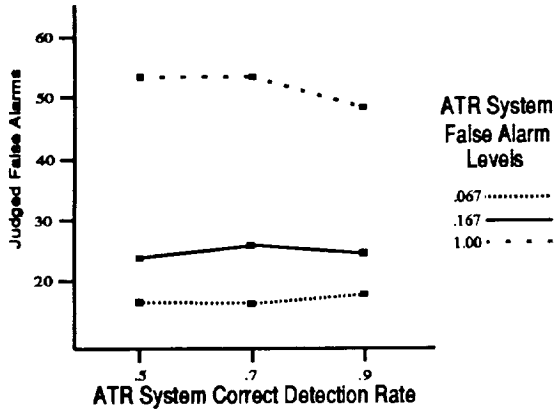

Figure 6 -- Subjects judgements of false alarms as a function of ATR system detection rates and false alarm levels in Experiment 2.

detection rate and the judged false alarms mirror those from Experiment 1. The judged detection rate data, presented in Figure 5, show a significant main effect of system detection rate, $F_{2,128} = 121.78$, $p < .01$, a main effect of system false alarm level, $F_{2,128} = 11.74$, $p < .01$, and a significant interaction between the two factors, $F_{4,128} = 12.41$, $p < .01$. As in Experiment 1, judged detection rate increased with increases in the represented system detection rate, and it increased with increases in the system false alarm level. The interaction is found in the inflated judged detection rate for the combination of high system false alarms and low system detection rates.

The data for judged false alarms, in Figure 6, show only a significant main effect of system false alarm level, $F_{2,128} = 327.66$, $p < .01$. Here, as in Experiment 1, there are substantial differences among the system false alarm levels but no consistent effect of system correct detection rate. These data also show that the reduction in system false alarm level included in Experiment 2 was clearly noticed by the subjects.

The judged tactical value results, shown in Figure 7, confirm the findings from Experiment 1. System correct detection rate had a significant effect on the subjects' judgement of tactical value, $F_{2,128} = 81.31$, $p < .01$. As the system's correct detection rate increased, so did the judged tactical value. The effect of system false alarm level was also significant, $F_{2,128} = 18.56$, $p < .01$, as was the interaction of false alarms with detection rate, $F_{4,128} = 9.30$, $p < .01$. As in Experiment 1, judged tactical value increased with decreases in system false alarm level with most of the effect occurring at the higher system detection rates.

One specific result is of particular interest. The two lower false alarm levels do not appear to differ from each other in judged tactical value. The overall means are 48.11 for .067 false alarms per square degree and 49.83 for .167 false alarms per square degree. At the higher two system detection rates, the two false alarm levels yield virtually identical tactical value judgements. This is true even though the judged false alarms clearly

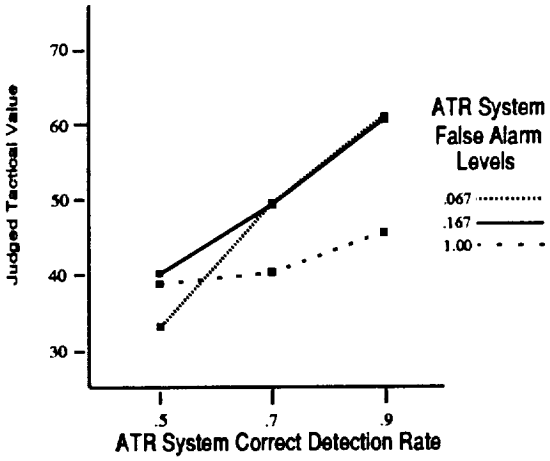

Figure 7 -- Subject judgements of tactical value as a function of ATR system detection rates and false alarm levels in Experiment 2.

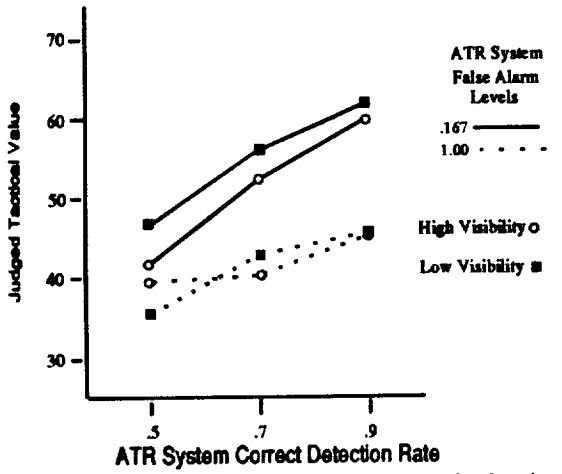

Figure 8 -- Subject judgements of tactical value as a function of ATR system detection rates, false alarm levels, and visible quality compared across Experiemnts 1 and 2.

showed a difference between the two system false alarm levels (see Figure 6).

The final aspect of the data to be examined here focuses on the effect of the change in visible quality between Experiment 1 and Experiment 2. In the first study, the infrared display used a 320 X 200 pixel image presented on a 20" monitor. The second study used a 640 X 400 pixel image shown on a 9" monitor.

The analyses here include only the sixteen subjects who participated in both experiments. Thus, the data of two subjects from Experiment 2 have been eliminated. In a preliminary analysis, we compared the three ten-trial blocks used at the beginning of the first session of Experiment 2 with the first ten trials of the corresponding conditions from Experiment 1. For all three of the judgements, there was no significant main effect of Experiment 1 vs. Experiment 2 and no interaction of this factor with the ATR system conditions represented. This finding supports the contention that subjects successfully re-anchored their scale use to be consistent with Experiment 1. In general, this is true, but a detailed examination of the data shows that individual subjects did change their use of the three scales from Experiment 1 to Experiment 2. Therefore, the following results should be treated as preliminary findings.

The tactical value data graphed in Figure 8 shows the comparison between the first and second experiments, and hence the comparison of a low visible quality to a higher visible quality. A repeated-measures analysis of variance on the 2 (false alarm levels) X 3 (correct detection rates) design common to the two experiments showed a significant main effect of system correct detection rate, $F_{2,30} = 25.73$, $p < .01$, a significant main effect of system false alarm level, $F_{1,15} = 20.38$, $p < .01$, and a significant interaction between visible quality and system false alarm level, $F_{1,15} = 4.89$, $p < .05$. The two main effects are consistent with the findings reported earlier. The interaction can be found in the fact that visible quality had no effect on the higher level of false alarms, while at the lower false alarm level, the high visibility condition received consistently lower tactical value ratings. This result is consistent with the suggestion that the higher visible quality allowed subjects to discern more false alarms using the visual evidence, leading to lower tactical value ratings

GENERAL DISCUSSION

These data focus on the acceptance of a particular type of system by a subgroup of that system's potential user population. For current Army rotorcraft pilots, an ATR system appears to have an obvious tactical value only if it functions at a correct detection rate of .7 or better coupled with a false alarm level of .167 per

square degree or better. Improving the correct detection rate appears to have greater value than decreasing the false alarm level. Also, for the type of ATR system represented here -- a detection system -- showing pilots a better quality of image does not seem to offer much of a payoff as measured by judged tactical value.

The qualifications of our conclusions begin to draw in many of the other issues that must eventually be considered. Other population subgourps must be considered. Other levels of ATR systems must be examined such as those that go beyond detection to provide some form of categorizing of the detected objects. Other detailed implementations must also be examined, especially those that push the required level of technological sophistication. An ATR system that provides a near real-time output may need different system accuracy properties. Finally, other measures must be taken and our measures must be used in other contexts so that we can better understand the limitations of subjective estimates of tactical value.

The preliminary findings reported by O'Kane, Blecha, Do-Duc, & Flaherty (1990) begin to make the needed extensions to our approach. In that study, subjects searched a wide field consisting of five horizontally overlapping static infrared images. Each image was about seven degrees wide by five degrees high, or roughly twelve times the size of the images used here. The subject's task was to search the set of images and mark the location of all user-identified real targets. In one condition, no ATR system was available, providing a baseline unaided search condition. Other conditions simulated ATR systems and varied the false alarm level.

Several of the O'Kane, *et al.*, findings are of interest here. First, the difference between unaided search and aided search appeared to be mainly in the time required to approach asymptotic performance. Time to asymptote for aided search was about two-thirds that for unaided search. Given the projected usage conditions for ATR systems, this is a necessary finding, and it is in need of further confirmation.

The second O'Kane, *et al.*, result of interest here is the way that the false alarm level affected search. At low false alarm levels, ATR-assisted search was both faster and more accurate that unaided search. That is, more real targets were marked, fewer false alarms were accepted as real targets, and less time was required when using the ATR system than when searching the image without an ATR. At high false alarm levels, though, roughly equal numbers of real targets were marked, and more false alarms were marked compared with the unaided baseline condition.

The O'Kane, *et al.*, manipulation of false alarms corresponds to .057 false alarms per square degree on the low end, and .171 per square degree at the higher level. This re-casting of their false alarm levels shows them to be comparable to the levels used here in Experiment 2, levels for which we found no effects. It would seem, then, that the search task detected an effect of false alarms that our tactical value ratings did not. Clearly, this is one possibility. Search time and detailed accuracy measures may be more sensitive than the subjective judgement used here.

Another possibility, though, is that the effective manipulation of false alarm levels lies more in the count of false alarms per image than in the engineering metric of number of false

alarms per square degree. If the number of false alarms per image is the effective metric, then the comparison would be between our .50 and 1.50 false alarms per square degree conditions and the O'Kane, *et al.*, 2 and 6 false alarms per image. In this comparison, both procedures show effects of false alarm level.

Two final points need to be made. Throughout this discussion, we have talked about ATR systems as though they had relatively static properties, as though they performed entirely consistently. That is not the case. A given ATR system may perform very well in certain usage contexts and poorly in other contexts. Depending on the terrain being surveyed, the types of targets, and the types of target-competitive clutter, the same system could operate at different system accuracy levels. Thus, in one environment, the ATR system could achieve a detection rate of .9 with fewer than .2 false alarms per square degree. In a different environment, the same system could fall below a detection rate of .7 and show a much higher level of false alarms. In the first case, the ATR system may have substantial tactical value to a pilot. In the second case, the pilot could opt to turn the system off.

The second point is that the use of an ATR system is likely to be mission sensitive. On a scout mission, it may be important that all real targets be detected regardless of the level of false alarms. In this mission, false alarms can be filtered out before the data on real targets is forwarded to attack units. On the other hand, during an attack mission, a large number of false alarms may only distract the pilot from the real targets and delay his decision to engage, thereby losing any advantage that the ATR system might otherwise provide. A failure to detect some of the targets during an attack mission may not be that important. As an example, using tactical knowledge, the experienced pilot is aware that where there are three tanks, there are likely to be six tanks, along with their supporting units. If the ATR shows three tanks, the system may have served its purpose of making the pilot aware of his tactical situation.

Acknowledgements

This work was supported by NASA Ames Contract NAS2-12849 to Bio-Dynamics Research and Development Corp. Dr. Nancy Bucher served as the contract monitor, and her contributions to the quality of this work are gratefully recognized. Further guidance of this work has been provided by Bruce Tenney and Lisa Maston of the Aviation Advanced Technology Directorate, Ft. Eustis, VA.

References

Friedman, N. The Vincennes incident. U.S. Naval Institute Proceedings, Vol. 115/5/1035, May, 1989, 72-79.

Green, D.M. and Swets, J.A. *Signal Detection Theory and Psychophysics*. New York: Wiley, 1966.

O'Kane, B., Blecha, W., Do-Duc, H., and Flaherty, R. Aided vs. Unaided Target Search. Paper presented at the U.S. Army Human Engineering Laboratory symposium on ATR MMI, September 11-12, 1990, Aberdeen, MD.

NASA
National Aeronautics and Space Administration

Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
NASA CR-177583 USAAVSCOM TR-90-A-009		

4. Title and Subtitle	5. Report Date
User Acceptance of Intelligent Avionics: A Study of Automatic-Aided Target Recognition.	May 1991
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Curtis A. Becker, Brian C. Hayes, and Patrick C. Gorman	A-91095
	10. Work Unit No. 505-61-51

9. Performing Organization Name and Address	11. Contract or Grant No.
Bio-Dynamics Research and Development Corp. 1000 Willagillespie Rd., Suite 200 Eugene, Oregon 97401	NAS2-12849
	13. Type of Report and Period Covered Contractor Report
12. Sponsoring Agency Name and Address Ames Research Center, Moffett Field, CA 94035-1000, and Aeroflight-dynamics Directorate, U.S.Army Aviation Research and Technology Activity, Ames Research Center, Moffett Field, CA 94035-1099	14. Sponsoring Agency Code

15. Supplementary Notes

Point of Contact: Nancy Bucher, Ames Research Center, MS 243-4, Moffett Field, CA 94035-1000
(415) 604-5161or FTS 464-5161

16. Abstract

User acceptance of new support systems typically has been evaluated after the systems are specified, designed, and built. The current study attempts to assess user acceptance of an Automatic-Aided Target Recognition (ATR) system using an emulation of such a proposed system. The detection accuracy and false alarm level of the ATR system were varied systematically, and subjects rated the tactical value of systems exhibiting different performance levels. Both detection accuracy and false alarm level affected the subjects' ratings. The data from two experiments suggest a cut-off point in ATR performance below which the subjects saw little tactical value in the system. An ATR system seems to have obvious tactical value only if it functions at a correct detection rate of 0.7 or better with a false alarm level of 0.167 false alarms per square degree or fewer.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
Man-machine interface User interface design Subjective value of expert systems	Unclassified-Unlimited Subject Category - 53

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	20	A02