The effects of the reliability of an automatic target recognition system on image analyst performance

Peerly Setter **Hadas Marciano** **Joel Norman**

Ergonomics and Human Factors Unit, University of Haifa
Mount Carmel, Haifa 31905, ISRAEL
psetter@univ.haifa.ac.il

Lt. Col. Michal Hovev

HFE Section
IAF

ABSTRACT

Motivation – To study the effects of the reliability of ATR (Automatic Target Recognition) designations on the performance of expert image analysts of SAR (Synthetic Aperture Radar) images.

Research approach – A psychophysical study of the performance of 12 expert analysts of SAR images.

Findings – Analyst performance was influenced by ATR reliability. Higher reliabilities yielded higher hit rates and higher false alarm rates, and low reliabilities the opposite results. This and a signal detection theory analysis indicate that ATR reliability affects the response criterion and not performance per se. (But see Discussion).

Research Implications – The fact that the reliability of items designated by the ATR system affected the criterion of the analysts has important implications. The tendency to mark more items that were designated by the ATR as being true targets should improve the overall performance of analysts working with state-of-the-art ATR systems (see Discussion).

Originality/Value – The research systematically manipulated the reliability levels of simulated ATR systems, and measured their influence on the performance of human analysts. In this context reliability rate means what percentage of the designated items by an ATR system are actually correct targets. Each ATR block was coupled with a similar non-ATR block, a design that aimed to extract the added value of the ATR system to the performance of the human analysts. In addition, a complete within subjects design was used. This procedure provided a good basis for comparing the different conditions in the experiment.

Take away message – While developing an ATR system, one should provide the image analysts with valid assessments of the system's reliability.

Keywords

Automatic system reliability, Over-trust/under-trust, SAR, ATR, Image interpretation.

INTRODUCTION

ATR is a computer algorithm that can recognize and identify targets, usually of military import. The assumption is that ATR can help human analysts performing image interpretation. (Kuperman and Sobel, 1993). SAR is an airborne long-range radar system yielding high resolution images. These images do not resemble optical images and are quite difficult to interpret. In the current research, the human factors aspects of ATR image interpretation was explored. Since ATR systems are not fully reliable, the human analyst must decide which of the target-like items is a true target and which is not. The analyst should carefully examine items designated by the ATR, but also not ignore items not designated by it.

Human Errors

The analysts can make several types of error. The first type is "misses". In the case of items that were not designated by the ATR, these are said to result from "attentional tunnelling", i.e., failing to scan items not designated by the ATR. This kind of error can occur even in a perfectly reliable ATR system since there is a possibility that the system missed true targets (for examples see, Davidson & Wickens, 1999; Wickens, 2001).

Other kinds of human error are related to the amount of trust the analyst ascribes to the ATR system. Since the currently available ATR systems are not perfect, there are situations where the system fails and designates non-targets as targets. If the analyst has developed over-trust in the ATR system, s/he is liable not to catch the error of the ATR system, leading to a false alarm (Parasuraman & Riley, 1997). On the other hand, if the analyst identifies failures of the ATR system, s/he might lose trust in the system, and completely stop relying on the system, even when it could be helpful (Riley, 1996).

METHOD

Subjects: 12 expert SAR image analysts from the Israeli Air Force.

Stimuli: Using the MSTAR SAR Database (see http://cis.jhu.edu/data.sets/MSTAR/) a bank of SAR images was prepared. These contained between 10 to 18 SAR items, some of which were targets and some distractors. These were planted in background images and served as stimuli in our experiment. The arrangement of the targets in the image was based on common combat doctrines.

Procedure: On each trial the analysts were presented with a SAR image. Using the left mouse button they marked each item that they thought was a target with a red + and then moved on to the next image. Each block of trials with ATR designations was presented at a different level of reliability (80%, 50%, and 33%) and the analysts were informed of that reliability level. Each block consisted of twelve randomly presented trials (12 images). In addition to the three blocks with ATR designations, the analysts were presented with three non-ATR blocks (mirror images). ATR and non-ATR blocks were alternated, and the block order was balanced over subjects. Red rectangles surrounding specific items served as ATR designations.

RESULTS

The performance measures were hit rate (the proportion of targets that were marked by the analyst), and false alarm rate (the proportion of distractors that were marked by the analyst). The Signal Detection Theory measure d' was also computed. In order to eliminate the effect of the relative difficulty of specific images the analyses were based on the differences between performance with and without ATR designations. An analysis of variance (ANOVA) was performed as a function of the ATR reliability and post-hoc Duncan tests ($p<0.05$) were performed when an effect was significant.

Hits rate differences: A significant effect of the reliability of the ATR system was found $F(2, 107)=10.13$, $p<0.0008$ (see Table 1). This effect was a result of the significant difference between each of the three reliability levels.

False alarm rate differences: A significant effect of the reliability of the ATR system was found $F(2, 107)=3.41$, $p<0.051$ (see Table 1). This effect was a result of the significant differences between the 33% reliability level and the 80% reliability.

d' differnces: Taken together the results for hit rates and false alarm rates indicate that the ATR system changed the decision criterion of the analysts instead of affecting their capability. This claim was supported by the d' analysis. No main effect for the d' differences as a function of the ATR reliability level was found. Namely, for all three reliability levels the d' differences were close to 0 (see Table 1).

Table 1: Hit rates, False alarm rates, and d' as a function of ATR reliability.

Reliability	Hit rate			False alarm rate			d'		
	ATR	No-ATR	Diff	ATR	No-ATR	Diff	ATR	No-ATR	Diff
80%	0.71	0.576	0.134	0.273	0.177	0.096	1.789	1.581	-0.208
50%	0.686	0.691	-0.0046	0.233	0.236	-0.003	1.427	1.489	0.062
33%	0.524	0.659	-0.135	0.208	0.258	-0.049	1.327	1.076	-0.251

DISCUSSION

ATR reliability affected the hit rates and the false alarm rates, raising them when it was high and lowering them when it was low. These findings indicate that ATR reliability influenced the analysts' criterion for marking an item as a target, making it more lax when the reliability was high and stricter when the reliability was low. On the face of it these findings appear to indicate that the analysts' ability to identify targets does not change with changes in the ATR reliability. While this might be the case, it can also be argued that the changes in the analyst's criteria on the basis of reliability can be of benefit.

Consider a more sophisticated ATR system that is capable of informing the analyst of the specific reliabilities, or levels of confidence, of each item in the image. Presumably the levels of confidence assigned by the ATR will be correlated with the quality of the specific item. In other words, we would assume that the images that receive higher ATR confidence levels will have a higher probability of being true targets. In such a case, the fact that the analysts viewing these high ATR confidence images lower their criterion cut-off point; i.e., become more lax in their willingness to call the item a target, will yield superior performance. There will be an increase in the hit rates without the concomitant rise in the false alarm rates, which is what we are seeking. Thus, the fact that the participant analysts tend to take the ATR confidence levels seriously and change their willingness to accept the ATR designations when the confidence is high will improve actual performance.

Davidson, H., & Wickens, C. D. (1999). *Rotorcraft hazard cueing: The effect of attention and trust.* (Technical Report ARL-99-5/NASA-99-1). Savoy, IL: University of Illinois, Institute of Aviation, Aviation Research Laboratory.

Kuperman, G. G. & Sobel, A.L. (1993). Systems engineering/human factors approach to human machine interface design for aided target acquisition. *Technical Report, United States Department of Energy (DE)*, 22p.

Parasuraman, R. & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230-253.

Riley, V. (1996). Operator reliance on automation: Theory and data. In Parasuraman & Moulou (Eds.), *Automation and Human Performance* (pp. 19-35). Mahwah, NJ: Lawrence Erlbaum Associates.

Wickens, C. D. (2001). Attention to safety and the psychology of surprise. Presented at the *11th International Symposium on Aviation Psychology*, Columbus, OH: The Ohaio State University.