COGNITIVE ASPECTS OF AUTOMATED TARGET RECOGNITION INTERFACE DESIGN: AN EXPERIMENTAL ANALYSIS

Prepared by:

Marvin S. Cohen, Bryan Thompson, and Jared T. Freeman

Cognitive Technologies, Inc.
4200 Lorcom Lane
Arlington, VA 22207
(703) 524-4331

FINAL TECHNICAL REPORT

14 November 1997

i

ABSTRACT

The focus of this research was on the intersection of cognitive and perceptual aspects of human target recognition performance, and on potential enhancements of the human-ATR interface. Three series of experiments were conducted with active duty Army pilots. Each study attempted to lay a scientific basis, and to test a practical methodology, for a promising ATR design application. The studies address the following issues in ATR-human interface design: (1) effective displays of target classification conclusions to support rapid verification and application to the mission (2) effective displays of target imagery to support rapid and accurate user verification of ATR conclusions, and (3) effective support for decision making processes that allocate user attention, decide where and how long to verify ATR conclusions, and determine which targets to engage. Our results suggest that: (1) ATR conclusions should be labeled at different levels of specificity for different types of vehicles; (2) enhancement of vehicle profile and selected vehicle details can improve speed and accuracy of visual recognition; and (3) engagement decision making is improved by techniques for quickly guiding user attention to images classified as high-confidence enemies, high confidence friends, or significant and uncertain.

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

1. INTRODUCTION

The Problem

The pilot of an attack helicopter may emerge above the trees or hills for only a minute or two to assess the battlefield situation before remasking. During that brief period, he must collect accurate and relevant information about potential targets while minimizing his own exposure to attack. The pilot must be able to detect and discriminate friends and foes in a highly target-dense, rapidly changing, and visually and electronically noisy environment; he must classify targets that are relevant to his mission (e.g., tanks versus armored personnel carriers versus anti-air artillery), and prioritize them for attack. After remasking behind trees or terrain, he must decide whether to unmask again in a different location to collect more information, or to pop up to engage a target.

Target recognition has become a crucible of success on the battlefield not only for helicopters, but for virtually every weapons platform. The reason for its importance lies both in the recent evolution of U.S. war-fighting doctrine and in the development of new sensor and weapon technologies. Exploitation of U.S. night-fighting capabilities, for example, degrades the quality of optical information available for recognition decisions by both helicopters and tanks. Utilization of stealth technology and rapid maneuver tactics constrains the information obtainable from communications and from active sensors like radar (which would alert the enemy to one's own presence), while compressing the time in which recognition decisions must be made. At the same time, increased enemy mobility and speed and improved enemy sensor and weapon ranges may reduce the time available for recognition; and information denial techniques (such as camouflage, stealth, electronic countermeasures, and tactical deception) increase the uncertainty that such decisions must resolve. Nine U.S. soldiers and nine British soldiers were mistakenly killed by U.S. aircraft during Operation Desert Storm. But accurate target identification had become a major U.S. military concern well before the Persian Gulf War (see, for example, *Defense News*, March 25, 1991), and it is easy to imagine scenarios in which it would have played a much more crucial role in determining the success of battle.

Current Approaches

One approach to target recognition is the development of cooperative identification systems, e.g., distinctive visual markers for friendly vehicles, or systems for the electronic exchange of codes among friendly aircraft and vehicles. The limitations of these devices are obvious: In a sophisticated battlefield, they both expose friendly units to detection and identification by the enemy, and may be exploited by enemy units who mimic the friendly codes. Current research and development interest centers primarily on so-called non-cooperative target recognition (NCTR) systems.

It is natural to think of non-cooperative target recognition as primarily a sensory or perceptual problem. From this point of view, improvement in recognition accuracy and speed will come by providing more and better target information to the human operator: i.e., (a) improved sensors and analyzers (e.g., infrared, electro-optical image enhancement, synthetic aperture radar, laser radar, and others), and (b) improved cockpit display technologies, e.g., higher resolution, color, digital interactive displays with symbolic overlays, and so on. Current work on the user-computer interface is leading to even more dramatic input/output technology: e.g., spatial data management, natural language understanding, voice I/O, object-oriented direct-

manipulation interfaces, and multi-media three-dimensional virtual environments. The downside of this approach (taken by itself) is that it may leave the operator overloaded in environments where the sheer number of targets overwhelms his ability to detect and recognize them.

A third approach to target recognition is to automate the human's role in interpreting the sensor outputs. Automated Target Recognition (ATR) devices are under development which aim to automatically detect, track, classify, and prioritize objects of interest in a sensor image. In many cases the goal of such systems is to reduce the role of the human to that of a passive observer (Toms and Kuperman, 1991). Despite considerable progress, the performance of ATR systems has not yet been sufficient to achieve this goal. Current systems often fail under novel conditions (e.g., of weather, time of day, target aspect, or target configuration), under degraded observation conditions (e.g., low contrast or high clutter), or in the face of enemy countermeasures. In order to reduce the number of missed targets to an acceptable level, false alarm rates must often be set intolerably high, once again overloading the human operator.

A final approach regards ATR systems and human operators as partners. At least in the near future, humans will fill the gaps in ATR performance and ATR systems will relieve human workload. From this point of view, target recognition is only one phase or component of the human's overall task: For example, the human must also decide how long to remain exposed, where to look and with what sensors, whether to look again, when to fire, where to fire, and with what weapons. Designs for human-ATR interaction must take this entire complex of activities into account. It is to this approach that we now turn.

A Cognitive Approach

What ATR's automate is only one part of a far more complex process. While an ATR may provide classifications of targets within sensor range of the user's platform, human interaction with an ATR is not simply a matter of reading off those conclusions from the ATR's display. Figure 1 places the simple act of reading off an ATR conclusion within a larger nexus of tasks that must be performed if human-ATR interaction is to be efficient and effective.

An ATR classification and at the image it applies to (i.e., the two lower boxes in Figure 1) must be seen as part of a larger process. This larger process (i.e., the three boxes along the top of Figure 1) includes deciding which images to look at, and how long to look at them. That decision in turn will be based on an awareness of relevant contextual parameters, such as the importance of the mission and the cost of different kinds of errors (e.g., the danger of encountering friendlies), the cost of delay (e.g., the increase in risk of being targeted while taking time to verify ATR conclusions), and the user's overall confidence in the ATR's accuracy. Finally, the outcome of this activity is a decision or series of decisions regarding engagement of likely targets. This decision, like the earlier decisions (what images to look at and how long to verify ATR conclusions), involves a balance of different kinds of costs (e.g., the cost of missing an enemy target vs. the cost of mistakenly engaging a friendly).

Awareness of context

Mission (targets, presence of friendlies). Cost of delay. ATR accuracy.

Allocate attention among Images

Which images to look at. How long to look.

Make engagement decisions

Balance costs of different kinds of errors

Look at ATR Classification of Image

Look at Image

Extract visual features. Verify ATR conclusion.
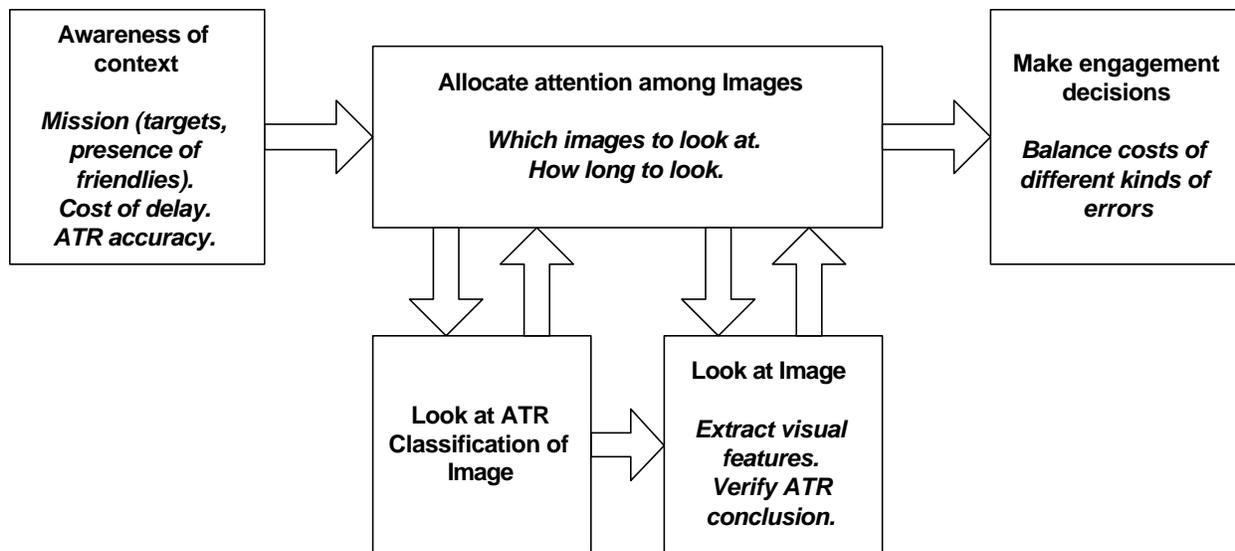
Figure 1. A schematic diagram of a human's interaction with an Automated Target Recognition (ATR) Device.

As Figure 1 demonstrates, improving the effectiveness of human-ATR interaction is not simply a matter of improving the accuracy of ATR conclusions. It will require an examination of all aspects of their interaction:

- effective displays of target classification conclusions,

- effective displays of target imagery to support user verification, and

- effective support for executive processes that allocate user attention, decide where and how long to verify ATR conclusions, and determine which targets to engage.

These three issues are the principle topics of the present research.

One morale of Figure 1 is that target recognition interweaves perceptual and cognitive components, and there is often no hard and fast separation between the two. A comprehensive approach must address both aspects, including how cognitive and decision making strategies exploit and direct perceptual processes. Thus, our research begins with what might be thought of as a cognitive issue – how users apply verbal categories to targets. Yet the process of verbal categorization is heavily influenced by underlying perceptual similarities, and verbal labels help to direct the ATR user's attention to relevant perceptual features. The next topic of research is the perceptual process itself, the sequence of features that is extracted by the visual system in order to classify a target. The final topic – how users handle targets about which the ATR is uncertain – focuses on decision making strategies for allocating attention among targets, in other words, the way in which cognitive processes direct perceptual resources.

Each topic has both a descriptive and a prescriptive (or applied) aspect, and the empirical studies are designed to address both, as far as possible. Figure 2 shows some of the potential prescriptive implications of the research, with respect to each component of the human-ATR interaction process. For example, we ask how users in fact verbally categorize targets, and explore the implications for how ATR's should label their conclusions. We ask how users perceptually process images, and explore the implications for how ATR's can visually enhance

images to facilitate user recognition. Finally, we ask how users allocate their attention among targets that vary in recognitional uncertainty, as a function of differences in the mission and in time stress, and explore the implications for how ATR's can direct user's attention where it is needed most.
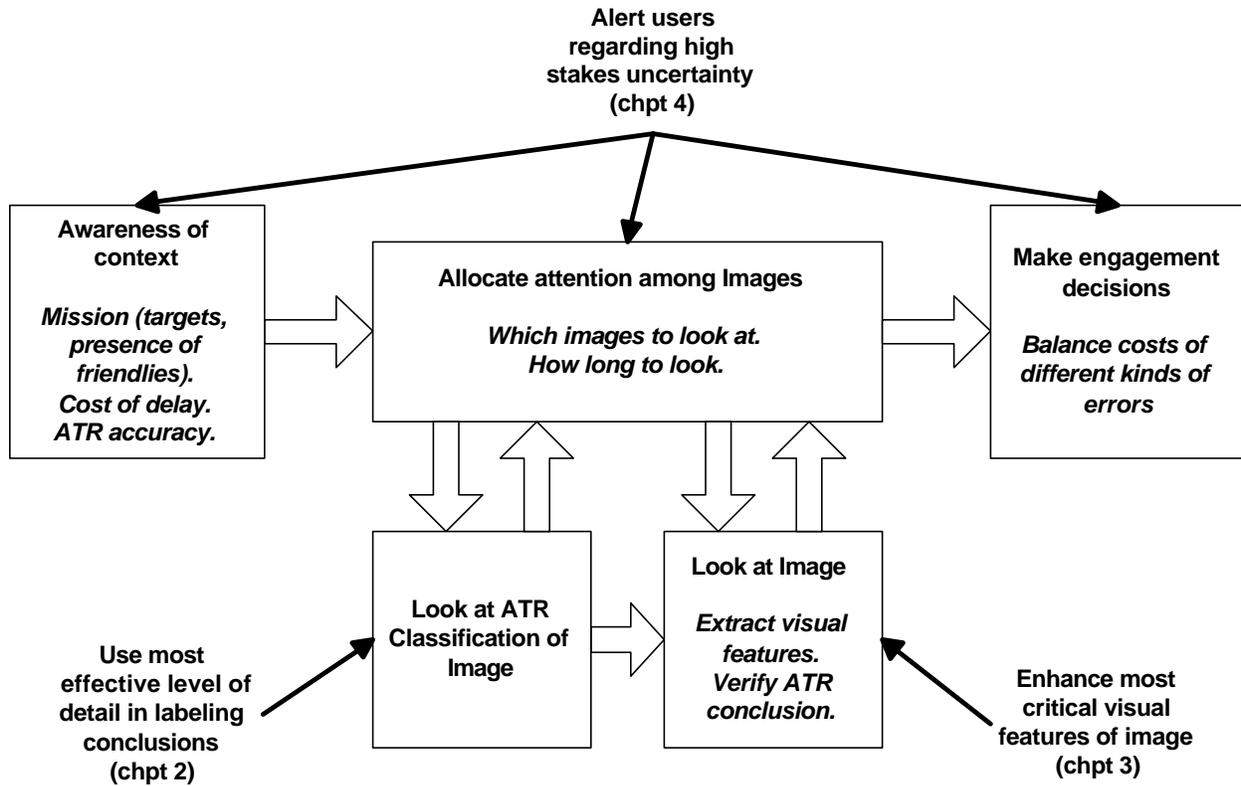
Figure 2. Questions for ATR design addressed by the present research.

More specifically, the present research addresses the following three sets of questions in the next three chapters:

1. *Verbal categorization:* What determines the label that operators apply to a target, i.e., the level of generality or specificity at which they choose to describe it? How do ATR operators represent the patterns of similarity and dissimilarity among different classes of targets and non-targets? How do different levels of categorization reflect this structure? What are the effects of typicality and familiarity? What are the implications for the way ATR conclusions should be reported?

2. *Visual processing:* How do operators represent and process visual data? When and how do they decompose images into parts? What is the sequence of features that they extract from a target image? Can ATR's enhance aspects of visual images that are key to human recognition, and thereby improve human recognition performance.

3. *Uncertainty handling:* How do operators handle uncertain conclusions? When do they attempt to resolve the uncertainty by collecting more data, and when do they simply accept an uncertain result? How do operators allocate their attention across a display, when targets vary in importance and uncertainty? What are the implications for the way

ATR's should report conclusions about which it is uncertain, and for how ATR's should direct user's attention in a display?

Certain common methodological principles have guided us in all three sets of studies:

- We have drawn on methods, theory, and results in cognitive and perceptual research.

- We have used active-duty helicopter pilots, with experience in target recognition, as participants in the studies.

- We have designed imagery materials and conditions to resemble actual target recognition events as much as possible.

- We have extracted implications for Automated Target Recognition (ATR) interface design, including both specific design recommendations and general methodological tools.

2. VERBAL KNOWLEDGE OF TARGETS

Introduction

This chapter explores the verbal organization of knowledge about targets. We begin with a deceptively simple question: How should an Automated Target Recognition device describe a target's identity? For example, should it refer to a particular target as a tracked vehicle, a tank, or as a T-62? Should it refer to another target as a wheeled vehicle, a truck, or a KRAZ? Most importantly, does the selection of a name *matter*, in terms of the pilot's efficiency in retrieving relevant perceptual and non-perceptual information about the target, and thus in verifying and acting upon an ATR conclusion? If it does matter, what empirical methods can be used to guide interface design?[1] A second, related question emerges at the end of the chapter: How is a pilot's verbalizable knowledge about targets organized? What kind of knowledge structure is reflected in the similarity relationships extracted from verbalized features? Answers to this question will help guide our exploration of visual processing in the next section.

Recent research findings on how people verbally categorize objects may be relevant to the way in which automated recognition systems can most effectively display their own conclusions to human users. A starting point for this research is a hypothesis regarding the purpose of categorization: *we categorize objects in order to enhance prediction and control*. If an object can be categorized based on observation of only some of its features, predictions can be made regarding additional features of the object and appropriate responses. Predictions might also be made regarding the features of other objects of the same type that may be encountered in the future. The usefulness of categorization thus depends on objects clustering in terms of shared features (Anderson, 1990) -- or, on correlational structure among features across objects. Categories tend to reflect this correlational structure, grouping together objects that share many features and distinguishing objects that share few features (Rosch et al., 1976).

If prediction is the purpose of categorization, then people should be more likely to use categories that lead to better predictions. Often categories can be arranged hierarchically: e.g., tracked vehicle, tank, T-62. Rosch et al. (1976) proposed that in everyday hierarchies (such as furniture, chair, armchair; animal, dog, collie) there is often an "intermediate" level at which prediction is most effective. Rosch referred to this as the *basic* level. Suppose people are asked to list the features that they associate with various categories. Objects that belong to the same basic-level category (e.g., chairs, tables, refrigerators; dogs, cats, mice) share many common features. Most of these properties disappear, however, at the more abstract levels in such a hierarchy (e.g., furniture; animal), where exemplars have far fewer common properties. On the other hand, moving to a more specific level does not produce a comparable increase in features. Most of the properties that collies share are also shared with other dogs.

A similar phenomenon appears to occur in hierarchies of battlefield classifications. Knowing that something is a tank tells an operator much more than knowing that it is a vehicle; but knowing that it is a T-62 may add relatively little additional information. *Tank* is thus likely to be a basic-level category in this hierarchy. (Exceptions will be noted below; for example,

---

[1] In this study, we assume the ATR is relatively certain in its identification of the target. In study 3, we look at the question of labeling when the ATR is uncertain regarding its conclusion.

where the task of discriminating friend from foe depends on a more detailed classification.) The basic level concept is illustrated in Figure 3, which depicts the increase in the number of features subjects recall concerning an object denoted with names of increasing specificity. The "knee" in the curve occurs at an intermediate level.
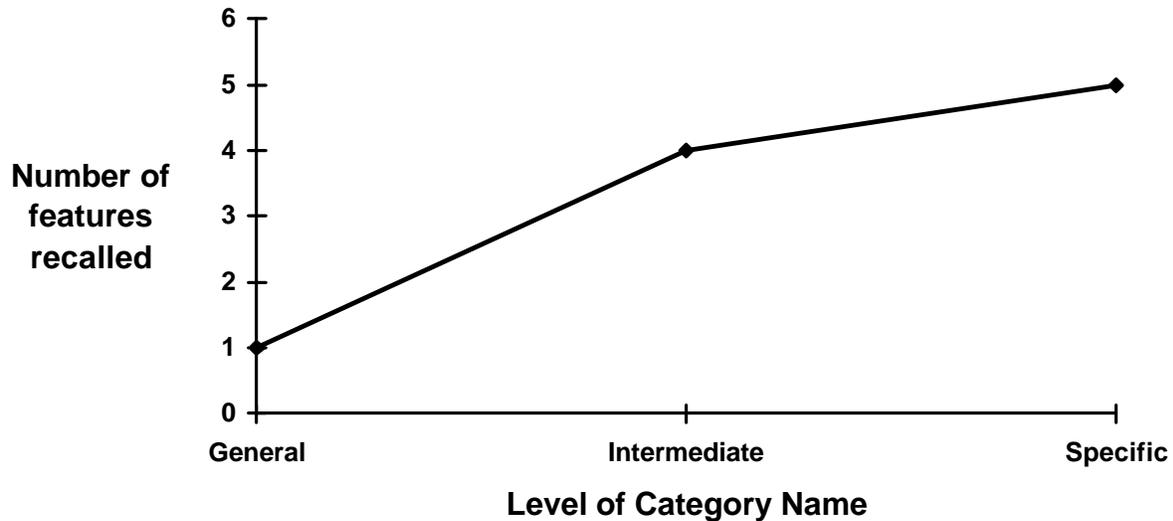
Figure 3 As the label for an object becomes more specific, the number of features associated with it may grow at a decreasing rate.

Informativeness is, of course, maximized with the most specific categories, which are associated with the most features. But there is a huge increase in the number of categories (e.g., the vast number of specific models of tanks) required to achieve rather small gains in the number of features. Rosch and her colleagues account for basic-level categories in terms of a balance between informativeness (many common features within the category) and effort (using as few category distinctions as possible). Corter and Gluck (1992) formalize these ideas in terms of a measure that trades off the predictability of features from the category label (informativeness) against the predictability of the category label from features (effort). Efficiency in naming implies maximizing informativeness for the effort expended. The most efficient category level is likely to be that intermediate level at which a "knee" occurs.[2]

Experimental data (Rosch et al., 1976, and others) confirm that people are more likely to use basic-level categories in spontaneous naming of objects. They are also faster in verifying that an object belongs to a basic category than to a more abstract or more specific category. (For

---

[2] The abscissa in Figure 3 appears to be merely ordinal, so it might seem inappropriate to plot these data as line charts rather than bar graphs, or to use the term "knee" to refer to their "negatively accelerated" shape. There is both a pragmatic and a theoretical reason for plotting such data as lines, however. In speaking of a "knee," we are referring to the difference between the change from general to intermediate and the change from intermediate to specific. This is, of course, analogous to "slope," and is far more easily visualized by a line graph than by a bar graph. The use of a line chart may also be theoretically meaningful. The "distance" between general and intermediate, and the "distance" between intermediate and specific, are not completely arbitrary. They are determined by the empirical fact that there are a finite number of short category labels available to describe a given domain, and that they can be arranged in a partial order in terms of inclusion relationships.

example, verifying that a picture of a robin is correctly described as *bird* is faster than verifying that the picture is correctly described as *robin* or *animal.*) Basic-level category names are the earliest to be acquired by children learning language (Anglin, 1983). Finally, experiments on learning to use artificially created objects and category labels confirm that the correlational structure of features influences the labels that subjects tend to acquire first and to use (Corter and Gluck, 1992).

Other observations suggest that the notion of a single, fixed basic level is oversimplified. Rosch et al. (1976) themselves had noted that the basic level depends on expertise; increasing familiarity with a domain (i.e., knowledge of more features) can shift the basic level to more specific categories. It is also likely that specific tasks requiring distinctions at a more refined level can shift the basic level to more specific categories (Cruse, 1977).

Joliceur, Gluck, and Kosslyn (1984) raised a more fundamental problem. They found that verification that an object belongs to a basic-level category is fast only if the object is a *typical* member of that category. Unrepresentative exemplars are more quickly identified by a more specific category name. For example, a robin is a typical bird, and is quickly verified to be a *bird.* But an ostrich is not a typical bird, and it is faster to verify that an ostrich is an *ostrich* than that an ostrich is a *bird*. Instead of a single basic level in each hierarchy of classification terms, each *object* may have its own favored level of categorization.

This result forces a revision of the basic-level concept, but it supports the original hypothesis, that categories are selected for efficient prediction. An atypical member of a category (such as an ostrich) does not share as many features with other members of the category (e.g., birds) as do typical members (robins, sparrows). For that reason, categorizing an ostrich as a *bird* is less useful than categorizing a robin as a *bird*. Calling an ostrich *bird* does not as support as many inferences from some features of the ostrich to other features of the ostrich. And learning about ostriches does not tell us as much about other birds as learning about robins.

In addition to atypicality, differences in familiarity and task relevance may also lead to the simultaneous use of different levels of categorization for members of the same general class. For example, people identify their acquaintances and colleagues (whose behavior they try to understand and predict in detail) by their individual names, while classifying others in more general terms, e.g., as customers or suppliers, or even more generally, as men and women.

Some authors have stressed the importance of shape for basic-level categories. Barsalou (1991) argues (a) that shape is more rapidly extracted during visual processing than other feature information, and (b) that most members of a basic-level category (e.g., birds) have the same shape. Members of higher level categories (e.g., animals) have many diverse shapes, while more specific categories (e.g., sparrow, robin) are associated with more detailed visual information than simply shape. This account links basic-level concepts to fixed features of visual processing. But it does not accommodate the apparent role of tasks and familiarity in determining the favored level of categorization. It does not explain why categorization may stop at the basic level (i.e., a shape-based categorization) in some cases, but proceed to a more specific level in others. More recently, Barsalou (1992) has suggested a compromise view: The immediate categorization response to an object may be determined by shape, while subsequent categorization is determined by informativeness and efficiency.

In what follows, the *basic* level of categorization refers to a level in a hierarchy of increasingly specific terms, after which there is a decrease in the rate of increase in the number

of associated features. The basic level is generally (but not necessarily always) preferred across all objects described by the hierarchy. However, we will use the term *favored level of categorization* to refer to the level that is preferred for a *particular object* at a particular time. Based on the research in this area, we expect the favored level for an object to be the same as the basic level for the object's category hierarchy unless: (1) the object is atypical of its category (as a Harrier, with its vertical takeoff capability, is atypical of aircraft), (2) the object is significantly more familiar than other objects in its hierarchy (as an F14 is highly familiar to military pilots), or (3) the task requires a more detailed prediction (as in the need to identify a specific type of jeep in which an enemy commander is thought to be riding). In these three cases, the favored level will tend to be more specific than the basic level. On the other hand, if a particular object is highly unfamiliar compared to other objects in the same class, the favored level for that object may be more general than the basic level for that hierarchy.

Implications for Display of ATR Conclusions. Should ATR's display recognition conclusions at the favored level for a particular object? We have mentioned research results on basic level categories, in which time to verify the application of a label to an image is faster for basic level terms. These results suggests advantages of basic (and possibly favored) level terms in ATR labeling.

We can imagine two strategies that users of an ATR might use in consulting and verifying its conclusions. (1) Users might look at the ATR label for an object first and then look at the image to verify the ATR's classification. After looking at the ATR label, the user must generate an image of the expected appearance of the object in order to compare it with the actual image. Verification will thus be faster for labels that are strongly associated with typical images of the object. (2) Alternatively, users might first look at the image in order to form their own conclusion, and then look at the ATR label for confirmation. In this case, users must generate a label from the image, and then compare that label with the label displayed by the ATR. Again, verification will be faster for labels that are strongly associated with images (this time, in the image-to-label direction). The verification task will thus serve as a test of the usefulness of ATR labeling at the favored level.

Method for All Experimental Tasks

These experiments were designed to determine the level of specificity with which an ATR should label the conclusions of sensor processing. We wished to learn, for example, whether an object identified by the ATR as a T-62 be labeled T-62 or tank? We hypothesize that the optimal level of specificity at which to label an object is a function of four factors: the efficiency of the name in aiding feature retrieval, the typicality of the object relative to the class of objects the name denotes, the familiarity of the named object and the task in which the named object is considered.

Design

We examined three levels of labeling for ten military vehicles (three tanks, three APC's, two trucks and two jeeps). Several experimental tasks provided data concerning three determinants of basic and preferred label levels discussed above: features, typicality and familiarity. (The influence of the task, or mission, was varied in a later series of experiments.) In the feature naming task, subjects generated a list of features they associated with a given vehicle label. Results of this task were used to specify the basic level of naming. In the familiarity task,

subjects rated their familiarity with labels and images of vehicles. In the typicality task, they rated the degree to which a label was representative of an image. Results of these tasks and the feature naming task were used to identify the preferred level of naming.

Convergent validation concerning the preferred label level was achieved using two additional tasks. In one of these tasks, we elicited the label with which subjects spontaneously named an object and we then tested to determine whether that name corresponded to the hypothetical preferred name. In the second task, subjects verified whether an object image and an object label matched. We then evaluated whether response times were fastest for the hypothetical preferred name The logic of this experimental series is illustrated in Figure 4.
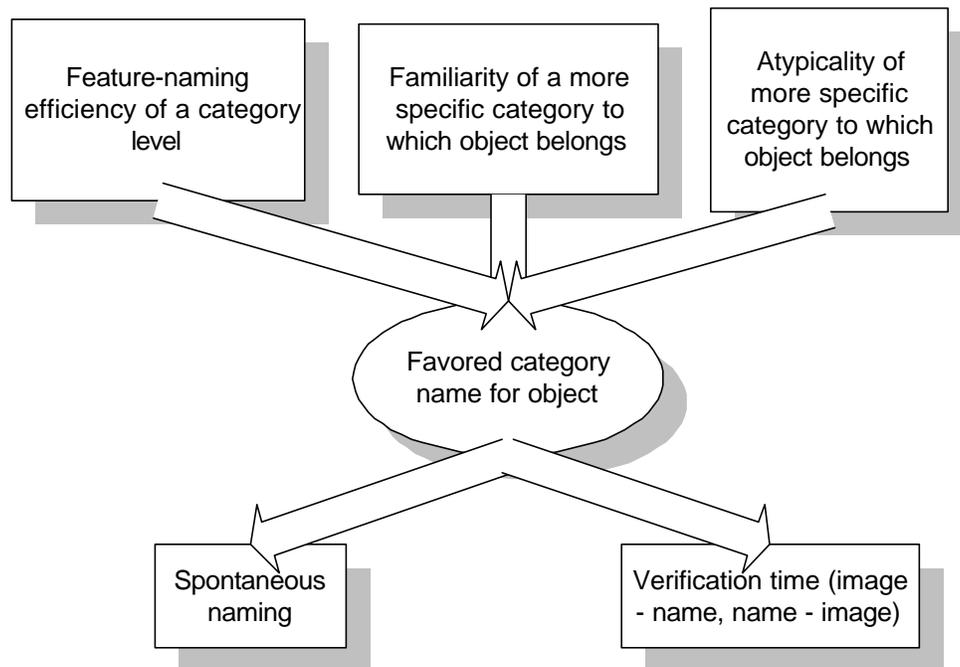
Figure 4: Five experimental tasks (in rectangles) provide convergent validity for the favored name.

The order of tasks was counterbalanced between subjects. The presentation of stimuli within each task was randomized by subject. Each subject performed most or all of the experimental tasks.

Participants

Data were collected from 17 active duty U.S. Army helicopter crewmen at Ft. Bragg, North Carolina. All subjects held the rank of Warrant Officers, class II. The officers had 9.6 years experience each of Army service, including major exercises and, in some cases, military campaigns. Subjects participated under orders from their commanding officers.

Materials

The images used in the experimental tasks were side views of ten vehicles: three tanks, three APC's, two trucks and two jeeps (see Appendix A for all experimental images). Images roughly filled a field of the screen that was 2 3/4" wide by 1 1/4" high..

General

Tracked Vehicle	Wheeled Vehicle

Intermediate

Tank	APC	Truck	Jeep

Specific

T-62
(CIS main battle tank)

BMP
(CIS infantry combat vehicle with gun)

BRDM
(CIS amphibious scout car, no gun)

ZIL
(CIS)

UAZ-469
(CIS)

T-55
(CIS main battle tank)

BTR
(CIS personnel carrier with gun)

KRAZ
(CIS)

**M-151**
(USA)

M-551
(USA light tank)
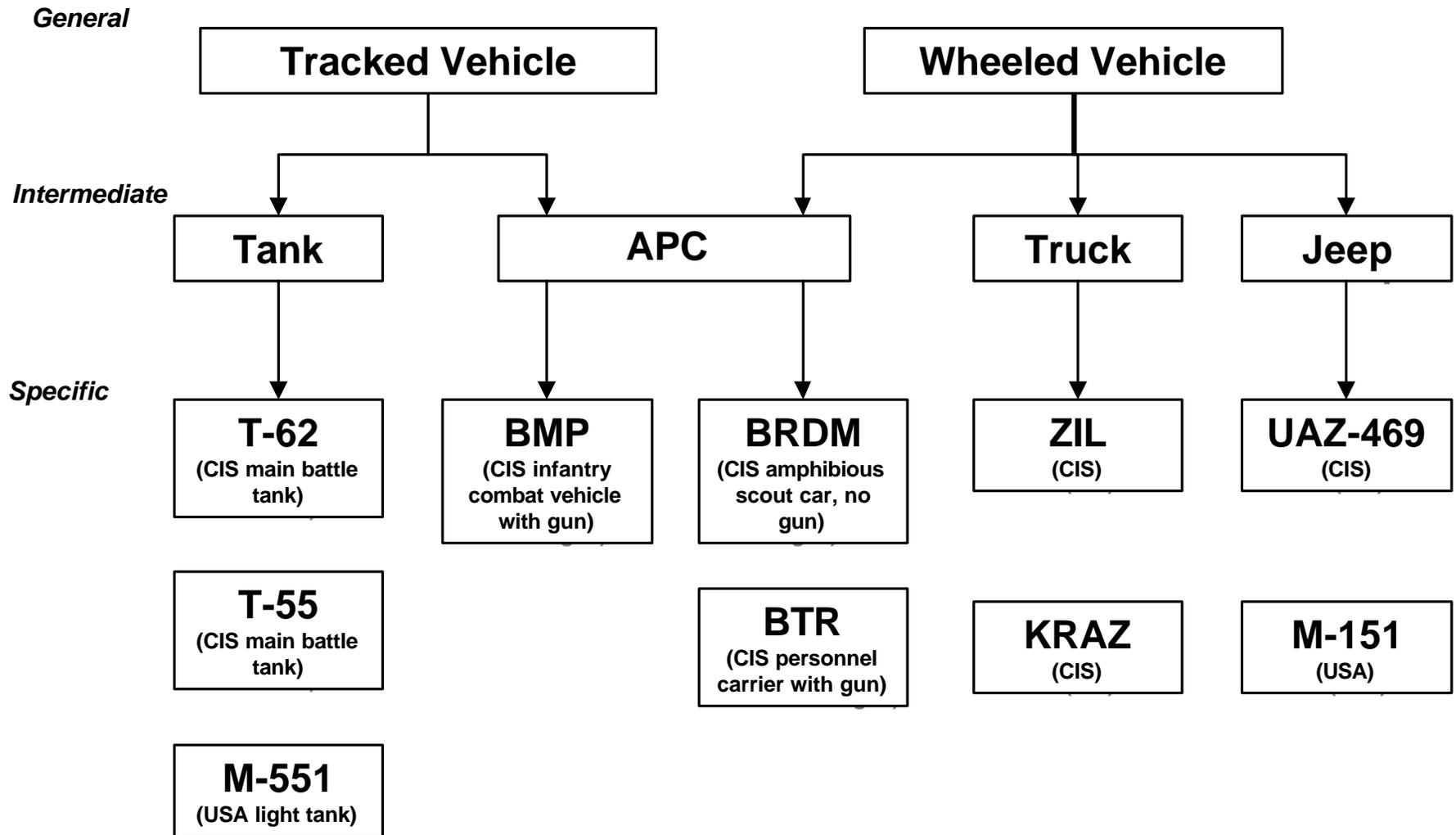
Figure 5. General, intermediate, and specific experimental labels used in the study of feature naming. (CIS indicates the Commonwealth of Independent States, the former USSR.)

Sixteen labels were also used as stimuli, as shown in Figure 5. The labels were selected so that each vehicle could be described by three labels of varied specificity. The most general labels were *tracked vehicle* and "wheeled vehicle." The labels of intermediate specificity were *tank*, *APC*, *truck* and "jeep." The most specific labels were the vehicle names (e.g., *UAZ*, *BMP*)

Apparatus

Stimuli for this experimental series were presented to subjects on a personal computer with a 90MHz Pentium processor (DEC XL 590) running the NeXTSTEP version of the UNIX operating system. CTI developed the software for administering experimental tasks on this platform. The system was capable of displaying static images of vehicles overlaid with dynamic, simulated FLIR noise (at more than 30 frames per second. However, static FLIR noise was used in the present set of experiments[3]. It could display textual stimuli (such as vehicle labels), textual instructions and textual and graphical prompts. The system also maintained detailed records of the experimental tasks executed by each subject, as well as response time data at a level of several milliseconds accuracy.

Procedure

Each subject individually performed the experimental tasks in a quiet classroom. The subject was informed that the study was intended to provide data for the design of Automated Target Recognition systems. He then completed a biographical questionnaire.

For each experimental task, the subject read instructions from the system display, after which the experimenter answered any questions. The subject then practiced the task, using images of common animals and corresponding labels, until he felt he understood the task and was proficient with the interface. The experimental task then commenced, using the stimuli described above. In the following, we provide details concerning the procedures for each task and the results for that task.

Specific Studies

Feature naming

Procedure

In the feature naming task, the subject viewed a vehicle label on the computer display and then typed a list of features associated with it. Subjects had unlimited time and free-text display space to generate each list of features. Each of the 16 labels was presented once in random order. All 17 subjects completed this task.

---

[3] FLIR noise was generated using NVSIM, a UNIX-based Thermal Imaging System Simulator developed by John Horger at the U.S. Army Night Vision & Electro-Optics Laboratory (NVEOL), Ft. Belvoir, VA. Vehicle images were produced by Mr. Horger in collaboration with Robert Lafollette and were provided by Dr. Barbara O'Kane, also of NVEOL.

Analysis

Of the features named by pilots, answers such as "no knowledge" and "unfamiliar" were removed from the data set. All other responses were retained, and were standardized using a simple attribute: value syntax.

Appendix B contains a list of all features attributed by any pilot to a label at any level of specificity. The object of this study was to examine the efficiency of different labels in conveying information about the objects named. Therefore, all features associated with a given label were considered part of the information it conveyed. The list in Appendix B thus includes features that may be visible on FLIR scope (e.g., turret shape: round, and brake temperature: hot) as well as non-visual features (e.g., handling: hard to drive, and mission: recon).Included in the latter category were responses that named the possible class(es) of the vehicle (e.g., vehicle type: APC or tank). (We shall focus on visible features only when we look at the process of visual recognition at the end of this chapter and in the next chapter).

The list in Appendix B specifies the features actually mentioned by pilots in response to a label. However, it is likely that information not mentioned by pilots may also be implicitly communicated by a label. One class of implicit information can be discovered by examining features associated with more general labels. For the analysis of informational efficiency, we assumed that more specific labels "inherited" information associated with more general labels that included them. Thus, for each subject, we added features named in response to general labels to the lists the subject generated for intermediate and specific labels; similarly, we added features named in response to intermediate and general labels to the lists made for specific labels.[4] If features were named at both levels, redundancies were eliminated. Inheritance of this sort is further justified by the results of the typicality rating task (described below), in which all images of vehicles were rated as typical of all the labels, at whatever level of generality, that correctly applied to them.

Results

Because of the analysis described above, it was mathematically necessary that at least as many features would be recalled for more specific labels as for levels above them in the hierarchy. However, it was not mathematically necessary that the increase in number of features, if any, would be larger going from general to intermediate labels than from intermediate to specific labels. In fact, the analysis, in which more specific levels "inherited" features from more general levels, would tend to work against this predicted result. We shall refer to the effect of interest as a "knee," defined as the change in number of features going from general to intermediate labels less the change in number of features going from intermediate to specific labels.

At the mean, subjects listed 2.114 features in response to general labels, 4.245 for intermediate labels and 4.747 for specific labels. The increase from general to intermediate levels

---

[4] The labels used do not, of course, form a perfect hierarchy, so some decisions were made regarding how to inherit features of APC's, which can be either tracked or wheeled. Specific labels for APC's inherited features belonging to either *wheeled vehicle* or *tracked vehicle*, depending on which was appropriate. The intermediate level, *APC*, inherited a fraction of the features from both *wheeled vehicle* and *tracked vehicle*, corresponding to the proportion of APC's that were wheeled or tracked, respectively.

was 2.131 (= 4.245 - 2.114); but the increase was only 0.502 for the difference between the intermediate and specific levels of labeling.

There was a knee in the increase in the number of features per label level for every individual vehicle type, as shown in Figure 6. The effect was significant at p < .05 in paired two-tailed t-tests for every vehicle except the T62 (for which it approached significance: $t_{13} = 1.681$, $p = 0.117$).

As can be seen in the upper left plot in Figure 6, the size of the knee effect varied as a function of the type of vehicle involved. The knee was smallest for the three tanks (M551 = 0.349, T62 = 0.474, T55 = 0.599) and largest for the three APC's (BMP = 1.932, BTR = 3.403, BRDM = 3.932). Jeeps and trucks were between these extremes, with the knee for trucks (ZIL 0.911 and KRAZ = 1.051) somewhat smaller than the knee for jeeps (M151 = 1.36 and UAZ-469 = 1.83).

When data were aggregated for the intermediate categories, the knee effect was statistically significant for each one: for APC's, $t_{11} = 6.358$, $p < .001$; for tanks, $t_{13} = 3.015$, $p = .010$; for trucks and jeeps, $t_{13} = 4.527$, $p = .001$). The size of the effect, as already noted, was largest for APC's, smallest for tanks, and intermediate for jeeps and trucks.
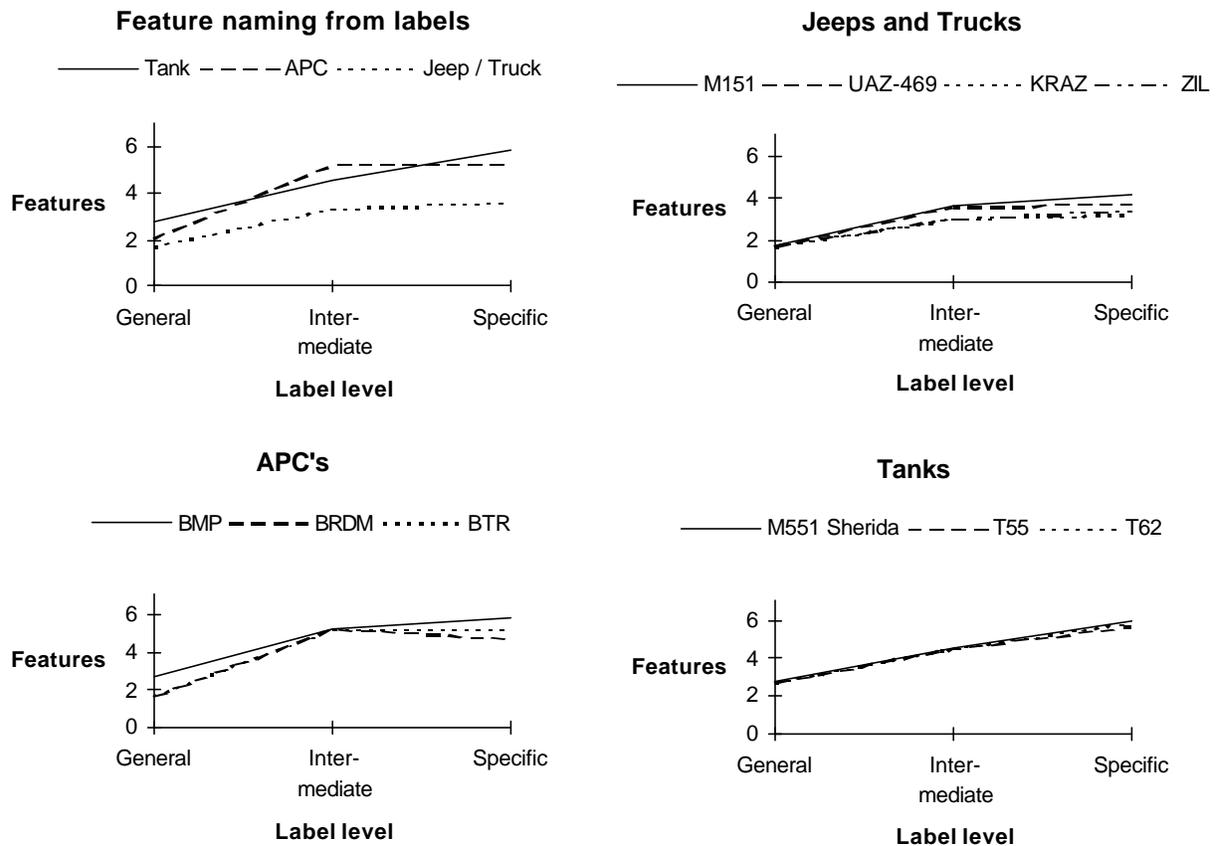

Figure 6: Increase in the number of features recalled as a function of specificity of the label. The upper left figure aggregates over intermediate categories. All other figures show data for specific vehicle types.

The summary plot in the upper left of Figure 6 also suggests that there were systematic differences in the absolute number of features named for the different vehicle types (APC, tank, and jeep / truck). This was indeed the case. In response to labels at the intermediate level (*APC, tank, jeep,* and *truck*) officers cited more features for APC's overall (mean = 5.375) than for tanks overall (4.500) (t_{15} = 3.656, p = 0.002) and more for tanks than for jeeps and trucks (3.313) (t_{15} = 3.09, p = 0.007). At the level of specific labels, fewer features were generated for APC's (5.354) than for tanks (5.813), but this difference was not reliable (t_{15} = 1.418, p = 0.177). However, the number of features cited for jeeps and trucks (3.598) was reliably lower than the number given for APC's (t_{16} = 4.116, p = 0.001) or tanks (t_{15} = 4.667, $p <$ 0.001). In addition, the label *tracked vehicle* reliably elicited more features (2.917) than did "wheeled vehicle (1.667) (t_{11} = 2.803, p = 0.017).

In sum, there was a decreasing benefit in number of features recalled as labels became more specific. The curves for all vehicles had a knee at the intermediate level. The effect was strongest for APC's and stronger for jeeps and trucks than for tanks. Finally, subjects generated more features in response to labels for APC's and tanks than for jeeps and trucks.

Familiarity

Procedure

In two tasks designed to elicit ratings of familiarity with specific vehicles, 16 subjects viewed either a specific label or an image of a vehicle and rated their familiarity with that vehicle on a scale ranging from "1 – extremely unfamiliar" to "9 – highly familiar."

Results

When specific labels were used as stimuli, participants indicated that they were highly familiar with APC's overall (mean = 8.500) and tanks (8.083), though this small difference in mean familiarity by intermediate vehicle type was reliable (t_{15} = 2.331, p = 0.034). Subjects gave much lower familiarity ratings for jeeps and trucks (3.125) than for either APC's (t_{15} = 15.768, $p <$ .001) or tanks (t_{15} = 14.768, $p <$ .001). The M151 was the most familiar member of the family of jeeps and trucks. (See Figure 7).

Familiarity with labels

Jeeps and Trucks


APC's

Tanks
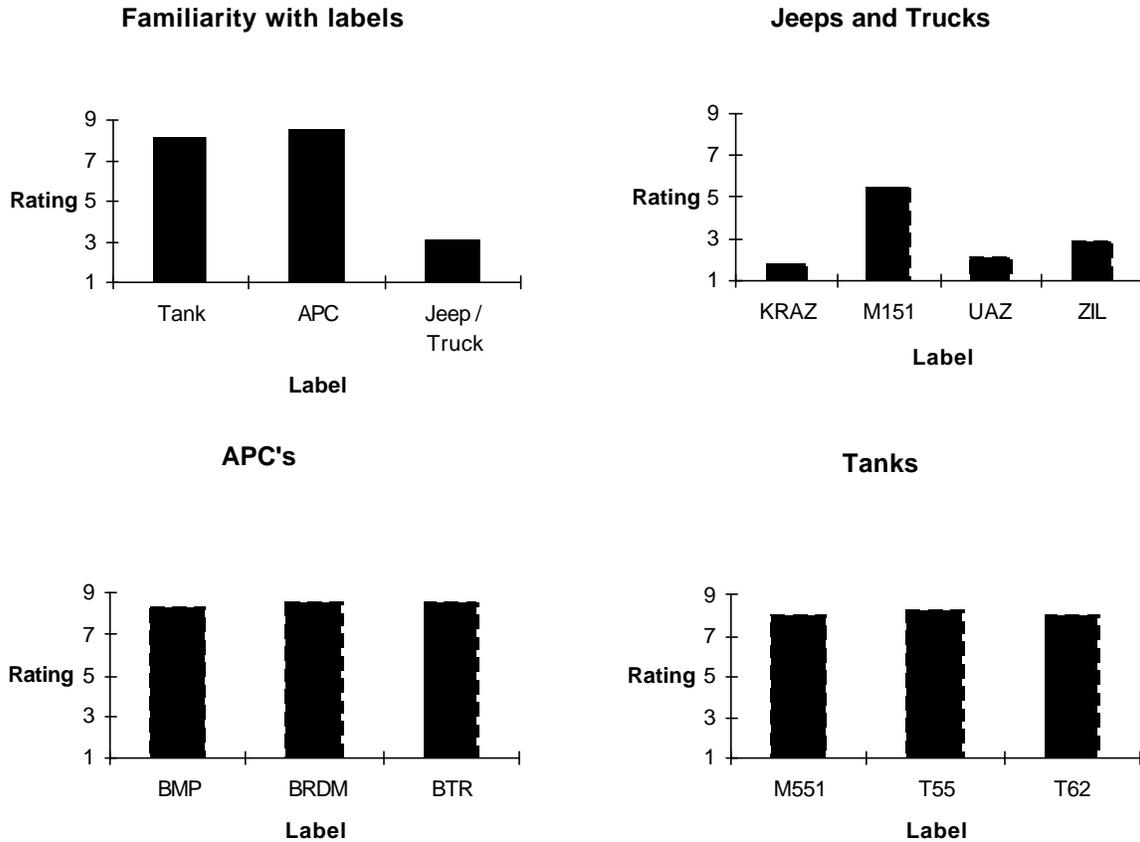
Figure 7: Ratings of familiarity in response to specific labels of vehicles. The upper left figure aggregates across intermediate categories, while other figures show data for specific vehicles.

As shown in Figure 8, ratings of familiarity with images showed a similar pattern. Ratings were higher for APC's overall (mean = 8.375) than for tanks (6.812) (t_{15} = 4.457, $p <$ .001) or for jeeps or trucks: t_{15} = 5.278, $p < 0.001$). Familiarity ratings tended to be higher for tanks than for jeeps or trucks (5.948), though not reliably so (t_{15} = 1.312, $p = 0.209$). However the absolute level of ratings of familiarity with images of all vehicle was moderate to high. By contrast, familiarity ratings for labels were relatively low for jeeps and trucks.In sum, subjects were moderately to highly familiar with APC's and tanks, whether the stimuli were labels or images. Images of jeeps and trucks were moderately familiar, but the names UAZ (a jeep), KRAZ and ZIL (trucks) were not.

Familiarity with images

Jeeps and Trucks


APC's

Tanks
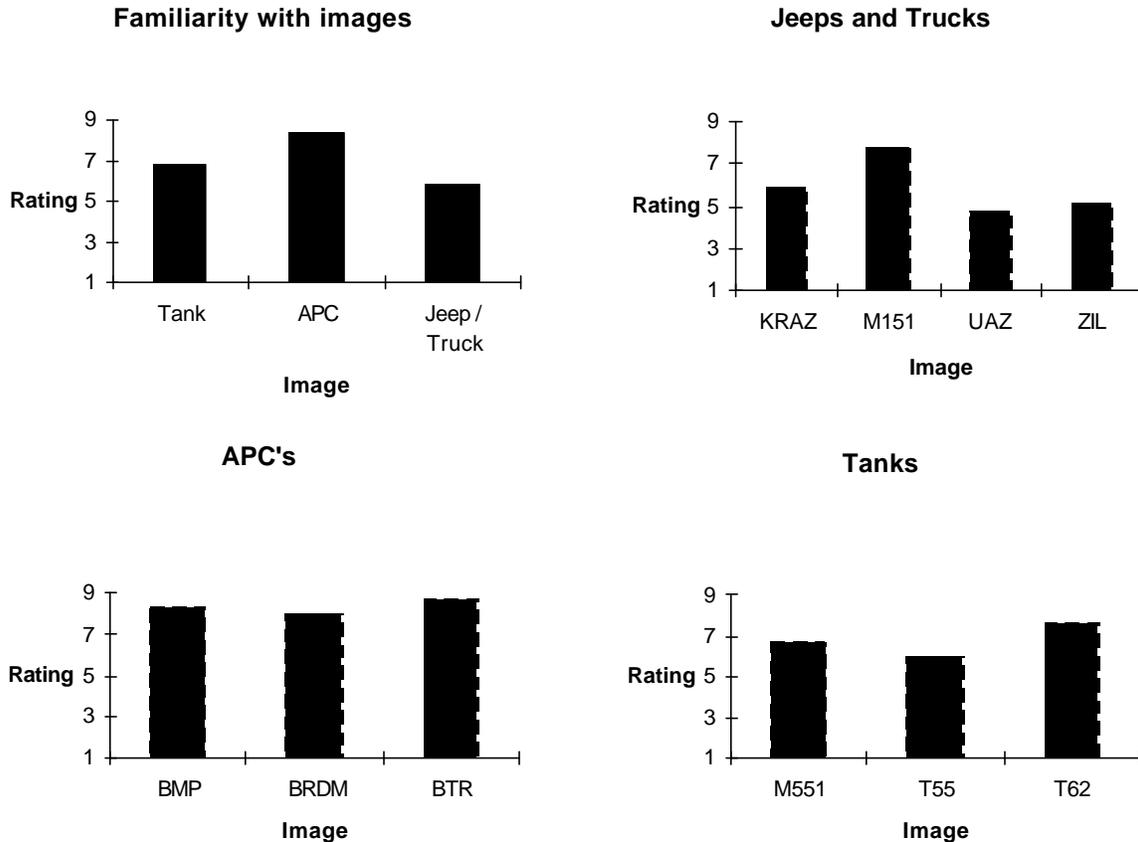
Figure 8: Familiarity ratings for images of vehicles.

Typicality

Procedure

In the typicality task, subjects viewed a vehicle label (general, intermediate or specific) for one second followed by an image of a vehicle. Subjects then rated the degree to which the image was typical of the label, using a nine-point scale defined as follows: "1 means not a typical," and "9 highly typical." All 16 labels were matched with all 10 images during the experimental task. However, only responses to correctly matched pairs of labels and images are analyzed below. (E.g., responses to an image of an M151 (jeep) paired with the label *tank* were discarded). Fifteen subjects performed this task.

Results

All images were rated as at least moderately typical of their labels, regardless of the level of label specificity. However, typicality ratings were lowest for the vehicles with which subjects were least familiar: the UAZ jeep (5.000), KRAZ (4.000) and ZIL (4.667) trucks and the T55 tank (5.133).

Subjects gave uniformly high typicality ratings to general labels paired with images of vehicles (mean = 7.953). Intermediate labels were judged less typical of images (7.403). Specific

labels elicited the lowest typicality ratings for tanks overall (6.555), and for jeeps and trucks (5.150). Specific labels elicited the highest typicality ratings for APC's (8.378). (See Figure 9).

The key finding from this task is that no images were atypical of their labels. There were reliable differences in typicality ratings for individual vehicles between label levels; however this is not immediately relevant[5].

Typicality (labels. then images)

——— Tank – – – APC · · · · · · Jeep / Truck

Jeeps and Trucks

——— M151 – – – UAZ · · · · · · KRAZ – · – · – ZIL

APC's

——— BMP – – – BRDM · · · · · · BTR

Tanks

——— M551 – – – T55 · · · · · · T62
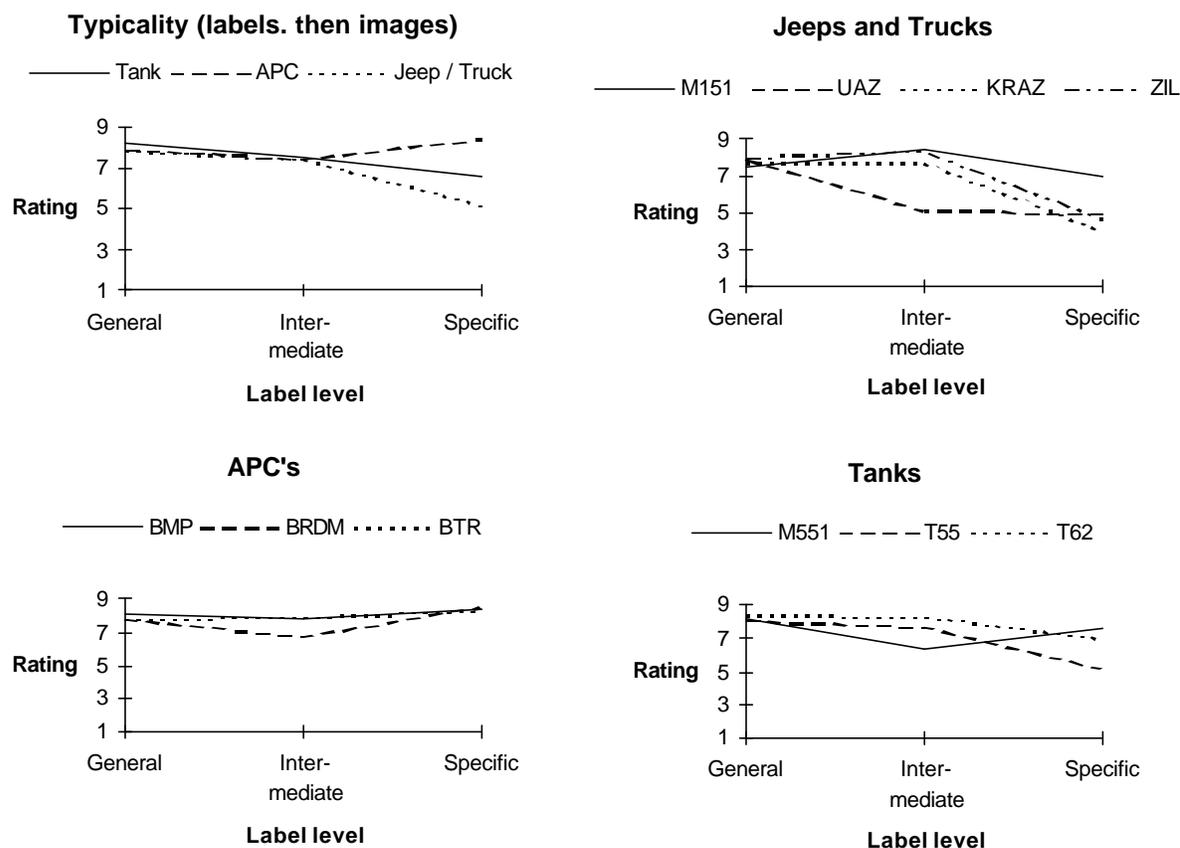
Figure 9: Ratings of how typical a vehicle image was with respect to correct labels of the vehicle designating classes at different levels of generality.

Discussion: Implications for Favored Level of Labeling

Two patterns of results emerged from the feature naming, familiarity rating, and typicality rating tasks: one in which the specific name was favored and the other in which the intermediate name was favored. (See Table 1.)

The feature naming data concerning APC's exhibited a distinct knee at the intermediate level of labels. Accordingly, the intermediate level is considered the basic naming level for these vehicles. Images of APC's were not atypical of general, intermediate or specific APC labels, and

---

[5] Subjects reliably judged the UAZ to be more typical of the general label *wheeled vehicle* than the intermediate label *jeep* ($t_{14} = 3.827, p = 0.002$). Subjects reliably rated images of several vehicles to be more typical of intermediate level names than specific ones. This was the case for the M151 ($t_{14} = 2.488, p = 0.026$), KRAZ ($t_{13} = 3.05, p = 0.009$), ZIL ($t_{14} = 3.325, p = 0.005$), T55 ($t_{14} = 2.385, p = 0.032$), and the T62 ($t_{14} = 1.887, p = 0.08$).

subjects rated all three APC's as highly familiar. (The validity of these ratings was supported by the relatively large number of features subjects associated with APC's). Under these conditions (distinct knee, no atypicality, high familiarity), it is the specific level of labeling for APC's that is favored according to the decision rules described on page 9.

The intermediate label for tanks (*tank*) was not necessarily the basic level. The increase in the number of features named between the general and intermediate levels roughly equaled the increase between the intermediate and specific levels. This produced a weakly articulated knee in the plot of frequency of feature names. No images of tanks were atypical of labels, and all of the tanks were highly familiar. Accordingly, the favored level of naming for tanks appears to be the specific level.

For jeeps and trucks, the intermediate label level (*jeep* or *truck*) was the basic level of naming, corresponding to the distinct knee in the graph of frequency of features. None of these vehicles was atypical of its label, specific labels for these vehicles were moderately familiar at best, and images of them were mostly moderately familiar. (The American M151 is arguably an exception). Thus, the basic level names at the intermediate level were also the favored names.

Table 1: Summary of findings for each of three experimental tasks.

Type of vehicle	Feature naming efficiency	Familiarity	Typicality	*Implications for favored level*
Jeep & Truck	Intermediate	Low	Medium or High	*Intermediate*
APC	Intermediate	High	High	*Specific*
Tank	Intermediate or specific	High	Medium or High	*Specific*

Verification

Procedure

The verification response task was meant to simulate an observer's attempt to verify an ATR' conclusion (i.e., a label specifying the classification of a target) by visually examining the image. In the verification task, the subject was asked to depress the left and right shift keys on the computer keyboard according to on-screen prompts at every trial. The system then presented the officer with a label for one second, followed by an image. The subject lifted one shift key to indicate that the label and image matched, or the other shift key to indicate that they did not match. (The key labels were counterbalanced between subjects. For half of all subjects the left key indicated a match and the right key a non-match. For the other half the labels were reversed). Subjects were asked to "answer quickly and accurately." Response times measures were accurate to several milliseconds. Each subject viewed all combinations of 10 images and 16 labels. All 17 subjects executed the task.

We also administered a verification task with the order of image and label reversed. In this task, participants viewed an image for one second, followed by a label, and lifted a key to

indicate match or mismatch. There were, however, insufficient data points in this task to support any statistical conclusions.

Results

Verification Response Times

Since more general categories (e.g., animal) convey less information than more specific categories (e.g., cat), they should be verified more quickly, other things being equal. Favored level categories, however, have an advantage in efficiency and/or familiarity, that may overcome this effect of information quantity. Our prediction was that favored labels would be faster, or at least no slower, to verify than more general category labels that were not favored. Based on the results of the previous studies (feature naming, familiarity, and typicality), we predicted that specific labels would be favored for APC's and tanks, and thus would be at least as fast as (rather than slower than) intermediate labels. We predicted that intermediate labels would be favored for jeeps and trucks, and thus would be at least as fast as (rather than slower than) general labels. Response times were analyzed for accurate responses to matching labels and images (i.e., "hits"). In the next section, we will address error rates.

As shown in Figure 10, there were different patterns of results for the different types of vehicles. For jeeps and trucks, response times were flat between general and intermediate labels, but rose significantly for specific labels. The difference between general and specific labels was marginally significant for only one of the four jeeps or trucks, the UAZ ($t_5 = 2.093$, $p = 0.091$). This difference was not significant when all jeeps and trucks were combined ($t_{15} = 1.036$, $p = 0.317$). On the other hand, the advantage of intermediate labels over specific labels was reliable or nearly so for both jeeps (M151: $t_5 = -3.202$, $p = 0.024$; UAZ: $t_1 = -7.957$, $p = 0.08$) and one of the two trucks (ZIL: $t_6 = -5.242$, $p = 0.002$). The advantage of intermediate over specific was also reliable when all jeeps and trucks were combined ($t_{16} = 2.950$, $p = 0.009$). Since intermediate labels covey more information, they would be expected, other things being equal, to require more time for verification than general labels. That this was not found is consistent with the prediction that intermediate labels are favored for jeeps and trucks.

For APC's, overall response times were rose slightly from general to intermediate and then flattened out between intermediate and specific. There was a significant increase in response times between general and intermediate labels for one APC, BRDM's ($t_{11} = 2.508$, $p = 0.029$). When all APC's are taken together, there was a non-significant trend for intermediate response times to be longer than general response times ($t_{14} = 1.742$, $p = 0.103$). On the other hand, there was no reliable difference in response times between intermediate and specific labels for any APC, or for all APC's together ($t_{13} = .496$, $p = 0.628$). Since the specific labels would be expected, other things being equal, to require more time to verify than the intermediate labels, this pattern of results is consistent with the prediction that specific response times are favored for APC's.

For tanks, as for jeeps and trucks, response times were flat between the general level, and the intermediate level, but rose for specific labels. The difference between general and intermediate general labels was non-significant for every tanks as well as when all tanks were combined ($t_{14} = -1.742$, $p = 0.103$). However, the advantage of intermediate over specific levels was reliable or nearly so for all types of tanks (T55: $t_6 = -4.489$, $p = 0.004$; T62: $t_{10} = -2.649$, $p = 0.024$; M551: $t_{12} = -1.504$, $p = 0.158$), and for all tanks together ($t_{16} = 3.225$, $p = 0.005$). Since

intermediate labels would be expected to take longer to verify than general labels, this result suggests that intermediate level labels were favored for tanks (as for jeeps and trucks). The result is not consistent with the prediction, based on pilot's claimed familiarity with specific tanks, that specific labels would be favored for tanks.
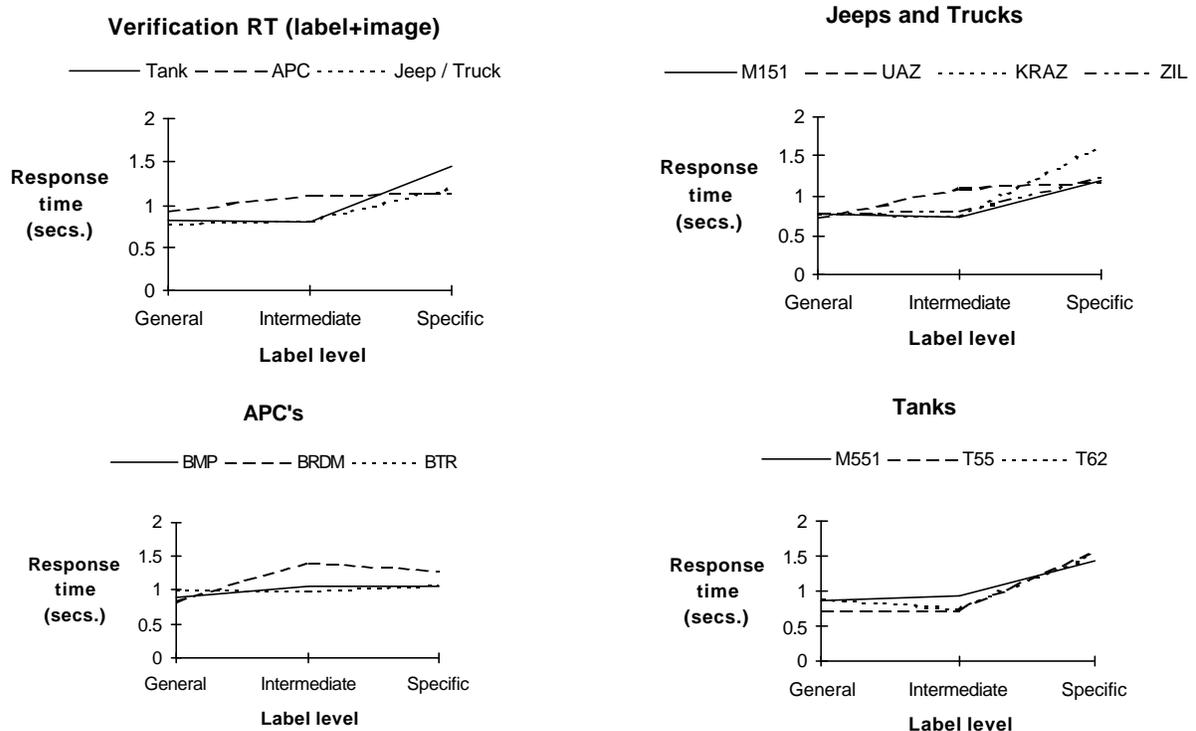

Figure 10: Response times to verify a match between a label and image. Upper left chart represents average times.

In sum, when subjects were presented with matched images and labels, response time did not increase monotonically as labels became more specific and thus conveyed more information. For jeeps, trucks, and tanks, participants were at least as fast to verify the match when intermediate labels were used as when general labels were used. For these types of vehicles, participants were faster for intermediate labels than when specific labels were used. Thus, for jeeps, trucks, and tanks the intermediate label was favored. For APC's, the pattern was different. There was a slight decrement in verification speed at intermediate compared to general labels, but no decrement in verification speed for specific labels compared to intermediate labels. Thus, the specific labels were favored for APC's. Response times for specific APC labels overall were faster than for specific labels for tanks and jeeps and trucks, but this difference was not statistically reliable.

Verification Error Rates

In the verification task, participants were told to respond both accurately and quickly. Thus, it is important to investigate accuracy in addition to response rates. It is possible that faster reaction times for favored labels arise from participants' willingness to accept higher error rates for those labels. Examination of error data, however, shows that this was not the case. As shown in Figure 11, results of the error analysis generally conformed to those from the response time analysis for this task.

Verification errors (label+image)

———— Tank – – – – APC ·········· Jeep / Truck

100%
80%
60%
% Misses 40%
20%
0%

General   Intermediate   Specific

Label level

Jeeps and Trucks

———— M151 – – – – UAZ ·········· KRAZ – · – · – ZIL

100%
80%
60%
% Misses 40%
20%
0%

General   Intermediate   Specific

Label level

APC's

———— BMP – – – – BRDM ·········· BTR

100%
80%
60%
% Misses 40%
20%
0%

General   Intermediate   Specific

Label level

Tanks

———— M551 – – – – T55 ·········· T62

100%
80%
60%
% Misses 40%
20%
0%

General   Intermediate   Specific

Label level

Figure 11. Percentage misses in the verification task, by label level and type of vehicle in the image.

We calculated the error rates corresponding to the hit response times reported above. Presentation of the image of a vehicle and a matching label consisted in an opportunity for error, i.e., for a "miss." The relevant error rates, therefore, are the proportion of misses to opportunities for a miss at each level of labeling for each specific vehicle.

For jeeps and trucks generally, misses rose dramatically at the specific level (KRAZ, t_{16} = -3.922, p = 0.001; M151, t_{16} = -2.384, p = 0.03; ZILL, t_{16} = -3.108, p = 0.007). The UAZ was the only jeep or truck for which the rise in error rates between the intermediate and specific label levels was not reliable. No differences between the general and intermediate level were significant, nor was this difference significant for all jeeps and trucks together. This pattern confirms the inference based on verification response times, that the intermediate level is favored for jeeps and trucks.

For tanks, misses also tended to rise from the intermediate to the specific level of labeling, just as response times rose. This rise was statistically reliable for the T55 (t_{16} = -2.746, p = 0.014), and represented a trend for the T62 (t_{16} = -1.376, p = 0.188) and M551 (t_{15} = -1, p = 0.333). For all tanks taken together, the advantage of intermediate over specific was also significant (t_{16} = -3.165, p = 0.006). The difference between general and intermediate was not significant for any tank, or for all tanks together. This pattern confirms our earlier conclusion that the intermediate level if favored for tanks in the verification task.

For APCs, misses were flat between the general and intermediate levels, and also between the intermediate and specific levels. In no case was the change in error rates for APCs reliable. This again fits the earlier conclusion that the specific level is favored for APC's.

Spontaneous naming

Procedure

The spontaneous naming task required subjects to view an image of a vehicle and quickly respond by typing a single name for that vehicle[6]. Each subject viewed each vehicle once. Sixteen subjects performed this task.

Analysis

Names generated by the participants were standardized and scored for accuracy using Jane's AFV Recognition Handbook (Foss, Christopher. (1992). Jane's AFV Recognition Handbook. Alexandria, VA: Jane's Information Group). Of 162 responses generated by the participants, 105 names were scored as accurate because they were correct, technically specific labels for the given image (e.g., "M551 Sheridan"), correct nicknames (e.g., "Sheridan"), or correct intermediate names (APC, tank, jeep (or the newer term, humvee) or truck). No participant offered a general name for any vehicle. Forty-four responses were scored as inaccurate because they denoted a vehicle other than the one depicted in the image (e.g., the name jeep given for a truck, or the name Scorpion given for an M551). The remaining responses were dropped from the analysis. Thus, six responses that did not verifiably belong to the depicted vehicle or another vehicle were dropped. Also dropped were seven responses that denoted vehicle functions (recon) or characteristics that were orthogonal to the naming system used in this study (foe, friend, and soft target).

Results

We predicted that participants would use the favored labels for each object, as determined by earlier results (feature naming, familiarity, typicality), that is, intermediate category labels for trucks and jeeps, and specific category labels for APC's and tanks.

As shown in Figure 12, the distribution of names by label level differed strikingly between vehicle types. As predicted, images of trucks and jeeps were uniformly referred to by their intermediate level names: *truck* or "jeep." Also as predicted, APC's were always labeled with specific model names (e.g., BMP, BTR-60) and never with the name APC. Tanks were referred to either as tanks (the intermediate label) or by specific names with roughly equal frequency. This was true whether the tally counted only accurate non-unique responses or accurate plus inaccurate non-unique responses.

In sum, subjects used only specific labels for APC's, intermediate labels for jeeps and trucks, and both intermediate and specific labels for tanks.

---

[6]Many subjects encountered the spontaneous naming task after completing other tasks in which the researchers' labels were presented. Thus, these data may be biased towards use of the same labels employed by the researchers.
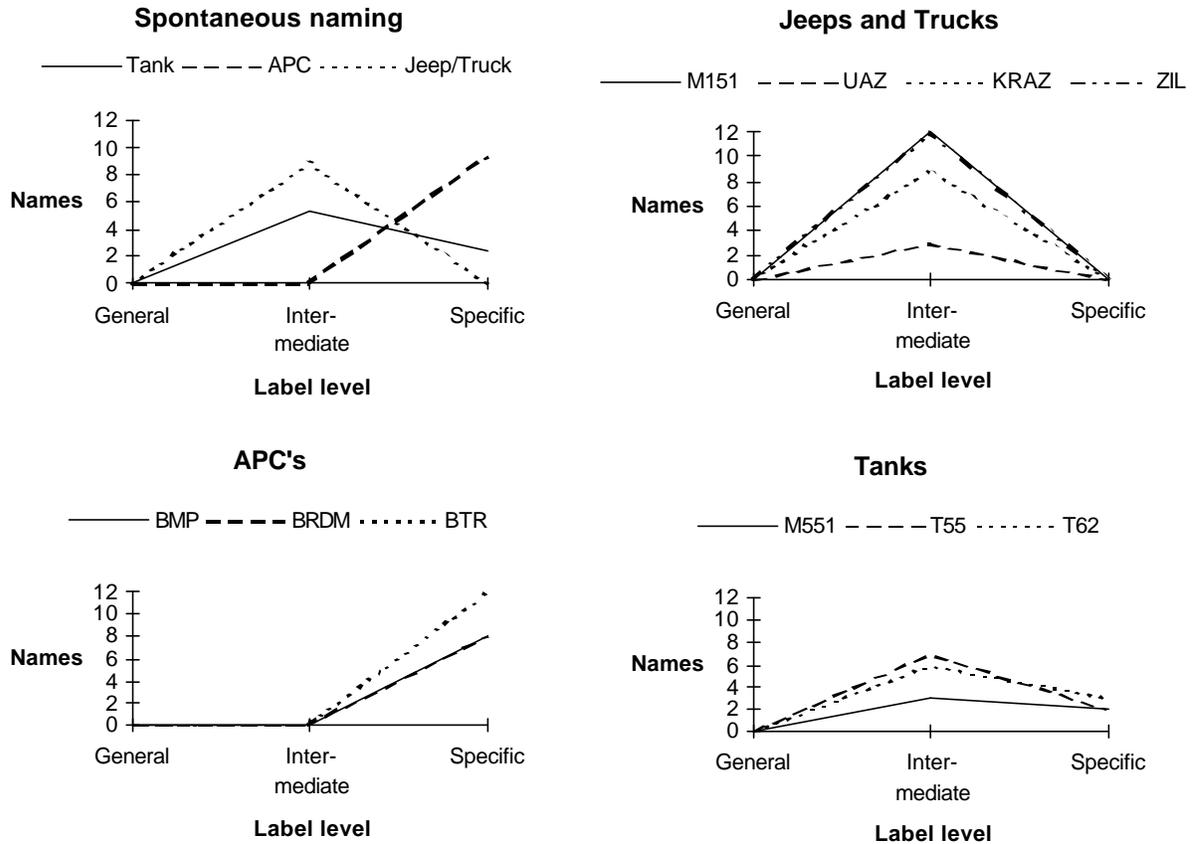
Figure 12: Responses at different levels of generality when spontaneously naming images.

Discussion

The results from the spontaneous naming task and the verification task are in agreement with each other. Participants spontaneously use specific names for APC's, and they recognized matches between specific labels and FLIR imagery of APCs rapidly (that is, about as quickly as they verified matches between images of APCs and intermediate or general labels).

Participants used more intermediate than specific names for jeeps, trucks and tanks. Similarly, they were faster to verify matches between intermediate names and images of those vehicles than they were to verify matches to specific names, and they were at least as fast verifying intermediate names as they were verifying general names.

In terms of our original predictions, based on feature naming, typicality, and familiarity, however, the results are mixed. Table 2 summarizes the results from all five studies. The verification and spontaneous naming tasks provided convergent validation for findings from the feature naming, typicality, and familiarity, for jeeps and trucks, and for APC's:

(1) For jeeps and trucks, feature naming efficiency, combined with lack of familiarity, suggested that intermediate names would be favored. Verification response times and spontaneous naming for jeeps and trucks confirmed this.

(2) Although feature-naming efficiency favored intermediate category labels for APC's, the pilots claimed high familiarity with specific APC models. This suggested that specific names

would be favored. Verification response times and spontaneous naming for APC's confirmed this.

(3) Feature naming efficiency for tanks was consistent with either intermediate or specific category labels. Pilots claimed high familiarity with specific tank models, and this suggested that specific labels would be favored (just as for APC's). However, verification response times were faster for intermediate labels of tanks, and although spontaneous naming utilized specific labels, it also utilized intermediate labels. The pilot's claimed familiarity with specific tank models was not backed up by speed of verification or exclusive use of specific names.

Table 2 Summary of the results of five experimental tasks designed to identify the favored level of labeling for various vehicles.

Type of vehicle	Feature naming efficiency	Familiarity	Typicality	Verification response time	Spontaneous naming	*Implications for favored level*
Jeep & Truck	Intermediate	Low	Medium or High	Intermediate	Intermediate	*Intermediate*
APC	Intermediate	High	High	Specific	Specific	*Specific*
Tank	Intermediate or specific	High	Medium or High	Intermediate	Intermediate or Specific	*Intermediate or specific*

Tanks were the sole exception to our original predictions. It is tempting to speculate that light might be shed on this exception by the one major variable that we did not manipulate in this study. We noted earlier that the favored level for a given object might be affected by the purpose or task for which the classification is being performed. Tanks are highly relevant for many attack helicopter missions, both as potential targets and as potential threats. Distinguishing specific types of tanks, e.g., in order to differentiate friend from foe, or to determine the degree of threat, may also be very important. This would account for the high degree of familiarity with specific tank images claimed by the participants. Nevertheless, the failure of verification response times, in particular, to reflect such familiarity, may have significant training implications.

Implications for ATR Design

We conclude that the labels used by ATR's to identify objects should vary by vehicle type. We can identify the following more specific implications of these findings for ATR interface design:

- Jeeps and trucks should be identified as jeeps and trucks. Intermediate terms for jeeps and trucks are most efficient, fastest to verify, and preferred in spontaneous naming.

- APC's should be labeled with model names (BTR, BMP). Specific terms reflect familiarity of the vehicles, are fastest to verify, and are preferred in spontaneous naming.

- Tanks should perhaps be labeled both as tanks and by model name (T-62, M-60), Specific and intermediate names are both efficient, and both are used in spontaneous naming. Specific terms reflect claimed familiarity of the vehicles, but they are not fast to verify.

Claims of familiarity and use of specific terms in spontaneous naming, despite slow verification times, may reflect the relevance of tanks in typical missions. Another solution to labeling tanks will be explored in the third set of experiments (Chapter 4 below), where we introduce a slightly more detailed type of intermediate labeling (as *enemy tank* and *friendly tank*) that is designed to retain the advantages of intermediate labels while reflecting mission requirements.

A variety of factors appear to influence the usefulness of different labels: Differing associations with features, differing familiarity, and mission relevance. The use of converging methods, such as spontaneous naming, verification response time, feature naming, and familiarity ratings can help identify the best labeling scheme for ATR conclusions.

Visual Features

In the next chapter, we will turn to visual processing of images. Before moving to that topic, we return here to the initial verbal categorization study, in which participants were asked to name features associated with category labels. A secondary goal of that study was to determine the features of images that participants deemed relevant in target recognition. While we must treat verbal accounts with caution, these feature lists may be useful in two ways: (1) They may help us interpret the results of the experiments on visual processing in the next chapter, and (2) they may illuminate the degree to which pilot's verbal organization of information abort vehicles correspond to their visual organization of the same vehicles, as examined in the next chapter.

As noted above, Appendix B contains a list of all features attributed by any pilot to a label at any level of specificity. The list includes both features that are visible and features that are not. In the present section, we narrow the focus to features that can be observed on FLIR scopes.

Table 3 summarizes the visible features named by pilots in response to specific labels, while Table 4 summarizes the visible features named in response to general and intermediate labels. These lists are derived from the master list of features in Appendix B by (1) dropping all clearly non-visible features, and (2) placing in the same summary category all features that pertain to the same part or aspect of the vehicle. For example, *weapon size: short, thick* and *weapons location: turret* are both placed in the *weapons* category; *tracks temperature: hot* and *tracks type: slack* are likewise in the *tracks* category; and *profile: tank-like* and *profile: slab-sided* are each in the profile category. In both tables, feature categories are listed in order of how frequently pilots mentioned them.

Both lists are characterized by a small group of frequently mentioned categories and a larger group of categories that are mentioned rarely. In Table 3, ten feature categories were mentioned five times or more, while 27 feature categories were mentioned less than five times (11 of which were mentioned by only one pilot in response to only one label). Even more strikingly, in Table 4 ten categories were mentioned five times or more, while 30 categories were mentioned less than five times (21 of which were mentioned by only one pilot on only one occasion). Moreover, the list of feature categories for specific labels and the list for general and intermediate labels are similar, though certainly not identical. In particular, seven of ten most frequently cited categories of features are common between the two lists: wheels, weapons,

profile, tracks, size, turret, and roadwheels.[7] In the following, we will focus on these seven common types of features.

It seems reasonable to group the seven most frequently used visual feature categories into four clusters, corresponding to different parts or aspects of the vehicle image:

- details of locomotion (i.e., wheels, tracks, and roadwheels)

- weapons

- turret

- the figure of the vehicle as a whole (profile and size)

These four clusters represent three parts of the image (track/wheel area, turret, and weapon) plus its overall figure.[8] Taken together, the four clusters (or, equivalently, the seven categories of features that make them up) account for 68% (169 out of 245) of the features mentioned by pilots in response to general and intermediate labels, and for 76% (264 out of 348) of the features mentioned in response to specific labels.

Perhaps even more importantly, each of the seven feature categories applies broadly across the spectrum of vehicles. The seven categories apply, on average, to eight of the ten specific labels, while the four clusters apply to 8.75 of the ten specific labels. The seven categories apply to 4.9 of the six general / intermediate labels, while the four clusters apply to 5.5 of the six general / intermediate labels. In other words, the four clusters tend to be all-purpose sources of discrimination regardless of the type of vehicles present in an image or the level of discrimination that is required. By focussing on the corresponding parts or aspects of an image, pilots may maximize their ability to recognize vehicles.

In sum, a qualitative analysis of data from the feature naming task indicates that pilots associate general, intermediate, and specific names of vehicles with overall vehicle figure (including profile and size – and possibly turret) as well as with more localized features (e.g., wheels and tracks, weapons, and turret). This finding suggests another look at related research (e.g., O'Kane, Biederman & Cooper, 1994) in which it is assumed that individual vehicle features (wheels, weapons, etc.) are the sole focus of cognitive processing during vehicle identification. Other research has supported the idea that the overall shape of an object may be perceived prior to more detailed features (Navon, 1977).

---

[7] The three top-ten categories in each table that are not shared with the other table tend to fall at the bottom of the top ten lists: i.e., skin, wheels or tracks, cargo area, nose, suspension, door.

[8] An argument could be made to merge the turret category into the cluster representing the figure of the vehicle as a whole. As examination of Appendix B shows, virtually all of the specific features in the turret category are coarse-grained; in particular, features such as the presence, size, or shape of the turret appear to be characteristics of a fairly rapid overall impression of the vehicle. By the same token, characteristics such as *weapons size: big* might also belong in the overall figure category. However, most (though not all) of the features in the weapons category require a more detailed look at the image beyond its overall impression (e.g., *weapons type: 100 mm*, or *weapons size: large gun with bore evacuator*). Our criterion for inclusion in the figure category is conservative, requiring explicit reference to profile or overall size of the vehicle.

Table 3:.Categories of -visual features named by pilots in response to specific labels.

Feature	Jeeps		Trucks		APCs			Tanks			Total
	UAZ	M151	KRAZ	ZIL	BMP	BRDM	BTR	M551	T55	T62	
weapons		4	1		13	9	7	10	12	13	69
turret			3		9	4	9	10	10	9	54
wheels	2	4	2	3	2	13	13	1	1	2	43
profile	1	1	2	2	7	4	4	9	5	5	40
tracks		3	1		10	1	1	5	4	6	31
size		4		1	1	2	2	3	3	2	18
roadwheels					1		4	3	6	3	17
nose					7	5	1			1	14
suspension				1	2			4	3	3	13
door							5				5
hand rails				1					1	2	4
heat source					1	1			1	1	4
antenna					1	1		1			3
chassis		1		1			1				3
searchlight								1		2	3
skin						1			1	1	3
camouflage						1				1	2
cargo area	1	1									2
cover		1		1							2
cupola							1	1			2
snorkel										2	2
splash guards									1	1	2
engine							1				1
orientation								1			1
sprocket								1			1
suspension rollers										1	1
track skirts										1	1

wheels or tracks					1						1
body		1									1
hatch							1				1
light									1		1
rifle port							1				1
sponson box								1			1
top					1						1
Total	4	20	9	10	56	42	51	51	49	56	348

This finding has implications for ATR design. It suggests that FLIR image processing algorithms should not necessarily enhance vehicle components at the expense of the overall vehicle profile or shape. We will test this prediction in the next section.

Table 4. Categories of visual features named by pilots in response to general and intermediate labels.

Feature	WHEELED	TRACKED	JEEP	TRUCK	APC	TANK	Total
wheels	11	3	11	11	2	2	40
weapons	1	5	1	1	9	16	33
profile	2	5	3	6	9	1	26
tracks		7			5	11	23
size	2	3	5	7	3	2	22
turret		2	1		3	9	15
skin	1	1	2	1	5	5	15
roadwheels	1	6				3	10
wheels or tracks					7		7
cargo area			1	4			5
antenna	1	1	1		1		4
cab	1			3			4
heat source	1	1				2	4
suspension		2	2				4
nose			3		1		4
camouflage			1			1	2
cover			1	1			2
roof			2				2
windshield			2				2
axles				1			1
chassis	1						1
cupola						1	1
door			1				1
engine			1				1
external fuel tank						1	1
gun port					1		1
hand rails		1					1
loaded				1			1

peep hole		1					1
searchlight						1	1
snorkel						1	1
sprocket						1	1
support rollers						1	1
top skin			1				1
window				1			1
body						1	1
brake				1			1
frame				1			1
headlight			1				1
sponson box						1	1
Total	22	38	40	39	46	60	245

Similarity based on Feature Naming

Undoubtedly, some of the features used by pilots to visually recognize vehicles are not easily verbalized, and perhaps not even accessible to conscious awareness. However, we were interested in the extent to which the visual features that they could verbalize resemble the features actually used in visual processing. At the very least, we can ask, to what extent are the verbalizable features by themselves capable of making the required discriminations among vehicles?

A multi-dimensional scaling analysis was conducted, in which the data were the number of visual features shared by each pair of specific vehicle labels. The solution was plotted in three dimensions. R-square for this solution, using the Guttman stress formula, was excellent ($r^2 = 0.99965$) and a plot of stress against dimensionality featured a pronounced knee at three dimensions, indicating that this was superior to a two-dimensional solution and not significantly better than a solution of higher dimensionality. Figure 13 shows the three-dimensional MDS solution.
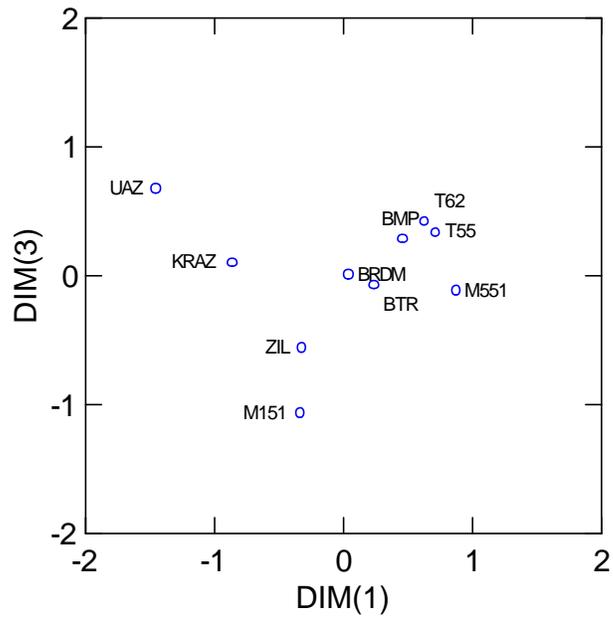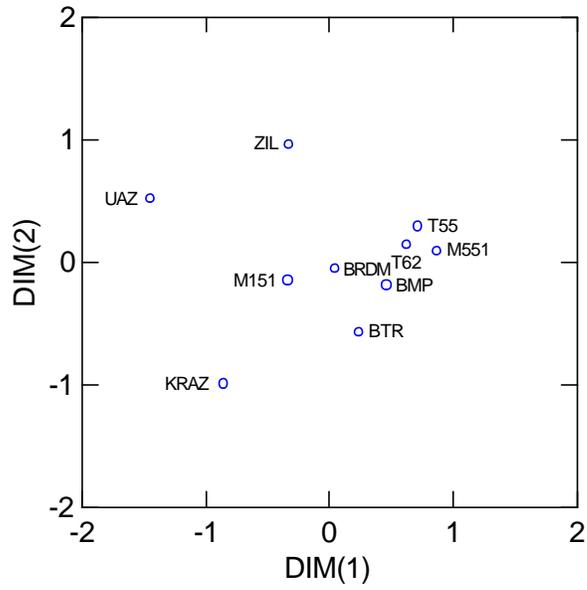
Figure 13. Three dimensional multidimensional scaling solution based on feature naming in response to specific vehicle labels.

From a qualitative point of view, the features explicitly named by pilots induce a reasonable amount of separation among the vehicles, suggesting that named features may be serviceable for recognition. Moreover, the structure of the distance relations in this space reflect plausible similarity relationships. In both charts, tanks tend to be close to other tanks, APC's to other APC's, and, to a lesser degree, jeeps and trucks to other jeeps and trucks. However, the similarity relationships in Figure 13 are not best characterized in terms of such conventional semantic relationships. Rather, the two tanks belonging to the former Soviet Union (T62 and T55) seem to form the core of a set of increasingly inclusive classes. A hierarchical clustering analysis based on the same feature naming data is shown in Figure 14. The clustering analysis reveals a sequence of nested categories rather than a semantic hierarchy. We have laid out the sequence of categories, together with hypothetical category labels, in Table 5. (No short expressions are available to describe two of the more general categories.)

Cluster Tree
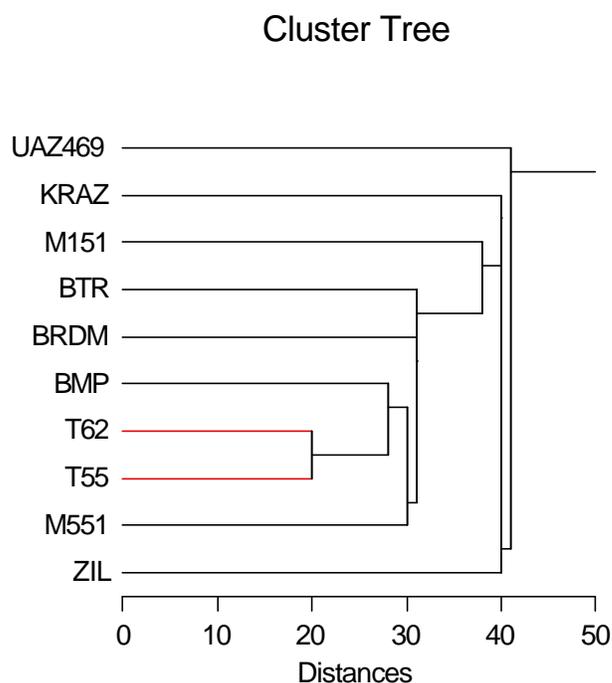
Figure 14. Hierarchical cluster analysis of specific vehicle labels based on feature naming.

These results are consistent with the military purpose of target recognition. Pilots are not botanists who attempt to recognize and classify every vehicle type in a taxonomic scheme. If that were the case, ,pilots would be expected to learn many features to discriminate jeeps from other jeeps, trucks from other trucks, APC's from other APC's, and jeeps, trucks, and APC's from one another. The result would be a structure more like Figure 5, with separate clusters corresponding to APC's, jeeps, and trucks, joined hierarchically in higher level categories like tracked vs. wheeled. By contrast, what is distinctive about Figure 14 and Table 5 is that all categories other than *enemy tank* are supersets of *enemy tank*. The pilot's knowledge of visual features is organized in service of the goal of distinguishing typical targets (enemy tanks) from everything else. This organization of feature knowledge would be suited to a decision making strategy of increasing refinement toward enemy tank. This might take the form of a sequence of processing

stages in which the initial impression of a vehicle is refined until the possibility of an enemy tank is eliminated, whereupon processing stops, or the possibility is established. In the next chapter, we turn to these and other questions related to visual processing.

Table 5. Sequence of increasingly inclusive categories suggested by hierarchical clustering analysis of feature naming.

Enemy tanks	T62	T55								
Enemy tracked / armored vehicles	T62	T55	BMP							
Tracked /armored vehicles	T62	T55	BMP	M551						
Armored vehicles	T62	T55	BMP	M551	BRDM	BTR				
?	T62	T55	BMP	M551	BRDM	BTR	M151			
?	T62	T55	BMP	M551	BRDM	BTR	M151	KRAZ	ZIL	
Vehicles	T62	T55	BMP	M551	BRDM	BTR	M151	KRAZ	ZIL	UAZ

A final question concerns the four categories of verbally named visual features that we identified earlier in this chapter. Assuming that they capture different foci of attention as visual processing progresses, what discriminations would be expected as attention shifts from one to the other – for example, from profile to wheels and tracks, and then from wheels and tracks to turret and to weapon? We performed multidimensional scaling and hierarchical clustering analyses based on each category of features separately, again measuring similarity by the number of verbalized features that two images shared. The results are included in Appendix B. They demonstrate that interesting and different separations among the images are obtainable from each of the clusters of features alone. One salient point stands out, however, reemphasizing the importance of profile. Profile is the only one of the feature categories that required a three dimensional similarity space to capture the differences among vehicles that it unveiled.

3. VISUAL PROCESSING OF IMAGES

Introduction

At least in the near-term, users of ATR systems will need access to sensor information in order to validate or supplement ATR conclusions. This chapter asks whether a user's visual processing of such sensor data, e.g., a FLIR display, can be facilitated by graphically enhancing aspects of a vehicle image. If so, which aspects of the imager should be enhanced: viz., the vehicle's overall profile, or specific details such as tracks, wheels, turret and gun? A key research issue is whether the critical aspects of images may be identified through investigation of stages of human visual processing. To answer this question, we tested a methodology for identifying the features of a stimulus that are extracted by the visual system, as function of time after the stimulus is presented. Our hypothesis was that the greatest improvement in human recognition performance would result when an ATR enhanced vehicle features that are extracted early in visual processing.

In a simple detection task, observers can reduce one kind of error (e.g., misses) by increasing the frequency of another (e.g., false alarms). Signal detection theory is a method for using performance data to distinguish such trade off decisions, which may be influenced by payoffs or biases, from the observer's underlying perceptual ability. Recognition tasks are far more complex than simple detection tasks. In recognition, potential tradeoffs among different kinds of confusion errors occur between every pair of classes to which a target may be assigned. For example, observers can always increase his chance of correctly identifying tanks if they are willing to accept an increase in the chance of misidentifying APC's as tanks. At the same time, observers can increase correct identifications of trucks relative to APC's by allowing an increase in the number of APC's misclassified as trucks. The observer's *ability* to discriminate can also vary between different classes, independently of biases: For example, the observer may be better at discriminating trucks from tanks and APRs than at discriminating tanks and APC's from one another.

In principle, signal detection theory could be extended to recognition with multiple categories. However, the analysis of experimental data in terms of a generalized SDT model is prohibitively complex (e.g., Broadbent, 1971). A variety of alternative techniques exist that under many conditions (e.g., assuming normal and symmetrical internal sensory dimensions) produce very close approximations to the SDT analysis. These techniques also accomplish the basic goal of distinguishing the effects of discriminability from the effects of decision criteria in multi-category tasks. Some of these alternative techniques also offer advantages in their interpretability in terms of underlying psychological processes.

The best known technique of this kind is the Biased Choice Model described by Luce (1956, 1977). According to this model, the probability of a response in the presence of a stimulus is a joint function of response bias and similarity to the stimulus associated with the response. More formally, the probability $p(i,j)$ of a response j in the presence of stimulus i is proportional to the bias $b(j)$ in favor of response j, and the similarity $h(i,j)$ between classes i and j. It is assumed that similarity is symmetric, that the similarity of a stimulus with itself is 1, and that the sum of the $b(i)$ is 1. The following equation describes response probabilities:

$$p(i,j) = \frac{b(j)\boldsymbol{h}(i,j)}{\sum_j b(j)\boldsymbol{h}(i,j)}$$

The biased choice model, however, has no straightforward interpretation in terms of underlying visual processing events. The approach we explore here is based on a model of multiple category recognition called the Informed Guessing Model (IGM) (Pachella, Smith, and Stanovich, 1978; see also, related models by Broadbent, 1971, and Townsend, 1971). IGM is a special case of the biased choice model, in the sense that it fits only a subset of the data that might be fit by the less restrictive biased choice model. According to the Informed Guessing Model, when a stimulus is presented, each of a set of possible perceptual events has a particular probability of occurring. These events might correspond to detections of relevant features of the stimulus, or use of prior evidence about the stimulus, and each such event, if it occurs, narrows down the class of possible responses to the stimulus. When observers must give a response, they guess from the current confusion set, i.e., the set of possible responses that have not yet been eliminated. Response biases and payoffs influence these guesses. If no information about the stimulus has been extracted, the confusion set from which the observer guesses consists of all the possible stimulus categories. If the confusion set has been narrowed down to one possible response, no guessing is necessary.
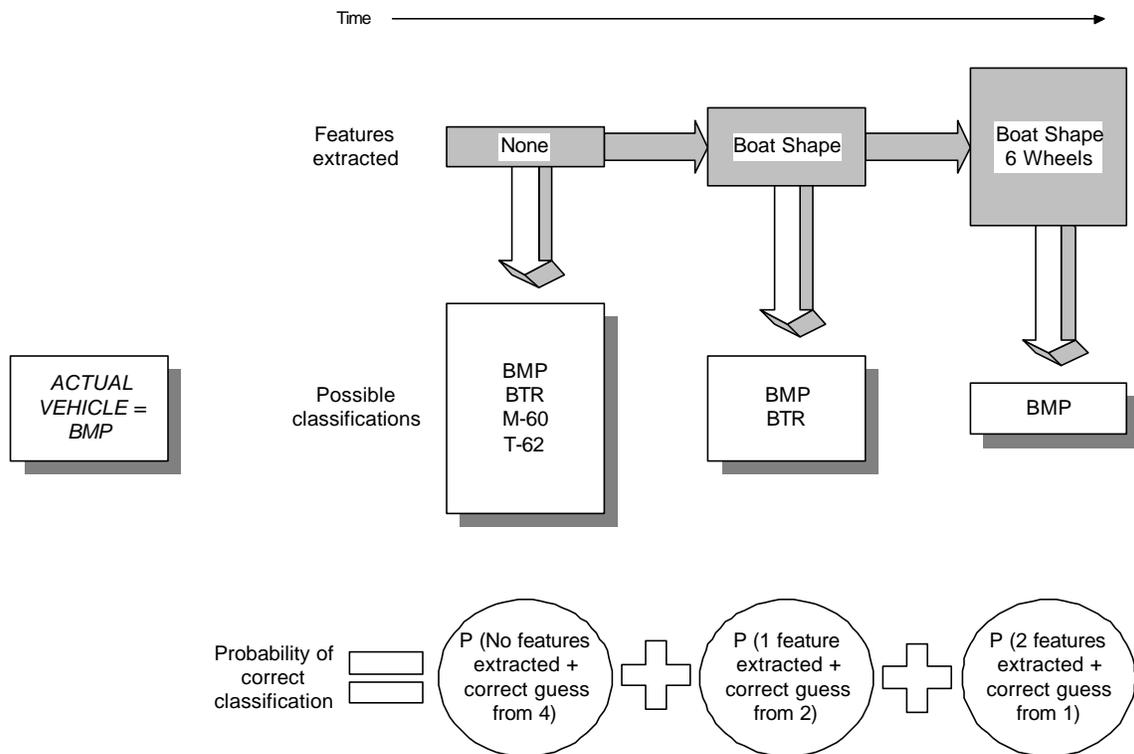

Figure 15. Illustrative sequence of processing stages as represented in the Informed Guessing Model.

Figure 15 shows how IGM represents visual processing. In this example, observers know that the image they see will either be a BMP, BTR, M60, or T62. On this particular trial, the actual image presented to the observer is a BMP. During the first stage of visual processing, however, observers have extracted no information from the stimulus, and their knowledge is

represented by the full confusion set (i.e., BMP BTR M60 T62). If the observer is forced to respond at this time, the chance of a correct response is simply the chance of guessing BMP from this set of four alternatives, and will reflect response biases, and the costs of different kinds of errors. At the second stage of processing, however, observers have extracted a single feature (the shape of the vehicle). This feature enables them to distinguish APC's (BMP, BTR) from tanks (M60, T62), but not to discriminate further within each class. In this example, since the actual stimulus is a BMP, the observer's knowledge is represented by the BMP BTR confusion set. The probability of a correct response is now equal to the chance of guessing BMP when the two alternatives are BMP and BTR. Finally, at the last stage of processing, the observer has extracted a second feature (e.g., number of wheels) that discriminates BTR's from BMP's. The observer's knowledge is now represented by the singleton set BMP, and there is no need of guessing. The total chance of a correct response at any time is the sum of the chances of being in each of these three states times the probability of a correct response in that state.

More technically, in a set of four stimuli, each belonging to a different class, the probability of response j to stimulus i is:

$$p(i,j) = \frac{B(j)}{B(i)+B(j)}\mathbf{x}(i,j) + B(j)g$$

$$p(i,i) = 1 - \sum_{i \neq j} p(i,j)$$

g is the probability that insufficient information was extracted from the stimulus to rule out any class of stimuli. $\mathbf{x(i,j)}$ is the probability that enough information was extracted only to narrow down the possibilities to i or j. $B(i)$ is the bias in favor of response i. Again, similarity is assumed to be symmetrical, i.e., the probability of any particular confusion set is the same regardless of which member of the set served as the stimulus. The sum of the bias parameters equals 1. Finally, the sum of the probabilities of all the confusion sets containing any particular stimulus class must equal 1:

$$g + \mathbf{x}(i) + \sum_{j \neq i} \mathbf{x}(i,j) = 1$$

Notice that the probability of a correct response can be directly represented as a sum of "true discriminations," $\mathbf{x(i)}$, "lucky" guesses from the confusion sets representing pairs of stimulus classes, and "lucky" guesses from the confusion set consisting of all stimulus classes:

$$p(i,i) = \mathbf{x}(i) + \sum_{j \neq i} \frac{B(i)}{B(i)+B(j)}\mathbf{x}(i,j) + B(i)g$$

This leads to a simple correction for guessing (and for bias), in which $p(i,i)$ is replaced by $\mathbf{x(i)}$ as a measure of the underlying ability to discriminate i from other stimuli. Maximum likelihood methods for assessing the parameters of the Informed Guessing Model are described in Pachella, Smith, and Stanovich (1978).

The Informed Guessing Model permits examination of the similarity structure among a set of stimuli, and investigation of how that structure might change over the time course of visual processing. We observed in the last section that observers can trade off different types of recognition errors. A similar tradeoff can occur with respect to speed and accuracy. Observers are able to reduce errors (of any kind) by spending more time to process a stimulus. Conversely,

observers can improve their reaction time to a stimulus by accepting a degradation of accuracy. Speed is particularly important in battlefield target recognition. A large number of targets, or the urgency of a potential threat, may prevent the pilot from processing stimuli to the maximum level of accuracy. The most significant contribution of ATRs may in fact be in situations where the operator does not have adequate time to fully process a stimulus.

A speed-accuracy operating characteristic is a plot of accuracy against the amount of time spent processing. One way to obtain such a characteristic is to impose a response time-window (Reed, 1976) or deadline (Pachella and Pew, 1968) on the observer, and to measure accuracy as the window or deadline is varied. A plot of overall accuracy (e.g., percent correct classifications) against time can help validate an ATR design. If accuracy with an ATR is as good as accuracy without an ATR at all response times, and better at some of the response times, then we can safely conclude that improvements in accuracy occur without increasing the time required for processing the stimulus.

Our use of the speed-accuracy methodology here is to gain insight into the qualitative time course of perceptual processing. By plotting the specific components of accuracy (e.g., the $x(i,j)$ and the $x(i)$) against time, we get a series of snap shots of processing, revealing which features of a stimulus set are extracted early and which are extracted late, and how recognition processing might be qualitatively affected by the introduction of an ATR. Such an analysis can be an important input into the design of the ATR interface.

Method

Design

There were two principal within-subjects variables: four display durations (adapted for each individual participant to elicit approximate accuracy rates of 80%, 65%, 50% and 35%), crossed with four vehicle types (BMP, BTR, M60 and T62). Three additional orthogonal within-subjects variables were utilized: two views of the vehicles (side and oblique), two vehicle orientations (left and right) and four image sizes (simulating approximately ranges of 5km, 3.5km, 3km and 2.5km).

Five kinds of image enhancement were varied both within and between subjects. In the unenhanced condition, stimuli resembled FLIR imagery as it currently appears on FLIR displays. Three enhancements heightened the contrast of details: tracks or wheels, gun and turret assembly, and entire vehicle, respectively. The fifth enhancement highlighted the vehicle's silhouette at the expense of all details, by lowering the contrast within the vehicle image. Nine subjects viewed only the unenhanced imagery. Ten subjects viewed the unenhanced imagery plus some enhanced images, as shown in Table 6.

Each subject performed approximately 700 trials. Subjects who viewed only the unenhanced FLIR imagery responded to each of 256 stimuli (4 stimulus durations x 4 vehicles x 2 views x 2 orientations x 4 sizes) two or three times (mean = 2.75). The remaining subjects responded to raw and enhanced imagery. For these subjects, image enhancement was incompletely crossed with the other five variables. These subjects responded to approximately 170 images in each enhancement category. The order in which images were presented was randomized over all images within subjects.

Table 6: Groups of participants, in rows, and the types of imagery to which they were exposed, i.e., unenhanced (raw FLIR), and various types of enhancements.

N	Unenhanced	Guns/turret	Tracks/wheels	Entire vehicle	Silhouette
9	*				
4	*	*	*	*	
5	*	*	*		*
1	*		*	*	*

Participants

Subjects were 19 U.S. Army helicopter pilots stationed at Ft. Bragg, NC[9]. Subjects were homogeneous with respect their rank: 18 of the 19 were commissioned warrant officers of the second grade; two were first lieutenants. Only two of the officers had combat experience. At the median, officers had 9 years of military experience and 3.8 years of experience in Army flight positions involving target recognition. At the median, the subjects had attended 4 military schools, and 2.5 military exercises.

Materials

Images in this study were produced by Mr. Horger in collaboration with Robert Lafollette and were provided by Dr. Barbara O'Kane, all staff of NVEOL. FLIR noise was generated using NVSIM, a UNIX-based Thermal Imaging System Simulator developed by John Horger. Image manipulations were produced by CTI using Adobe Photoshop.

The primary materials were four vehicle images, BMP, BTR, T62 and M60, as shown in Figure 16. Based on results from the feature naming study (see previous chapter), these vehicles differ in a variety ways. Figure 16 illustrates only two such discrimination possibilities: one based on a global feature (e.g., boat-like vs tank shape) and the other based on a detailed feature (e.g., number of wheels).

Apparatus

Stimuli were presented on an Intel Pentium personal computer running the NextStep operating system. The presentation software developed by CTI for the first set of experiments was used in this study as well.

---

[9]One of 20 subjects was dropped, leaving 19 subjects, because the experimenters accidentally allotted him less time to respond to stimuli (.475 seconds) than they did for other subjects (.625 seconds).

DETAILED FEATURE	Boat Shape	Tank Shape
6 Wheels	BMP	M-60
4 Wheels	BTR	T-62

Figure 16. Four unenhanced stimuli used the visual processing study, their labels, and two properties that might distinguish them.

Procedure

At the beginning of each session, the experimenter briefly queried the subject concerning his familiarity with the four vehicles used as stimuli. All subjects expressed high familiarity with these vehicles.

The session began with a 300-trial practice run. Pilot trials had indicated that the accuracy of most novice subjects reached asymptote within 300 trials. Experienced subjects were expected to attain proficiency on the system at least as rapidly.

The practice session was followed by 320 calibration trials, which were used to assess the subject's accuracy at each of eight stimulus display durations. The eight presentation durations were the same for all participants during both the practice and calibration trials: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.75, and 1 second. At the conclusion of the 320 calibration trials, the system computed the subject's accuracy at each of the eight stimulus durations, and the experimenter selected four stimulus display durations to use during the experimental trials. As indicated above, the durations were selected in order to elicit accuracy rates of approximately 80%, 65%, 50%, and 35%. The actual stimulus durations used in this study ranged from .0075 to 1.25 seconds.

The subject then executed approximately 700 experimental trials with the four selected stimulus durations. The experimental session ended at the two-hour mark.

Display stimulus-response mapping	S depresses all four keys	Blank screen	Variable duration stimulus	Bell sounds, stimulus erased	S lifts one response key	Bell if S was early (< .075") or late (>.7")
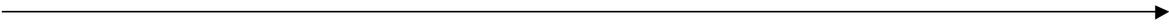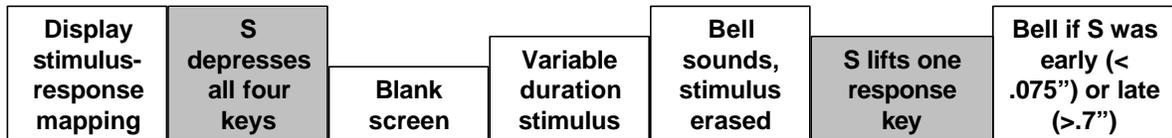
Time

Figure 17. Sequence of events on each trial.

The sequence of displays and participant responses in each trial is shown in Figure 17. First, the system displayed the mapping of vehicle names to four response keys (the shift and control keys on the left side and the right side of the keyboard). This mapping was constant for any given subject but varied between subjects. The subject depressed all four keys to indicate readiness; at this time, the screen blanked, and a stimulus image appeared in the center of the screen. After a variable period of time, chosen from among the four display durations, a bell sounded and the system simultaneously erased the stimulus. The subject then attempted to name the previously displayed target by lifting a response key. Subjects were given a window of .625 seconds within which to respond. If the subject responded within less than .075 of the erasure of the image, or after more than .7 after the erasure, the system sounded a bell to indicate that the response was early or late. Thus, each subject received feedback concerning the timeliness of responses. There was no feedback concerning accuracy. The system paused between trials for 0.5 seconds.

Results: Visual Processing

The analysis of the time course of visual processing focuses on responses by nine officers who received images in raw FLIR format, but saw no enhanced imagery.

We fit parameters from the Informed Guessing model (Pachella, Smith, and Stanovich, 1978) to the data. That model provides a correction for guessing in which the empirical frequencies of correct responses and confusion errors are used to estimate the underlying ability of the subjects to discriminate the stimuli (as opposed to guessing correctly by chance). Specifically, the model was used to generate the probability $x(i)$ of extracting enough information to make a correct identification of each individual vehicle type (e.g., the ability to respond BMP given BMP), the probability $x(i,j)$ of extracting only enough information to discriminate any pair of the stimuli from the others and then guessing between the two (e.g., if the subject is in the BMP BTR confusion set, he knows the stimulus is not a T62 or a M60, but must guess between BMP and BTR), and the probability g of extracting no information and being required to guess from the set of all four stimuli. Table 7 provides the estimated parameters. These numbers represent the underlying perceptual achievement of the observer, after correcting for guessing.

We will first discuss accuracy results for different types of vehicles, then focus on the time course of perception, examining a series of models that might explain that time course. Finally, we will turn to the features extracted at each stage.

Table 7. The probability of correctly identifying an image or of being in dyadic or total confusion at each of four subject-specific response deadlines varying from shortest (1) to longest (4).

Confusion set	Probability of being in the specified confusion set as a function of the time available for processing (shortest = 1, longest = 4)				
	1	2	3	4	Mean
BMP	.118	.298	.496	.645	0.39
BTR	.164	.405	.619	.769	0.49
M60	.177	.322	.436	.554	0.37
T62	0	.104	.332	.436	0.22
BMP BTR	.135	.145	.130	0	0.10
BMP M60	0	0	0	0	0.00
BMP T62	.122	.184	.139	.121	0.14
BTR M60	0	0	0	0	0.00
BTR T62	0	0	0	0	0.00
M60 T62	.156	.284	.304	.278	0.26
BMP,BTR,M60,T62	.624	.373	.226	.164	0.35

Recognition of Individual Vehicle Types

Figure 18 shows the probability of correctly identifying each individual type of vehicle (adjusted to remove guesses) as a function of the time available for perceptual processing. Time is represented ordinally, in terms of the set of four stimulus durations created for each participant. Thus, interval 1 represents the shortest interval for each participant.

Subjects were most accurate identifying the BTR (49% accuracy over all response conditions). Subjects identified the BMP (39%) and the M60 tank (37%) with roughly equal accuracy. Subjects were least accurate in identifying the CIS T62 tank (22%).

These results, which show faster recognition of APC's than tanks, are consistent with findings described in the last chapter. Reaction times to verify the association of an image with a specific label were faster for APC's (e.g., BTR, BMP) than for tanks (e.g., T62, M60). In spontaneous naming of images, subjects were more likely to use specific rather than intermediate level labels for images of APC's than for images of tanks. They were also more likely to be inaccurate in the use of specific labels for tanks than for APC's.
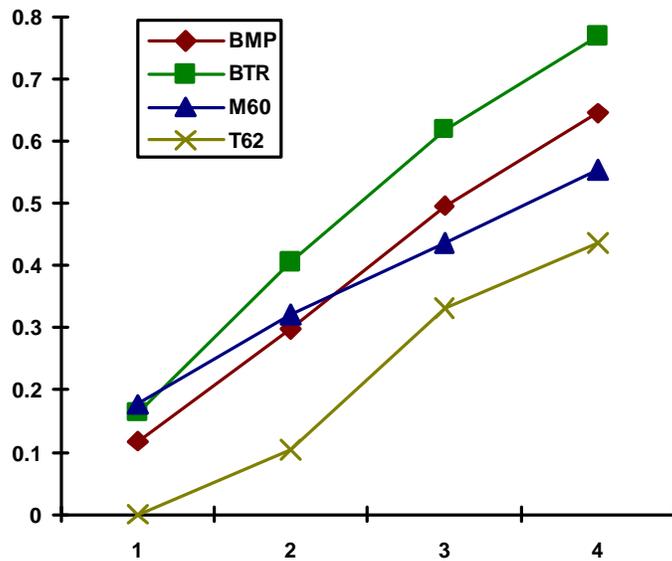
Figure 18. Probability of extracting enough information to correctly identify specific vehicle types, as a function of time available for processing (ordinal representation of stimulus duration).

Stages of visual processing

What happens as the time available for processing increases? What is the sequence with which information is extracted as the image is processed? We will now explore how the Informed Guessing model can provide insight into the sequence of processing stages that underlies the growth of accuracy with time in Figure 18. For each stimulus, IGM can describe three basic stages of processing: a pure guessing state in which no information has been extracted, a set of possible states in which the stimulus has been partially processed, and a state in which the stimulus has been fully processed. For example, suppose the stimulus actually presented is a BMP. At some early time, the observer's knowledge is characterized by the full confusion set: BMP BTR M60 T62. Later, if discrimination is successful, it will be characterized by the singleton set, BMP. What happens in between is our primary interest.

It is important to understand that the IGM itself places almost no constraints on what the transition from ignorance to recognition is like. For example, any and all dyadic confusion sets containing BMP can serve as transition stages between complete ignorance (BMP BTR M60 T62) and recognition (BMP). In fact, it is also possible that no dyadic confusion set plays a role, and that the transition from BMP BTR M60 T62 to BMP occurs in a single feature-extraction step. The only mathematical constraint imposed by IGM is that the probabilities of all the confusion sets containing a given vehicle type add to 1.0. Thus, at any given time, discrimination parameters for the following sets, corresponding to a BMP as stimulus, must add to 1.0: BMP BTR M60 T62, BMP BTR, BMP M60, BMP T62, and BMP. Similarly, the parameters for the following sets, corresponding to a T62 as stimulus, must also add to one: BMP BTR M60 T62, BMP T62, BTR T62, M60 T62, and T62. Similarly, parameters for confusion sets containing BTR and M60, respectively, must also add to 1.0. As shown in Table 8, these constraints are reasonably well satisfied by the present data, indicating that the IGM parameters provide a good fit.

Table 8. Sum of probabilities for confusion sets containing a given vehicle type, by stage of processing. (Predicted values are 1.0.)

	Stage 1	Stage 2	Stage 3	Stage 4
BMP	0.999	1	0.991	0.93
BTR	0.923	0.923	0.975	0.933
M60	0.957	0.979	0.966	0.996
T62	0.902	0.945	1.001	0.999

Any pattern that is found in the IGM parameters beyond the required summations to 1.0 is thus not dictated by IGM, but sheds light on the actual sequence of visual processing.

Figure 19 shows that early in processing, there is a high probability (62%) that no information has been extracted and all responses are guesses (indicated by the confusion set BMP BTR M60 T62). The figure shows that the probability of being in the pure guessing state decreases with processing time. At the same time, there is an increase in the probability that enough information will be extracted to discriminate a pair of vehicles from the other vehicles, but not from one another: i.e., the partially processed state.

Table 7 helps us understand what the partially processed states are. Of the six possible pairs of four vehicles (dyadic confusion sets), the probability was zero for three pairs at all stimulus durations. This indicates the lack of psychological reality of these confusion sets; i.e., no perceptual features were extracted that discriminated just these pairs of vehicles from the other two, but not from one another. (These vehicles might be confused with one another if the subject was still in the pure guessing state, however.) For three of the pairs, the probability of the dyadic confusion set was non-zero. For each of these latter pairs, the probability showed a similar pattern: an initial increase followed by a decline (Figure 19). The increase represents the first step of processing when features that discriminate this pair from the other vehicles (but not from one another) are being extracted. The decline represents the second step of processing when additional features are being extracted that discriminate the members of the pairs from one another: i.e., the fully processed state.
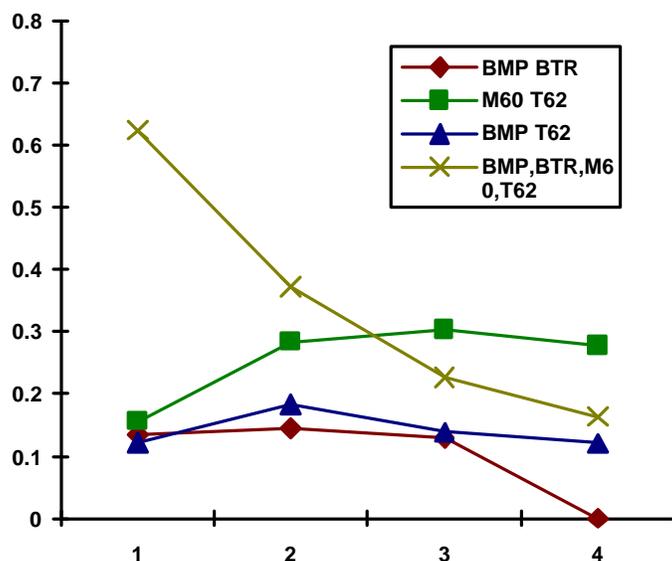
Figure 19. Probability of extracting enough information to correctly distinguish pairs of vehicles from the others but not from one another, as a function of time available for processing (ordinal representation of stimulus duration). Also shown is the decline in the probability of extracting no information.

Visual Processing Models

Changes in perceptual discrimination over time, as revealed by the pattern of parameter change over time in Figure 18 and Figure 19, can provide the basis for the development of more detailed models of visual processing. In this section, we will consider a series of models that specify the processing steps underlying the changes in discrimination. The first two models are unsuccessful, but the reasons for their failure are illuminating, and each of them becomes a part of a third model, which is successful in fitting the data.

Model 1. The data are accounted for in part by a simple hierarchical model of feature extraction over time. Two of the dyadic confusions sets in Figure 19 (BMP BTR and M60 T62) are complementary to one another. This suggests that there is an intermediate state of processing in which enough information has been extracted to discriminate BMP's and BTR's from M60's and T62's, but not enough information to discriminate BMP's from BTR's, or M60's from T62's. In other words, observers in this stage can discriminate tanks from APC's, but not tanks from one another or APC's from one another. Figure 20 is an tree depicting this hypothesized sequence of processing stages.

Immediately after the image of a particular vehicle is presented, all observers are in the full confusion set (BTR BMP M60 T62). After any specified amount of time, depending on the actual identify of the vehicle in the stimulus image, this model implies that observers can be in any one of three situations: They may remain in the full confusion set, they may be in the appropriate dyadic confusion set (BTR BMP if the stimulus is a BTR or a BMP; M60 T62 if the stimulus is a M60 or a T62), or they may be in the appropriate singleton set. There is a chance **p** that observers have extracted an initial feature that enables them to distinguish the two dyadic sets from one another. This discrimination causes them to leave BTR BMP M60 T62, and places

45

them in either BMP BTR or M60 T62, depending on what the stimulus actually is. Since the only way observers can get out of the full confusion set is by making this initial discrimination, the chance of still being in the full confusion set is 1 - **p**.

There is also a chance, at any given time, that observers will have made both the initial discrimination and a second discrimination. The second discrimination enables them to distinguish individual vehicle types. This may, but need not, involve different features for the BTR BMP confusion set and for the M60 T62 confusion set. Thus, we introduce two parameters for this second discrimination: **q** is the chance that observers will, by this time, have discriminated BTR's from BMP's, given that they have discriminated BTR BMP from M60 T62. **r** is the chance that observers will have discriminated M60's from T62s, given that they have discriminated BTR BMP from M60 T62. The chance of being in a singleton set is the product of the probabilities of the branches on the path leading to that set. For example, the chance of identifying an image of a BTR as a BTR is **pq**. The chance of identifying an image of a T62 as a T62 is **pr.**

Finally, the chance of being in a dyadic confusion set at any given time is the probability of arriving in that set multiplied by the probability of not leaving it. For example, there is a **p** chance of getting to the BTR BMP set, given that the stimulus is a BTR or BMP, and a chance **q** of then leaving it for the appropriate singleton. Thus, the chance of being in the BTR BMP set is **p**(1-**q**).
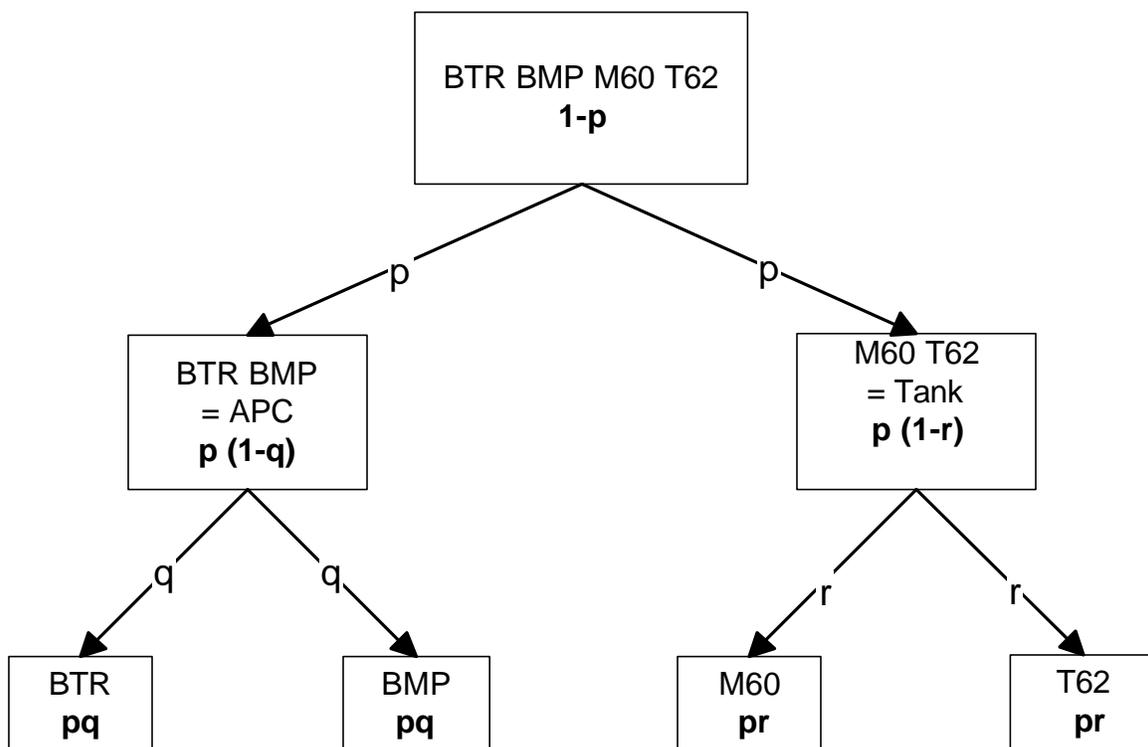

Figure 20. First hierarchical model of perceptual discrimination. Probabilities are conditional on presentation of the image of a particular vehicle, and apply only to nodes representing confusion sets that contain that vehicle.

There are three parameters in this model (**p**,**q**,**r**), which must be fit for each of the four temporal stages separately. At each stage, the three parameters can be estimated from a subset of

the discrimination parameters in Table 7 (also in Figure 18 and Figure 19) at that stage. The model can then be tested by deriving predictions for the remaining subset of parameters at the same stage. The parameter estimation process is quite straightforward. For example, it can be seen from Figure 20 that the value of **p** is 1 minus the discrimination parameter for the full confusion set. Once **p** is known, we use the discrimination parameter for the dyadic confusion set BTR BMP to solve for **q**, and the confusion parameter for the dyadic confusion set M60 T62 to solve for **r**. We can use the resulting estimates to predict the discrimination parameters for the singleton sets, BTR, BMP, T62, and M60. At each temporal stage, therefore, we have a test with 3 degrees of freedom (7 categories– 3 estimated parameters – 1 (since there is a constraint that all probabilities must sum to 1) = 3).

Figure 21 compares predicted and actual parameter values for the four singleton sets, i.e., the expected and actual probabilities of recognizing individual vehicle types. From a qualitative point of view, the predictions behave in a regular manner, i.e., the probabilities for recognizing an individual vehicle type parallel one another and increase in a regular way with time for all four vehicle types. However, the model fit is not perfect. If the model were correct, all points would fall on the dashed line, which they do not. Moreover, the deviations seem to be systematic.

Model predictions are based on the requirement that confusion sets based on the same perceptual discriminations should be equal in probability. The deviations of the IGM parameters from the processing model in Figure 21 can be interpreted in the light of that assumption. First, the probability of discriminating a BTR should equal the probability of discriminating a BMP, since the same information extraction steps account for both. However, in Figure 21 actual BTR probabilities are consistently higher than actual BMP probabilities. Second, the probability of discriminating an M60 should equal the probability of discriminating a T62, since the same information extraction steps account for both of these also. However, in Figure 21 actual M60 probabilities are consistently lower than actual T62 probabilities. Finally, we may recall that this model accounts only for two of the three confusions sets that received significant probability in Table 7; it does not account for the BMP T62 confusion set at all.

Model 2. The model parameters depicted in Table 7 (and Figure 19) suggest that there is another possible intermediate state. This state involves extracting enough information to discriminate T62's and BMP's from the other vehicles, but not from one another. However, the complementary confusion set, BTR M60, receives no probability in Table 7,. This implies that the information extracted places observers in the BMP T62 confusion set if presented with a BMP or T62, but is sufficient to discriminate specific vehicle types if presented with BTR or M60. Because of this direct path to recognition of BTR's and M60's, this model might predict their higher than expected recognition rate in Figure 21. A model that represents this alternative picture of visual processing stages is depicted in Figure 22.
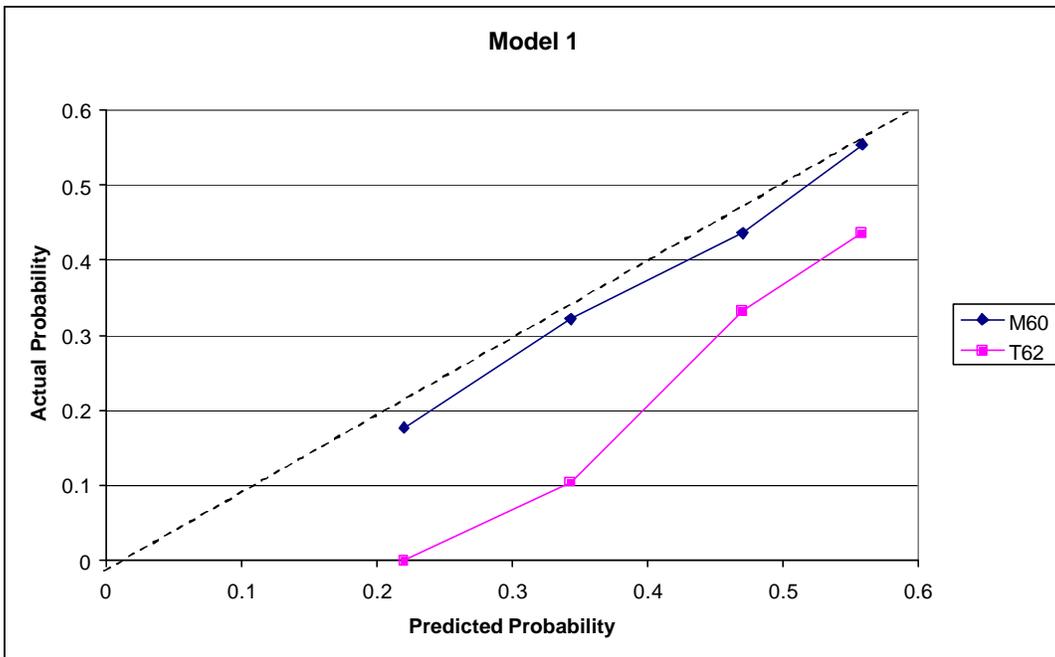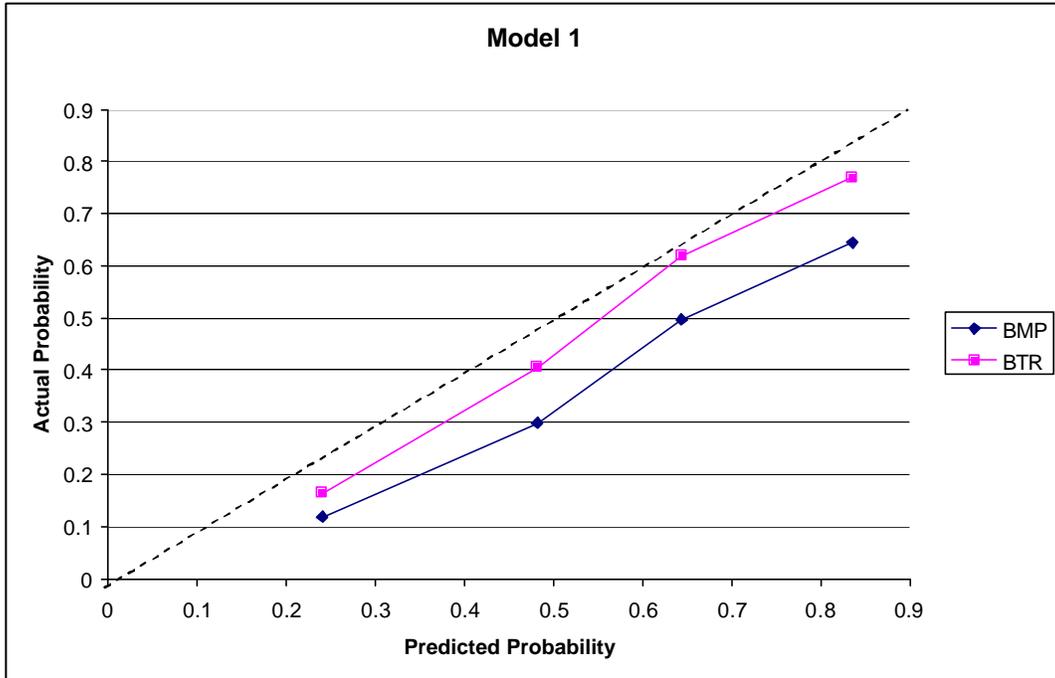
Figure 21. Comparison of predicted and expected probabilities of recognizing individual vehicle types according to Model 1.
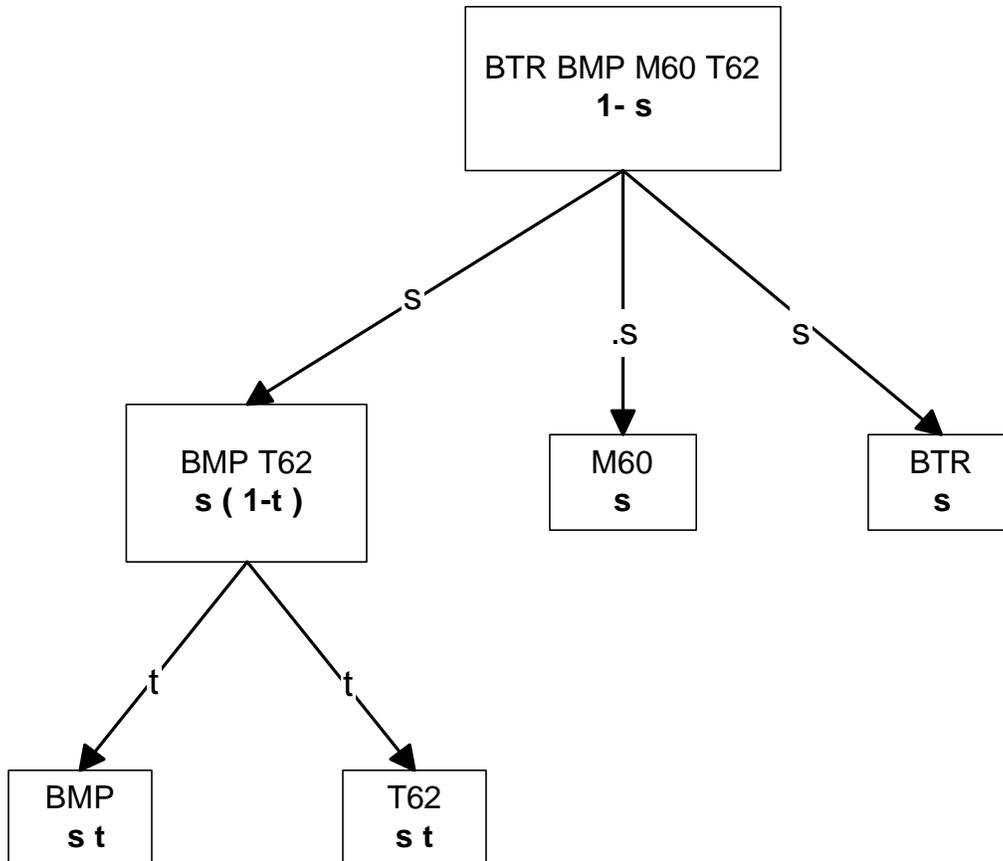
Figure 22. A second hierarchical model of perceptual discrimination.

This model involves only two parameters: **s** is the chance of the initial information extraction step, which may lead to the dyadic confusion set BMP T62, or to the singletons M60 and BTR, depending on the actual stimulus that is presented; **t** is the chance of the further information extraction step required to discriminate BMP's from T62's, given that the first step has already taken place. These parameters can be estimated separately for each temporal stage in the same way as the parameters for model 1. For example, it can be seen from Figure 22 that the value of **s** is 1 minus the discrimination parameter for the full confusion set. Once **s** is known, we use the discrimination parameter for the dyadic confusion set BMP T62 to solve for **t**. We can use the resulting estimates to predict the discrimination parameters for the singleton sets, BTR, BMP, T62, and M60. At each temporal stage, we have a test with 3 degrees of freedom (6 categories – 2 estimated parameters – 1 (for the constraint that all probabilities must sum to 1) = 3).

Figure 23 compares predicted and actual parameter values for the four singleton sets, i.e., the expected and actual probabilities of recognizing individual vehicle types. As for Model 1, the predictions behave in a fairly regular manner from a qualitative point of view, i.e., the probabilities for recognizing an individual vehicle type increase in a regular way with time for all four vehicle types. Once again, however, the actual parameters deviate systematically from those predicted by the model
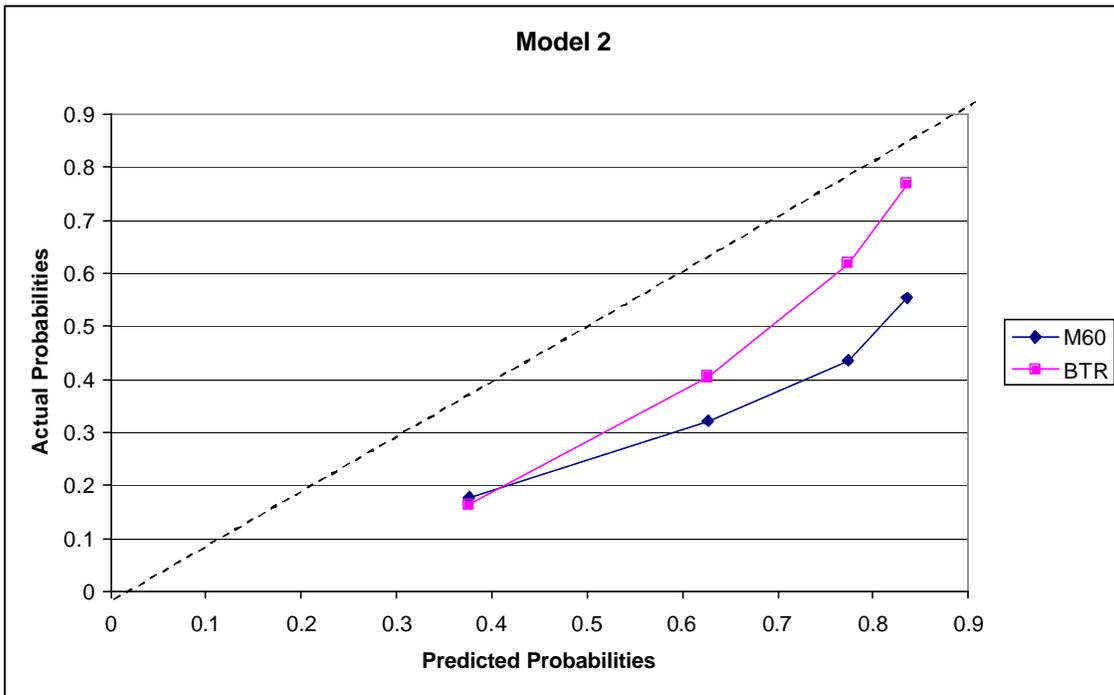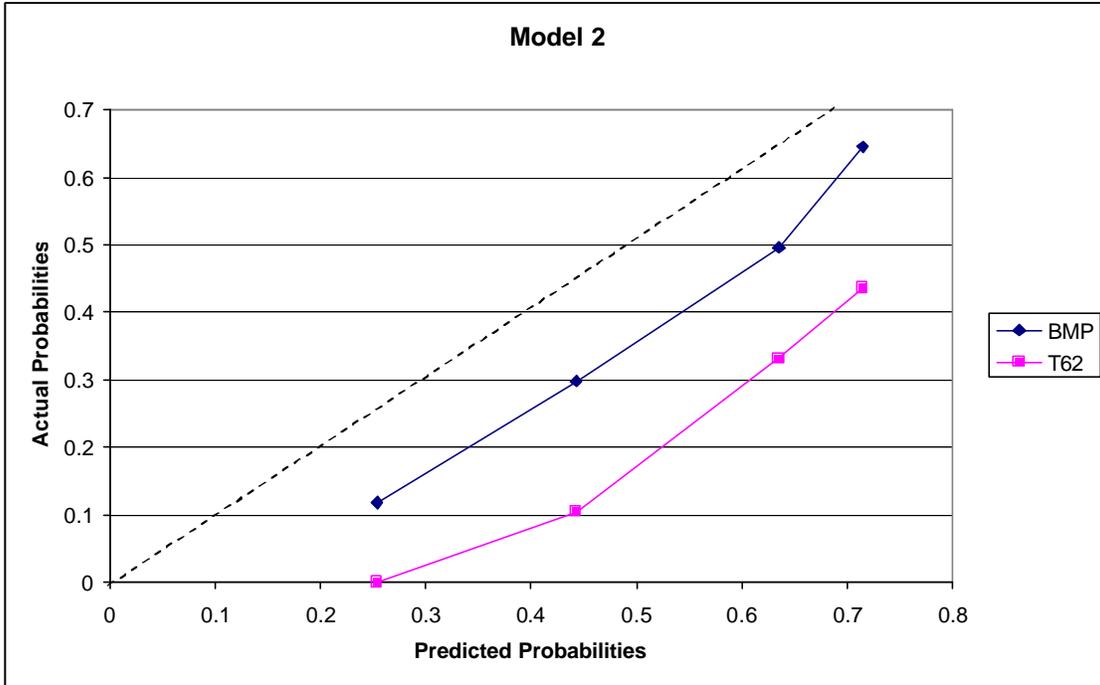
Figure 23. Comparison of predicted and expected probabilities of recognizing individual vehicle types according to Model 2.

This model predicts that the probability of discriminating a BMP should equal the probability of discriminating a T62, since the same information extraction steps account for both. However, Figure 23 shows that the actual chance of recognizing a BMP is systematically higher than the chance of recognizing a T62. The model also predicts that the probability of

discriminating an M60 should equal the probability of discriminating a BTR, since a single information extraction step accounts for discriminating the BMP T62 confusion set from BTR's and M60's as specific vehicles. This also is not the case. The actual chance of discriminating a BTR tends to be higher than the actual chance of discriminating an M60. Finally, this model does not account for two of the confusion sets that receive significant probability in Table 7.

Model 3. A more complex model, which mixes models 1 and 2, accounts for all features of the observed data. In the mixed model, shown in Figure 24, visual recognition can proceed along either of two processing pathways. One pathway, like Model 1, involves first discriminating APC's from tanks, and then APC's from one another or tanks from one another. The other pathway involves first discriminating M60's from BTR's and from BMP's or T62's, and then, if necessary, discriminating BMP's and T62's from one another. Any vehicle type can be recognized in either of these two ways. The predicted probability of recognizing a vehicle type is the sum of the probabilities of the two pathways that could lead to its recognition, as indicated at the bottom of Figure 24. For example, the chance of discriminating a BTR equals the sum of the probability of recognizing it via the BTR BMP confusion set ($\mathbf{pq}$), as in the first model, and the probability of recognizing it directly from the BTR BMP M60 T62 confusion set ($\mathbf{s}$), as in the second model.

Despite its larger number of parameters, the mixed model yields testable predictions for the probabilities of recognizing three of the four singletons. According to this model, an observer can be in any one of eight different perceptual states after receiving a stimulus (the eight different confusion sets represented by the boxes in Figure 24). Parameters from Table 7 for five of these sets (the full set, the three dyadic sets, and any one of the singleton sets) must be used in order to predict the remaining three singleton sets. In the following, we will use the probability of recognizing a BTR for the purpose of generating predictions for the other singletons.

First, inspection of Figure 24 shows that the difference between the probability of discriminating a BTR and the probability of discriminating a BMP (which are equal in Model 1, since both are APC's) should equal the probability of being in the BMP T62 confusion set (which was introduced by Model 2). Putting the same thing mathematically, the mixed model predicts that the chance of recognizing a BMP ($\mathbf{pq} + \mathbf{st}$) will be equal to the chance of recognizing a BTR ($\mathbf{pq} + \mathbf{s}$) minus the chance of being in the BMP T62 confusion set ($\mathbf{s}(1\text{-}\mathbf{t})$).

Similarly, Figure 24 shows that the difference between the probability of discriminating a BTR and the probability of discriminating an M60 (which are equal in Model 2, since they are both discriminated in the first step of visual processing) should equal the probability of the M60 T62 confusion set minus the probability of the BTR BMP confusion set (which were introduced in Model 1). Again, expressing this mathematically, the chance of recognizing an M60 ($\mathbf{pr} + \mathbf{s}$) should equal the chance of recognizing a BTR ($\mathbf{pq} + \mathbf{s}$) plus the chance of being in the BMP BTR confusion set ($\mathbf{p}(1\text{-}\mathbf{q})$) minus the chance of being in the M60 T62 confusion set ($\mathbf{p}(1\text{-}\mathbf{r})$).

A third relationship is also be expected based on Figure 24. The difference between the chance of discriminating an M60 and the chance of discriminating a T62 (which are equal in Model 1, since both are tanks) should equal the probability of being in the BMP T62 confusion set (which was introduced in Model 2). This relationship, however, is already implied by the two prediction that we have just derived. Mathematically, the chance of recognizing a T62 ($\mathbf{pq}+\mathbf{st}$) is the probability of recognizing an M60 ($\mathbf{pq} + \mathbf{s}$) minus the chance of being in the BMP T62 confusion set ($\mathbf{s}(1\text{-}\mathbf{t})$).
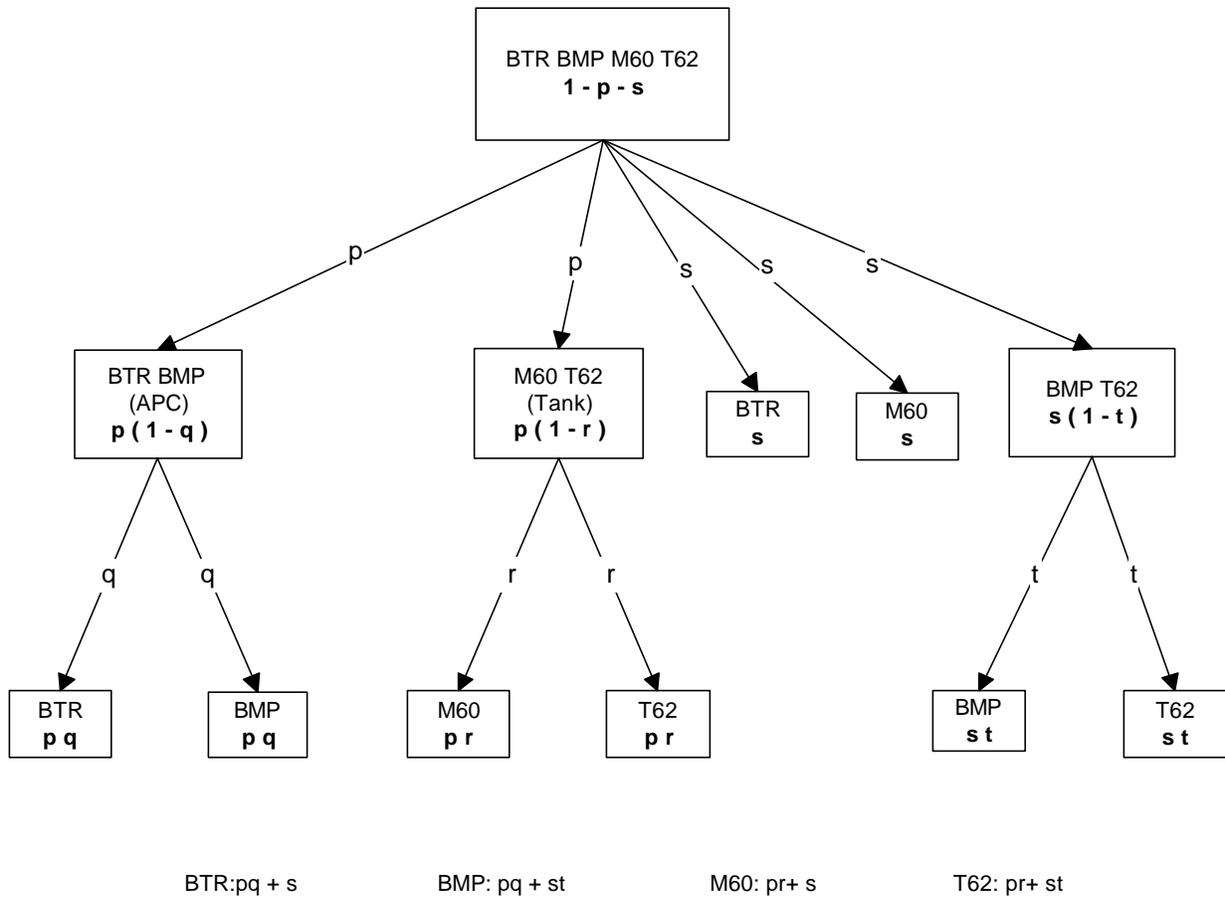
Figure 24. A probabilistic mixture of the two hierarchical models.

At each of the four temporal stages, Model 3 yields a test with 2 degrees of freedom: 8 categories minus the 5 categories used to generate the predictions, minus 1 (since all probabilities must sum to 1). Figure 25 shows that the mixed model fits the data exceedingly well. All points on the three curves fall near the predicted line. Moreover, the mixed model accounts for all of the qualitative features of the data that were not captured by one or the other of the two simpler models. These qualitative features are reflected in the relative locations of points for each of the singletons in Figure 25; for greater clarity, the singleton parameters are also shown in Table 9.

As shown in Table 9, the chance of discriminating a BTR is greater than the chance of discriminating a BMP at every temporal stage (both actually and as predicted by the mixed model), contrary to the prediction of the first model, in which they are equal. The mixed model explains this by the possibility of direct recognition of BTR's (as in the second model). Also as shown in Table 9, the chance of discriminating an M60 is greater than the chance of discriminating a T62 at every temporal stage (both actually and as predicted by the mixed model), contrary to the prediction of the first model. The possibility of direct recognition of M60's in the mixed model (as in the second model) explains this pattern.

Table 9. Probability of recognition predicted by Model 3 for BMP, M60, and T62, at each temporal stage of visual processing. Actual parameter values for BTR are also shown.

|	Stage 1	Stage 2	Stage 3	Stage 4
BTR (actual)	0.164	0.405	0.619	0.769
BMP	0.042	0.221	0.480	0.648
M60	0.143	0.266	0.445	0.491
T62	0.021	0.082	0.306	0.370
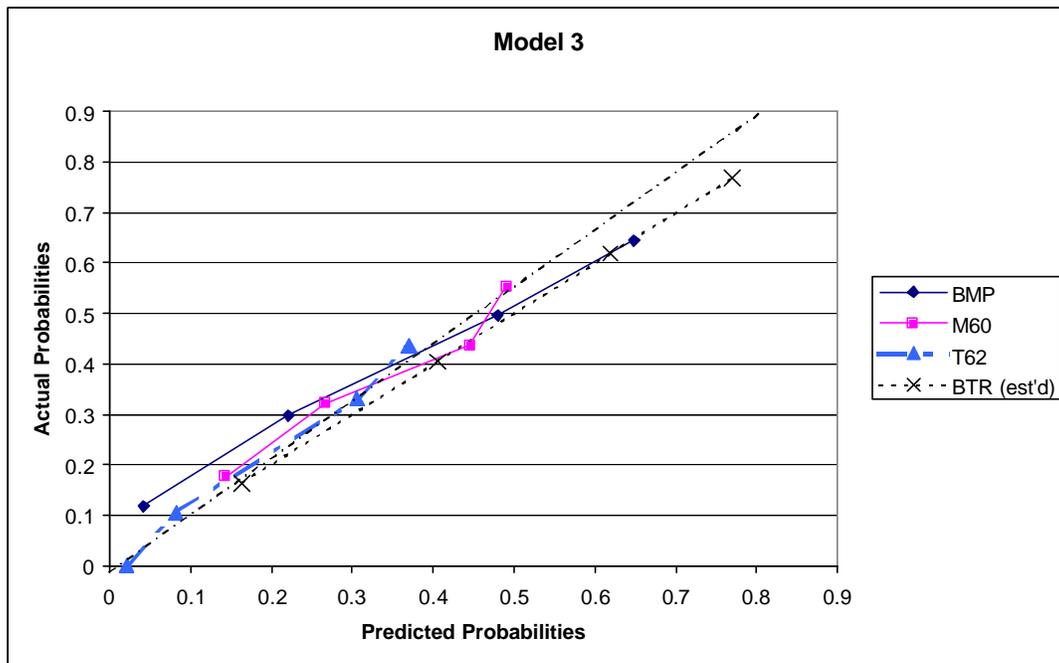
Figure 25. Comparison of predicted and actual parameter values for the Model 3.

Another pattern evident in Table 9 is the superiority of BTR recognition to M60 recognition and the superiority of BMP recognition to T62 recognition at every temporal stage. This is contrary to the predictions of the second model, according to which these probabilities should be equal. By contrast, the mixed model can accommodate this finding, as does the first model, by setting the parameter **q** for APC recognition higher than the parameter **r** for tank recognition. In addition, the mixed model makes specific predictions based on the relationship of **q** and **r**. The probability of recognizing a BTR (**pq** + **s**) minus the probability of recognizing an M60 (**pr** + **s**) is equal to **p** (**q** - **r**). Thus, a positive value for this difference implies that the probability (**q**) of the information extraction step that discriminates specific APC's is greater than the probability (**r**) of the step that discriminates specific tanks. Similarly, the probability of recognizing a BMP (**pq** + **st**) minus the probability of recognizing an T62 (**pr** + **st**) is also equal to **p** (**q** - **r**). The mixed model thus predicts that the latter difference will exactly equal the former difference at every temporal stage of processing. Table 10 shows that these values are positive at every temporal stage, and are identical.

Table 10. The mixed model predicts that values in each column will be equal.

	Stage 1	Stage 2	Stage 3	Stage 4
BTR – M60	0.021	0.139	0.174	0.278
BMP – T62	0.021	0.139	0.174	0.278

These aspects of the mixed model are qualitatively consistent with the studies reported in the previous chapter. Those studies showed that specific terms are more favored for APC's than they are for tanks. In the study on spontaneous naming, for example, subjects were more likely to use the intermediate label (*tank*) for images of tanks than to use specific labels (e.g., *T62*); but they were more likely to use specific labels of APC's (e.g., *BTR*) than to use the intermediate label (*APC*). Similarly, in the experiment on verification reaction times, participants were slower to verify tank images paired with specific labels (like *T62*) than they were to verify tank images paired with intermediate labels (*tank*). But they were just as fast verifying APC's with specific labels (like *BTR*) as verifying APC's with intermediate labels (*APC*). Such results regarding verbal labels may reflect the superior perceptual discriminability that is enjoyed by this set of APC's over this set of tanks.

Features Underlying Processing Stages

The model in Figure 24 involves two possible sequences by means of which vehicles can be recognized: In one sequence, subjects extract enough information to discriminate tanks from APC's, then extract additional information to discriminate members of those sets from one another. In another possible sequence, subjects first extract enough information to discriminate BTR's and M60's, and to discriminate both from BMP's and T62's, but not the latter from one another. They then extract information to discriminate BMP's from T62's.

What information is extracted at each of those stages? The feature naming study reported at the end of the previous chapter, and the set of multidimensional scaling and hierarchical cluster analyses based on them, provides a set of possibilities. Figure 26 indicates some of the features or categories of features verbalized by participants in the feature naming study that could have been used to make the relevant visual discrimination in the present study.

The first step in the first processing model can be accomplished readily by several of the feature categories. For example, tanks tend to cluster with other tanks, and APC's with other APC's, from the point of view of wheels and tracks, turret, and weapon, but not in profile. The second step night also involve a number of different categories of features. For example, APC's might be distinguished from one another by tracks (BMP) versus wheels (BTR), by height, by number of wheels, or by different characteristic profiles of the turret area. Among tanks, the M60 can be distinguished from the T62 by the presence of a relatively large, square turret, by hot tracks, by height or by number of wheels. The latter are for the most part subtle details rather than global features of shape. They may be particularly difficult to discern in small images. As a result, they may be available more slowly for decision making, explaining why they are extracted later in processing and also possibly accounting for the greater difficulty of discriminating specific tanks compared to discriminating specific APC's. The superior discriminability of APC's from one another, compared to tanks, may be due to the existence of more features

capable of discriminating them, or to the higher salience of the tracked versus wheeled distinction.

The second processing model presents an almost opposite case. The BMP and the T62 are clustered together, while the two APC's (BMP and BTR) and the two tanks (T62 and M60) are distinguished early in processing. Examination of the multidimensional scaling and hierarchical classification results in Appendix A shows that the from the point of view of wheels and tracks, turret, and weapons, this result is most unlikely. The non-profile categories (wheels and tracks, turret, and weapon) all place the two APC's *maximally close* to one another. It seems unlikely that the most difficult discrimination would be made first in the processing sequence. Moreover, in the turret and weapon categories, the T62 and the BMP are far apart instead of clustered together as they are in the second model. For wheels and tracks, T62 and BMP are closer, since both have cold tracks (contrasted with wheels for BTR and hot tracks for M60), but they are still not in the same maximally related set.

There is only one category of features that efficiently produces the confusion sets predicted by the second visual recognition model, and that is profile. (For example, the BTR's height is low, and the M60's is high, permitting these two vehicles to be readily discriminated; the BMP and T62 are of similar intermediate height.). The similarity space and clusters generated by profile almost perfectly match the clusters generated by the second model. BMP and T62 are maximally close, and share the same initial cluster in the hierarchical analysis. Moreover, the BTR and the BMP are highly separated.

Perhaps the most salient aspect of Figure 26 is the likelihood that profile, in either the narrow sense used here or in a broader sense that includes the overall shape of the turret and the size of the weapon, plays an important role in recognition. It is clear that this concordance of verbally stated features and stages of perceptual processing should only be regarded as heuristically suggestive. However, the pattern of recognition stages depicted in Figure 26 confirm the importance both of overall vehicle profile and specific vehicle features, and is thus consistent with the feature naming analysis at the end of the last chapter. In the next section, we test the importance of profile in recognition more directly, by examining the effects of selective enhancements of aspects and parts of images.
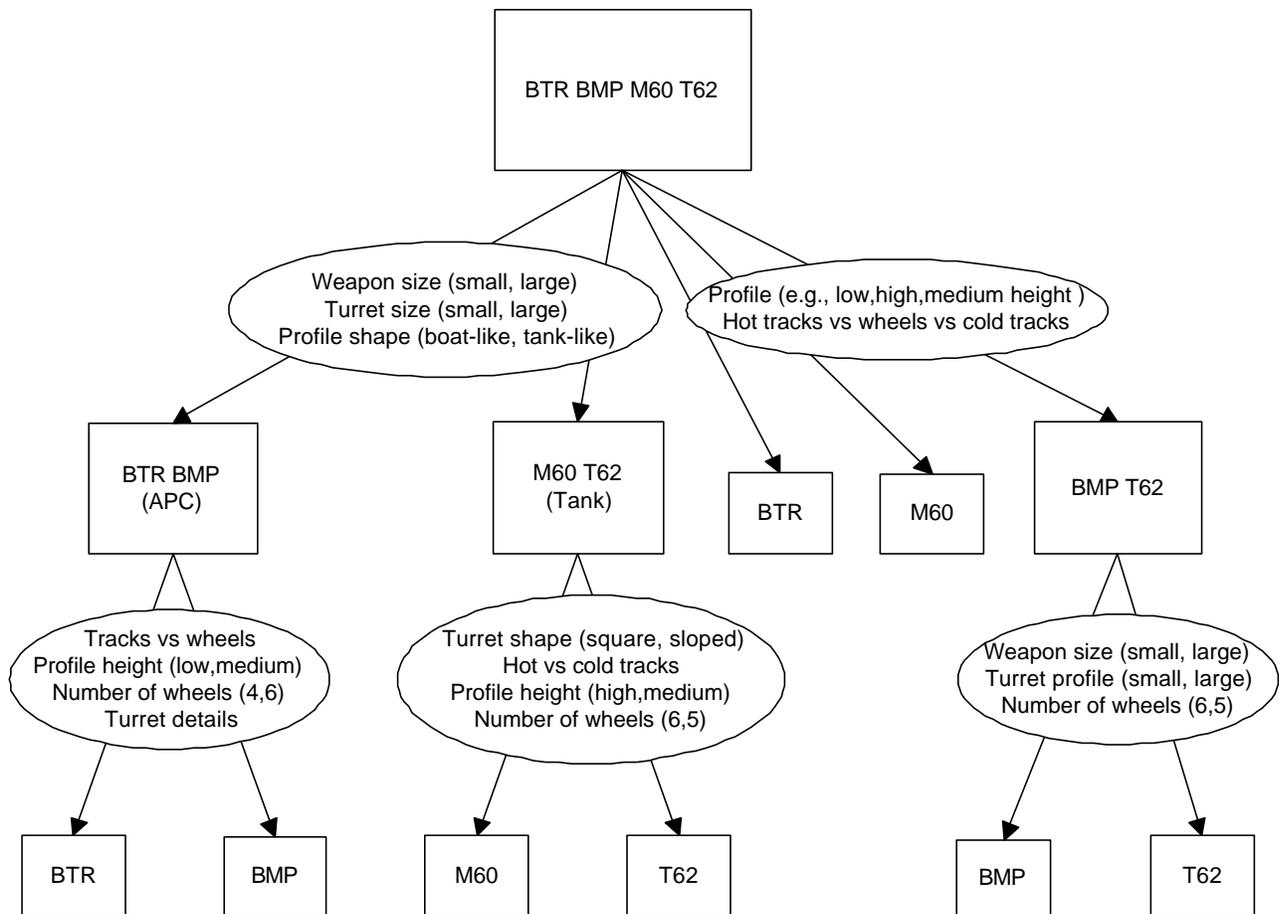
Figure 26. A model of the features extracted in perceptual processing.

Image Enhancements

Introduction

If observers use features of images to make rapid visual discriminations, enhancing those features may improve recognition performance. Enhancement of images may also, however, slow down recognition, or increase errors, if, for example, the features that are enhanced are irrelevant for the required discrimination, are not easily utilized by human observers, and/or distract attention from features that are more useful. In this section, we explore the potential of selective visual enhancement of images to improve recognition performance. If properly designed, such enhancements may help users of ATR's verify ATR conclusions more rapidly and accurately.

A second goal of studying image enhancements is to test hypotheses about the features actually used in visual recognition. If enhancement of the region or aspect of an image that contains a particular feature or features reduces confusion errors (e.g., tanks vs. APC's, or M60's vs T62's), it is likely that the enhanced feature or features play a role in the relevant discriminations.

Design

There were four types of image enhancement: (1) guns and turret, (2) tracks and wheels, (3) the entire vehicle, and (4) the vehicle silhouette. Figure 27 provides an example of the effects of two very different kinds of enhancement on the image of an M60 tank. In one image, details of the tracks and wheels have been heightened, by enhancing contrast in that area of the image. In the other image, all internal details have been suppressed in order to heighten the contrast of the vehicle silhouette against the background. Appendix C shows all the enhancements for all vehicles, at a single, oblique viewing angle, and in a scale that is closer to the true size of the simulated FLIR images that were presented.
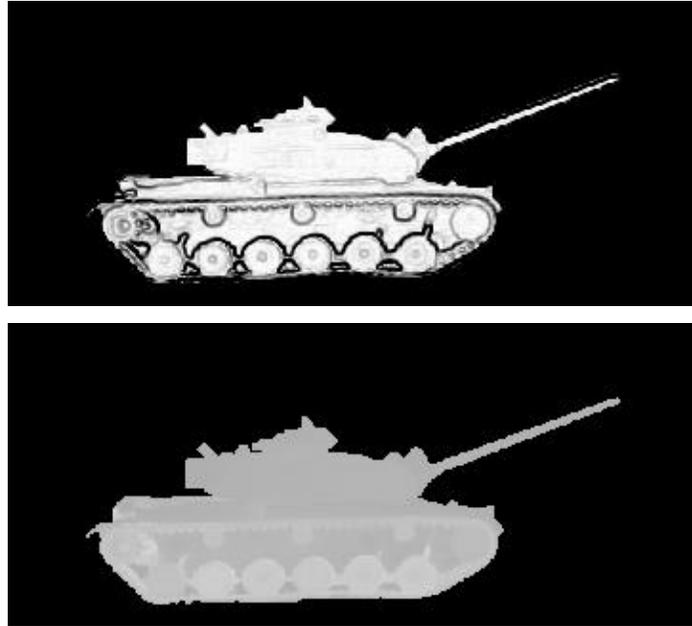
Figure 27. Examples of enhanced images. Top: Enhanced tracks and wheels. Bottom: Enhanced silhouette.

As already noted, 9 participants received only unenhanced FLIR imagery, while 10 participants received both unenhanced imagery and various combinations of enhanced imagery in different blocks of trials (see Table 6). Each participant, in each enhancement condition to which that participant was exposed, saw an average of about 42 presentations of each of the four vehicles types (BMP, BTR, T62, M60). These presentations varied in size, angle, and range of the vehicle and in stimulus window duration. Participants who viewed only enhanced imagery saw about four times as many instances for each vehicle type (i.e., about 44 instances of a vehicle per stimulus duration). Because of the smaller number of data points, we did not to fit the Informed Guessing Model to the enhanced imagery data. Recall that high error rates were intentionally induced by imposing short stimulus durations.

Results

Overall Error Rates

Figure 28 shows the effect of image enhancement on overall recognition error rates. The overall effect of the enhancement manipulation on recognition accuracy was highly significant ($\chi^2_4 = 135.69$; $p < .001$). In addition, all differences among specific manipulations in Figure 28

are statistically significant ($p < .001$), except the difference between tracks/wheels and gun/turret, and the difference between unenhanced and entire vehicle.
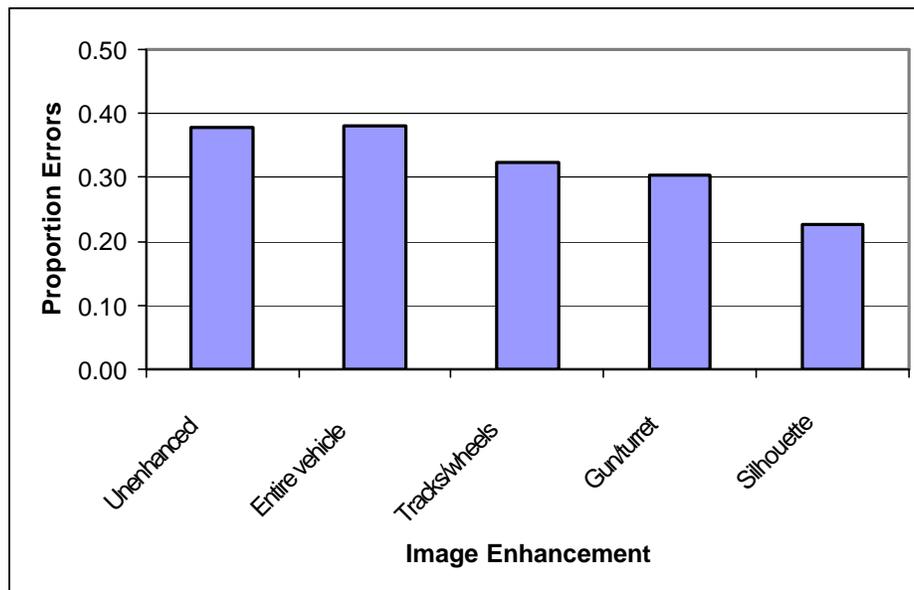
Figure 28. Errors rates in different image enhancement conditions.

38% of the responses to unenhanced images were incorrect. Enhancing tracks and wheels reduced the error rate to 32% ($\chi^2_1 = 16.48$; $p < .001$), while enhancing gun and turret reduced the error rate to 30%. ($\chi^2_1 = 35.61$; $p < .001$). Based on these results, it might be expected that enhancing the entire vehicle, which is equivalent to enhancing *both* track/wheels and gun/turret, would also improve recognition performance. However, enhancing the entire vehicle did not affect performance. In fact, there was a statistically insignificant 1% decrement in performance associated with full enhancement.

Perhaps the most surprising result is the effect of silhouette enhancement. This condition represents a strategy diametrically opposed to the other enhancement manipulations. Rather than making details of the image easier to perceive, it obscures them in order to heighten the contrast between the overall profile of the vehicle and its background. Silhouette enhancement had the largest effect on recognition performance, reducing errors to 23%, a 65% improvement in comparison to unenhanced imagery. This was a statistically significant improvement over the 30% error rate associated with the next most effective manipulation, i.e., enhancement of gun/turret ($\chi^2_1 = 20.31$; $p < .001$).

It might be supposed that different enhancement methods would be optimal depending on what the vehicle actually is. For example, enhancing guns/turrets might be the best way to support recognition of a T62 or M60, while enhancing track/wheels might be the best way to help observers discriminate a BTR from other possibilities. As shown in Table 11, the data do not support this hypothesis. Silhouette enhancement produced the best recognition performance for all four types of vehicles.

Table 11. Proportion errors for each enhancement condition, broken down by the actual type of vehicle present in the image.

	Unenhanced	Entire vehicle	Tracks/ wheels	Gun/ turret	Silhouette
BMP	0.32	0.25	0.25	0.19	0.18
BTR	0.32	0.31	0.29	0.25	0.19
M60	0.40	0.46	0.40	0.38	0.37
T62	0.46	0.49	0.37	0.39	0.31

Other factors might also be expected to influence the optimal type of image enhancement. For example, different features might be relevant depending on the angle at which the vehicle is viewed, the distance at which it is viewed, and the amount of time available for viewing it. We examined the effect of the different enhancements under each of these variations. As shown in Table 12, Table 13, and Table 14, respectively, the silhouette enhancement was superior no matter how the data were broken down. Silhouette enhancement outperformed part or whole vehicle enhancement, as well as the unenhanced condition, whether the vehicle was viewed from the side or obliquely, whether the vehicle was at short range or long, and whether the vehicle was viewed for a relatively short or a relatively long period. (The latter manipulation compared the two shortest stimulus durations for each participant with the two longest stimulus durations for each participant.)

An interesting possibility is that enhancement may partially neutralize the effects of range or time constraints on recognition. This appears to be the case. Silhouette enhancement produced a larger reduction in errors for distant vehicles than for closer vehicles (Table 13). Similarly, it reduced errors more for shorter stimulus exposures than for longer stimulus exposures (Table 14).

Table 12. Proportion errors for each enhancement condition, broken down by the angle at which the vehicle is viewed.

	Unenhanced	Entire vehicle	Tracks/ wheels	Gun/ turret	Silhouette
oblique	0.40	0.47	0.34	0.30	0.22
side	0.37	0.35	0.32	0.32	0.23

Table 13. Proportion errors for each enhancement condition, broken down by the range at which the vehicle is viewed (i.e., image size).

	Unenhanced	Entire vehicle	Tracks/ wheels	Gun/ turret	Silhouette
long range	0.51	0.49	0.43	0.41	0.31
short range	0.33	0.37	0.29	0.28	0.20

Table 14. Proportion errors for each enhancement condition, broken down by the amount of time available for viewing the image.

	Unenhanced	Entire vehicle	Tracks/ wheels	Gun/ turret	Silhouette
shorter windows	0.50	0.45	0.41	0.36	0.27
longer windows	0.28	0.36	0.25	0.27	0.19

Confusion Errors

How did the enhancement manipulations affect the pattern of confusion errors among the four vehicles? Even though we did not fit the Informed Guessing Model to these data, analysis of the confusion errors may shed light on the way different enhancements interact with visual processing.

Figure 29 shows the result of multidimensional scaling of the four vehicle types based on confusion data from the five imagery enhancement conditions. Each of these represents a zero-stress fit (based on Kruskal's loss function) within a two dimensional Euclidean space. The distance between any two vehicles represents their dissimilarity, or lack of confusability, under the conditions of that particular image manipulation. (Note that the orientation of the vertical and horizontal dimensions of this space are arbitrary.)

With unenhanced imagery (upper left, Figure 29), the two tanks (M60 and T62) are relatively close to one another, while the two APC's (BTR and BMP) are more easily distinguished both from one another and from the tanks Enhancement of the entire vehicle (upper right of Figure 29) did not change this qualitative pattern.

Two manipulations enhanced only a part of the image: gun/turret and tracks/wheels (middle row, Figure 29). These had quite similar and pronounced effects on the confusability relations among the vehicles. In particular, the tanks moved farther apart from one another, while the APC's moved closer to one another and to the tanks. The most pronounced qualitative change, however, was that all the discriminations now seem to coverage on a single dimension, For the track/wheels condition, the four vehicles are virtually lined up along the horizontal dimension, with no differences at all along the vertical dimension. For gun/turret, similarly, there is only a little variation along the vertical. It is as if the observers were using less, rather than more, information in the part-enhancement manipulations.
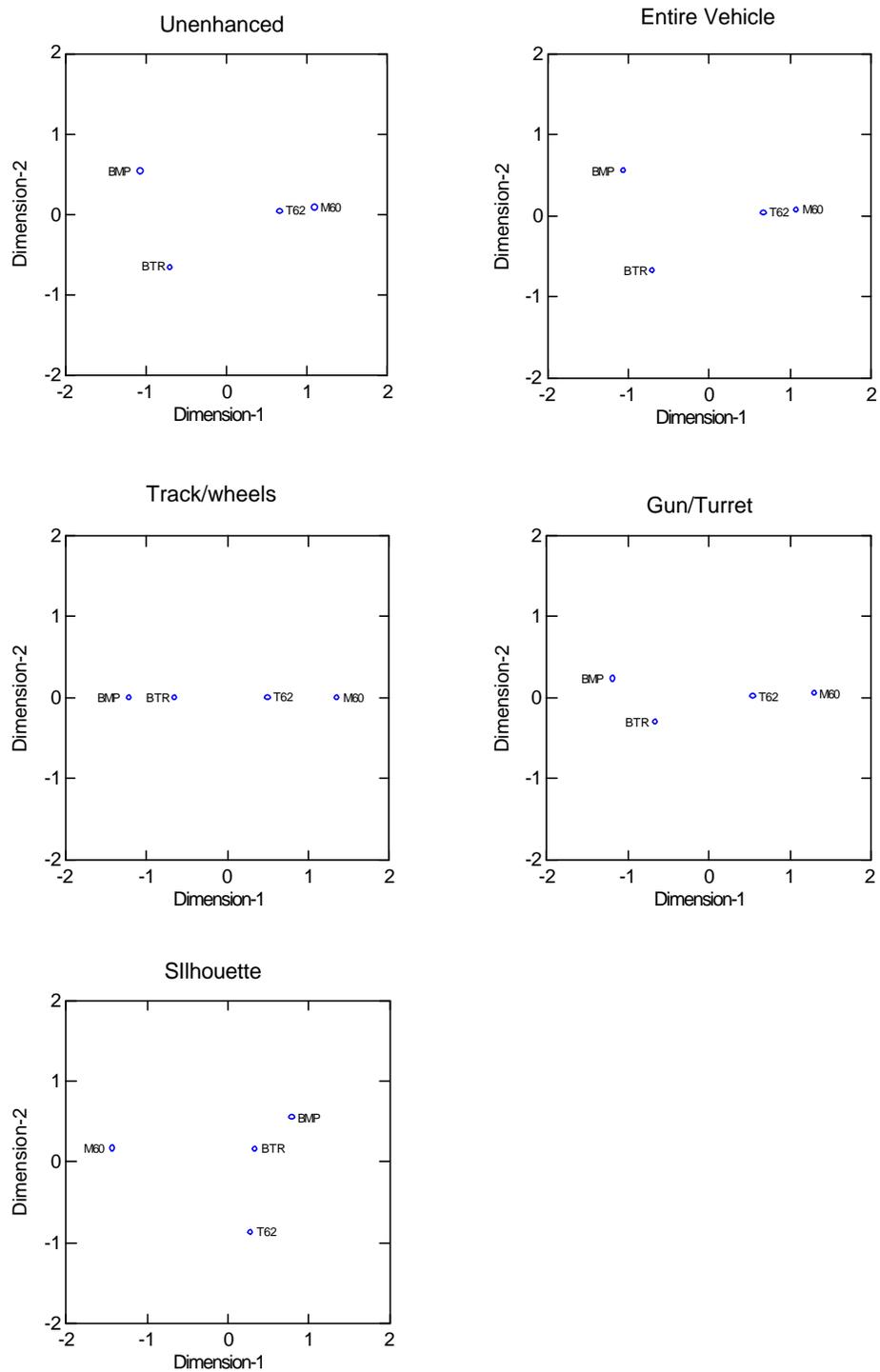
Figure 29. Multidimensional scaling of four vehicle types based on confusion errors in different image enhancement conditions.

The effect of silhouette enhancement (bottom, Figure 29) is dramatically different. There is no reduction in spread along either dimension. Rather than compressing the spatial representation, silhouette enhancement reorganizes it. The tanks are now easily distinguished from one another and from the APC's, while the APC's have moved slightly closer together. Presumably, silhouette enhancement leads observers to use different features of the vehicles for recognition. The corresponding improvement in the overall error rate (as described in the preceding section) suggests that the features induced by silhouette enhancement may sometimes be more effective.

The silhouette-based similarity space contrasts in richness with spaces based on the other enhancements. This result, which is based on confusions in rapid visual recognition, is reminiscent of our finding in the first chapter, that profile-based *verbalized* similarity required more dimensions for its representation than any of the other three categories of features.

Discussion and Implications for ATR Design

The results confirm the importance of profile in rapid visual recognition of objects. Within the range of conditions that we tested, silhouette enhancement was significantly superior to two forms of part enhancement (tracks and wheels, and gun and turret), as well as to their combination. In fact, enhancing all the details in the figure appeared to degrade performance in comparison to the unenhanced condition, under some circumstances. Both conditions in which only parts of the image were enhanced consistently improved recognition performance. However, silhouette enhancement was superior to them under all circumstances. Ironically, silhouette enhancement, which suppressed details within the image, produces more differentiation and a more complex similarity structure, than the part enhancement conditions, in which internal details of the image were selectively heightened.

An implication for ATR design might be to enhance images by heightening the contrasts of their silhouettes against the background, even at the expense of some internal features. This conclusion, of course, only applies to the conditions of this study: rapid recognition of vehicles viewed from the side and oblique angles. Additional research will be required to determine if the results extend to longer viewing periods or to front angles of view.

4.   DECISION MAKING UNDER UNCERTAINTY

Introduction

We argued in the Introduction that recognition does not take place in a vacuum. It occurs in the service of other decisions, such as selecting a target or avoiding threats, upon which battlefield success depends. Selection of an appropriate target in particular usually depends on accurate and timely classification. A major concern of this set of experiments will be to test hypotheses about how observers perform and how their performance can be improved in this larger context.

The third set of experiments broadens the scope of the inquiry in several ways: (1) It introduces significant uncertainty about the accuracy of ATR classifications. (2) It introduces a mission (to engage enemy tanks) in which the costs of an error depends on the correct target classification (e.g., mistakenly engaging a friendly vehicle is worse than mistakenly engaging an enemy jeep). (3) It introduces multiple possible targets that may be classified and either engaged or not engaged. And (4) it imposes a more realistic time constraint, in seconds rather than fractions of a second, and applied to recognition and engagement decisions for multiple rather than single targets.

In this chapter, as in the previous two chapters, the concern is both descriptive and prescriptive. On the descriptive side, we will ask how three factors influence decision making performance: uncertainty about ATR conclusions, time constraints, and costs of errors. On the prescriptive side, we will investigate strategies by means of which ATR's might support user performance under conditions of time stress, high stakes, and uncertainty. One issue is whether the ATR should attempt to draw the user's attention to targets for which the ATR is uncertain and for which the costs of an error are high (e.g., the target could be a friendly vehicle or it could be an enemy tank). A second issue is how an ATR should label its conclusions when they are uncertain. We will compare a labeling system based on the *favored* labels explored in the first set of studies (Chapter 2), to a variety of other approaches, including one that adapts dynamically to the degree of uncertainty in the conclusion.

Both the descriptive and prescriptive aspects of this experiment are guided by a model of trust in decision aids (Cohen, Parasuraman, Serfaty, & Andes, 1997). The model attempts to specify conditions under which decision aid users should take time to verify an aid's conclusion, and when they should act immediately (e.g., to engage a target). We describe that model in the remainder of this section and in Appendix D.

<u>A Model of the Verification Decision</u>

The user of an ATR, like any decision aid user, sometimes must choose between immediately acting on an aid recommendation (by accepting it, modifying it, or rejecting it), on the one hand, and collecting more information about the recommendation, on the other hand. Immediate action is often irreversible, like engaging a target, but it may also be provisional, such as deciding not to engage the target. Before engaging a contact labeled by an ATR as an enemy target, pilots may want to verify the classification for themselves. Verification in the case of an ATR includes looking carefully at the image of a target. The decision to verify or act immediately comes up again, as the user decides how long to continue examining the image before making a final decision.

A convenient tool for representing this kind of decision, and for building benchmark models of verification decisions, is the decision theoretic concept of *value of information* (VOI) (Cohen & Freeling, 1981; Raiffa & Schlaifer, 1961; LaValle, 1968). A simple formula for value of information, as applied to verification decisions by ATR users, is the following:

Value of verification information =
Sum over all observational outcomes that would change the user's subsequent decision
[probability of the observational outcome *
change in expected payoffs due to the change in decision
– cost of time spent making the observation]

The user should verify the ATR's conclusion, and continue doing so, as long as this value is greater than zero. Value of information is a significant improvement over other information measures, such as entropy reduction, which measure the sheer "quantity" of information without taking into account the reason why information may be of value, i.e., its actual role to support decision making. (See Appendix D for a more detailed derivation of this rule.)

Dynamic Constraints on Verification Decisions

Despite their generality, measures based on the value of information have limitations. Primarily, these limitations flow from the requirement that the possible observational consequences of verification be specified explicitly in advance (Cohen & Freeling, 1981). A significant advantage of interactive versus fully automated systems may be the human ability to handle novel and unexpected situations. In these cases, the results of human intervention may not be known ahead of time. There are several, closely related problems:

Visual recognition. The verification process may be very straightforward, yet the potential observations cannot be anticipated. For example, the user of a target identification aid can verify identification of an image as a hostile tank simply by looking at the image, yet it might be very difficult to specify in advance all the relevant details that the user might see.

Critical thinking. The verification process itself may be less straightforward in some situations. For example, conflict between an aid's recommendation and their own or others' conclusions, may prompt a process of critical thinking, in which users look for an explanation of the differing recommendations. Resolution of the conflict may take the form of discovering unreliable assumptions that were implicit in the soldiers' conclusions or the aid's. It is virtually impossible to make all assumptions explicit in advance. Key assumptions may come into focus only when they lead to problems, such as conflicting recommendations (Cohen, Freeman, & Thompson, 1997).

Novel situations. More generally, new issues to investigate may spring up as a result of unique or unusual circumstances, or due to the pattern of ongoing verification results. Just as novel situations may not be anticipated by the designer of a decision aid, so they may not be anticipated by the training designer.

Information interdependence. The value of one piece of information may be very low when considered by itself, but high when considered in the context of other possible observations; or the significance of one piece of information may be unclear until other pieces of the puzzle are obtained. The decision of whether or not to verify the

first feature must therefore take into account the possibility of continuing the verification process to include the other features.

Fortunately, these difficulties can be surmounted without giving up the essence of the value of information approach. We will describe a simple framework for deriving benchmark models of verification performance, without specifying all possible observations. The framework can, therefore, be applied to situations where previously learned or explicitly identified patterns may be insufficient to guide decisions about user-aid interaction. This framework will thus apply even when verification involves visual recognition of unanticipated patterns, critical thinking that ferrets out hidden assumptions, creative problem solving in novel situations, and interlocking or reinforcing pieces of information. The solution is to derive necessary conditions, or *constraints*, that must be satisfied if any verification at all is to be of value. If the situation does not satisfy these constraints, verification cannot be worthwhile, regardless of the number of unmodeled potential observations and insights. These constraints need not be static, but may change dynamically as a the situation itself evolves.

Rather than attempting to model individually all the observations that could be made during verification, we will assume that *perfect information* is obtained. A simplified model can be obtained by assuming that verification will produce observations that are perfectly correlated with outcomes that determine the appropriate reliance decision. If verification is not worthwhile under this assumption, then it cannot be worthwhile under more limited conditions. It turns out that these constraints can be expressed relatively simply, in terms of current trust in the aid by itself, the costs of verification, and the potential affect of verification on payoffs. In particular, users should accept an aid recommendation without verification if:

trust > 1 - cost of verification / the cost of incorrectly accepting the aid recommendation

If the aid conclusion is binary (e.g., classification of a contact as friend *or* foe), users should reject an aid recommendation without verification if:

cost of verification / the cost of incorrectly rejecting the aid recommendation > trust

Users *may* choose to verify if neither of the above is the case, i.e.:

1 - cost of verification / the cost of incorrectly accepting the aid's recommendation > trust > cost of verification / the cost of incorrectly rejecting the aid's recommendation

Trust in these equations refers to the chance of a successful aid recommendation. (See Appendix D for a more detailed derivation of these constraints.)

Figure 30 represents a benchmark model based on these constraints. Trust in the ATR (i.e., the probability that this particular ATR conclusion is correct) is shown on the vertical axis, ranging from no trust (0) to complete trust (1.0). Time is plotted along the horizontal axis. Thus, the long-dotted line shows that confidence in the aid begins relatively low; in fact, if further verification was not possible, this user would choose to reject the aid conclusion. However, trust increases in this example with time spent verifying the conclusion. This might happen, for example, if the image of a vehicle that the aid has classified as an enemy tank in fact has features associated with a T62. Of course, confidence could also have declined as new evidence was considered, for example, if the image of the vehicle turned out to have features associated with a friendly APC.

upper bound = 1 - cost of delay / cost of
incorrectly accepting aid's conclusion

1.0

Trust
(Probabiliity
Aid
Conclusion
is Correct)

Accept aid
conclusion

Verify
conclusion

Reject aid
conclusion

0

lower bound = cost of delay /
cost of incorrectly rejecting aid's
conclusion
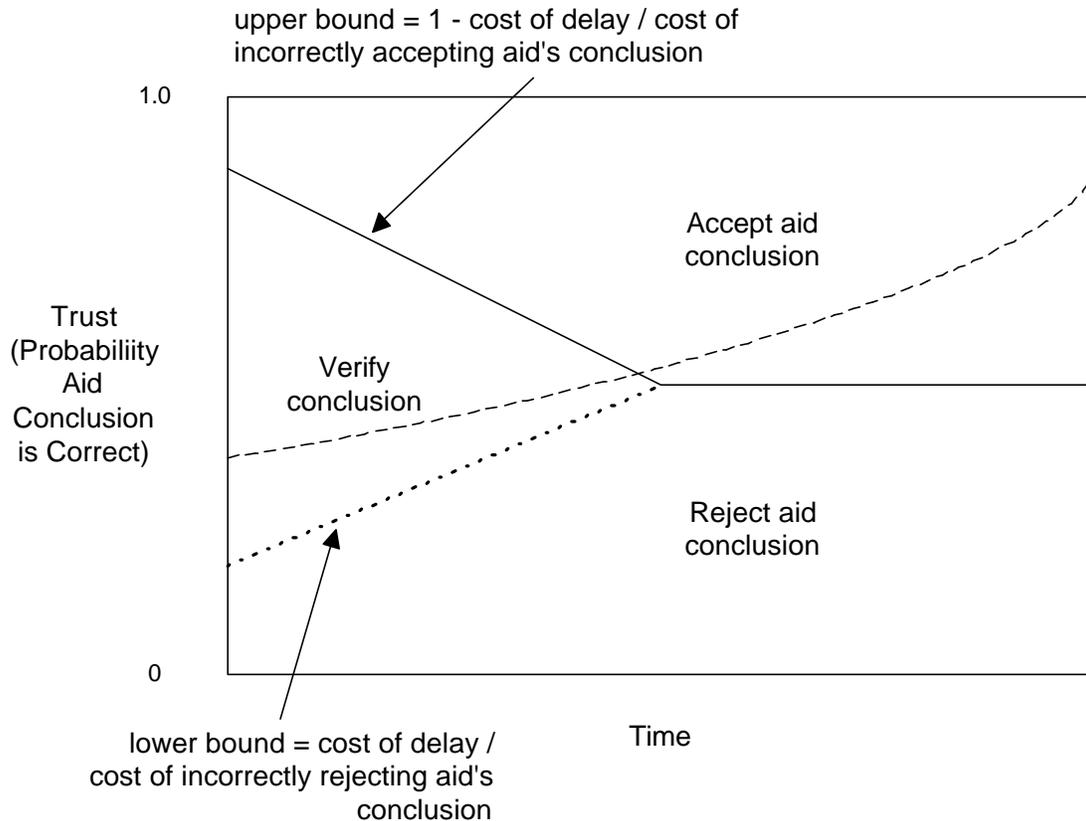
Time

Figure 30. Benchmark model for deciding when to accept, reject, or take time to verify a decision aid's conclusion. Trust is represented by the long-dotted line.

At any point in time, the vertical dimension is divided into two or three regions. If trust in the aid's conclusion falls in the upper region, the user should simply accept the conclusion (e.g., engage the target), without taking further time for verification. If trust in the aid's conclusion falls in the lower region, the user should reject the aid's conclusion without taking further time. (For example, a target identification aid concludes that a vehicle is an enemy tank, but the user is reasonably sure based on visual identification that the target is a friendly.) If trust is neither high nor low, but falls in the intermediate region, then it may be worthwhile for the user to take more time to decide what to do. In Figure 30, the user's initial level of trust warrants further verification of the aid's conclusion. After a while, however, the user's confidence in the aid has increased enough to enter the upper region, where its conclusion should be accepted. At this point, the user should stop thinking and act.

What determines reliance decisions in this model? This surprisingly powerful representation has only three key variables: uncertainty, time stress, and stakes.

1. *Uncertainty* pertains principally to the *resolution* of the trust assessment, i.e., the proximity to zero or one of the probabilities discriminated by the user. The less resolution in the user's assessment of trust, the more likely that a calibrated assessment will fall in the middle region of Figure 30, and the user will tend to utilize more time before making a decision. The resolution of a trust assessment is influenced by the *completeness* of the user's knowledge of conditions that affect system performance. The more complete the knowledge of relevant features of the domain, situation, task, and system, and the more reliably these features are

observed on a given occasion, the closer the user's initial trust assessments will come to zero or one. Informed users will be able to assess the value of aid recommendations more quickly.

3. Time stress is represented by the cost-of-delay parameter in the equations determining the upper and lower bounds (see Figure 30). As time stress increases, the upper boundary moves down and the lower boundary moves up, reducing the size of the intermediate region where verification might be appropriate. For example, the risk of being targeted by an enemy may increase with the time spent unmasked, e.g., to verify the enemy's identity. Time stress can also increase due to neglect of other tasks, which grow increasingly urgent. When the cost of delay is great, action is more imperative, even with relatively low-resolution levels of trust. The cost of delay need not be constant, but may itself be a function of time. Figure 30 illustrates such a case, in which the cost of each further moment of delay is higher than the one before, until finally no further delay is justified. When the upper and lower bounds meet due to increasing costs of delay, the user must act, regardless of the level of trust.

2. *Stakes.* The locations of the boundaries in Figure 30 depend on the relative costs of being in each of the three regions. We have already considered the cost associated with being in the middle region; time stress affects the upper and lower bounds symmetrically, driving them closer together as it increases. By contrast, there are two different kinds of stakes, corresponding to the costs of mistakenly accepting or rejecting the aid's conclusion, respectively. These two kinds of costs affect the two bounds independently. To think about stakes, the user simply asks, regarding whatever action he or she is about to take, *what are the consequences if I am wrong?* The more severe the consequences of a mistake, the more difficult it is to clear threshold for taking the corresponding action (i.e., to get into the upper or lower region).

The upper bound increases (and the upper region gets smaller) with the cost of incorrectly *accepting* the aid's recommendation. For example, suppose that a target identification aid recommends engagement of a contact, and the user is considering accepting this recommendation. For the sake of argument, suppose that the user and the aid are wrong and that the contact is not an appropriate target (e.g., it is a friendly vehicle or an enemy non-target). Stakes are defined simply as the average difference in the expected value of engaging such a contact and not engaging it. For example, incorrectly engaging a contact is likely to be more costly, the higher the proportion of friendlies among the non-targets. Thus, increasing the number of friendlies in the area will raise the upper bound, making it harder for trust to clear the threshold for acting on the aid's recommendation to engage.

By contrast, the lower bound decreases (and the lower region gets smaller) when there is an increase in the cost of incorrectly *rejecting* the aid's recommendation. For example, suppose again that the aid recommends engagement, but the user this time is leaning away from engaging. For the sake of argument, suppose that the user is wrong in rejecting the aid's recommendation, and that the contact is in fact an appropriate target. Stakes are defined as the average difference in value between not engaging such a contact and engaging it. (This is parallel to the definition of stakes for the upper bound, except that we now assume the contact is an appropriate target.) For example, the cost of failing to engage a target are higher the more threatening the target is to one's own platform or to other friendly assets; the cost is also higher if the value of the threatened assets is higher. Thus, a bigger threat or more valuable assets to be protected can reduce the lower bound, making it harder for distrust to clear the threshold for rejecting the aid's recommendation to engage.

Illustrative Scenarios

Benchmark models can be used to set up a series of experimental scenarios in which different decisions are appropriate, as a function of variations in stakes, time stress, and uncertainty about ATR accuracy. Figure 31 through Figure 34 show a set of scenarios generated by systematically manipulating two of the three key variables — time stress and stakes. For this example, we have kept trust constant, at .4 chance that the aid's classification conclusion is correct. For the purpose of the example, we assume that the aid has classified a contact as an appropriate target for engagement (e.g., an enemy tank). Stakes are varied for the upper bound only, by manipulating the mix of friendlies and enemy non-targets, thus affecting the expected cost of a mistaken engagement. Time stress is varied by manipulating the rate of increase in the danger of being targeted, as the pilot spends more time unmasked.
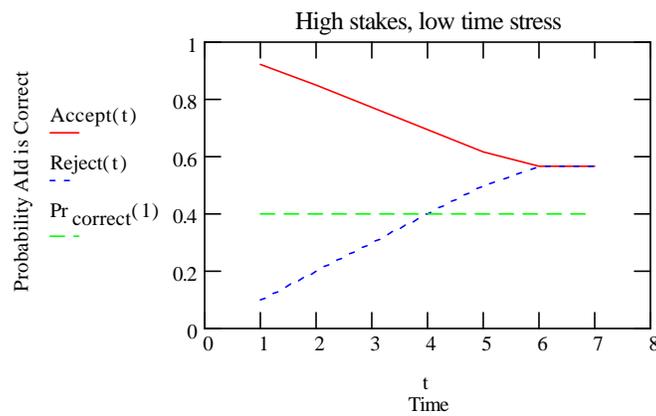
Figure 31. Scenario in which there is a large proportion of friendlies relative to enemy non-targets, producing high stakes of incorrectly accepting the aid's recommendation to engage. The probability of being targeted by enemy platforms is low, but increases with time. Trust in the ATR conclusion is highly uncertain, at .4. The result is a significant amount of time (from time 1 to time 4) spent verifying the aid's recommendation to engage. Finally, the cost of remaining unmasked leads to a decision (in this case, not to engage).

In these scenarios, the user must decide not only what to do — i.e., whether to engage or not to engage a contact — but how long to wait before doing it. In two of the scenarios (Figure 32 and Figure 34), the appropriate action is to accept the aid's recommendation and engage, while in the other two (Figure 31 and Figure 33), the appropriate action is to reject the aid's recommendation and not to engage. The appropriate time spent verifying the aid's recommendation varies from 3 units (in Figure 31) to 1 unit (in Figure 32 and Figure 33) to 0 units (in Figure 34). Scenario variations of this kind might be useful in a training context, as well as the present experimental context. Pilots might be evaluated and given feedback on both of the two dimensions, correct engagement decisions and appropriate strategies for verifying ATR recommendations. Exercises of this kind can help maintain skills in the primary task, while enhancing the ability to interact effectively with the ATR.

Low stakes, low time stress
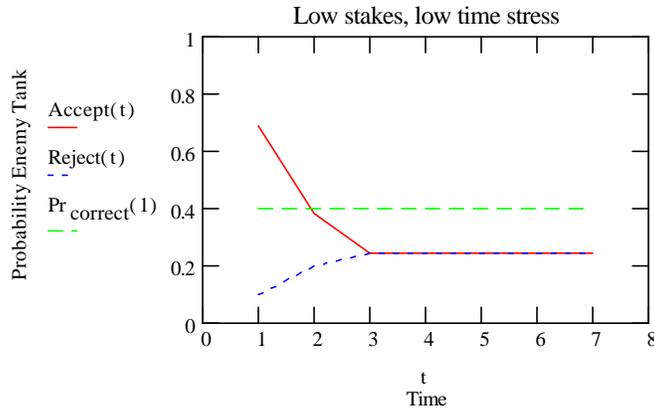
Figure 32. Scenario in which the low proportion of friendlies relative to enemy non-targets leads to a low threshold for engagement. Even though time stress is low (as in the previous example), less time is spent verifying the aid's recommendation (from time 1 to time 2) because of the low cost of an error. A relatively quick decision is made to engage.

High stakes, high time stress
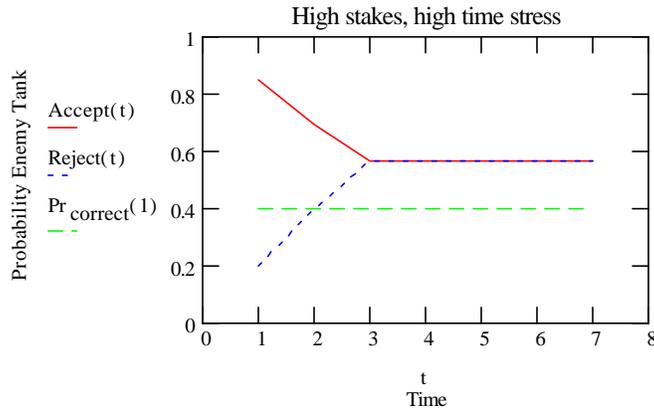
Figure 33. Scenario in which the cost of a mistaken engagement is high, due to a high proportion of friendlies. However, time stress is also high, due to a rapid increase in the chance of being targeted with time spent unmasked. This results in a relatively early decision, in this case not to engage.

In the above examples, the upper and lower bounds were independent of trust in the aid's conclusion, and trust remained constant. As Figure 35 illustrates, however, neither of these conditions is necessary. In this example, trust begins, as before, at .4. However, in verifying the aid's recommendation to engage, the user finds evidence that supports the aid's identification of the contact as hostile. Thus, confidence in the recommendation to engage increases to above .8. As the user becomes increasingly convinced that the contact is hostile, there is also a rise in the perceived chance of being targeted, and as a result, time stress increases along with trust. The result is a somewhat earlier decision to engage the target, as compared with Figure 31.

Low stakes, high time stress

Probability Aid is Correct

Accept(t)

Reject(t)

$Pr_{correct}(1)$

t
Time
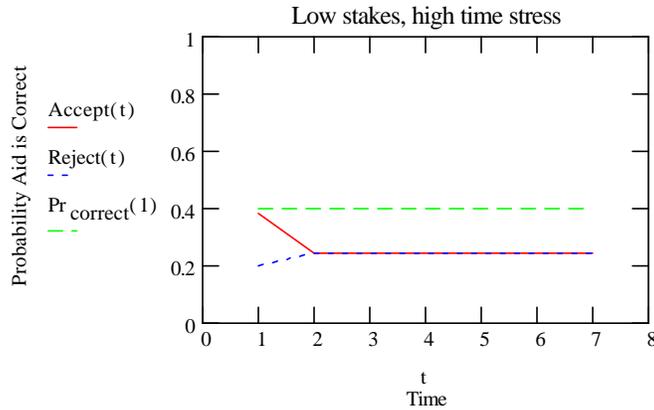
Figure 34. Scenario in which the cost of a mistaken engagement is low (due to low proportion of friendlies) and time stress is high (due to rapidly increasing chance of being targeted). The result is no time spent verifying aid's recommendation, and an immediate decision to accept the recommendation to engage.

High stakes, increasing time stress

Probability Aid is Correct

Accept(t)

Reject(t)

$Pr_{correct}(t)$

t
Time
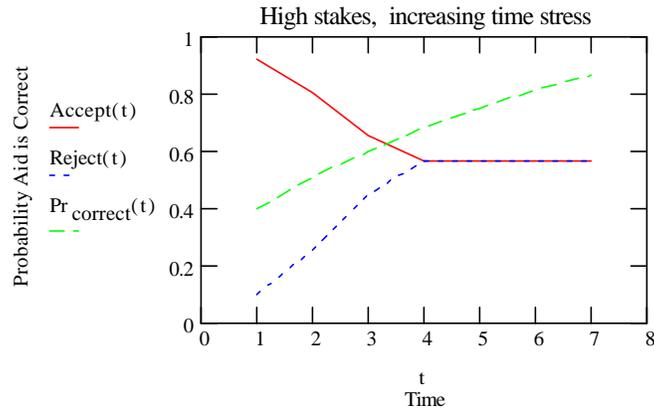
Figure 35. Scenario in which trust in the aid's identification of the contact as hostile increases, bringing with it an increase in time stress due to the expectation of being targeted. The result is a somewhat earlier decision to engage than in Figure 31, which is otherwise based on the same underlying parameters.

ATR Support for User Verification

In this study, we will investigate two ways that an ATR might foster effective user verification of its conclusions. One design concept involves alerting users when they are likely to be in the region where verification is appropriate. Such alerts might be based on ATR uncertainty, the costs of an error, and available time. The other design concept involves enhancing user performance when it is appropriate to verify, by adapting labels to ATR uncertainty. Instead of exposing users to detailed ATR conclusions when those conclusions are highly uncertain, the ATR might use more general labels (e.g., armored vehicle rather than tank or APC), to avoid misleading users when their independent judgment is particularly critical.

Method

Task

In this experiment participants attempted to recognize and engage enemy tanks, under varying degrees of time stress and with varying numbers of friendly vehicles in the area. Figure 36 outlines the basic task in this experiment, for all conditions. The vehicles were presented in 3 by 4 grids of 12 masked images. Pilots pressed a function key corresponding to a cell in order to unmask it and view the vehicle it contained. They then decided whether to engage that vehicle immediately or to wait. In either case, the 3 by 4 grid was redisplayed, and pilots could then view another image. Engaged vehicles were marked by an "X" and images that pilots had observed without engaging were marked with an "O" on the grid. Pilots could return to any unengaged image at any time, re-examine it and decide again whether to engage it. Each block of trials consisted of 10 grids, i.e., 120 images. Blocks were presented to participants as separate missions with varying levels of stakes and time stress.

System displays 3 by 4 grid of cells (+ labels &/or color coding).	S depresses function key to select a cell.	System displays simulated FLIR image in selected cell.	S indicates by lifting key whether to engage target or not engage	System redisplays original grid, but flags cell to indicate engaged or observed.	Ss examine or re-examine any cells they wish	Display indicates time up.

Time

Figure 36. Sequence of events in experimental task. Shaded boxes represent actions by pilots.

Design

Seven independent variables were manipulated in the study, as summarized in Table 15. Two within-subjects variables, stakes and time stress, were associated with blocks, i.e., multiple grids containing multiple images. Three within-subjects variables were manipulated within blocks – ATR accuracy, vehicle range, and vehicle viewing angle. The two prescriptive variables – search guidance and labeling rule – were manipulated between subjects. We will discuss each of these variables in turn.

Stakes (within-subjects, between blocks). All participants received the same basic mission: to find and engage enemy tanks. However, in some blocks participants were told they were flying a close air support mission, in which friendly vehicles were mixed with enemy vehicles, while in other blocks, participants were told they were flying a deep interdiction mission, in which friendly ground vehicles were relatively rare. In the high stakes condition (close air support), engagement of a vehicle other than an enemy tank was likely to be a fratricide. In the low stakes condition, engagement of a vehicle other than an enemy tank was likely simply to destroy some other enemy vehicle (e.g., truck, jeep, or APC). Thus, in high stakes conditions, a mistaken engagement was likely to be more costly than a mistaken engagement in low stakes conditions. Table 16 shows how the proportion of friendlies changed as a function of the stakes variable.

Table 15. Variables manipulated in the third experiment, with their levels and type of variation.

	Stakes	Time stress	ATR accuracy	Vehicle range	Vehicle viewing angle	ATR search guidance	ATR labeling rule
Levels	High, Low	High, low	High, low	Close, Medium, Far	Side, Front	Color, No color	1:Preferred 2: Adaptive 3. Control 4. None
Variation	Within S Between blocks	Within S Between blocks	Within S Within blocks	Within S Within grids	Within S Within grids	Between S	Between S

Table 16. Proportion of different types of vehicles as a function of the stakes variable.

	High Stakes Close Air Support	Low Stakes Deep Interdiction
Enemy tanks	.40	.50
Enemy non-tanks	.26	.40
Friendly vehicles	.34	.10

Participants received more high stakes trials (56 grids, 672 images) than low stakes trials (10 grids, 120 images).

Time stress (within-subjects, between blocks). In the high time stress condition, pilots had 30 seconds to search each grid of 12 vehicles and make engagement decisions. In the low time stress condition, pilots had 60 seconds per grid. These times were determined with the aid of a flight instructor and standards officer at Ft. Bragg, and were designed to be realistic and (for the high time stress condition) challenging. Time stress was implemented in two ways: In their mission instructions preceding a block, pilots were told how much time they would have for each grid. In addition, a countdown clock displayed the time remaining to search each grid, and the grid was removed when the designated time expired.

Of the 17 participants in this experiment, 15 received all combinations of time stress and stakes. One participant did not receive any high stress missions, and one did not receive the high stress, low stakes missions.

Vehicle viewing range (within-subjects, within blocks). Viewing range was simulated by varying the size of the image, and was associated with an image's location in the matrix. The bottom row of vehicles was viewed at a simulated distance of 2500-3000 meters, the middle row at 3000-3500 meters, and the top row at 3500-4000 meters. All displayed images were at roughly the same actual viewing distance from the pilots.

Vehicle viewing angle (within-subjects, within blocks). Individual images varied in the angle of the vehicle to the observer. Viewing angles could be either side r front. The side angle

exposed some features to the observers that were not as clear from the front angle, e.g., number of wheels, size of gun, overall profile, and certain FLIR hot spots.

ATR accuracy (within-subjects, within blocks). For a given grid, a basic ATR accuracy parameter could be set either high or low. This parameter corresponded to the user by the ATR of sensor or intelligence data that was not available to the pilots. It thus introduced a lack of correlation (hence, a degree of complementarity) between ATR accuracy and human accuracy.

Regardless of how the ATR chose to report its conclusions, it went through the same internally simulated "recognition" process. For each image, this process generated a probability distribution across all the specific vehicle labels (BTR, M60, UAZ, etc.). The ATR then "recognized" the image by randomly selecting a specific vehicle label based on this probability distribution. The correct specific label always had the largest chance of being selected, and the likelihood of other specific labels was calculated based on the similarity of the corresponding vehicles to the vehicle in the image. The ATR then chose a label to *report* its conclusion (which might be at the specific, intermediate, or general level) based on both the internally generated recognition process and the labeling rule (which we discuss below).

Probabilities for different specific level conclusions were determined by a formula based on both similarity and guessing (something like the Informed Guessing Model described in the previous chapter). Thus, for each of several dimensions of similarity among the vehicles, there was a parameter reflecting the chance that the ATR would extract that dimension and no others, thus being unable to discriminate the vehicles that had the same value on that dimension. The ATR would then guess at the correct vehicle type from the candidates that shared the relevant property. For example, suppose the actual image were a T62. Let p_0 be the probability that the ATR will extract only enough information to determine that the image represents a vehicle, let p_1 be the probability of extracting only the dimension armored/not-armored, and let p_2 be the probability of extracting only the dimension wheeled/tracked. (The p_i are normalized so that they sum to 1.0.) In the experimental image set, there are a total of 14 vehicles, 9 of which are armored (including the T62) and 8 of which are tracked (also including T62). If these were the only dimensions of similarity, the chance of misidentifying the T62 as an M60 (a tracked armored vehicle) would be $p_0 (1/14) + p_1 (1/9) + p_2 (1/8)$; the chance of misidentifying the T62 as a BTR (a wheeled armored vehicle) would be $p_0 (1/14) + p_1 (1/9)$, and the chance of misidentifying it as a KRAZ (a wheeled non-armored vehicle would be $p_0 (1/14)$. The probability of correctly identifying the T62 as a T62 would be: $p_0 (1/14) + p_1 (1/9) + p_2 (1/8) + p_3$, where p_3 is the chance of extracting enough information to uniquely identify a specific vehicle type.

Dimensions actually used in the similarity algorithm were: locomotion (hot tracks, slack tracks, or wheels), the vehicle type (APC, tank, jeep or truck) and the level of armor (armored or not armored). The effect of these particular dimensions was a high proportion of ATR confusions among tanks and among APC's, and a moderate degree of confusion between tanks and APC's, but relatively fewer confusions between tanks and jeeps or trucks. These dimensions were intended to represent a plausible ATR process, which overlapped with, but was not identical with, human similarity judgments.

The largest term in the formula determining ATR recognition probabilities was the term reflecting the similarity of a vehicle to itself (p_3 in the example above). This term was based on both the ATR accuracy parameter and on features of the specific image that might plausibly

affect the ATR's classification success. A base self-similarity of 100% was discounted to varying degrees based on three factors.

$$Self\text{-}similarity = Orientation * Range * ATR$$

Orientation reflects the angle of view of the vehicle in the image. There was no discounting for side views of the vehicle (i.e., *orientation* = 1), while *orientation* = 0.7 for front views. Similarly, there was no discounting for close (2.5 K) vehicles, while *range* = 0.93 at 3K, and *range* = 0.85 at 3.5K. *ATR* reflects sensor or intelligence information sometimes available to the ATR but not to the user. *ATR* = 1.0 when this information is available to the ATR, and *ATR* 0.7 when it is not. These variations in *ATR* constituted the high and low ATR accuracy conditions, respectively, and were randomly determined for each grid.

Based on the simulated recognition process, ATR recognition accuracy, at the specific vehicle level, could vary from 93% to about 35% on any given image. Although there was considerable variation in ATR accuracy from image to image within an ATR accuracy condition (due to variations in viewing angle and distance), the ATR accuracy parameter significantly affected the average accuracy in a given grid. In particular, if all combinations of the other two factors were experienced equally in the grid, the average accuracy would be 74% in high ATR accuracy grids, and 51% in low ATR accuracy grids. Though it was intended that all combinations of the three factors be applied with equal frequency, a disproportionate number of images were inadvertently presented in the low ATR accuracy condition and in the front view. These images would be difficult for both the simulated ATR and for the human observer.

Once recognition was completed at the level of a specific vehicle label, the ATR chose a label to report its conclusion according to the prevailing labeling rule (discussed below). For example, if the labeling rule indicated that intermediate terms should be used for jeeps and trucks, and the ATR recognized (correctly or incorrectly) a vehicle as a UAZ, the displayed label would be *jeep*, which is the corresponding intermediate term.

When it reported a label, the ATR also reported its confidence in that label (e.g., *Jeep 82%*). The reported confidence reflected the ATR's simulated understanding of conditions that affected its accuracy (i.e., orientation and range of the vehicle in the image, and the ATR's access to special sensor or intelligence data). Thus, reported confidence was not influenced by whether recognition was correct or not on a particular occasion, since the ATR would have no way of knowing this. Reported confidence was always the same as would have been reported had the conclusion arrived at by the ATR been the correct one. This prevented observers from using the pairing of confidence and label as a cue to the correctness of a particular ATR classification. For labels above the specific level, confidence was obtained by summing confidence in the ATR's specific-level conclusion and the chances of confusing the vehicle referred to by the ATR's conclusion with all others contained in the relevant higher-level category.

The order of the within-subjects conditions was counterbalanced between participants. ATR accuracy conditions, viewing angle, and range were crossed with all other within-subjects conditions (i.e., stress and stakes) for all participants

Search guidance (between-subjects). As noted above, images of vehicles were not visible to pilots until selected for observation. Ten of the 17 pilots received guidance in allocating their attention among the 12 images on the grid. Search guidance took the form of color coding the masks concealing each image. The following scheme was used:

- Red: The ATR is at least 90% certain that the vehicle in the cell is an enemy tank.

- Blue: The ATR is at least 90% certain that the vehicle in the cell is friendly armor.

- Yellow: There is conflicting and inconclusive evidence. There is at least a 25% chance that the vehicle in the cell is an enemy tank, and at least a 25% chance that it is friendly armor. Of necessity, neither probability can be as high as 80%.

- Grey: Other conditions of ATR certainty and vehicle type, such as unarmored vehicles or low confidence assessments by the ATR.

The yellow cells were designed to alert users to images for which (1) there was uncertainty, and (2) the cost of an error was high because the uncertainty involved potential confusion between a target (enemy tank) and a friendly. Yellow thus signaled images for which there might be a particularly high payoff for the user's attention. This application of color codes was driven by a model of the verification decision (described earlier in this chapter and in Appendix D), used to predict when extended viewing by the pilot was most likely to be worthwhile.

Seven pilots received no search guidance. For these pilots, all 12 cell masks in each grid were colored gray.

Labeling rule (between-subjects). The mask concealing an image might also contain a label representing the ATR's classification of the vehicle in that cell. The same label continued to be displayed after the observer selected a cell for observation. Four different schemes for labeling vehicles were compared:

Rule 1. Preferred labels: In this condition, the system displayed the favored labels identified in the first set of experiments, with one modification. Recall that in the first set of experiments, we employed feature-naming, familiarity ratings, typicality ratings, recognition latencies, and spontaneous naming to determine what level of specificity seemed optimal for different vehicle types. The label favored for trucks and jeeps was intermediate, namely *truck* or *jeep*. The favored names for APC's were highly specific (e.g., *BTR*). However, pilots were ambiguous with respect to the preferred label for tanks. Intermediate labels (e.g., *tank*) and specific ones (e.g., *T62*) were favored in different experimental tasks. We hypothesized that mission relevance was responsible for the results that favored specific labels for tank. In this study, we adopted a set of labels that remained at the intermediate level, but which captured additional mission relevant information: The labels *enemy tank* and *friendly tank* were used in lieu of *tank* or *M60* and *T62*. In Figure 37, the names used for Label Rule 1 are shaded gray.

Rule 2. Adaptive labels: This rule used the favored labels as indicated in Rule 1 except where there was both significant uncertainty and a real possibility of fratricide (i.e., high stakes) The goal was to prevent unreliable ATR labels from unduly influencing pilots in situations where the pilot's own recognition abilities might be crucial. Labels differed from Rule 1 when the ATR identified a vehicle as armor but could not distinguish whether it was an enemy tank or friendly armor at an 80% level of confidence. In these cases, the ATR ascended the classification hierarchy above the preferred label until it found a more general label that it could assert with at least 80% confidence (and would thus be less likely to lead the pilot astray). For example, if the preferred rule 1 label was "enemy tank," but it was associated with a confidence of only 52%, the ATR might substitute the label "armor" with a confidence of 83%. For participants who received search guidance, this adaptation occurred in the same cells that were coded yellow.

Rule 3: Contrast labels. This scheme was designed to be a reasonable approach to naming vehicles, while at the same time offering a contrast with the scheme of naming at the favored level. In some cases, this rule applied labels that were more general than the favored level. For example, the more general term *light vehicle* was used for jeeps and trucks instead of the favored intermediate labels *jeep* or *truck*. The more general label "enemy APC" was used instead of the favored specific labels such as *BMP*. In other cases, this rule used labels that were more specific than the favored level. For example, labels like *T62* were used for tanks in place of the intermediate label *enemy tank* and "friendly tank." Thus, rule 3 supported several interesting contrasts with the putatively favored naming scheme in rule 1. It is important to stress that rule 3 was by no means a straw man; it offers more specific labels for tanks, which are central to the pilot's missions both experimentally and in real-world settings, and it requires less specific labels for jeeps, trucks, and APC's, which are not typically central to the pilot's missions. Contrast labels are shaded black in Figure 37.

Rule 4: No labels. In this condition, no ATR conclusions regarding the classification or identity of the vehicle was provided. This condition, when combined with the condition in which there is no color-coded search guidance, provide a baseline representing the absence of an ATR.

For rules 1, 2 and 3, the ATR displayed its numerical confidence in the label when an image was unmasked. ATR confidence ratings were perfectly calibrated with respect to ATR accuracy. For example, ATR labels were correct 80% of the time across all images for which the ATR's displayed confidence was 80%. As already noted, the ratings varied from about 93% down to about 42% as a function of range, viewing angle, and the ATR accuracy parameter.

The two between-subjects variables were incompletely crossed such that no participant had labeling rule 3 (the contrast rule) in the no-color condition. The distribution of pilots in the between-subjects conditions is shown in Table 17.

Table 17: The distribution of participants in the between-subjects conditions.

Label Rule	Description	Search: Color	Search: No color	Total *n*
1	Favored labels	3	2	5
	No adaptation to uncertainty			
2	Favored labels, except that:	2	3	5
	As uncertainty increases, ATR labels become more general			
3	Not favored labels	3	0	3
	No adaptation to uncertainty			
4	No ATR	2	2	4
	Total participants	10	7	17
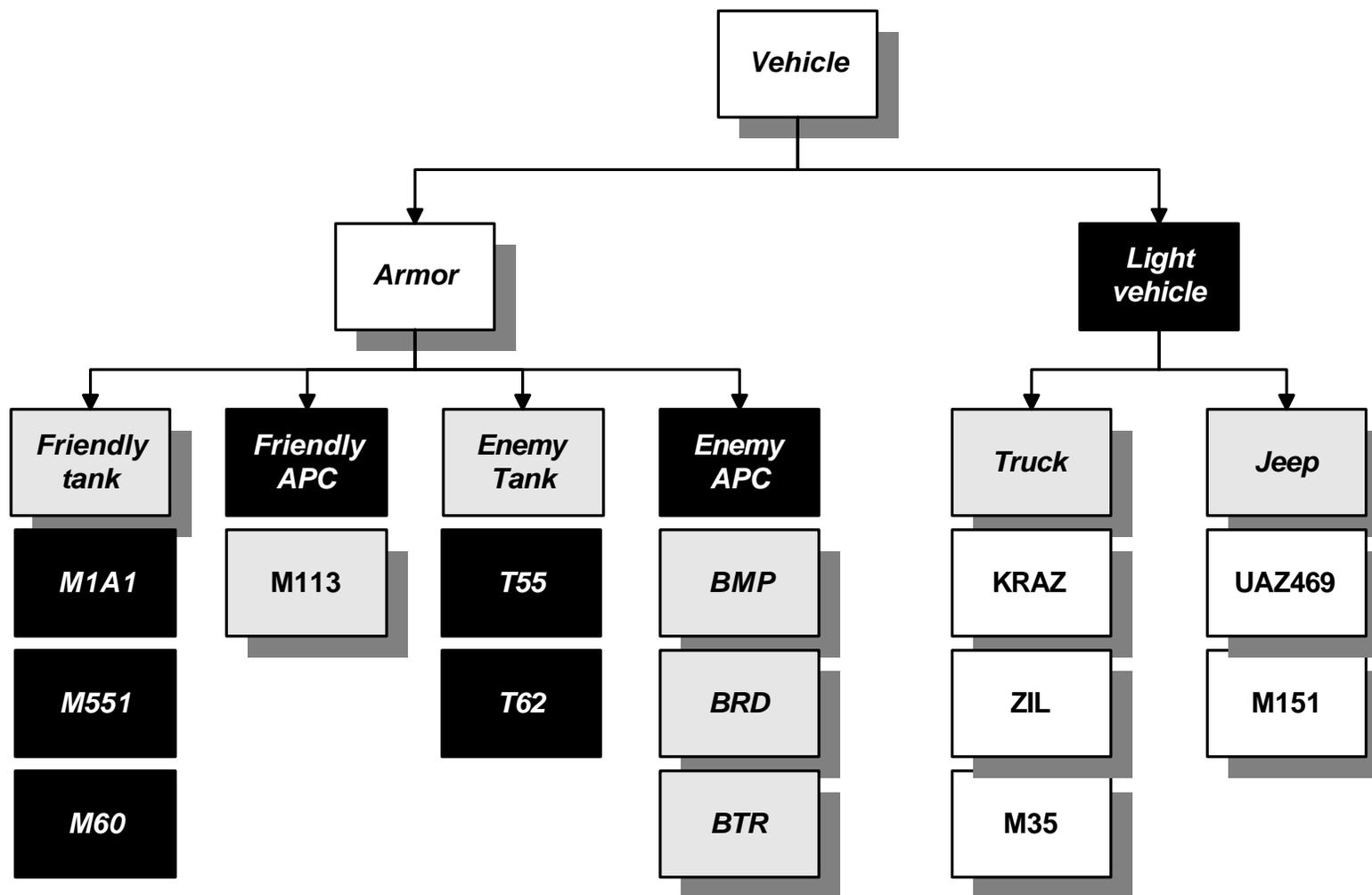
Figure 37. Eighteen labels and 14 images used in the third experiment. Labels used in favored labeling condition (Rule 1) are shaded gray. Labels used in the control condition (Rule 3) are shaded black. The 14 images correspond to the specific model names.

Materials

Front and side images of 14 vehicles were used as stimuli along with 18 descriptive labels. The vehicles included tanks, APC's, jeeps and trucks. A programming error toughened the task somewhat for all participants by presenting front views of vehicles in 85% of all cases and side views 15% of the time, rather than presenting vehicles in each view an equal proportion of times. Figure 37 indicates the images used and the corresponding, accurate labels.

The images used in this study were produced by John Horger in collaboration with Robert Lafollette and were provided with the cooperation of Dr. Barbara O'Kane, all staff of NVEOL. FLIR noise was generated using NVSIM, a UNIX-based Thermal Imaging System Simulator developed by Mr. Horger.

Apparatus

As in the first experimental series, stimuli were presented on an Intel Pentium personal computer running the NextStep operating system and software developed by CTI. The previously described software was enhanced to support masked and unmasked presentation of multiple images, as well as presentation of color-coding and ATR labels and confidence levels.

Procedure

After a brief introduction by the experimenter, each participant completed a biographical information form. The pilot then read printed instructions concerning the experimental task. These instructions included an overview of the four possible missions (crossing stakes and time stress), and introduced the experimental system interface (e.g., the grid, how to unmask images, how to engage vehicles). These general instructions were followed by additional instructions that differed somewhat for the different between-subjects conditions, e.g., explanations of color coding and labeling rules corresponding to the different combinations of those conditions.

Immediately prior to each block of trials, the participant received specific mission instructions. For all missions, these instructions indicated that the participant was to engage enemy tanks and avoid other vehicles, especially friendlies, that there were sufficient munitions to engage all enemy tanks plus some extra rounds and that a constant amount of time would be provided for each new grid of 12 vehicles. Different enemy orders of battle were provided for the deep interdiction and close air support missions. Appendix E contains a complete set of written instructions provided to participants in all conditions of the this experiment.

Participants began their first mission by practicing on five grids of images (i.e., 60 images), and began each subsequent mission by practicing on two grids of images. Each pilot participated in the experiment for approximately two hours.

Dependent Variables

Two kinds of dependent variables are of interest. One class pertains to how well the mission is performed. These measures look at the overall outcome of ATR-pilot collaboration, not at how the outcome was achieved. The bottom line is: Are targets engaged and are non-targets not engaged? Two kinds of errors are possible. In signal detection terminology, engaging a target might be referred to as a "hit," analogous to detection of a signal in noise, and failing to engage a target would be a "miss." Engaging a non-target would be referred to as a "false

alarm," while not engaging a non-target would be a "correct rejection." These variables appear in the last column of Table 18. Another dimension of overall ATR-pilot performance is the latency of target engagements, that is, how long it takes from the first appearance of a grid containing a particular target to the decision to engage that target.

The second class of dependent variables focuses on *how* outcomes were achieved, and in particular, on the pilot's role *vis a vis* the ATR. What actions did the pilot take to influence the overall hit and false alarm rates? Table 18 outlines the possible pilot interventions, and shows how each of them contributes to the measures of overall performance discussed in the previous paragraph. The table can be viewed as an event tree, with a series of branching possibilities. First, either a particular image represents a target (enemy tank) or it does not. If it does represent a target, the ATR can either recognize it as such (a "hit") or fail to recognize it as a target (a "miss"). If the image does not represent a target, the ATR can either recognize it as a non-target (a "correct rejection") or mistakenly recognize it as a target (a "false alarm").

Just as the ATR can be thought of as a target detector, the pilot can be thought of as a detector of ATR errors or shortcomings. This conceptualization allows us to isolate the specific contribution of the pilot, over and above cases where the ATR correctly classifies an image as a target or non-target and the pilot accepts that classification. In other words, what does the pilot contribute besides engaging targets designated by the ATR and not engaging ATR designated non-targets? Measures based on this concept are shown in the *pilot interventions* column of Table 18.

The pilot's role, when the ATR believes an image to be a target (the top and bottom rows of Table 18), is to look for possible ATR false alarms, i.e., images that the pilot believes are in fact non-targets. In this verification process, the pilot may be successful or unsuccessful. If the image really is a target, the pilot should accept the ATR classification as a target and engage it. In signal detection terminology, this response is a "correct rejection" of the possibility of an ATR false alarm. If the image is not really a target, the pilot should reject the ATR classification and not engage. In signal detection terminology, this response is a "hit" with regard to an ATR false alarm. Similarly, if the image really represents a non-target, the pilot's role is to look for possible ATR misses. This, too, may be successful or unsuccessful. If the image is in fact a target, the pilot should override the ATR classification and engage (a pilot "hit" regarding an ATR miss). If the pilot instead accepts the ATR classification and does not engage, we have a pilot "miss" regarding an ATR miss. On the other hand, if the image is really a non-target, the pilot should accept the ATR classification and not engage (a "correct rejection" of a possible ATR miss). If instead, the pilot overrides the ATR classification and engages, we have a pilot "false alarm" regarding a possible ATR miss. Since the probability of a miss is one minus the probability of a hit, and the probability of a correct rejection is one minus the probability of a false alarm, we will utilize only four of the eight pilot intervention measures referred to, as indicated in Table 18.

Use of signal detection theory concepts not only helps us isolate the pilot's contribution to overall performance, it also enables us to analyze how independent variables such as stakes, stress, and ATR accuracy influence the pilot. For example, if a manipulation increases the probability of hits (i.e., pilot detections of ATR errors) but also increases the corresponding rate of false alarms, the affect can be attributed to a change in response bias; for example, the pilot may simply be more or less inclined to engage. On the other hand, if a manipulation increases hits without also increasing false alarms, it can be attributed to a change in sensitivity; for example, the pilot may be examining images more closely.

Table 18. Dependent variables pertaining to engagement accuracy. Italicized measures in the last two columns are used in the study.

Actual vehicle	ATR Action	Pilot intervention	Overall ATR-pilot outcome
Target (enemy tank)	Correctly recognize as target ("hit")	Correct acceptance of ATR engagement ("correct rejection" re possible ATR false alarm)	*Correct engagement of target ("hit")*
		Incorrect override of ATR engagement ("false alarm" re possible ATR false alarm)	Incorrect non-engagement ('miss")
	Incorrectly recognize as non-target ("miss")	Incorrect acceptance of ATR non-engagement ("miss" re ATR miss)	Incorrect non-engagement ('miss"))
		Correct override of ATR non-engagement ("hit" re ATR miss)	*Correct engagement of target ("hit"*
Non-target	Correctly recognize as non-target ("correct rejection")	Correct acceptance of ATR non-engagement ("correct rejection" re possible ATR miss)	Correct non-engagement of non-target ("correct rejection")
		Incorrect override of ATR non-engagement ("false alarm" re possible ATR miss)	*Incorrect engagement of non-target ("false alarm")*
	Incorrectly recognize as target ("false alarm")	Incorrect acceptance of ATR engagement ("miss" re ATR false alarm)	*Incorrect engagement of non-target ("false alarm")*
		Correct override of ATR engagement ("hit" re ATR false alarm)	Correct non-engagement of non-target ("correct rejection")

Analysis

Overall performance. A multivariate analysis of variance was performed utilizing two within-subjects variables (rule and search guidance) and three within-subjects / repeated

measures (stakes, time stress, and ATR accuracy). Three measures of overall ATR-pilot performance were used: probability of correctly engaging a target (hits), probability of incorrectly engaging a non-target (false alarms), and latency to correct engagement of targets. When multivariate effects were significant, univariate analyses of variance were performed on all three measures separately. In addition, the probability of incorrectly engaging a friendly vehicle was analyzed separately. The latter could not be included in the multivariate ANOVA because insufficient data were available in low stakes conditions, where friendlies were relatively rare.

Pilot interventions. A similar multivariate analysis of variance was performed, utilizing two within-subjects variables (rule and search guidance) and three within-subjects / repeated measures (stakes, time stress, and ATR accuracy). Two measures of pilot intervention were used: the probability of correctly overriding an ATR non-engagement recommendation (hits with respect to ATR misses), and the probability of incorrectly overriding an ATR non-engagement recommendation (false alarms with respect to ATR misses). When multivariate effects were significant, univariate analyses of variance were performed on both measures separately. Other pilot interventions (correctly overriding an ATR recommendation to engage, and incorrectly overriding an ATR recommendation to engage) could not be analyzed formally due to empty cells (e.g., in labeling conditions where the ATR was not specific enough to recommend engagement).

Results

Handling Uncertainty

We will look first at how pilots handled uncertainty under varying conditions of time stress, stakes, and ATR accuracy, without respect to the two ATR interventions (labeling rule and search guidance). We will then explore how the two interventions influenced that performance. In both cases, we will be interested both in pilot interaction strategies, and the overall outcomes for ATR-pilot performance.

Stakes. Figure 38 shows that high stakes improved pilot intervention performance with respect to vehicles labeled as non-targets by the ATR. High stakes led to a slight increase in correct overrides by pilots, and a sharper decrease in incorrect overrides. In a MANOVA, which looked simultaneously at the chance of correctly and of incorrectly overriding ATR non-engagement recommendations, stakes had a significant effect (Hotellings $F_{2,9} = 7.15834$; $p =.014$; same significance level for Pillais and Wilks tests). When we look at the two measures separately, only the chance of incorrectly overriding non-engagement decisions approached significance ($F_{1,27} = 4.11$; $p =.070$).

We might expect that manipulations of stakes would change the pilot's response bias with respect to engagement. In particular, increasing the proportion of friendlies might reduce the pilot's tendency to engage. The reduction in incorrect overrides of non-engagements supports this interpretation. However, the increase in *correct* overrides of non-engagements contradicts that view. If the pilot's tendency to engage had decreased, there should have been fewer, not more, decisions by the pilot to engage when the ATR suggested otherwise. In terms of signal detection theory, Figure 39 shows that increasing stakes produced improvements in both false alarms and hits rather than a tradeoff. This pattern on the two intervention measures implies that

changes in stakes influenced the sensitivity parameter, rather than the bias parameter. Increasing stakes improved the quality of the pilot's decision making.
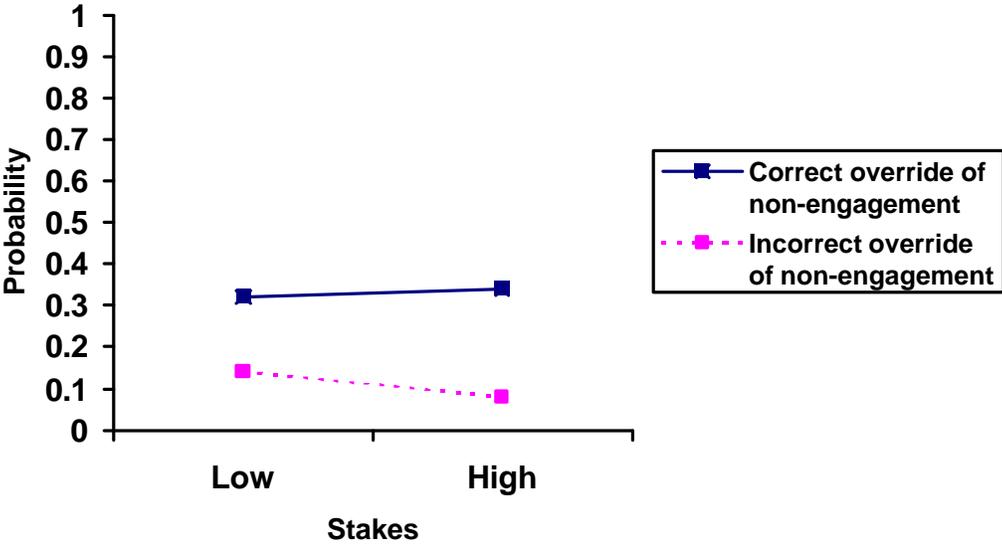
Figure 38. Effect of two levels of stakes (number of friendlies) on pilot override decisions.
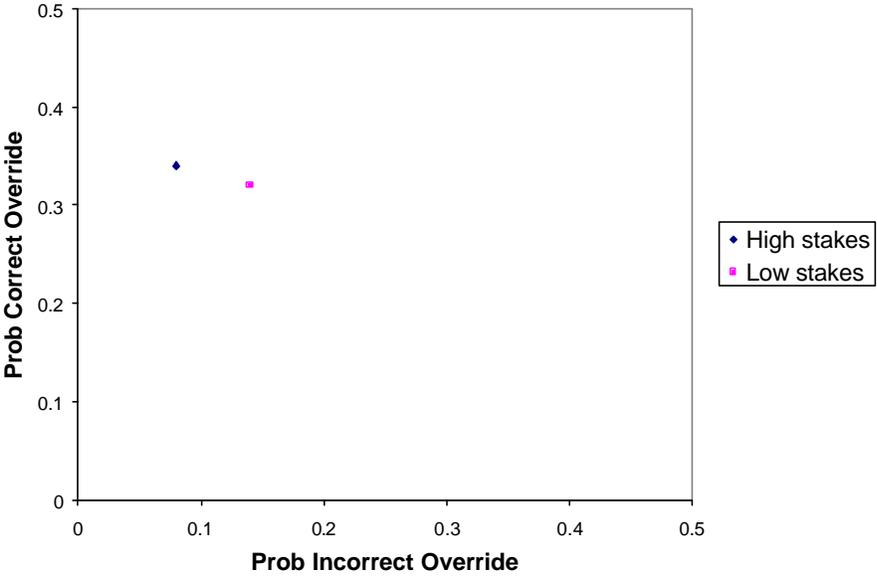
Figure 39. Probability of correct and incorrect override of ATR non-engagement recommendation at the two different levels of stakes.

The effect of stakes on pilot actions carried over weakly to overall outcomes, as Figure 40 shows. The effect of stakes approached significance in a univariate test on the probability of incorrect engagements (false alarms) ($F_{1,10} = 4.01$; $p = .073$), but was not significant in the MANOVA (Hotellings $F_{3,5} = 2.56463$; $p = .168$; same significance level for Pillais and Wilks tests).
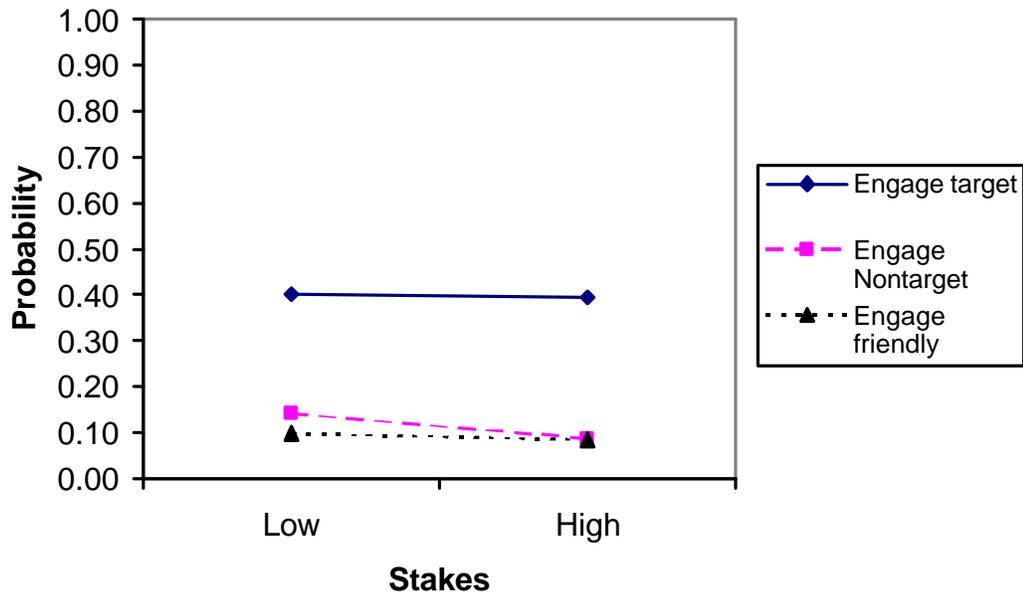
Figure 40. Effect of two levels of stakes (number of friendlies) on overall outcomes.

Time stress. Decreasing the amount of time available to view each grid had predictable effects on all measures. The MANOVA on pilot interventions was highly significant (Hotellings $F_{2,9} = 11.19440$; $p = .004$; same significance level for Pillais and Wilks tests). As time stress increased, there were significant decreases in both correct ($F_{1,10} = 11.95$; $p = .006$) and incorrect ($F_{1,7} = 18.21$; $p = .002$) overrides of ATR non-engagements. As Figure 42 suggests, increasing time stress reduced the tendency to engage, whether correctly or incorrectly. By contrast with the effects of stakes, there is no evidence that stress affects the quality of the decisions that are made.

Time stress also had a significant effect on overall ATR-pilot performance, as shown in Figure 43. The MANOVA was highly significant (Hotellings $F_{3,5} = 14.26195$; $p = .002$; same significance level for Pillais and Wilks tests). Univariate effects on accuracy measures were also highly significant. Time stress reduced the probability of correctly engaging targets ($F_{1,10} = 25.34$; $p = .001$), but at the same time also reduced the chance of incorrectly engaging non-targets ($F_{1,10} = 18.51$; $p = .002$). The effect on engaging friendlies was marginal ($F_{1,10} = 3.65$; $p = .085$).

Not surprisingly, time stress had a highly significant effect on the latency of engaging targets ($F_{1,7} = 45.12$; $p < .001$). This does not necessarily mean that pilots are faster engaging targets; it is probably an artifact of the shorter time window in high time stress conditions. More interestingly, there was a significant interaction in the effects of times stress and stakes on target engagement latencies ($F_{1,7} = 12.87$; $p = .009$; the MANOVA reflected only a trend: Hotellings $F_{3,5} = 3.39360$; $p = .111$; same significance level for Pillais and Wilks tests). Figure 44 shows that high stakes reduced target engagement latencies more under low time stress than under high time stress.
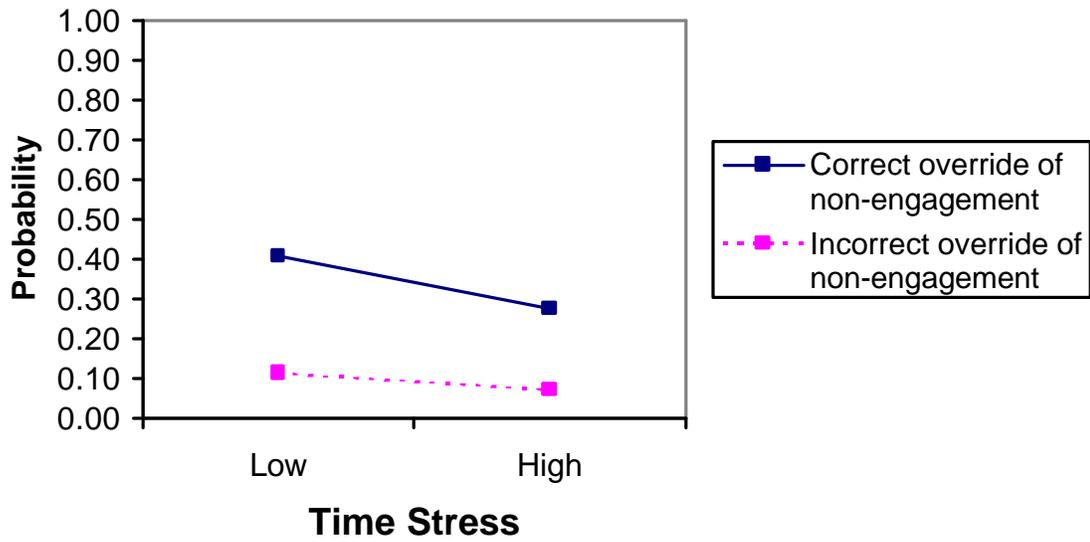
Figure 41. Effect of two levels of time stress on pilot override decisions.
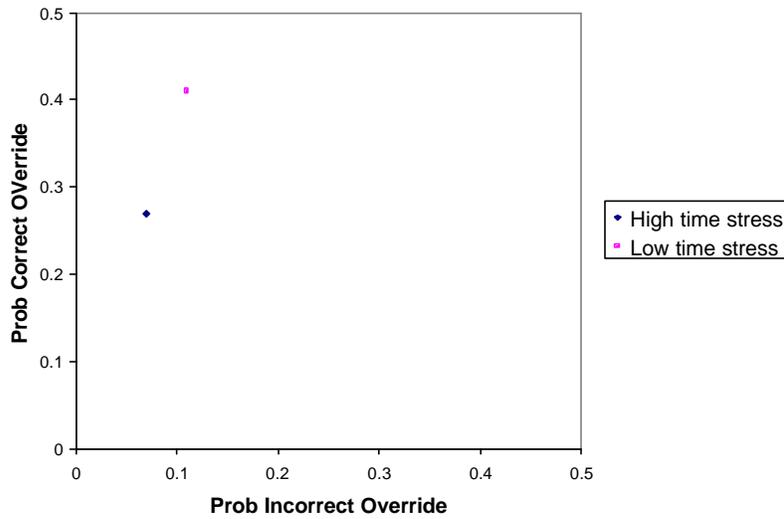
Figure 42. Probability of correct and incorrect override of ATR non-engagement recommendation at the two different levels of time stress.
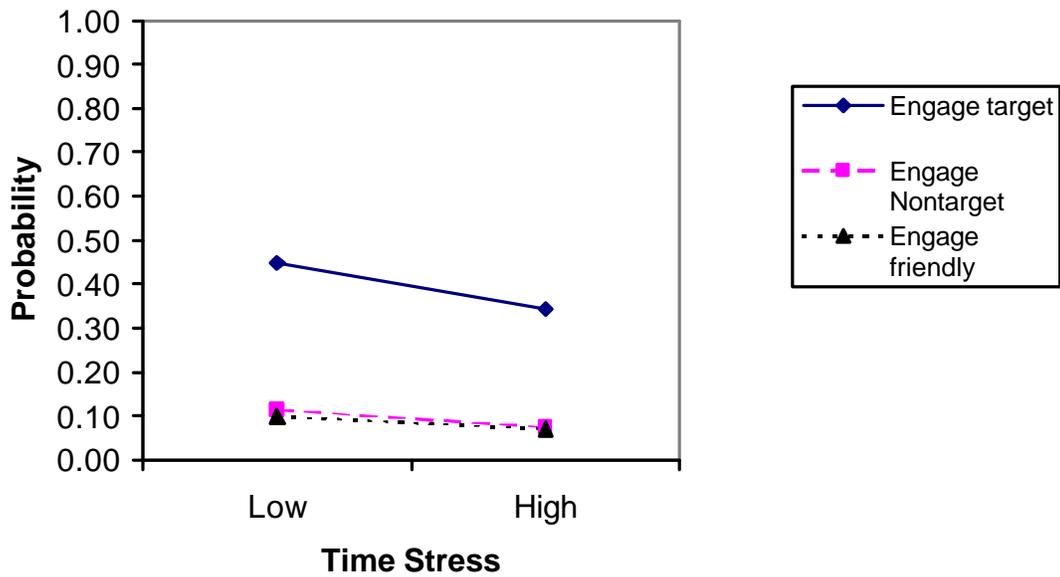
Figure 43. Effect of two levels of time stress on overall outcomes.
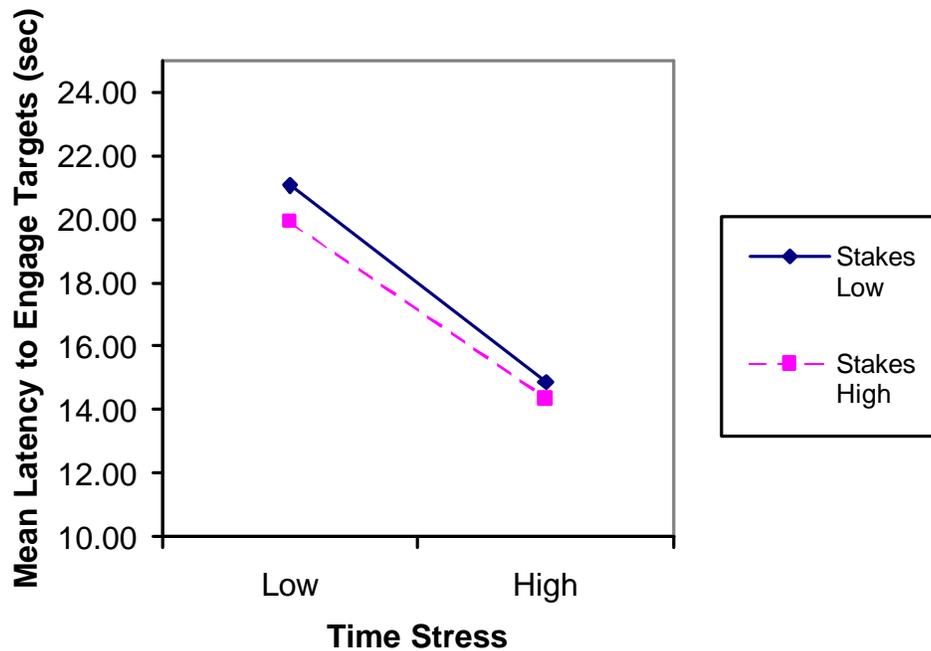
Figure 44. Interaction of time stress and stakes on latency to engage targets.

ATR accuracy. Increasing ATR accuracy increased the pilot's bias for engagement, whether correctly or not. Increasing ATR accuracy had small, but significant effects on override

decisions (Hotellings $F_{2,9}$ = 3.59151; p =.071; same significance level for Pillais and Wilks tests). As Figure 45 shows, higher ATR accuracy led to increased probability of correct override of ATR non-engagement decisions ($F_{1,10}$ = 7.82; p =.019), and a trend for more incorrect overrides of ATR non-engagement decisions ($F_{1,10}$ = 3.12; p =.108). Figure 46 shows that the effect of ATR accuracy is a tradeoff between increasing hits and accepting more false alarms. It thus lends itself to treatment as a bias, rather than a sensitivity effect.

Increased ATR accuracy tended to produce a net beneficial effect on overall performance (Hotellings $F_{3,5}$ = 3.24489; p =.119; same significance level for Pillais and Wilks tests). ATR accuracy significantly improved the chance of engaging targets (hits) ($F_{1,10}$ = 16.49; p =.002), and also improved the latency of doing so ($F_{1,7}$ = 4.72; p =.066).

Recall that one effect of high stakes was an increase in the quality of pilot decision making. This improved decision making may counter the effect of bias induced by high ATR accuracy. There was a highly significant interaction between stakes and ATR accuracy in their effects on pilot override of ATR non-engagement recommendations (Hotellings $F_{2,9}$ = 11.86330; p =.003; same significance level for Pillais and Wilks tests). As shown in Figure 48, under low stakes, increasing ATR accuracy has the effect we have already noticed: an increase in the chance of overriding non-engagement recommendations, both correctly and incorrectly. Under high stakes, however, as shown by the flatter curves in Figure 49, both effects of increasing accuracy disappear. First, accuracy brings about a significantly smaller increase in correct overrides under high stakes than low stakes ($F_{1,10}$ = 13.33; p =.004). This is because high stakes increases correct overrides in the low accuracy condition. Second, high stakes eliminates the rise in incorrect overrides due to high ATR accuracy ($F_{1,10}$ = 4.45; p =.061).
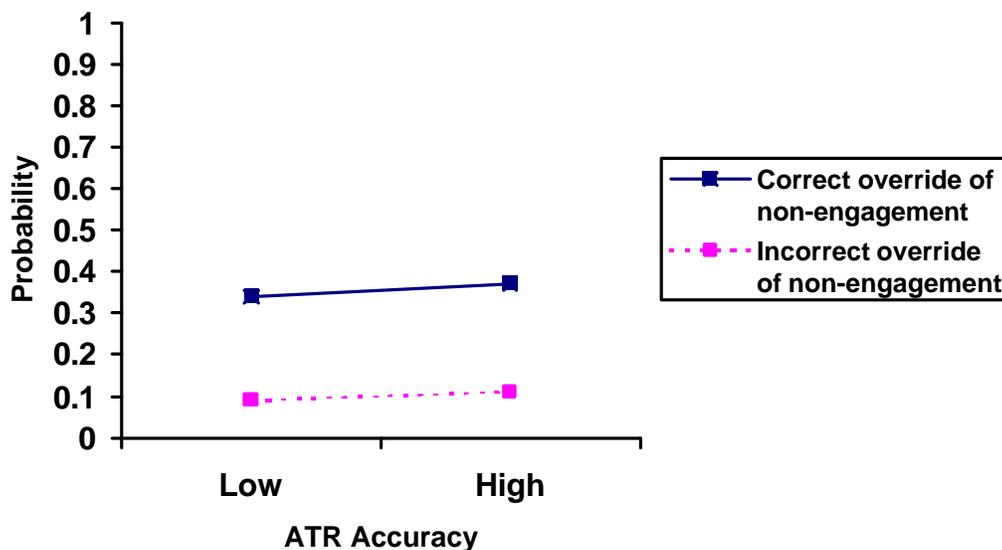
Figure 45. Effect of two levels of ATR accuracy on pilot override decisions.
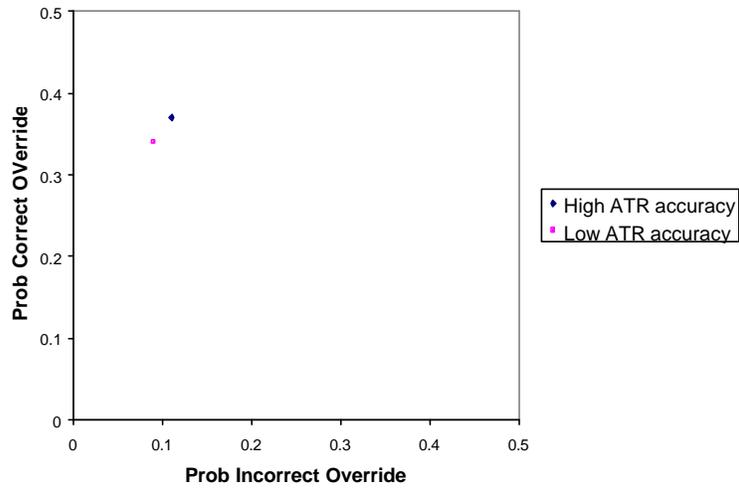
Figure 46. Probability of correct and incorrect override of ATR non-engagement recommendation at the two different levels of ATR accuracy.
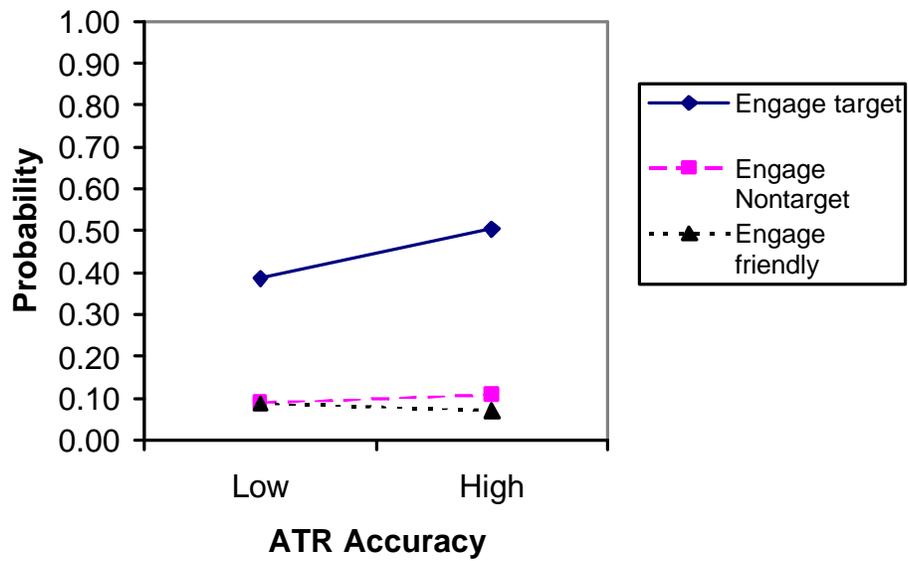
Figure 47. Effect of two levels of ATR accuracy on overall outcomes.
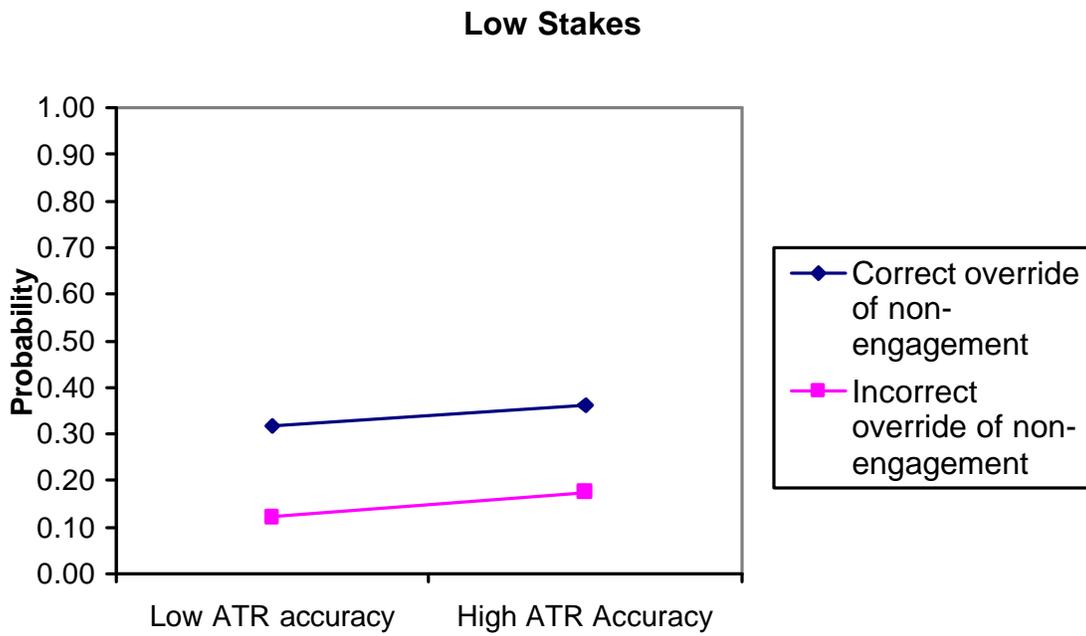
Low Stakes

Figure 48. Effect of ATR accuracy on pilot overrides under low stakes.
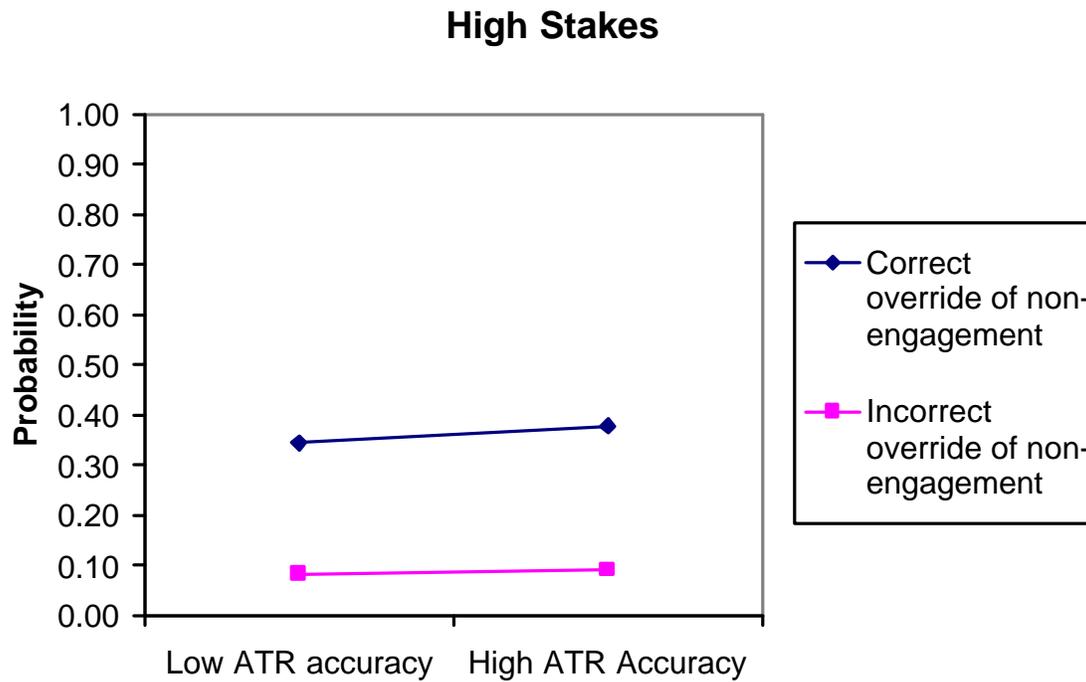
High Stakes

Figure 49. Effect of ATR accuracy on pilot overrides under high stakes
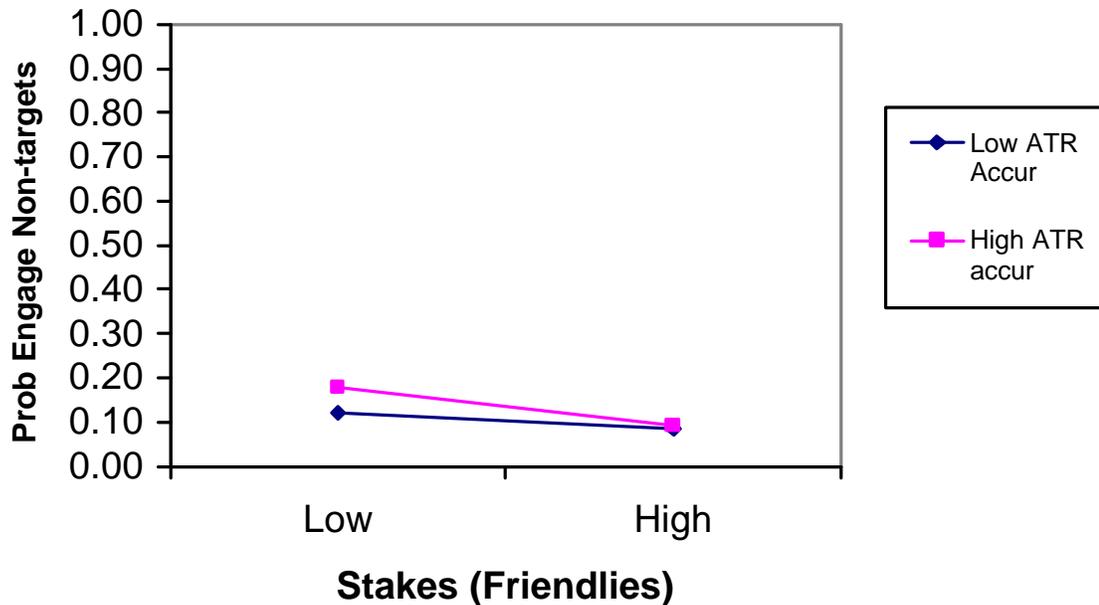
Figure 50. Interaction of stakes and ATR accuracy on engagement of non-targets.

Figure 50 suggests that the interaction of stakes and ATR accuracy extends to overall pilot-ATR outcomes. The interaction of stakes and accuracy is only a trend (univariate $F_{1,10} = 2.74$; $p = .129$; Hotellings $F_{3.5} = 2.81684$; $p = .147$; same significance level for Pillais and Wilks tests). However, under high stakes the tendency to engage non-targets at high ATR accuracy, virtually disappears. When it really matters, pilots take enough care in decision making to wipe out the biasing effect of ATR accuracy.

ATR Support for Verification

Search guidance was introduced to the ATR design to steer pilots' attention to images about which the ATR was uncertain and for which the cost of an error was high. The model of the verification decision presented in the introduction to this chapter and in Appendix D, predicts that such guidance would be beneficial only under conditions of relatively low time stress. When time is available, pilots can take advantage of the guidance in order to examine problematic images more carefully. Guidance would not be useful under conditions of high time stress. It might even be harmful if it distracted pilots from acting to engage enemy tanks about which the ATR was relatively certain.

A similar prediction is made by the verification model for the effects of labeling rules. Favored labels, as determined by the first series of studies (Chapter 2 above), should outperform other labels under most conditions. The adaptive rule presents more general labels under conditions of uncertainty and high cost of errors, in order to permit independent pilot

determination of vehicle classification. This rule should be most useful under conditions of low time stress, when pilots are able to make such determinations.

We will look first at the effects of search guidance, then at the effects of labeling rules.

Search guidance. Alerting pilots to problematic images significantly reduced the probability of fratricide, as shown in Figure 51 (univariate $F_{1,10} = 7.47$; $p = .021$; MANOVA not significant). Search guidance had no discernable influence on the probability of engaging targets.

Was search guidance more effective under conditions of low time stress? There is evidence, shown in Figure 52, that search guidance led pilots to take more time to think before an engagement under low time stress, but did not lead them to take more time under high time stress. There was a marginally significant interaction of search guidance and time stress in their effects on the latency of target engagements (univariate $F_{1,7} = 4.07$; $p = .083$; MANOVA not significant).

The principle effect of search guidance, as noted above, was a reduction in the probability of a fratricide. If pilots were making better use of search guidance under time stress, then we would expect the reduction in fratricides to be greater under low time stress than under high. This is precisely what is observed in Figure 53, though the effect is not significant.
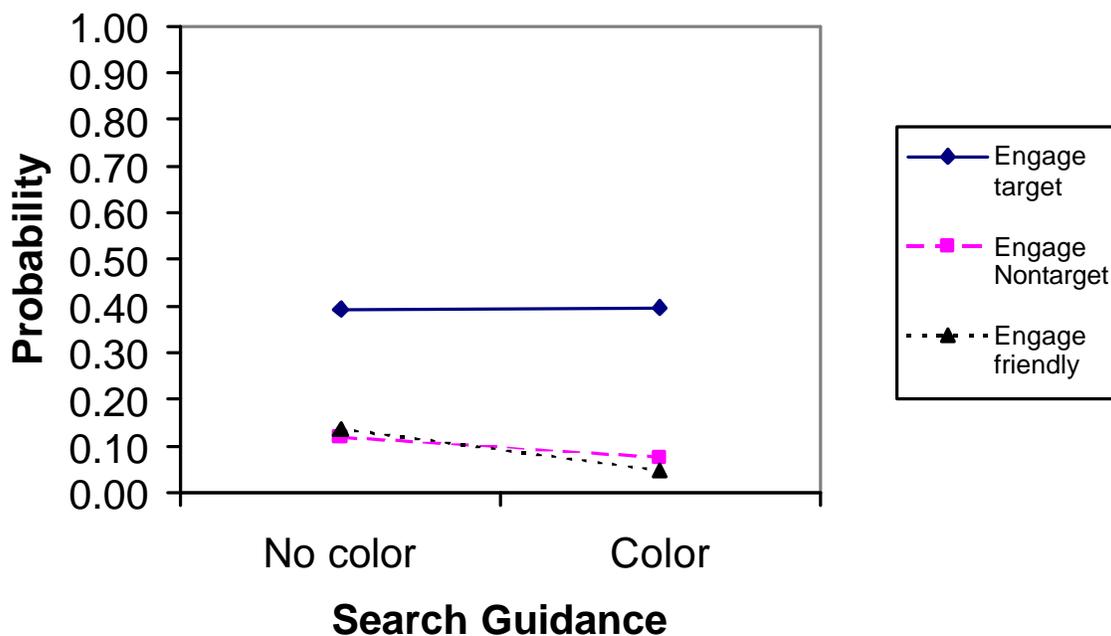
Figure 51. Effect of search guidance on overall outcomes.
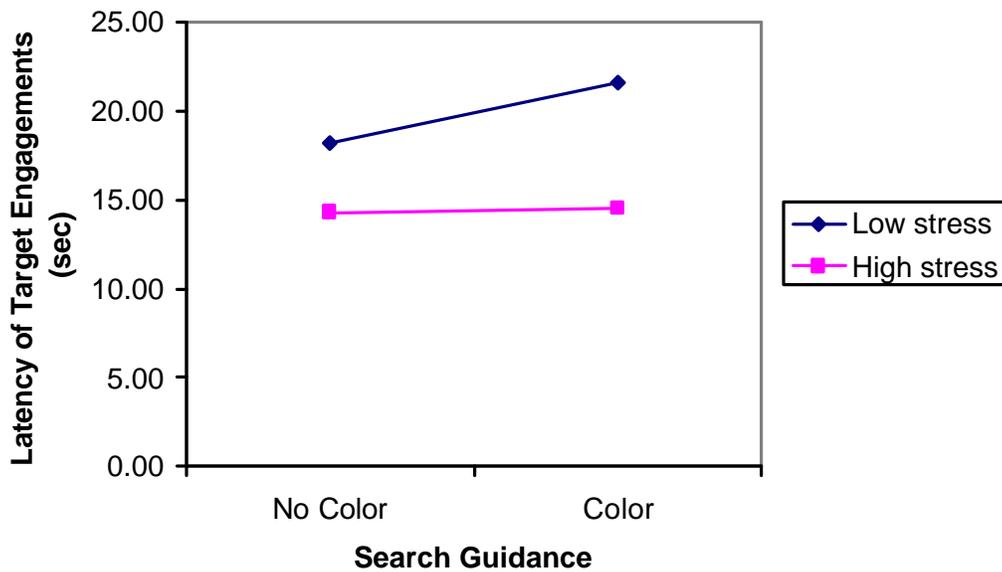
Figure 52. Effect of search guidance on latency of target engagements under two levels of time stress.
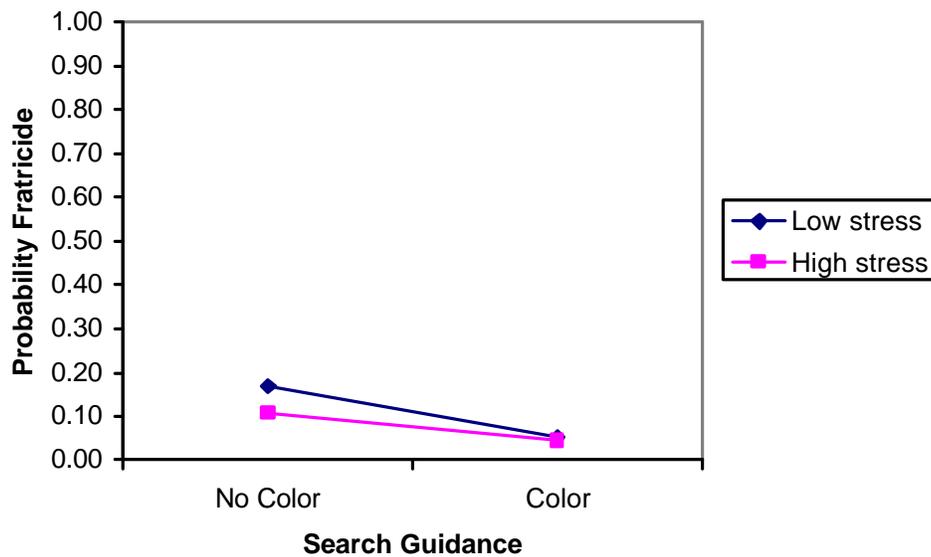
Figure 53. Effect of search guidance on engagements of friendlies under two levels of time stress.

Other support for the relative advantages of search guidance under low stress comes from examination of pilot interventions. Figure 54 shows that under high time stress search guidance may have been harmful, while under low time stress it may have been helpful. Under high stress search guidance led to a non-significant decrease in correct overrides of ATR engagement recommendations, while under low workload it led to an increase in correct overrides of

engagement recommendations. By contrast, search guidance reduced incorrect overrides of engagement recommendations under both high and low time stress (Figure 55).

Correct Override of Engagements
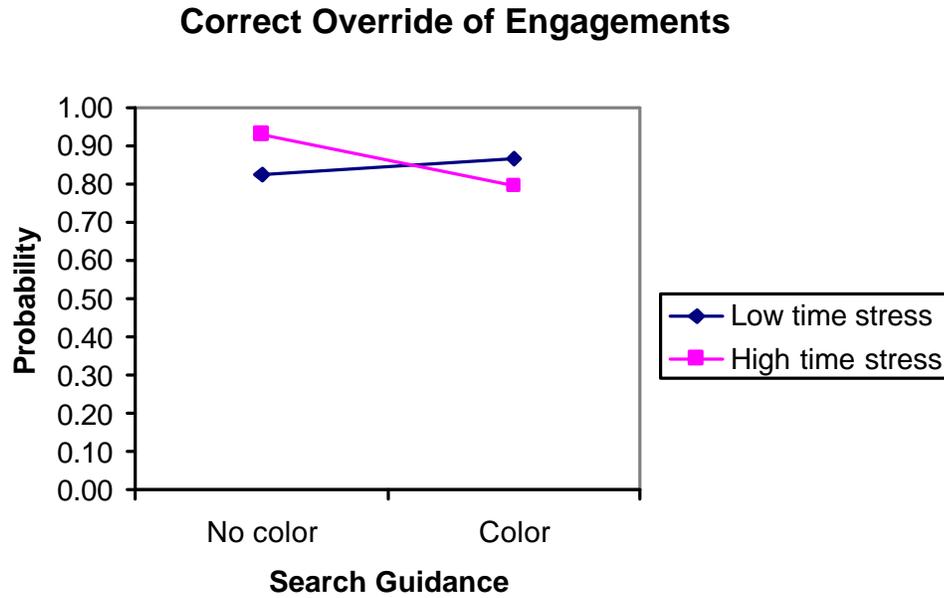
Figure 54. Effect of search guidance on the pilot's correct override of ATR engagement recommendations under two levels of time stress.

Incorrect Override of Engagements
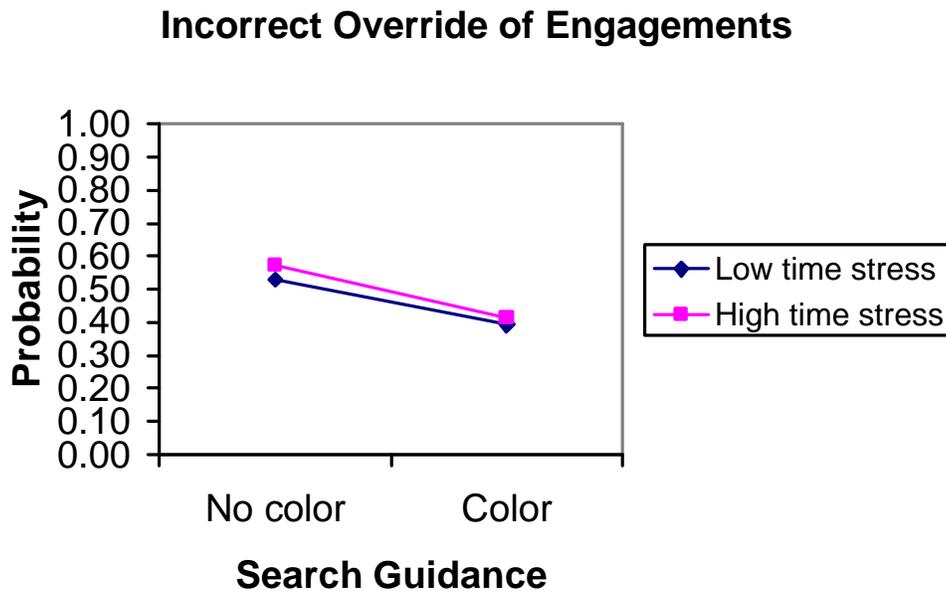
Figure 55. Effect of search guidance on the pilot's incorrect override of ATR engagement recommendations under two levels of time stress

Search guidance led to consistently better performance, regardless of the measure, across all different types of vehicles. As Table 19 shows, there was little effect overall on the probability of engaging enemy tanks. However, search guidance reduced the chance of engaging a friendly tank from 17% to 6%, and the chance of engaging a friendly APC from 6% to 1%.

Table 19. Effect of search guidance on ATR-pilot performance as a function of vehicle type. Asterisk marks better performance.

		Search Guidance	
Measures	**Vehicle Type**	**No Color**	**Color**
Engage targets	**APC**		
	Jeep		
	Tank	0.39	*0.40
	Truck		
Engage non-targets	**APC**	*0.11	*0.11
	Jeep	0.05	*0.03
	Tank	0.17	*0.06
	Truck	0.05	*0.04
Engage friendlies	**APC**	0.06	*0.01
	Jeep	0.04	*0.02
	Tank	0.17	*0.06
	Truck	0.07	*0.00

In order to understand better how search guidance works, Table 20 breaks down the results according to the color code that was displayed. For example, the first row represents the chance of engaging an image that is in fact a target, given that it is represented in the ATR grid by the symbol at the top of the column (i.e., gray, red, blue, or yellow in the search guidance conditions, and nothing at all in the no search guidance condition).

Not surprisingly, alerting the pilot to likely enemy tanks (by means of a red color) increased the chance of engaging enemy tanks from 39% in the no guidance condition to 81%. Moreover, the average time required to engage an enemy tank was cut in half, from 16.60 seconds in the no guidance condition to 8.58 seconds when it was marked red.

When images of enemy tanks were labeled yellow, to warn of significant, high-stakes uncertainty, the chance of engagement was, of course, reduced. The most remarkable thing about Table 20 is that the reduction was not greater. The rate of engagement for enemy tanks that were marked yellow was 47% - still a better level of performance against these targets than in the no search condition! Inviting pilots to examine an image with special care does not prevent them from eventually recognizing it as a target and engaging it. Moreover, latencies for engaging targets marked yellow were no longer than the latencies in the no guidance condition.

Table 20. Effect of different types of search guidance on various measures of performance. Asterisks indicate types of guidance which outperformed the no-search-guidance condition on a particular measure.

Measures	Search Guidance				
	No Search Guidance	Low stakes (gray)	Enemy (red)	Friend (blue)	Uncertain & high stakes (yellow)
Engage targets	0.39	0.31	* 0.81	0.17	* 0.47
Engage non-targets	0.12	* 0.07	* 0.00	* 0.05	* 0.14
Engage friendlies	0.14	* 0.04		* 0.03	* 0.13
Latency to engage targets	16.60	22.07	* 8.58	21.60	* 16.44

Similarly, the chance of engaging a friendly, as well as the chance of engaging non-targets in general, was lower regardless of the color code applied to the image, when compared to the no guidance condition. Interestingly, mistakenly marking a non-target as an enemy resulted in no incorrect engagements. Again, marking non-targets as yellow, indicating the chance that they were targets, did not increase the chance of incorrect engagements compared to the no guidance condition.

Labeling Rules. There was a tendency for the effectiveness of different labeling rules to vary with time stress (Pillais $F_{9,21}$ = 1.88467; p =.111; slightly lower significance levels for Hotellings and Wilks tests). In particular, there was a marginally significant interaction of labeling rule and time stress on the probability of engaging a non-target ($F_{3,10}$ = 3.69; p =.051). As shown by Figure 56, the favored labeling rule outperformed the other three rules under both levels of time stress. Most interestingly, all labeling rules except the favored rule produced worse performance under high time stress than under low time stress. The favored rule was absolutely consistent across variations in time stress. It should be noted also that the adaptive rule performed worst of all under both levels of time stress. It may be better for the ATR to give the pilot its best guess regarding target classification even when it is uncertain.

Figure 57 shows a similar, but non-significant interaction with respect to the probability of engaging a target. All labeling rules except the favored rule produced worse performance under high time stress than under low time stress. Once again, only the favored rule was unaffected by time stress. Moreover, under high time stress, the favored rule produced the best performance. Under low time stress, the adaptive rule and the contrast rule provided the highest probability of engaging targets. However, performance under both of these rules plummeted when time stress increased.

Engaging Non-Targets
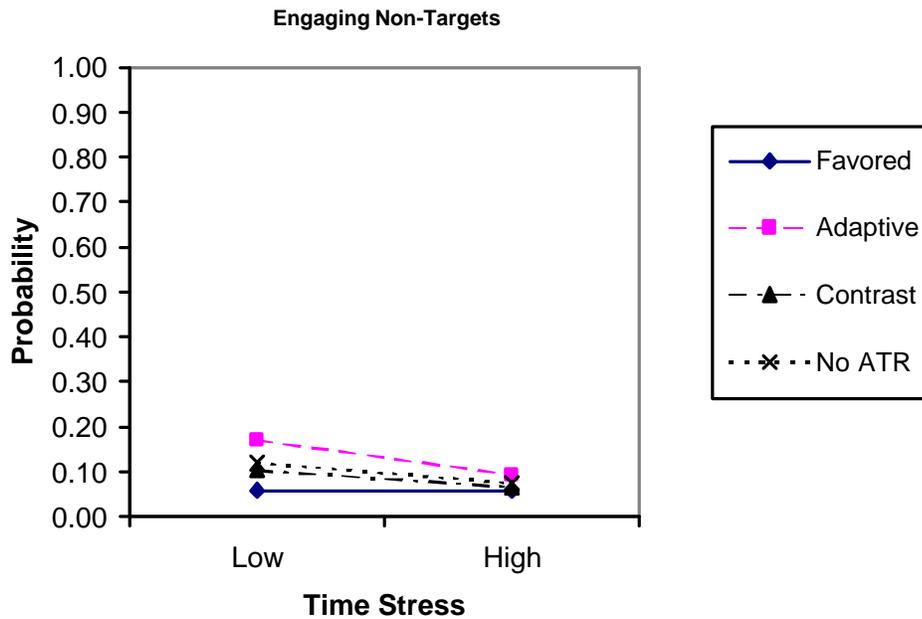
Figure 56. Effect of time stress on engagement of non-targets, for different labeling rules.

Engaging Targets
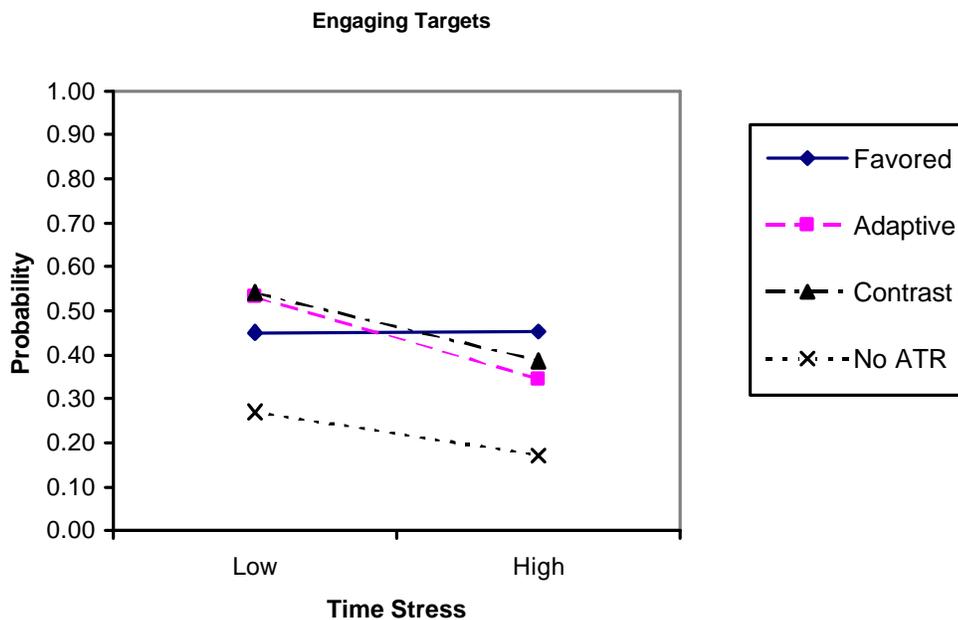
Figure 57. Effect of time stress on engagement of targets, for different labeling rules.

It should not be overlooked that the most dramatic impression in Figure 57 is the advantage of *any* labeling rule over no labels at all, under both conditions of time stress. The presence of an ATR is evidently beneficial for engaging targets, regardless of the fine points of

how it labels its conclusions. Figure 56, and the associated statistical tests, suggest, however, that the mode of labeling makes more of a difference in avoiding the engagement of non-targets.

The favored rule consistently outperforms the other labeling rules across all vehicle types, as shown in Table 21. Once again, there was little difference across rules in the chance of engaging a target. However, the chance of engaging a non-target or a friendly tended to be lower for the favored rule than for any other rule, for all non-target vehicle types. In this regard, the adaptive labeling rule was often no better, or even worse than, no labeling at all. The contrast rule performed relatively well across the different vehicle types, outperforming both the adaptive rule and the no-label rule across all vehicle types and measures.

Table 21. Effect of different labeling rules on different measures of performance, as a function of actual vehicle type in the image. Asterisks indicate best performance on a particular measure for a given vehicle.

Measures	Vehicle Type	Labeling Rule			
		Favored	Adaptive	Contrast	No labels
Engage targets	APC				
	Jeep				
	Tank	0.45	0.44	* 0.46	0.22
	Truck				
Engage non-targets	APC	* 0.07	0.14	0.10	0.11
	Jeep	* 0.01	0.06	0.04	0.07
	Tank	* 0.07	0.15	0.09	0.10
	Truck	* 0.01	0.06	0.04	0.06
Engage friendlies	APC	* 0.01	0.06	* 0.01	0.03
	Jeep	* 0.01	0.03	0.02	0.05
	Tank	* 0.07	0.15	0.09	0.10
	Truck	* 0.00	0.07	* 0.00	0.04

The favored rule tends to outperform other rules with respect to accuracy. How does it compare in terms of speed? Figure 58 suggests that it does quite well. Although the differences are not significant, the favored rule was associated with faster reaction times than any other rule for engaging targets. The contrast rule, which leads to a relatively high accuracy for engaging targets, falls down in the speed category. The contrast rule requires more time for target engagement than any other rule, including no labeling at all.

In sum, although results are mostly in the form of trends, the favored labeling rule seems to outperform others in both speed and accuracy across a range of vehicles and time stress conditions.
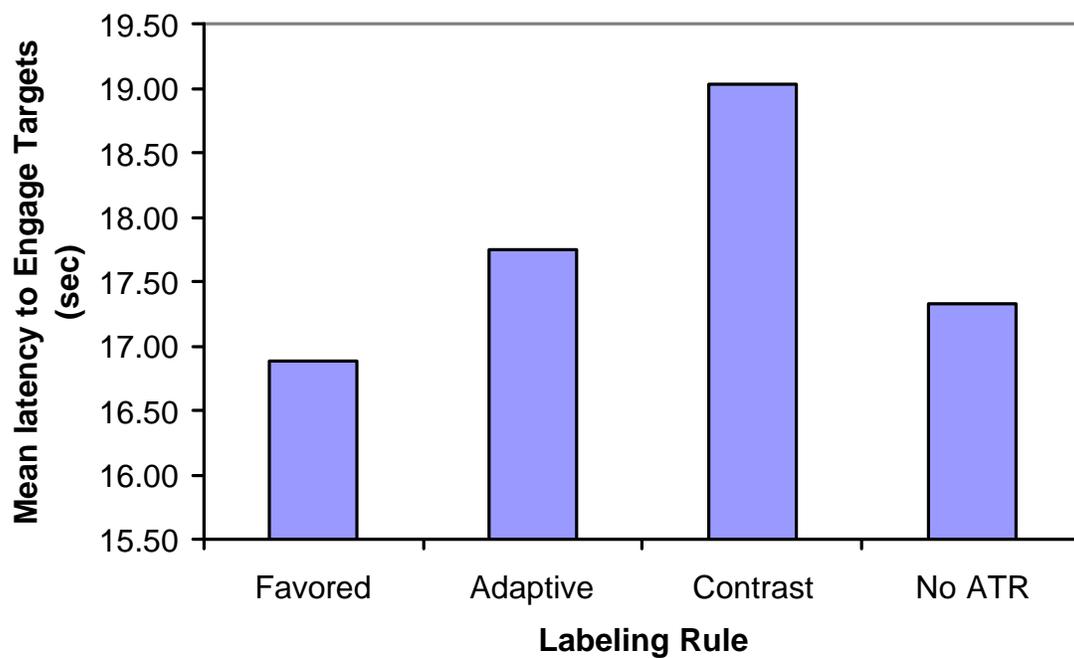
Figure 58. Effect of labeling rule on time required to engage a target.

5. SUMMARY AND CONCLUSIONS

The focus of this research has been on the intersection of cognitive and perceptual aspects of human target recognition performance, and on potential enhancements of the human-ATR interface. The premise of the research was that improving the effectiveness of human-ATR interaction is not simply a matter of improving the accuracy of ATR conclusions. It will require an examination of all aspects of human-ATR interaction, including: (1) effective displays of target classification conclusions so that human users can quickly grasp the implications of an ATR conclusion for the mission and, if necessary, verify it by examining the image, (2) effective displays of target imagery to support rapid and accurate user verification of ATR conclusions, and (3) effective support for decision making processes that allocate user attention, decide where and how long to verify ATR conclusions, and determine which targets to engage.

Three series of experiments were conducted with active duty Army pilots. Each study attempted both to lay a scientific basis, and to test a practical methodology, for a promising ATR design application. The studies address the following issues in ATR-human interface design: the appropriate labeling of ATR classification conclusions, the most effective display of image data, and the support of user intervention in the case of uncertain ATR conclusions. Figure 59 summarizes the practical results of the three phases of this research, and places them within the overall context of ATR-user interaction that was introduced in Chapter 1.

To explore the most effective method for ATR labeling of classification conclusions, we conducted a series of studies on human verbal organization of knowledge regarding military vehicles. These studies included feature naming in response to labels, typicality and familiarity ratings, spontaneous naming, and verification of image and label match. The results supported the notion that labels at different levels of specificity would be most effective for different types of vehicles. In particular, the most effective labels for APC's were specific type names, such as *BTR* and *BMP*. The most effective labels for light vehicles were intermediate terms, such as *jeep* and *truck*. For tanks, both intermediate and specific terms were effective in different types of tasks, and the best labels appear to be a modified intermediate term, such as *enemy tank* and *friendly tank*. Feature naming data were analyzed for insight into the similarity structure characterizing verbal feature knowledge. The resulting structure suggested a structure optimized for the task of discriminating enemy tanks from other types of vehicles, rather than designed to produce accurate classifications of all types of vehicles. Four major categories of visual features were identified: profile, weapons, turret, and wheels/tracks.

To explore the most effective displays of target imagery, we investigated stages in the visual processing of an image. A response deadline method was used to obtain a snapshot of visual processing at varying times after presentation of the image. Confusion matrices collected at different points in time were analyzed to determine the feature information that had been extracted by that time. A model of visual processing in terms of successive elimination of possibilities, and guessing among the surviving candidates, provided an excellent fit to the data. The most plausible features extracted at several points in the model were aspects of the vehicle profile. This hypothesis was further tested by comparing visual recognition performance with selectively enhanced images. Two enhancement conditions involved details in a particular part of the images (i.e., wheels/tracks or turret/weapon); one enhancement condition included all details of the image; and a final condition enhanced the vehicle's profile, by heightening its silhouette together and suppressing internal detail. The surprising result was that silhouette or profile

enhancement produced more accurate recognition for all vehicles, at all ranges, and regardless of the amount of time available for visual processing. Moreover, enhancement of all parts of the vehicle produced significantly inferior performance compared to selective enhancement of only part of the vehicle.
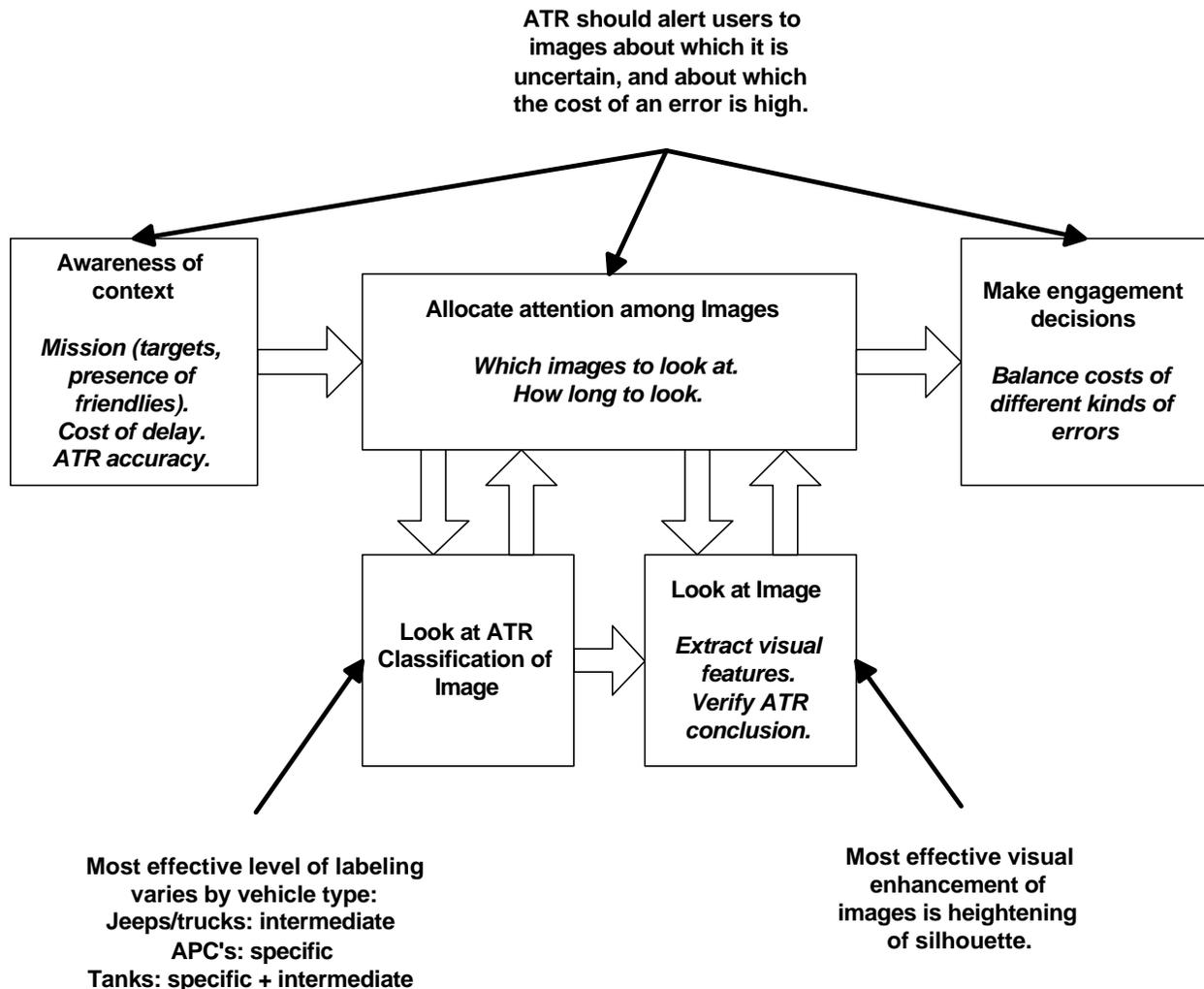
Figure 59. Implications of the three sets of studies for ATR design.

The final study explored methods for supporting user decision making about how best to allocate attention among images. Users had to recognize and engage enemy tanks in a field of diverse types of vehicles, under varying conditions of time pressure, presence of friendless in the area, and ATR classification accuracy. Two ATR design concepts were tested. One used color coding to guide user attention to images that were confidently classified as enemies, confidently classified as friends, or about which the ATR was uncertain whether they were enemy targets or friendly vehicles. A second ATR design concept involved a comparison of different rules for labeling ATR conclusions: One was based on the labels identified as optimal in the first series of studies; another rule adapted the label of an image about which the ATR was uncertain by

adopting a more general description; a third rule contrasted with the optimal labels but appeared to fit mission requirements; and a final condition involved no ATR conclusions at all. Search guidance by means of color coding significantly reduced the probability of fratricide, and produced more accurate engagement decisions for all types of vehicles. The labeling rule defined in the first set of studies likewise provided superior accuracy and speed for all types of vehicles and under all conditions of time stress.

The results of these studies are: (1) improved understanding of basic component processes in human-ATR interaction (i.e., verbal classification, visual recognition, and decision making); (2) specific ATR design concepts to enhance human-ATR interactions (i.e., optimized labels for ATR conclusions, optimized enhancements for ATR imagery, and guides for user allocation of attention to ATR images and conclusions); and (3) a set of tested empirical methods, models, and paradigms for generating additional understanding and ATR interface design concepts.

6. REFERENCES

Anderson, J. R. The adaptive character of thought. Hillsdale, NJ: Lawrence Erlbaum Associates, 1990.

Barsalou, L.W. Deriving categories to achieve goals. In G.H. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory (Volume 27). New York: Academic Press, 1991, 1-64.

Broadbent, D.E. Decision and stress. New York: Academic Press, 1971.

Cohen, M.S., and Freeling, A.N.S. 1981. *The impact of information on decisions: Command and control system evaluation* (Technical Report 81-1). Falls Church, VA: Decision Science Consortium, Inc.

Cohen, M. S., Parasuraman, R., Serfaty, D., & Andes, R. C. Trust in Decision Aids: A Model and A Training Strategy. Arlington, VA: Cognitive Technologies, Inc. 1997.

Corter, J.E., and Gluck, M.A. Explaining basic categories: Feature predictability and information. Psychological Bulletin, 1992, 111(2), 291-303.

Cruse, D.A. The pragmatics of lexical specificity. Journal of Linguistics, 1977, 13, 153-164.

Foss, Christopher F. (1992). Jane's AFV Recognition Handbook. Alexandria, VA. Page xii).

Joliceur, P., Gluck, M., and Kosslyn, S.M. Picture and names: Making the connection. Cognitive Psychology, 1984, 16, 243-275.

LaValle, I.H. 1968. On cash equivalents and information evaluation in decisions under uncertainty. *American Statistical Association Journal, 63*:252-290.

Luce, R.D. The choice axiom after twenty years. Journal of Mathematical Psychology, June 1977, 15(3).

Luce, R.D. The choice axiom after twenty years. Journal of Mathematical Psychology, June 1977, 15(3).

Navon, D. Forest before trees: The precedence of global features in visual perception. Cognitive Psychology, 1977, 9, 353-383.

O'Kane, B., Biederman, I., & Cooper, E. E. (1994). Shape features for target identification derived from thermal imagery. 19th Army Science Conference Proceedings, June, 1994.

Pachella, R.G., and Pew, R.W. Speed-accuracy tradeoff in reaction time: The effect of discrete criterion times. Journal of Experimental Psychology, 1968, 76, 19-24.

Pachella, R.G., Smith, J.E.K., and Stanovich, K.E. Qualitative error analysis and speeded classification. In J. Castellan (Ed.), Cognitive Science III. Hillsdale, NJ: Lawrence Erlbaum Associates, 1978, 169-198.

Reed, A.V. List length and the time course of recognition in immediate memory. Memory and Cognition, 1976, 4, 16-30.

Raiffa, H. & Schlaifer, R. 1961. Applied statistical decision theory. Cambridge, MA: The M.I.T. Press.

Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., and Boyes-Braem, P. Basic objects in natural categories. Cognitive Psychology, 1976, 8, 382-439.

Tabachnick, B.G. & Fidell, L. S. (1989). Using Multivariate Statistics. NY: Harper Collins Publishers, Inc.

Toms, M.L., and Kuperman, G.G. Sensor fusion: A human factors perspective. Dayton, OH: Logicon Technical Services Incorporated, September 1991.

Townsend, J.T. Theoretical analysis of an alphabetic confusion matrix. Perception and Psycho-physics, 1971, 9, 40-50.

APPENDIX A: IMAGES FOR FIRST SET OF EXPERIMENTS

APC

TANK

BMP
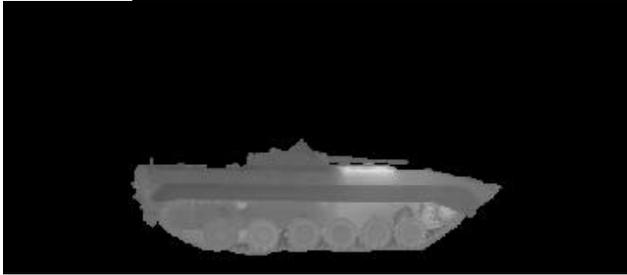
T62

BRDM

T55

BTR

M551

TRUCK

JEEP

ZIL

UAZ
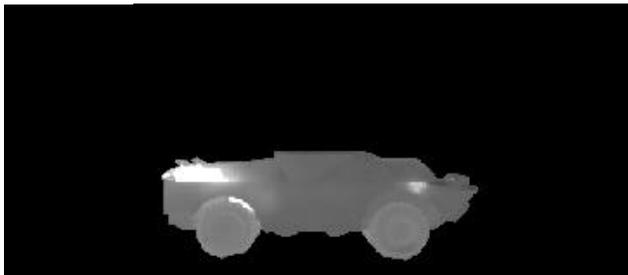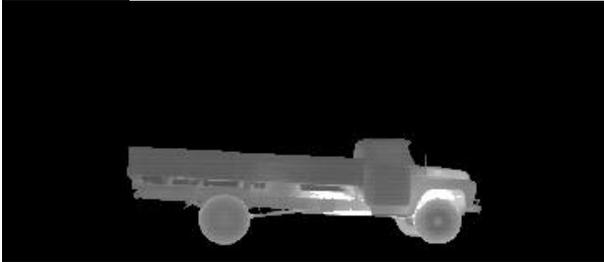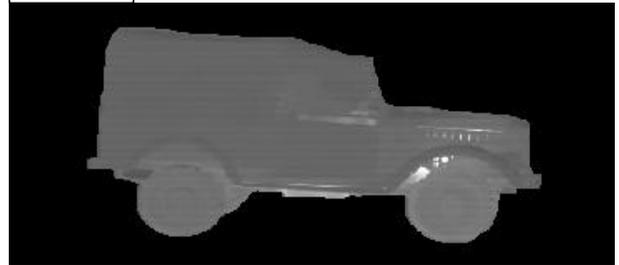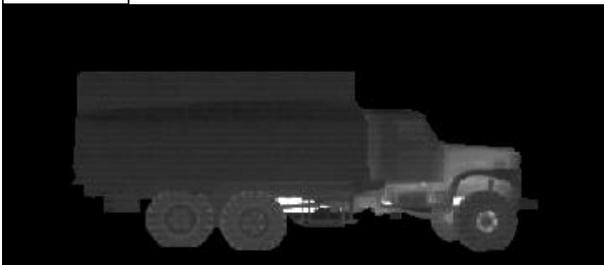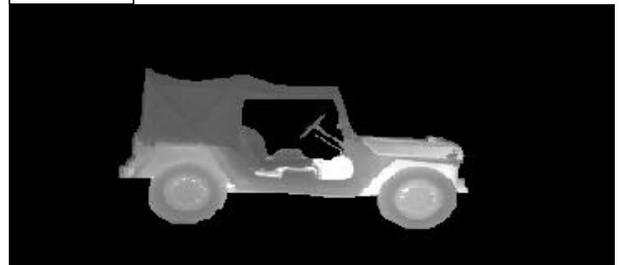
KRAZ

M151

Table 22. Non-visual features generated in the feature naming task in response to vehicle labels.

	WHEELED	TRACKED	APC	TANK	JEEP	TRUCK	BMP	BRDM	BTR	M551	T55	T62	M151	UAZ	KRAZ	ZIL	Total
age: old								1		2	5	1					9
cargo or crew count: 4					1							1	1				3
comfortable: no					1												1
handling: hard to drive													1				1
mission: c2			1														1
mission: opfor									1								1
mission: personnel carrier	1		6		3	2	2	3	2								19
mission: recon							1	2									3
mission: target practice										1							1
mission: utility vehicle	1					1							1	2	1	1	7
mobile:	1	1								1							3
owner: airborne										3							3
owner: enemy								1								1	2

owner: friend										2						2
owner: korea											1		1			2
owner: mrr and below							1									1
owner: soviet												1			1	2
owner: third world										1						1
range: short				1												1
rolls:	1				1											2
roof convertible: yes					1	1						1				4
speed: fast	1				2											3
speed: faster than tracked	1															1
speed: slow		1		1												2
target priority: secondary				1								1				2
vehicle type: apc	2	2					2	1	2					1	1	11
vehicle type: apc or tank		1														1
vehicle type: armament		1			1		1				1	1			1	6
vehicle type: brdm	1							1								2

																	Grand Total
vehicle type: btr	1								2								3
vehicle type: german half-track		1															1
vehicle type: jeep	2													2		1	5
vehicle type: m151					2												2
vehicle type: tank		3	1				1			3	6	7					21
vehicle type: tracked apc								1									1
vehicle type: truck	2					2	1					1		1		2	9
vehicle type: wheeled apc							1		1								2
versatile:		2			1												3
weapons type: 3k range				1													1
windshield type: folding													1				1
Grand Total	6	4	7	4	10	4	4	7	3	11	6	4	5	3	2	2	82

Table 23 Visual features generated in the feature naming task in response to vehicle labels.

	WHEELED	TRACKED	APC	TANK	JEEP	TRUCK	BMP	BRDM	BTR	M551	T55	T62	M151	UAZ	KRAZ	ZIL	Total
antenna:	1	1	1		1		1	1		1							7
axles number:						1											1
body size vs weapons size: small body & large weapon													1				1
body temperature: hot				1													1
brake temperature: hot						1											1
cab shape:	1					1											2
cab:						2											2
camouflage:				1	1			1				1					4
cargo area size:						2											2
cargo area size: long						1											1
cargo area size: small													1				1
cargo area:					1	1								1			3
chassis:	1								1				1			1	4
cover:					1	1							1			1	4

cupola location: center									1						1
cupola size: large										1					1
cupola:				1											1
door location: between two wheels									1						1
door location: mid-body									1						1
door location: rear									1						1
door location: sides									1						1
door size: large					1										1
door: none									1						1
engine location: rear									1						1
engine:					1										1
external fuel tank:				1											1
frame temperature: hot						1									1
gun port:			1												1

Attribute																Total
hand rails location: turret											1					1
hand rails:		1								1	1				1	4
hatch count: multiple									1							1
headlight size: big					1											1
heat source:	1	1		2			1	1		1	1					8
light size: big										1						1
loaded: yes or no						1										1
nose shape:					1			1	1							3
nose shape: boat-like								1								1
nose shape: flat			1	2												3
nose shape: low slope							1									1
nose shape: pointed							5	1								6
nose shape: round								1								1
nose shape: sharp slope								1								1
nose shape:							1				1					2

																	Total
sloped																	
orientation: position relative to battle area										1							1
peep hole:		1															1
profile:	2	3	3	1	3	2	1	2	3	3	2	2			1	2	30
profile: boat-shaped							1										1
profile: boxy			4			1		1	1					1			9
profile: car-like								1									1
profile: dump-truck shape						1											1
profile: flat		1															1
profile: high										2	1						3
profile: long															1		1
profile: low		1					5			1	3			1			11
profile: plain						1											1
profile: proportion of cab						1											1
profile: short										2							2
profile: slab-sided			1														1

																	Total
profile: tall			1										1				2
profile: tank-like														1			1
rifle port count: many												1					1
roadwheels count:			1														1
roadwheels count: 5													2	2			4
roadwheels count: 6									1								1
roadwheels count: 8												2					2
roadwheels location:			1														1
roadwheels location: center														1			1
roadwheels manufacture: stamped															1		1
roadwheels shape:			1														1
roadwheels size:			1														1
roadwheels size: large												2		2	1		5

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total
roadwheels spacing: area between	1														1
roadwheels:		2		3						1	1	1			8
roof material: canvas					1										1
roof: none					1										1
searchlight location: on gun										1					1
searchlight location: right of gun												1			1
searchlight:				1								1			2
size:		2	1	1	1	4	1			1				1	12
size: large				1		1	1		1	1		1	1		7
size: length						2							1		3
size: long		1							1						2
size: personnel length	1														1
size: small	1		2		4			1		1	3		2		14
size: wide												1			1
skin material: armor		1	1	3				1			1	1			8
skin material:			1	2											3

heavy armor																	
skin material: light				2	2												4
skin material: light amor				1													1
skin material: thin	1																1
skin temperature: hot						1											1
snorkel: present				1								2					3
splash guards: present											1	1					2
sponson box location: all around				1													1
sponson box location: behind turret										1							1
sprocket location: front & rear										1							1
sprocket:				1													1
support rollers:				1													1
suspension												1					1

rollers: none															
suspension type: christi						1			2	1	2				6
suspension:	2			2		1			2	2	1			1	11
top shape: flat						1									1
top skin: canvas				1											1
track skirts:											1				1
tracks shape:						1									1
tracks size: small						1									1
tracks spacing:										1					1
tracks temperature: hot	1														1
tracks type:	1														1
tracks type: slack											1				1
tracks: present	5	5	11			8	1	1	5	3	5	3		1	48
turret location: center									1						1
turret location: forward						1	1	3	1	1					7

turret shape:		2					1	1		4
turret shape: crab-shaped						1				1
turret shape: cup-like			1							1
turret shape: flat			1						2	3
turret shape: flat with gun						1				1
turret shape: long									1	1
turret shape: low			1		1					2
turret shape: middle							1	1		2
turret shape: oval							1	1		2
turret shape: round							4	4		8
turret size:		1								1
turret size: high						1				1
turret size: large						2				2
turret size: small	1		2	2	3		1	1		10

turret: present	2		6	1		3	1	2	3	1	1				20
turret: absent		2													2
weapons bore evacuator location: in center of gun tube										1	3				4
weapon size: short thick barrel									1						1
weapons location:	1		1			1									3
weapons location: top		1						1							2
weapons location: turret						1	1		1						3
weapons muzzle location: 1/3rd from barrel										1					1
weapons muzzle size: large muzzle											1				1
weapons muzzle size: short									1						1
weapons											1				1

muzzle supressor:														
weapons shape:			1						1	1				3
weapons shape: machine gun angle up							1							1
weapons size:	1				1									2
weapons size: big			1											1
weapons size: gun longer than tank										1				1
weapons size: large barrel			3		1					1				5
weapons size: large gun with bore evaucator at end									1					1
weapons size: short barrel					1		1	3						5
weapons size: small		1			1		1							3
weapons size: small gun						1								1
weapons type:										1				1

weapons type										Total
100mm gun										
weapons type: 100mm gun w muzzle break							1			1
weapons type: 105mm smooth bore gun								1		1
weapons type: 105mm to 125mm gun	1									1
weapons type: 105mm with bore evacuator							1			1
weapons type: 12.7mm machine gun	1		1	1				1		4
weapons type: 120mm								1		1
weapons type: 12mm machine gun					1					1
weapons type: 152mm gun						1				1
weapons type: 30mm gun									1	1
weapons type:			1							1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
30mm or .50mm															
weapons type: 7.62mm machine gun											1				1
weapons type: 85mm gun										1					1
weapons type: aa								1							1
weapons type: ADA		1													1
weapons type: anti-tank or saggars			1												1
weapons type: atgm							1								1
weapons type: grenade-launchers				1											1
weapons type: gun				1				1	1						3
weapons type: gun barrel				1								1			2
weapons type: howitzer												1			1
weapons type: machine guns				1						1					2

weapons type: main gun				2						1	1	1	5
weapons type: mid-sized main barrel w/ bore evacuator in front											1		1
weapons type: missile				1									1
weapons type: saggars								1					1
weapons type: small			1				1	1					3
weapons type: small caliber on top							1						1
weapons type: small cannon			1										1
weapons type: small gun										1	1		2
weapons type: small machine gun				1									1
weapons type: smooth bore											1		1
weapons: present	1	2	3	1	1		2	1	1		1	1	14

121

weapons: absent						1		1							2
wheels clarity: distinguishable		1													1
wheels count:	2			1		1	1			1	1	1			8
wheels count: 4					5	1		4	1			1			12
wheels count: 4 plus center wheels									1						1
wheels count: 4-18						1									1
wheels count: 4x4					1			1							2
wheels count: 6									2						2
wheels count: 8									4						4
wheels count: 8x8									1						1
wheels count: many						1									1
wheels location:	1														1
wheels or tracks count:			1												1

																	Total
wheels or tracks: present			6				1										7
wheels shape: flat	1																1
wheels size:	1																1
wheels size: big								1									1
wheels size: large									1								1
wheels size: small								1									1
wheels spacing:	1	1		1			1		1			1					6
wheels type: rubber	1																1
wheels: present	4		2		5	7	5	3					3	2	2	3	36
wheels: absent		1															1
windshield location: high					1												1
window: present						1											1
windshield: absent					1												1
Grand Total	30	46	47	60	43	41	61	46	56	54	56	64	23	6	11	13	657

123

Multidimensional Scaling and Cluster Analysis based on Major Categories of Verbalized Features

APPENDIX C: IMAGE ENHANCEMENTS

BMP	M-60
BTR	T-62
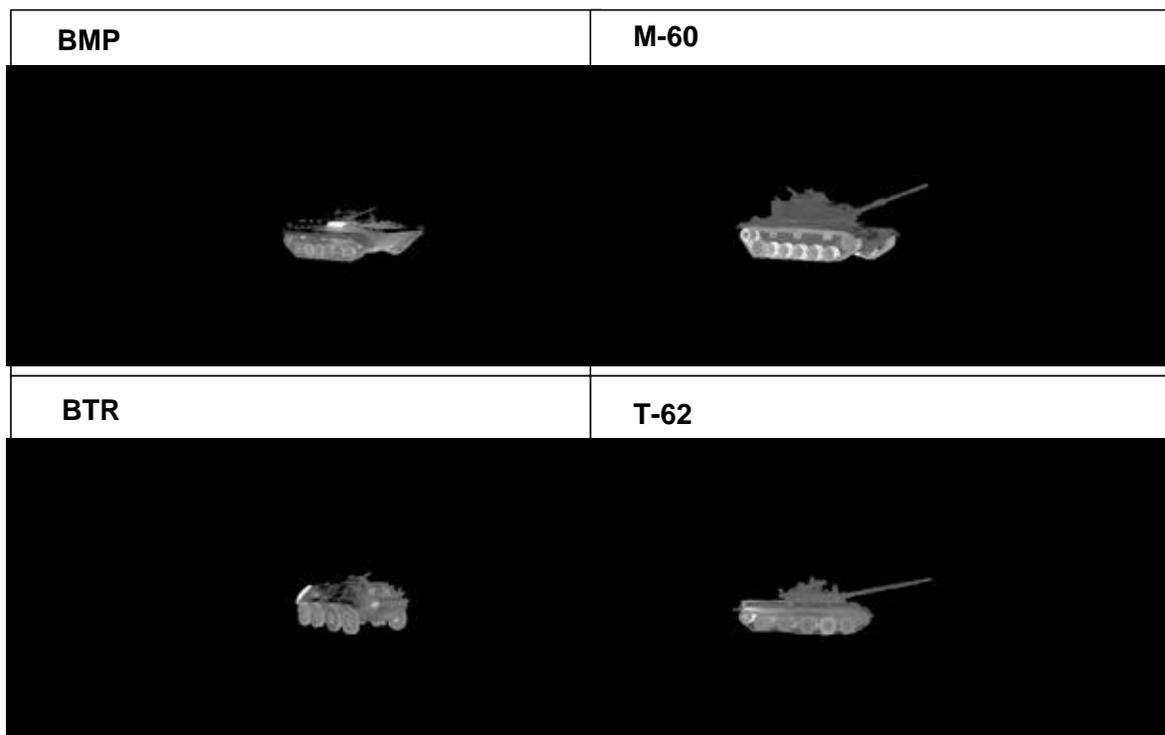
Figure 60. Unenhanced images.

BMP	M-60
BTR	T-62
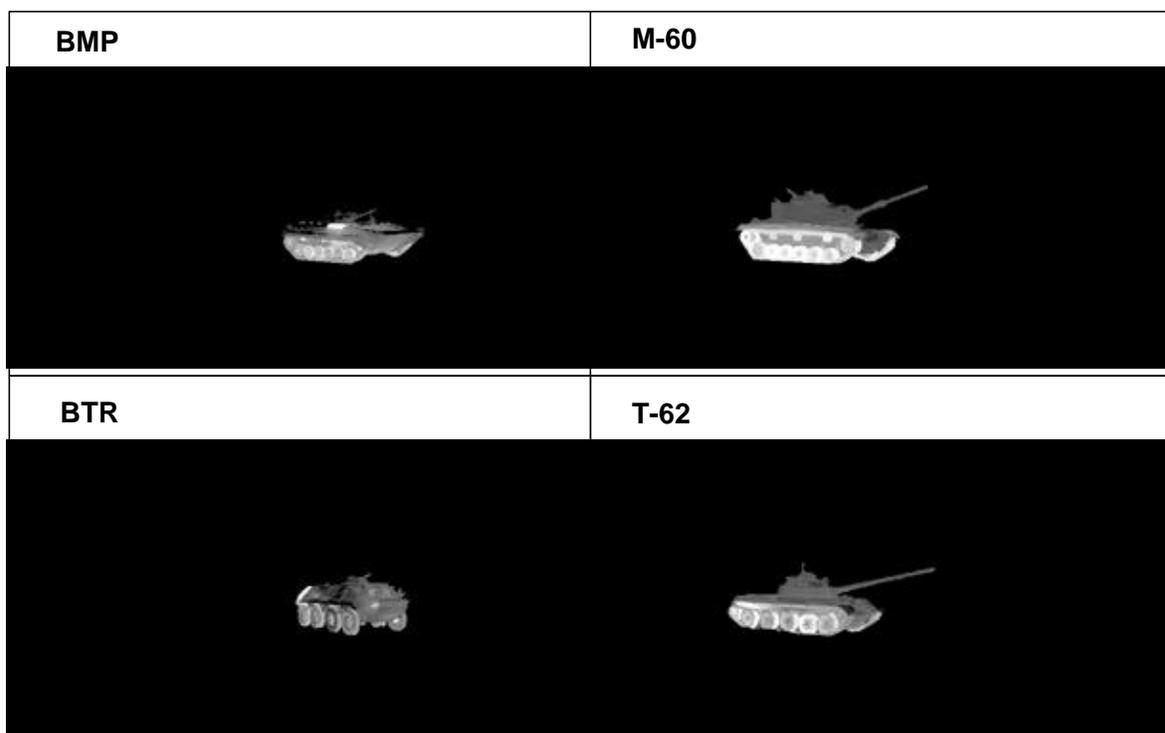
Figure 61. Images with enhanced wheels and tracks.

BMP	M-60

BTR	T-62
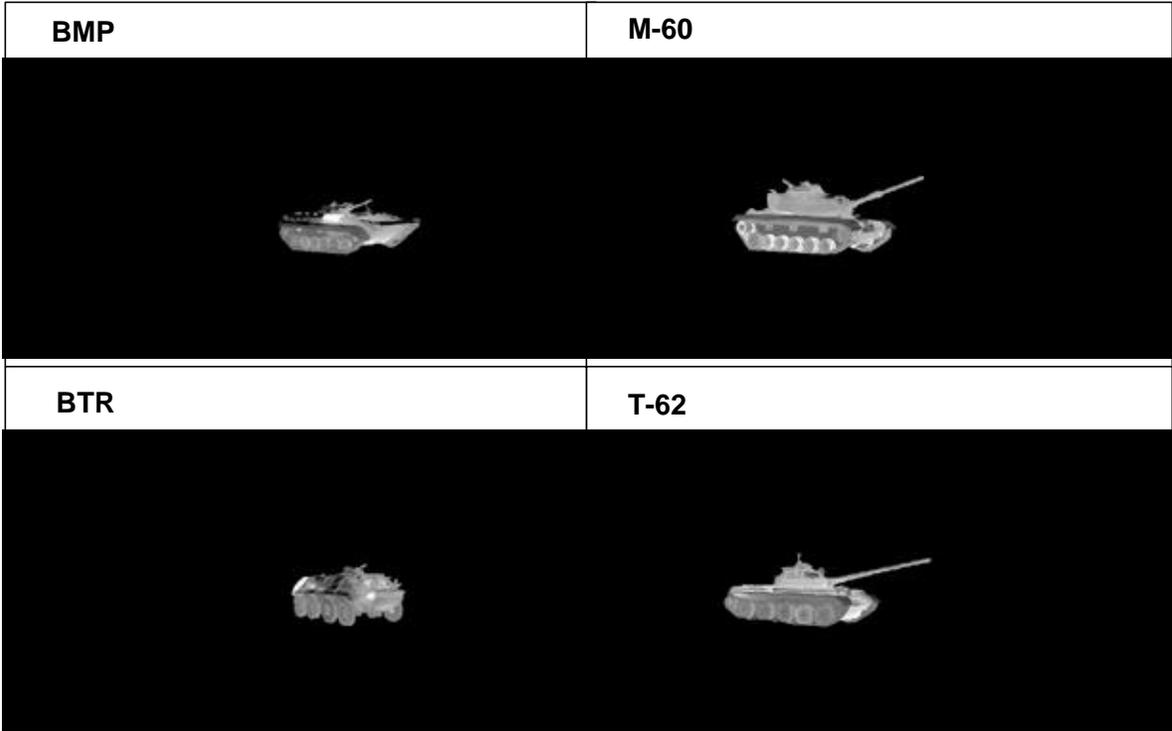
Figure 62. Images with enhanced turret and gun.

BMP	M-60

BTR	T-62
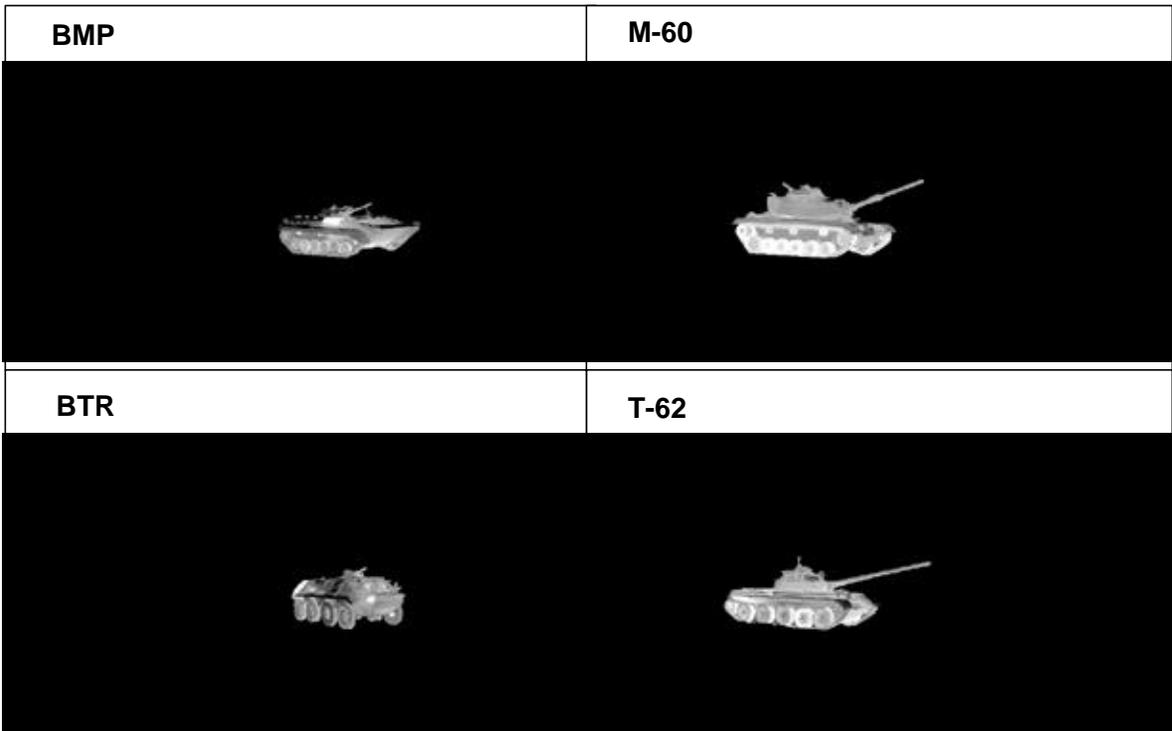
Figure 63. Images with entire vehicle enhanced.

BMP	M-60

BTR	T-62
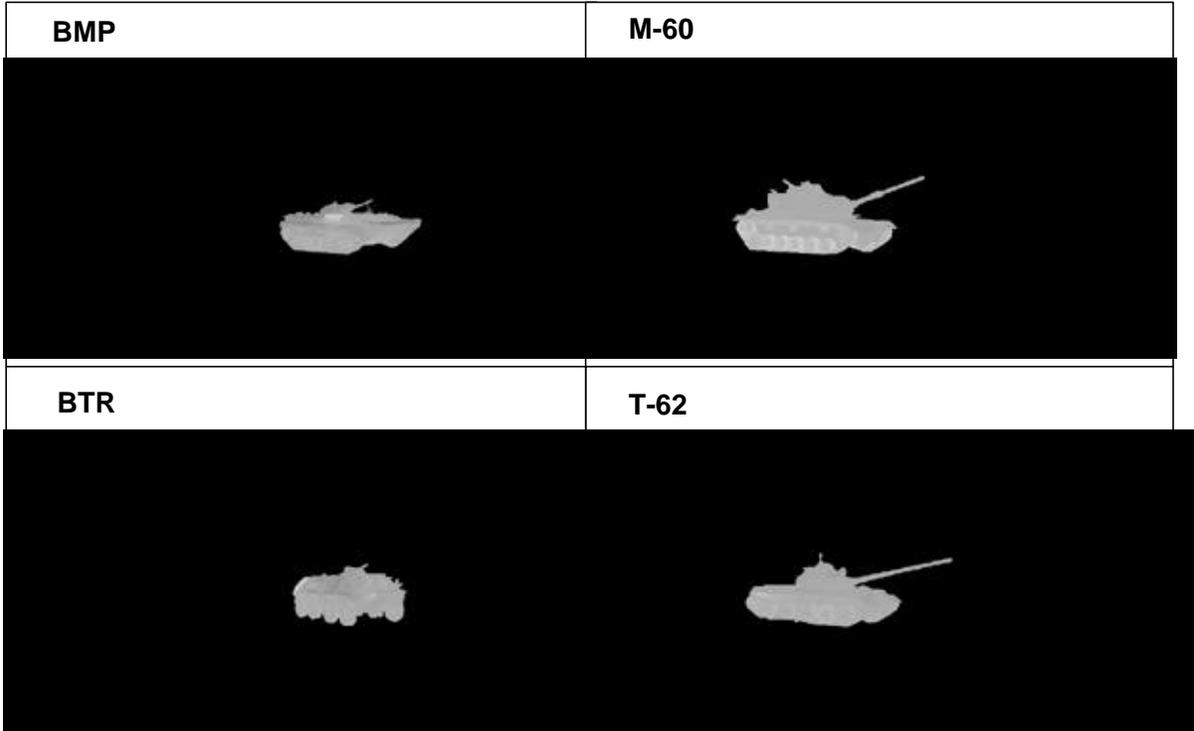
Figure 64. Images with silhouette enhanced.

APPENDIX D: A MODEL OF VERIFICATION DECISION

In this appendix, we will describe a model of the verification decision more generally, and a bit more formally. The model is an adaptation of the decision theoretic measure of value of information, as presented, for example, in Raiffa & Schlaifer (1961, pp. 79-92) and explored in Cohen & Freeling (1981). There are three main components of the model: User options for final dispensation of an aid recomme ndation, user options for verifying an aid recommendation, and the outcomes of those choices.

Value of Information Applied to Verification Decisions

Suppose the decision aid has arrived at a conclusion or recommendation r, e.g., regarding the classification of a contact. $a_1(r)\ldots a_m(r)$ are a set of options from which the user will eventually make a choice for final dispensation of r (e.g., accept the recommendation, modify the recommendation in any of various ways, reject the recommendation and accept a known alternative). Note that the a_i refer to user interactions with the aid, such as "accepting," "rejecting," or "modifying"; the actual action corresponding to $a_i(r)$ is a function of r, as indicated by the notation, since, for example, accepting a recommendation to engage is quite different from accepting a recommendation not to engage. For purposes of this model, all actual actions that might be adopted by the user (even options not yet identified or generated) will be represented as "modifications" of the aid recommendation. Note also that r may include more than one aid recommendation, possibly ranked in order of confidence (e.g., a set of possible classifications of a contact, in order of probability), and that user options may include, for example, adopting the aid's second-ranked recommendation. In this section, we can suppress the reference to r for simplicity, since the aid's recommendation is given at the time of the verification decision, and designate the user's interaction options as $a_1\ldots a_m$.

In the verification decision the user faces a choice among options: v_0, representing the decision not to verify the recommendation, and $v_1,\ldots v_n$, representing the available options for making observations, requesting information, performing additional analyses, identifying or creating new options, etc. If users do not verify (v_0), they collect no further information (which we will represent by the dummy observation z_0), and they immediately choose among the $a_1\ldots a_m$ for the current aid recommendation(s) r. If users choose a verification option v_i, they will observe one member of the set, $z_{i,1}\ldots z_{i,p}$, of mutually exclusive and exhaustive possible observational outcomes of the chosen verification process v_i. For example, let v_i = looking at the size of the gun on a vehicle image. The outcomes of this observation might be: $z_{i,1}$= large gun, and $z_{i,2}$ = small gun. Each z_i is thus a different random variable, representing the possible outcomes of a different verification option (e.g., looking at a different feature of the image).

The $s_1\ldots s_p$ are an exhaustive and mutually exclusive set of states of the world that determine the success of whatever action a_k the user has selected (e.g., s specifics the true classification of the vehicle). A complete path through the decision tree consists of the user selecting a verification option, v_i , observing some outcome $z_{i,j}$, selecting a final action a_k, and experiencing an outcome determined by s_g. The user's preferences are represented by a utility or value function on such paths through the decision tree: $\mathbf{u}(v,z,a,s)$. A key premise of this analysis is that this utility function can be decomposed into two additive parts, $\mathbf{u}(v,z,a,s) = \mathbf{u}_v(v,z) + \mathbf{u}_f(a,s)$, corresponding to the cost of the verification process and the value of the terminal action, respectively (Raiffa & Schlaifer, 1961, pp. 79-81). The final utility function, $\mathbf{u}_f(a,s)$, is the utility

of the outcome produced by terminal action a under conditions s (for example, the value of accepting the recommended classification of a contact as an enemy tank and engaging the contact, in the situation where it is in fact a friendly APC). The verification utility function, $\mathbf{u}_v(v,z)$, reflects the cost in time and risk of performing the observations or analyses represented by verification process v. $\mathbf{u}_v(v,z)$ is independent of the actual outcome z of the process. Moreover, because $\mathbf{u}_v(v,z)$ will typically be a negative quantity, it is convenient to deal with cost instead of utility:

$$\mathbf{c}(v_i) = - \mathbf{u}_v(v_i, z_i)$$

The costs of different verification options can vary considerably. For example, the cost of not verifying, $\mathbf{c}(v_0)$, is zero. However, the cost in time and risk of making the observations required to identify subtle features of an image might be quite high.

A decision process can be represented by a decision tree, which contains chance nodes to represent events not under the control of the decision maker and decision nodes to represent choices that are under the decision maker's control. There is an expected utility $\mathbf{EU}$ associated with every node that the user can arrive at by a combination of choices and chance events. Expected utility at chance nodes is represented by an expectation (or probability-wighted averaging) operator $\mathbf{E}$. For example, the final expected utility of adopting verification process v_1, observing $z_{1,1}$, and then selecting action a_3 is:

$$\mathbf{EU}_f(v_1, z_{1,1}, a_3) = \mathbf{E}_{s \,|\, v1, z1, 1} \; \mathbf{u}_f(a_3, s)$$

$$= \dot{\mathbf{a}}_i \; \mathbf{p}(s_i \,|\, v_1, z_{1,1}) \; \mathbf{u}_f(a_3, s)$$

This is the probability-weighted average of the final utility function $\mathbf{u}_f(a_3, s)$, over possible states of the world s, given that the verification option v_1 will be selected, observation $z_{1,1}$ will be observed, and final action a_3 will be chosen (assuming that s is a discrete and finite variable). (For the moment we are ignoring the other part of the utility function, $\mathbf{u}_v(v_{1,1}, z_{1,1})$, representing costs.)

Expected Utility at decision nodes is represented by the $\mathbf{max}$ operator, assuming that the decision maker makes the choice that will maximize expected utility. For example, after performing verification process v_1 and observing $z_{1,1}$, the expected utility of the final decision, is:

$$\mathbf{EU}_f(v_1, z_{1,1}) = \mathbf{max}_a \, \mathbf{E}_{s \,|\, v1, z1, 1} \; \mathbf{u}_f(a, s)$$

This represents the expected utility of whatever action a has the largest final utility, $\mathbf{u}_f(a, s)$, averaged over possible values of s, given $z_{1,1}$ and v_1.

The final expected utility of a verification option, say v_1, is the average, over its possible observational outcomes z_1, of the expected value of the final action:

$$\mathbf{EU}_f(v_1) = \mathbf{E}_{z1 \,|\, v1} \, \mathbf{max}_a \, \mathbf{E}_{s \,|\, v1, z1} \; \mathbf{u}_f(a, s)$$

In the special case of the option not to verify, v_0, no information is actually collected; thus, the expected utility of v_0 is simply the result of selecting the action a that maximizes utility averaged over values of s:

$$\mathbf{EU}_f(v_0) = \mathbf{E}_{z0 \,|\, v0} \, \mathbf{max}_a \, \mathbf{E}_{s \,|\, v01, z0} \; \mathbf{u}_f(a, s)$$

$$= \mathbf{max}_a \, \mathbf{E}_s \; \mathbf{u}_f(a, s)$$

The value of information ($\mathbf{VOI_f}$) for any verification option, say v_i, is simply the difference between the final expected utility of v_i and the final expected utility of immediately selecting an action:

$$\mathbf{VOI_f}(v_i) = \mathbf{EU_f}(v_i) - \mathbf{EU}_f(v_0)$$

$$= \mathbf{E}_{zi \mid vi} \, \mathbf{max}_a \, \mathbf{E}_{s \mid vi,zi} \, \mathbf{u_f}(a,s) - \mathbf{max}_a \, \mathbf{E}_s \, \mathbf{u_f}(a,s)$$

Let us define a' as whatever action the decision maker would choose if unable to verify; in other words, a' is the member of $a_1 \dots a_m$ (e.g., acceptance, modification, or rejection of the aid conclusion) that maximizes average utility over the possible values of s *without* knowledge of z. Thus:

$$\mathbf{VOI_f}(v_i) = \mathbf{E}_{zi \mid vi} \, \mathbf{max}_a \, \mathbf{E}_{s \mid vi,zi} \, \mathbf{u_f}(a,s) - \mathbf{E}_s \, \mathbf{u_f}(a',s)$$

Value of information is the utility expected to be gained by waiting to observe the value of z rather than acting immediately on the currently favored action a'. We can represent $\mathbf{E}_s$ as a weighted average of its values over possible values of z:

$$\mathbf{VOI_f}(v_i) = \mathbf{E}_{zi \mid vi} \, \mathbf{max}_a \, \mathbf{E}_{s \mid vi,zi} \, \mathbf{u_f}(a,s) - \mathbf{E}_z \, \mathbf{E}_{s \mid z} \, \mathbf{u_f}(a',s)$$

We can drop v_1 if we assume that the verification process itself does not affect the probabilities of situations, i.e., the probability of a situation s given that z is true is the same whether or not z was in fact observed; thus:

$$\mathbf{VOI_f}(v_i) = \mathbf{E}_{zi} \, \mathbf{max}_a \, \mathbf{E}_{s \mid zi} \, \mathbf{u_f}(a,s) - \mathbf{E}_z \, \mathbf{E}_{s \mid z} \, \mathbf{u_f}(a',s)$$

$$= \mathbf{E}_{zi} \, [\, \mathbf{max}_a \, \mathbf{E}_{s \mid zi} \, \mathbf{u_f}(a,s) - \mathbf{E}_{s \mid z} \, \mathbf{u_f}(a',s) \,]$$

This formulation is of special interest, and is the source of the verbal statement in the text. It represents the average over the possible observations z, of the difference between the utility expected from the option that would be selected given knowledge of z and the utility expected from the currently preferred option. (Notice that verifying and not verifying differ simply in the sequence of expectation operators $\mathbf{E}$ and maximization operators $\mathbf{max}$. When an $\mathbf{E}$ operator precedes a $\mathbf{max}$, it means that the user will have resolved the uncertainty represented by $\mathbf{E}$ by the time the user makes the decision represented by $\mathbf{max}$. When $\mathbf{E}$ follows $\mathbf{max}$, it means the user will not have the information when making the decision.)

Thus far, we have neglected the cost $\mathbf{c}(v)$ of verification process v. Users choose a verification option by maximizing expected utility (or trust) with respect to the *total* utility function $\mathbf{u}(v,z,a,s)$. Given the additive decomposition of that function, the chosen option will be the v_i such that:

$$\mathbf{EU}(v_i) = \mathbf{max}_v \, \mathbf{E}_{z \mid v} \, \mathbf{max}_a \, \mathbf{E}_{s \mid v,z} \, \mathbf{u}(v,z,a,s)$$

$$= \mathbf{max}_v \, \mathbf{E}_{z \mid v} \, \mathbf{max}_a \, \mathbf{E}_{s \mid v,z} \, [\, \mathbf{u_f}(a,s) - \mathbf{c}(v) \,]$$

$$= \mathbf{max}_v \, [\, \mathbf{E}_{z \mid v} \, \mathbf{max}_a \, \mathbf{E}_{s \mid v,z} \, \mathbf{u_f}(a,s) - \mathbf{c}(v) \,]$$

In the special case of no verification, v_0, cost is zero, so total utility is equal to final utility:

$$\mathbf{EU}(v_0) = \mathbf{EU_f}(v_0) = \mathbf{max}_a \, \mathbf{E}_s \, \mathbf{u_f}(a,s)$$

Users select a verification option by maximizing $\mathbf{EU}(v_i)$.

Dynamic Constraints on Verification Decisions

Dynamic constraints on verification can be derived by starting with the definition of expected value of information ($\mathbf{VOI_f}$) for verification process v_i in the previous section:

$$\mathbf{VOI_f}(v_i) = \mathbf{E}_{zi} [\max_a \mathbf{E}_{s\,|\,zi}\, \mathbf{u_f}(a,s) - \mathbf{E}_{s\,|\,z}\, \mathbf{u_f}(a',s)]$$

As before, a' is the option a that maximizes $\mathbf{E}_s\, \mathbf{u_f}(a,s) = \mathbf{E}_z\, \mathbf{E}_{s\,|\,z}\, \mathbf{u_f}(a,s)$; hence, a' is preferred before verification. Since it does not include costs, the maximum that $\mathbf{VOI_f}$ can take is the expected value of *perfect* information ($\mathbf{VOPI_f}$), when the observation z provides full foreknowledge of the situation s. In that case, we can substitute s for z and simplify:

$$\mathbf{VOPI_f}(v_i) = \mathbf{E}_s [\max_a \mathbf{u_f}(a,s) - \mathbf{u_f}(a',s)]$$

We now define a generalization of $\mathbf{VOPI_f}$ that we will call the expected value of *partial (but) perfect* information ($\mathbf{VOPPI_f}$). In this case, the observation z may, but does not necessarily, provide full knowledge of which situation s is the case. At the least, however, z provides sufficient knowledge of which final action a should be adopted. In other words, z will tell the user to *correctly* accept a' or else *correctly* select from among the remaining a. We will define S as a partition of the situations s that is just fine enough to guide selection of an a; i.e., the same final action a_i is appropriate in all s with the value S_i. We designate S' as the set of all situations s in which a' remains (correctly) preferred after verification, and S" as the complementary set of situations S" = { S_i: $S_i \neq$ S'}, in which some other action $a_i \neq$ a' is (correctly) preferred. By assumption, z will provide perfect knowledge that the true situation s is either in S' or in some other S $\subseteq$ S". Since knowledge of S exhausts the possibilities for changing the final decision a, $\mathbf{VOPPI_f}$ is quantitatively the same as the value of perfect information ($\mathbf{VOPI_f}$). But $\mathbf{VOPPI_f}$ allows for the possibility that z will not discriminate further among the situations s, even though these further distinctions may affect utility, $\mathbf{u_f}(a,s)$. (For example, suppose a target identification aid recommends engagement of a contact. Acceptance of this recommendation is correct if the contact is in fact an enemy tank (S'), and incorrect if it is a friendly truck *or* an enemy truck (S"). The distinction between a friendly truck (s_1) and enemy truck (s_2) is irrelevant for the engagement decision, and the aid may not discriminate them; however, the distinction has an enormous impact on the cost of a mistaken engagement.) Assuming that verification v_i provides perfect information regarding a, we can substitute S for z in the first equation above:

$$\mathbf{VOPPI_f}(v_i) = \mathbf{E}_S [\max_a \mathbf{E}_{s\,|\,S}\, \mathbf{u_f}(a,s) - \mathbf{E}_{s\,|\,S}\, \mathbf{u_f}(a',s)]$$

$$= \mathbf{p}(S') [\max_a \mathbf{E}_{s\,|\,S'}\, \mathbf{u_f}(a,s) - \mathbf{E}_{s\,|\,S'}\, \mathbf{u_f}(a',s)] + \mathbf{p}(S'') \mathbf{E}_{S\,|\,S''} [\max_a \mathbf{E}_{s\,|\,S''}\, \mathbf{u_f}(a,s) - \mathbf{E}_{s\,|\,S''}\, \mathbf{u_f}(a',s)]$$

When S' is the case, a' is the preferred option, and $\max_a \mathbf{E}_{s\,|\,S'}\, \mathbf{u_f}(a,s) = \mathbf{E}_{s\,|\,S'}\, \mathbf{u_f}(a',s)$. So, the first term on the right-hand side in the above equation is zero, and the value of verification focuses on S", which is the complement of S':

$$\mathbf{VOPPI_f}(v_i) = \mathbf{p}(S'') \mathbf{E}_{S\,|\,S''} [\max_a \mathbf{E}_{s\,|\,S''}\, \mathbf{u_f}(a,s) - \mathbf{E}_{s\,|\,S''}\, \mathbf{u_f}(a',s)]$$

Verification process v_i cannot be inappropriate if $\mathbf{VOPPI_f}(v_i) < \mathbf{c}(v_i)$, i.e., if the best that verification can accomplish is less than its cost. It follows algebraically that verification process v_i is inappropriate, if:

$$\mathbf{p}(S'') < \mathbf{c}(v_i) / \mathbf{E}_{S\,|\,S''} [\max_a \mathbf{E}_{s\,|\,S''}\, \mathbf{u_f}(a,s) - \mathbf{E}_{s\,|\,S''}\, \mathbf{u_f}(a',s)]$$

or equivalently, since $\mathbf{p}(S'') = 1 - \mathbf{p}(S')$,

$$p(S') > 1 - c(v_i) / E_{S|S''} [\max_a E_{s|S''} u_f(a,s) - E_{s|S''} u_f(a',s)]$$

$p(S')$ is trust in the currently preferred option a', i.e., it is the chance of a situation in which a' is successful. Thus, $p(S')$ is trust in the overall user-aid interaction (which resulted in a preference for a'). The inequality above shows how such trust constrains the appropriateness of verification: If trust is high enough so that the constraint is satisfied for all verification options v_i, the currently preferred action should be accepted without verification.

If the aid's conclusion is binary (e.g., identification as appropriate target or not-appropriate target), we can derive the somewhat simpler constraints stated in the text. In that case, there are only two final options: a' and a''. One of these must be acceptance of the aid recommendation, and the other rejection. Moreover, there are only two values of S, viz., S', implying that a' is appropriate, and S'', implying that a'' is appropriate. Finally, there is only one basic verification strategy, which involves seeking evidence bearing on both a' and a''. Different verification options v_i may still differ in the specific sequence of observations or information requests; we shall refer to the verification sequence with the highest expected utility as v_1. We can now simplify the inequalities above:

$$p(S'') < c(v_1) / E_{s|S''} [u_f(a'',s) - u_f(a',s)] \text{ , and, equivalently,}$$

$$p(S') > 1 - c(v_1) / E_{s|S''} [u_f(a'',s) - u_f(a',s)]$$

If a' (the option that is preferred prior to verification) happens to be *acceptance* of the aid's conclusion, then $p(S')$ is equivalent to *trust in the decision aid alone*. Then the second inequality is the first constraint in the text. If this constraint is satisfied, the aid's conclusion should be accepted (and its negation rejected) without verification. On the other hand, if the currently favored option a' happens to be *rejection* of the aid's conclusion, then $p(S'') = 1 - p(S')$ is trust in the aid, and the first inequality above gives us the second constraint in the text. If this constraint is satisfied, the aid's conclusion should be rejected (and its negation accepted) without verification. For binary conclusions, we can combine the two constraints. We will fix a' as acceptance and a'' as rejection of the aid's recommendation, making necessary switches in the first inequality. Then, verification is permissible only if:

$$1 - c(v_1) / E_{s|S''} [u_f(a'',s) - E_{s|S''} u_f(a',s)] > p(S_1) > c(v_1) / E_{s|S'} [u_f(a',s) - E_{s|S'} u_f(a'',s)].$$

This is the source of the upper and lower bounds in Figure 30. Note that the costs of different kinds of errors appear in the denominators of each constraint (viz., the cost of performing a' when a'' would be appropriate in the denominator on the left, and vice versa on the right), and that these costs depend on the distribution of s within S_1 and S_1. Thus, for example, the relative proportion of friendly vehicles and enemy trucks among non-targets will influence the threshold for acceptance of an identification friend-or-foe conclusion. As the ratio of friendly vehicles increases, the amount of trust required for engagement increases.

We can find the conditions at which the upper and lower bound meet, eliminating the possibility of further verification. First, for brevity, let:

$$\text{cost of incorrectly rejecting aid's conclusion} = E_{s|S'} [u_f(a',s) - E_{s|S'} u_f(a'',s)] = c_r$$

$$\text{cost of incorrectly accepting aid's conclusion} = [E_{s|S''} [u_f(a'',s) - E_{s|S''} u_f(a',s)] = c_a.$$

When the two constraints are equal, we have:

$$1 - c(v) / c_a = c(v) / c_r ,$$

where $c(v)$, as usual, is the cost of delay or risk associated with verification. It follows that verification is no longer appropriate when:

$$c(v) = c_r\, c_a\, /\, (c_r + c_a)$$

At this value of cost, the two constraints will have converged on the following value:

$$1 - c(v)\, /\, c_a = c(v)\, /\, c_r = \quad c_a\, /\, (c_a + c_r)$$

Not surprisingly, this is the trust criterion that would determine which action should be taken if verification were not possible. In other words, the expected utility of a' is greater than the expected utility of a" if and only if:

$$p(S') > \ c_a\, /\, (c_a + c_r)$$

APPENDIX E: INSTRUCTIONS FOR UNCERTAINTY EXPERIMENT

Overview

In this experimental task you will view FLIR images of vehicles at various ranges and decide whether or not to engage them.

The vehicles will appear on a 4 by 3 grid, in which vehicles in the lowest row are closest to you. (Note the range figures on the left of the grid.) The grid will appear as follows:

```
  000

          +--------+--------+--------+--------+
          |        |        |        |        |
  500     |        |        |        |        |
          |        |        |        |        |
          +--------+--------+--------+--------+
          |        |        |        |        |
  000     |        |        |        |        |
          |        |        |        |        |
          +--------+--------+--------+--------+
          |        |        |        |        |
  500     |        |        |        |        |
          |        |        |        |        |
          +--------+--------+--------+--------+
```

You will not be able to see the image of a vehicle until you select the cell in which it appears. You can select cells and examine the vehicles in any order you choose.

The 105-minute experimental task will be broken into four different missions. On each mission, your job is to engage enemy tanks and to avoid engaging other vehicles, especially friendlies. However, the missions vary in two ways:

Some missions will be deep interdiction, some will be close air support. In deep interdiction missions, there will be more enemy vehicles relative to friendly vehicles than in the close air support missions. For each mission, intel will provide estimates of the number of vehicles of various types you may encounter

The missions will also vary in how much time you can spend examining potential targets for engagement. A clock on screen will indicate how much time you have left to explore a grid.

You will receive new mission instructions at the start of each mission.

Examining and engaging targets

1. Select a cell in the grid by pressing the corresponding function key as shown below. The cell you select will be highlighted.

000

F9	F10	F11	F12

500

F5	F6	F7	F8

000

F1	F2	F3	F4

500

2. After selecting a cell, depress both shift keys and hold them down. When you do this, the vehicle image in the cell will appear.

3. Release the *right* shift key to engage this target. Release the *leFt* shift key to take no action.

4. The grid will reappear. If you engaged the vehicle (by lifting the right shift key), an "X" will appear in its cell, and you will not be able to view it again. If you took no action (by lifting the left shift key), an "O" will appear in its cell. You may view this vehicle again, if you wish.

Grid color coding

An automated target recognition (ATR) system colors the cells in the grid to help you explore the vehicles.

Red — If the ATR is at least 90% certain that the vehicle in the cell is an enemy tank, the cell is colored red.

Blue — If the ATR is at least 90% certain that the vehicle in the cell is a friendly armored vehicle, the cell is colored blue.

Yellow — If the ATR doesn't know, but the vehicle in the cell could be either an enemy tank or friendly armor, it colors the cell yellow.

Grey — All other cells are colored gray. (For example, vehicles classified by the ATR as friendly and enemy jeeps and trucks, as well as enemy APC's).

Remember, the ATR is not perfect. The colors are based on 90% confidence, and will be wrong about 10% of the time.

Cell labels (Rules 1 & 3)

In each cell, an automated target recognition (ATR) system will display its best assessment regarding the vehicle in that cell.

When you select an image for viewing, the ATR will also display its degree of confidence in this assessment.

For example, "Tank 90%" means that the ATR is correct 90% of the time in identifying tanks under the prevailing conditions, and that it thinks the present image is a tank.

Cell labels (Rule 2)

In each cell, an automated target recognition (ATR) system will display its best assessment regarding the vehicle in that cell.

When you select an image for viewing, the ATR will also display its degree of confidence in this assessment.

For example, "Tank 90%" means that the ATR is correct 90% of the time in identifying tanks under the prevailing conditions, and that it thinks the present image is a tank.

The ATR will not provide an assessment unless it is 90% confident.

For example, if it is only 60% confident that a vehicle is a tank, but 90% confident it is armored, it will display: "Armored 90%".

Mission instructions for deep interdiction

Your mission is to engage enemy tanks and to avoid engaging other vehicles, especially friendlies.

This is a deep interdiction mission.

You have enough munitions to engage all enemy tanks on this mission, plus some extra rounds.

You will view several grids to complete this mission. You will have the same amount of time to complete each grid.

These are the types of vehicles that intelligence has identified in the immediate area where you will be flying:

Vehicle		Count
Enemy tank	T62	188
Enemy tank	T55	188
Enemy APC	BMP	75
Enemy APC	BRD	75
Enemy APC	BTR	75
Enemy Jeep	UAZ469	30
Enemy truck	ZIL	23
Enemy truck	KRAZ	23
Friendly tank	M60	23
Friendly tank	M1A	23
Friendly tank	M551	15
Friendly APC	M113	8
Friendly Jeep	M151	3
Friendly truck	M35	4

When you are ready to begin, please tell the experimenter.

Mission instructions for close air support

Your mission is to engage enemy tanks and to avoid engaging other vehicles, especially friendlies.

This is a close air support mission.

You have enough munitions to engage all enemy tanks on this mission, plus some extra rounds.

You will view several grids to complete this mission. You will have the same amount of time to complete each grid.

These are the types of vehicles that intelligence has identified in the immediate area where you will be flying:

Vehicle		Count
Enemy tank	T62	150
Enemy tank	T55	150
Enemy APC	BMP	45
Enemy APC	BRD	45
Enemy APC	BTR	60
Enemy Jeep	UAZ469	15
Enemy truck	ZIL	15
Enemy truck	KRAZ	15
Friendly tank	M60	75
Friendly tank	M1A	75
Friendly tank	M551	38
Friendly APC	M113	38
Friendly Jeep	M151	16
Friendly truck	M35	16

When you are ready to begin, please tell the experimenter.