ARMY RESEARCH LABORATORY

Soldier-in-the-Loop Target Acquisition Performance Prediction Through 2001: Integration of Perceptual and Cognitive Models

by Barry D. Vaughan

ARL-TR-3833 **July 2006**

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer’s or trade names does not constitute an official endorsement or approval of the use thereof.

DESTRUCTION NOTICE—Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5425

ARL-TR-3833 **July 2006**

Soldier-in-the-Loop Target Acquisition Performance Prediction Through 2001: Integration of Perceptual and Cognitive Models

Barry D. Vaughan
Human Research and Engineering Directorate, ARL

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
July 2006		

4. TITLE AND SUBTITLE

Soldier-in-the-Loop Target Acquisition Performance Prediction Through 2001: Integration of Perceptual and Cognitive Models

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Barry D. Vaughan (ARL)

5d. PROJECT NUMBER

62716AH70

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

U.S. Army Research Laboratory
Human Research and Engineering Directorate
Aberdeen Proving Ground, MD 21005-5425

8. PERFORMING ORGANIZATION REPORT NUMBER

ARL-TR-3833

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Modeling Soldier-in-the-loop target acquisition performance is necessary for the development of improved sensors, more effective training methods, and better war game simulations. Accurately modeling this performance requires a detailed understanding of the environment, how a sensor responds to the environment, how it displays information to an observer, and how the observer employs that information to acquire a target. The first three requirements have been met; the fourth requirement, however, has not yet been achieved. Attempts to model the observer's visual and decision-making processes have been compromised by the analysis of the scene, based on physical parameters alone rather than how the visual system interprets the scene. Models based on such scene-derived factors have had limited success.

This report takes a two-pronged approach to how future models can be improved by the sensible integration of human visual processing. One prong concerns basic research from the perceptual psychology community. Over the last few decades, this research has generated a detailed theoretical understanding of visual processing and decision making, based on visual information. The other prong concerns important models, modeling frameworks, and scene metrics from the military target acquisition community. Particular attention is paid to issues of clutter, the extendibility of the Johnson criteria, classical and neoclassical search frameworks, the selection of methods and performance metrics, and existing Night Vision and Electronic Sensors Directorate models. Issues related to the validation of target acquisition models are also discussed.

(abstract continued on next page)

15. SUBJECT TERMS

clutter; conspicuity; target acquisition; target acquisition model; visual search

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Barry D. Vaughan
Unclassified	Unclassified	Unclassified	SAR	135	19b. TELEPHONE NUMBER (*Include area code*) 410-278-3324

Abstract (continued)

Existing target acquisition models tend to base performance on (a) one-dimensional (1-D) metrics defining the amount of information in the target (e.g., resolvable bar cycles, contrast, area, size, perimeter, speed of motion) and how that information correlates to level of performance in a target acquisition task (i.e., detection, classification, recognition, and identification), (b) search processes that are unrealistic (e.g., that assume random eye movements), and (c) 1-D metrics to define the whole scene (clutter) or regions of the scene (e.g., clutter, conspicuity, attractiveness). These tendencies fail to account for known human behavior, although models incorporating them may be insensitive to the details of human performance because they predict ensemble rather than individual performance.

Phenomena from perceptual psychology known to affect target acquisition are reviewed in terms of how target acquisition models do and do not account for them. Such factors include motion, color, and visual transients. Basic models of visual search are included as guides for how target acquisition models may incorporate some of these factors.

Visual selective attention is recommended as a means for the theoretically meaningful inclusion of psychologically important factors into target acquisition modeling.

INTENTIONALLY LEFT BLANK

Contents

List of Figures

List of Tables

INTENTIONALLY LEFT BLANK

1. Purpose, Objectives, and Scope

This technical report is part of a technology program annex (TPA) with the U.S. Army Materiel Systems Analysis Activity (AMSAA) that defines its purpose and objective and outlines particular topics of interest as follow.

1.1 Purpose

This TPA defines the proposed responsibility of the U.S. Army Research Laboratory's (ARL) Human Research Engineering Directorate in support of the AMSAA to perform human response-based activities that will provide improved search and target acquisition analysis tools, techniques, and methodologies.

1.2 Objectives

ARL proposes to establish a methodology development program that emphasizes the description and definition of the human processes of search and acquisition of military targets in realistic backgrounds and the relationship between them.

1.3 Scope

1.3.1 Topics of Interest

This review will survey relevant research in target acquisition and highlight the state of the art in modeling particular aspects of performance including those of (a) the target: target type, number, signature variation, cues (e.g., glint, muzzle flash), and representation, (b) the target-acquisition environment: effects of background and foreground, local and global environmental variation, type of environment (e.g., tropical, jungle, desert), day versus night viewing, and clutter, (c) sensor parameters: field of view (FOV), resolution, and stereoscopic versus non-stereoscopic, and (d) type of search: FOV, field of regard (FOR), time required to search, detect, recognize, and identify targets. Additional topics of particular interest are as follow:

Particular attention will be paid to the Johnson criteria, and to the ACQUIRE and Night Vision and Electronic Sensors Directorate (NVESD[1]) models since they or portions of them are used by AMSAA in current simulation efforts (e.g., Mazz, 1998). These models also serve as the basis for ongoing attempts to integrate additional scene and observer parameters such as motion (e.g., Meitzler, Kistner et al., 1998), multiple observers (Rotman, 1989), scene obscurants (Rotman, Gordan, & Kowalczyk, 1989), clutter (Tidhar et al., 1994), and multiple targets (Rotman, Gordan, & Kowalczyk, 1989) and selective visual attention[2]. As such, it is important to know the limitations and theoretical extensibility of the models.

[1]NVESD is part of the U.S. Army Research, Development, and Engineering Command's Communications and Electronics Research, Development, and Engineering Center.

[2]The author of this report is involved in ongoing research into the role of selective visual attention in target acquisition. One goal of the research is to determine if ACQUIRE's performance can be improved by the inclusion of attention parameters. ACQUIRE is not an acronym.

1.3.2 Perceptual Psychology and How It Can Inform Target Acquisition Modeling

The greatest theoretical advances to understanding visual search processes have occurred in the reductionistic environments of academic perception laboratories. The resulting models and theories may be of limited direct applicability to military target acquisition scenarios. However, they constrain models and inform the reader about known visual phenomena relevant to target acquisition. Current models from the perceptual literature are discussed in terms of their generalizability to the battlefield.

1.3.3 How Performance is Measured

Different models use different measures as predictors of performance (e.g., response time, observer sensitivity [d′], false detection percentage, probability of detection, etc.). Models may not be directly comparable in that the dependent measures (a) do not necessarily map onto each other in a well-defined way, and (b) may not exchange predictably as observer and scene parameters change. To the extent possible, models are discussed in terms of how these various dependent measures may be differentially affected by parameter changes.

1.3.4 Issues Related to the Validation and Testing of Models

The author of this report made no attempt to instantiate the models in software or hardware in order to evaluate them head to head. There is a brief discussion of issues related to the validation of models and the need for a robust data set to perform laboratory studies of models before field trials.

The scope of the review includes non-classified literature from the defense and the academic communities that relate to the acquisition of ground targets. Although target acquisition models date back several decades (see Greening, 1974, for a review of early efforts), this review focuses on identifying the state of the art in modeling and discusses only classic models that have broken new ground and are still of theoretical interest. Models from the perceptual psychology literature are also discussed for their role in promulgating new theoretical ideas that may or may not be generalizable to real-world target acquisition.

2. Introduction

Before target acquisition models can be discussed, it is important to define terms that appear throughout this report. Bliss pointed out in 1974 that no clear standards existed for what is specifically meant by the term “target acquisition.” Since then, models and theories of how targets can be acquired from various disciplines (e.g., machine vision, perceptual psychology, military simulation, electro-optical design) have proliferated. However, there remains an absence of standards for basic terms.

In 1990, the Quadripartite Working Group on Army Operational Research proposed standard definitions that are used in this report when we discuss target acquisition models. Some definitions from that working group are

- Target Acquisition

All those processes required to locate a target image whose position may be uncertain and to discriminate it to the desired level (detection, classification, recognition, identification). The target acquisition process includes the search process at the end of which the target is located and the discrimination process at the end of which the target is acquired. This definition assumes that a time-dependent search process is involved. However, target acquisition may involve the discrimination of a target whose position is known ahead of time. Such a static process is assumed to be the same as the discrimination stage of the above-defined target acquisition process.

- Search

The process of visually sampling the search field in an effort to locate or acquire targets.

- Discrimination

A process in which an object is assigned to a subset of a larger set of objects, based on the amount of detail perceived by the observer, and the application of knowledge of those details sufficient to afford such an action.

- Detection

The perception of an object image (which may be a target image) as being present at a particular location and distinct from its surroundings.

- Classification

The determination of whether a detected object is a member of a particular set of possible targets or non-targets (e.g., wheeled versus tracked vehicles).

- Recognition

The determination that a target belongs to a particular functional category (e.g., a tank, a truck, an armored personnel carrier, etc.).

- Identification

The most detailed level of discrimination of particular relevance for military target acquisition, as discussed shortly (e.g., a T-72, T-62, M1, or M60 tank).

Inherent in these definitions of the processes involved in target acquisition are the ideas that first, information must be extracted from the scene and second, that the Soldier in the loop must be able to use such information to make an appropriate decision. (In some cases, the decision made

must be that the information in the scene is insufficient even for detection. In such cases, the decision made by the Soldier is the declaration that no target is present.) In addition to information-related constraints, the Soldier must have both the perceptual capability to perceive and the cognitive ability to understand the information in order to employ it. Although this fact may seem obvious, modeling the observer's decision-making process is no simple feat.

The goal of this technical report is to provide an overview of the literature relevant to the modeling of the human in the loop in target acquisition. Figure 1 highlights the flow of information in the target acquisition process from the visual information in the scene through any optical or electro-optical sensor systems to the human visual system and finally, to the observer's decision-making processes. This report highlights the difficulties associated with target acquisition, which arise from each of these levels, with particular emphasis on the last three elements in which the human observer is given a scene, either optically or electro-optically, from which he[3] attempts to extract information and acquire a target.

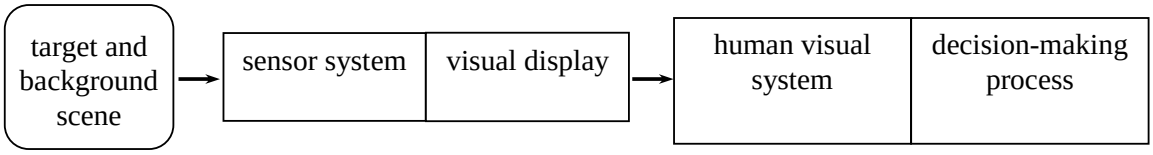


Figure 1. The flow of information in human-in-the-loop target acquisition.

Before we detail the complexity associated with the elements of the human-in-the-loop target acquisition process and how they influence modeling the target acquisition process, it is useful to briefly say why the human is in the target acquisition loop to begin with. Although research into automatic target recognition (ATR) and aided target recognition proceeds at a rapid pace, current ATR systems lack sufficient accuracy and flexibility to allow them to take over the process of target acquisition from humans (e.g., Dudgeon, 1998). The deficiencies of ATR become particularly apparent when they are called upon (a) to perform acquisition tasks when the space of possible targets is large, and (b) when non-visual factors such as situational context, experience, and judgment must be taken into account before an action is taken regarding a potential target. Therefore, the human observer must be available to make the final decision regarding action (or inaction) in the target acquisition situation.

Given that the human remains firmly in the loop for the foreseeable future, as the decision maker and as the actual acquirer of potential targets, it is imperative to understand the factors known to have an impact on Soldier performance in real-world target acquisition performance. Table 1 lists several such factors, broken into their effects on the visual display of the scene in which target acquisition is to be performed, and their effects on the decision-making process of the observer (from Howe, 1993).

[3]The male gender pronoun "he" is used throughout this technical report in order to facilitate readability.

Table 1. Factors known to affect performance of human in the loop.

Locus of Effect of Factors	Factors
Visual display of scene	Target type, size, shape, contrast with immediate background, motion, shadow, masking by background elements, camouflage, scene clutter, transient cues. Environmental visibility, cloud cover, sun angle, diurnal and seasonal variation, atmospheric scattering, illumination level, field of view.
Decision making of observer	Training, motivation, experience, expectations for possible targets, stress, concurrent task load, visual acuity, search pattern, fatigue, field of regard, attentional set.

In addition to factors in table 1, there are factors that depend on the sensor system being used. For instance, although table 1 may suffice to encompass factors relevant to an observer viewing a scene with the unaided eye, additional factors such as display resolution, phosphor decay rates, sensor temporal and spatial resolution, atmospheric turbulence and scattering, and target emittance and temperature must be added in order to account for performance variability when one is viewing FLIR (forward-looking infrared radar) imagery. Various models may take such factors into account (or fail to do so at their peril) when we are attempting to predict Soldier performance with various electro-optical devices.

Because no single model can possibly include all factors known to influence target acquisition performance, models will account for some of the factors and ignore others for theoretical reasons. (Such an approach, this reviewer would argue, is the only likely way these factors will ever be understood with the depth necessary to model them.)

The observer factors listed in table 1 may become especially acute, given the increasing demands placed on the individual Soldier by technology. Soldiers are called upon to use ever-more sophisticated sensor systems and will therefore be forced to deal effectively with an ever-increasing amount of information about the scene. In addition to the increasing cognitive and sensory demands placed on the Soldier by technology, potential enemies also use improvements in camouflage, concealment, and deception (CCD) technology to better hide themselves. Therefore, it seems obvious that any understanding of the human in the loop must account for observer variables and how they interact with factors influencing the display of visual information to the observer.

2.1 The Goals of Target Acquisition Modeling

There are several reasons why it is desirable to predict target acquisition performance. These reasons include

1. Better Soldier training

Training is costly and time consuming. Learning why Soldiers perform as they do and understanding the influences that experience, knowledge, and expectations have on acquisition performance may allow for better and more efficient training of Soldiers. For example, if a particular kind of terrain is known to cause problems in tank identification, then training may focus on providing more experience with the particular target-terrain interaction.

2. Reduced fratricide

Current weapon systems are accurate and lethal at ranges that often far exceed the identification range of the Soldier controlling the weapon. Misidentification may therefore lead to missing an enemy or firing on a comrade. Understanding when and why such misidentifications occur may inform the development of better sensor systems or training in order to reduce those errors.

3. Improved sensor systems

Sensor systems that provide the image to the Soldier in the loop cannot be evaluated properly unless we know what aspects of the sensor display (i.e., the rendered scene) have an impact on Soldier performance. Also, a functional model of the human in the loop will allow for sensors to be evaluated before production, thus reducing costs while increasing Soldier effectiveness.

4. More effective CCD techniques

The flip side of knowing the circumstances in which particular targets will be difficult to acquire will allow the Army to take advantage of those situations in order to make detection of our own forces more difficult.

2.2 Approach of the Author and Format of Review

Models of theoretical or historical importance are included to paint a relatively complete picture of the current state of target acquisition modeling with respect to the domain specified in the TPA. Major models are classified along a set of five dimensions (described next) and discussed. Theoretical details of the models are discussed in terms of the aspects of the scenes and observer variables accounted for, dependent measures predicted, and possible theoretical and empirical shortcomings. As mentioned previously, the author did not attempt to instantiate any of the models for a direct comparison. Rather, the literature reviewed in this report is described and critiqued in terms of agreement with empirical findings[4] and with theoretical understanding of human visual processing.

3. Model Description Scheme

A five-dimensional descriptive framework is outlined. The inherent strengths and weaknesses of models at various points of the dimensions are discussed. All models reviewed in detail are given scores along the dimensions.

[4]Since no experiments were done by the author, the empirical tests of most models come from the respective authors themselves or from third parties who instantiated and tested the models directly.

In order to make sense of the wide array of literature about target acquisition performance models, it is useful to rate each model, based on dimensions describing aspects of its function and the domain over which it may be used. Five dimensions were selected[5], based on Greening (1974):

1. optical/objective …… cognitive/subjective

This dimension refers to the locus of the observer's information processing. That is, does the observer make his decision on the basis of the visual information in the scene or on his subjective interpretation of what he perceives the visual percept to be? Modeling the former is straightforward in that all the information used to make the decision is readily available to the modeler. Modeling the latter is more problematic because inferences must be made about the cognitive processing that the observer performs to reach a decision.

2. reductive …… comprehensive

This dimension expresses the possible extremes of approach in terms of how much of the target acquisition process is to be accounted for by the model. (This dimension correlates highly with the generalizability of the model.) Reductive models are easy to support or disprove since they make testable predictions. However, such models lack sufficient detail to extend their predictions to real-world situations. Comprehensive models take many factors into account but may suffer from a combinatorial explosion of possible interactions and may be difficult to verify; tests of such models may lack sufficient statistical power to tease apart the effects of one or another factor.

3. target-centered …… situation-centered

This dimension expresses the range of information given in the scene that the subject can use to aid in acquiring the target. For example, a purely target-centered scene may contain a tank parked on a uniform texture field (i.e., no information in the scene guides searches for the target except the target itself). At the other extreme is a scene containing mountainous terrain and a number of roads upon which a target must travel. In this case, the roads guide the search for the target to such an extent that the target may become immediately apparent. Purely target-centered models exist primarily in studies of perceptual psychology and psychophysics or as a means of testing specific predictions about factors affecting performance. Situation-centered models, on the other hand, are more realistic but must make more assumptions about the cognitive processes underlying acquisition performance.

4. physiological …… empirical

This dimension refers to the degree to which the model is based on human visual physiology or on curve fits to previously collected empirical data. Between the two extremes lie models that base their performance predictions on known human psychophysics. Such psychophysical

[5]Note that no attempt was made to demonstrate the orthogonality of these dimensions.

models may rely either on psychometric functions (i.e., be more empirical) or on the physiology that underlies the psychometric functions (i.e., be more physiological). Models that are more physiological have the potential of being applicable to a greater variety of situations, although the models typically have more parameters to "tweak" to make them work, and the values of those parameters may not have strong theoretical underpinnings.

5. individual …… ensemble

This dimension refers to whether the model attempts to (or is able to) predict performance for an individual observer or an ensemble of observers. Although this dimension may at first glance appear to be a simple dichotomy, the breakdown is not so clear. For example, it would be a simple matter for an individual performance-based model to predict ensemble performance by processing groups of individuals, but it may or may not be possible for an ensemble-based model to step down to performance prediction at the level of the individual. The implications of this asymmetry come into play in terms of the inclusion of observer variables in that ensemble models typically assume the presence of "trained military observers" (e.g., O'Kane, 1995) and allow little theoretical room for the addition of individual factors.

4. Basic Types of Models

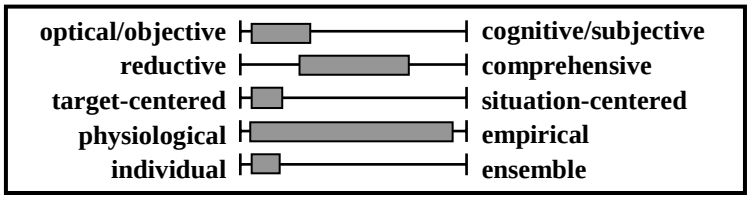
Although literally hundreds of models have emerged over the years, the bulk of the models reviewed in this report fall into a few basic classes. These classes are discussed.

This review of the literature divides the space of existing models into four broad types, as determined by the underlying processes that the model assumes drive performance. The classes are

1. Models based on physiology and empirical human psychophysics,
2. Models based on non-physiological feature extraction,
3. Models based on theoretical constructs and scene descriptions/metrics, and
4. Models based on largely atheoretical fits to empirical data.

We mention where each type of model lies along the five dimensions listed. Examples of such models are given, and the strengths and limitations of such models are discussed. It will be clear that there are models that do not fit neatly into one type but contain characteristics of several types. In such cases, the classification is based on the information purported to be used by the observer to make a decision.

1. Models based on physiology and empirical human psychophysics:



These models base their performance predictions on how the human visual system is known to respond to simple stimuli. That is, the models take what is known about vision from physiological studies of the visual system (e.g., Hubel & Wiesel, 1962, 1968; Campbell & Robson, 1968) and psychophysical studies of how physical stimuli determine overt perception and performance (e.g., Nachmias, 1981) and apply this knowledge to the acquisition of targets in the real world.

This category is the broadest in this report, largely because of the theoretical distance between physiology on one hand and psychophysics on the other. The reason why they have been grouped together is that both attempt to extend knowledge of how the visual system responds to simple stimuli (as determined by studies of visual physiology of psychophysics) and to militarily relevant stimuli. Also, physiological models are constrained in that they must conform to known psychophysics, so although two models within this category may process the visual information within a scene very differently (one by analyzing it with physiologically based filters and transforms; the other by appealing to psychometric functions), their result may be identical.

There are numerous examples of this type of model (e.g., British Aerospace ORACLE[6] model, Georgia Tech Vision [GTV], Wilson's Spatial Vision model, and the cortex transform-based distortion metric). These models tend to be some of the most complex of all target acquisition models because their bases in physiology and psychophysics allow them to incorporate many factors known to influence human perception so long as the effects of the factors are adequately understood.

There are also models in this class that base perception on the interpretation of the output of physiological mechanisms. Models of this kind treat the pieces of interpreted information as "features" or components of objects and background elements in the scene. Typically, these models are geared toward a basic understanding of the visual system and do not constitute full-scale models of target acquisition. Examples of these models include MIRAGE[7] (Watt & Morgan, 1985), MIDAAS[8] (Kingdom & Moulden, 1992), and various vision models by

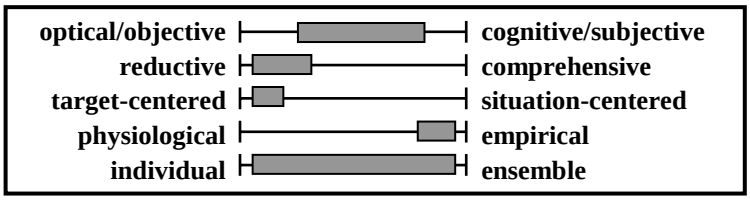
[6]ORACLE is not an acronym.

[7]The acronym MIRAGE is nothing short of a description in and of itself: "Multiple Independent filters of various sizes and with both signs, half-wave Rectified before Averaging. The resultant signals are Gated between adjacent zeroes for the Extraction of the primitive code."

[8]MIDAAS stands for Multiple Independent Descriptions Averaged Across Scale

Grossberg and colleagues (e.g., Grossberg, 1997; Grossberg, Mingolla, & Ross, 1994). These feature-based models are quite distinct from the second category of models.

2. Models based on non-physiological feature extraction:



These models base their predictions on the extraction of specific features from a scene rather than on an observer's ability to extract simple visual information. As was the case for physiology-based feature-extraction models in the previous category, the extracted features are assumed more likely to be properties of the visual signatures of military targets than of non-target elements in the scene. However, unlike the previous class, the selection of the features themselves in these models is not based on how the human visual system is known to function. Instead of appealing to simple physical stimuli such as oriented line segments (the output of early cortical visual processing [see Hubel, 1988, for an excellent review of this early work]) as the features of interest, these models assume that visual processing depends on more complex representations not having a direct correspondence to early visual processing.

Examples of such models include the edge-based 2½-dimensional representation (Marr, 1982; Marr & Hildreth, 1980), recognition by components (RBC) theory (Biederman, 1987), object symmetry (Rosenfeld, Wolfson, & Yeshurun, 1995), Guided Search models (Wolfe, 1994b; Wolfe & Gancarz, 1996), search by recursive rejection (SERR) (Humphreys & Muller, 1993), texture-based search (Nothdurft, 1991), and Feature Integration Theory (Treisman & Gelade, 1980; Treisman & Sato, 1990).

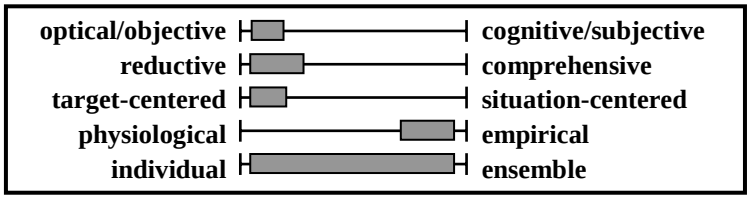
It is interesting to note that this class of models contains the greatest preponderance of thinking from perceptual psychology. The reason is that perceptual psychology has traditionally attempted to speak of the visual world in terms of objects (e.g., Duncan, 1984), groups (Vecera & Farah, 1994), surfaces (Nakayama & He, 1994), and features based loosely on visual physiology such as T- and L-junctions (Biederman, 1987), color (Theeuwes, 1995), etc. Much progress has been made in understanding human visual search by the use of this reductionistic technique, and some of the most theoretically sophisticated information-processing models of vision are based on such a breakdown of the scene.

The strength of the non-physiological feature approach is that the models have good agreement with human performance in the laboratory setting. The models can also more readily use the information required for discrimination judgments because they ostensibly concern the features that the visual system employs to form such judgments and because the models arise from the

perceptual psychology community where models of judgment and decision making are well developed.

The primary limitation of these models is obvious. Because they were developed in the laboratory where stimuli are reduced to their presumably most basic forms, there is little evidence that most models can be applied at all to visual processing of real-world stimuli. The primary reason for the lack of generalizability is that the real world cannot simply be reduced to a set of basic stimuli. (If it can, nobody has yet figured out what they are!) Some attempts to try to bridge the gap between the lab and the field have been made with limited success (e.g., Wolfe, 1994a).

3. Models based on theoretical constructs and scene descriptions:



These models also base their predictions of performance on the presence within the scene of information of a particular type. In these models, however, the information does not take the form of specific features or combinations of features but rather, a less theoretical form. Generally speaking, the more such information is present at the target location, the greater the probability or possible level of acquisition. The constructs used by the models are typically one-dimensional metrics such as conspicuity (e.g., Toet, 1996), number of resolvable cycles, N, of a bar pattern (i.e., a square wave) on a target (Johnson, 1958), or complexity (e.g., Tidhar et al., 1994). Such metrics may apply to the location of the target only or they may apply to the entire scene. For example, unidimensional clutter metrics can be global (relating to the entire scene) or local (relating only to a small region).

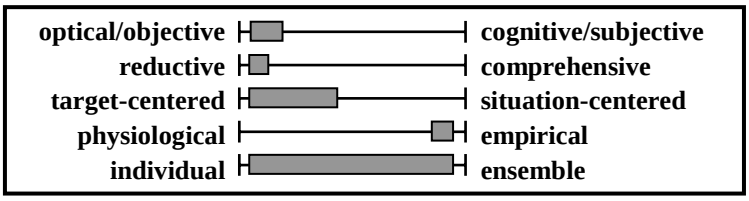
The logic underpinning these theories is that more information about a target should allow a greater proportion of observers to be able to acquire it. Most of the models and metrics based on these constructs are used for predicting ensemble performance. Examples of models in this category include the Johnson-criteria-based models from NVESD, FLIR92 (Scott & D'Angostino, 1992) and ACQUIRE (Tomkinson, 1990), the Bailey/Rand search model (Bailey, 1970), metrics of clutter and its inverse, conspicuity (e.g., Toet, 1996), and models of target distinctiveness (Ahumada & Beard, 1996).

The strength of these models comes from their simplicity and robustness. The metrics often used (e.g., resolvable detail, clutter) have stood the test of time and are widely used as predictors of performance. Clutter, for example, is known to influence performance very strongly and in many ways (Akerman, 1993a & b). In addition, new models of this sort are still being created and have predictive validity (e.g., Overington's 1982 disk discrimination metric; Bijl &

Valeton's 1998a triangle orientation discrimination metric). These new metrics are discussed in greater detail shortly.

The primary weakness is based on the facts that the hypothesized constructs are derived solely from the scene and that the models are designed around ensemble performance rather than individual performance. As such, there may be limited opportunity to add observer variables also known to influence performance.

4. Models based on largely atheoretical fits to empirical data:



There is a relatively uncommon class of models that predicts performance almost entirely by fitting empirical performance data from previous studies to a set of parameters measured or controlled in those studies. Models in this category tend to be older (e.g., Bishop & Stollmack, 1968, and Poe's model [see Bailey, 1970, for a discussion of Poe in relation to other models]).

Empirical models have few strengths. Their fundamental shortcoming is the lack of theory underlying the selection of parameters and the functions that the parameters are to fit. As such, although a curve fit through a set of data points for one study may be quite good, the curve will not be generalizable to experiments with different parameters. Even worse, the model might not be able to fit data with the *same* parameters because the way that the parameters mapped onto performance in one study may not take into account any third variables that actually drive performance or modulate the effects of parameters. Thus, even though the situation would *seem* to be identical to the first study, in reality, it may be quite different.

5. Classic Modeling Concepts

Most models make some common underlying assumptions or are based on a few fundamental phenomena. This section of the review discusses those assumptions as they have been incorporated into many models. Of particular interest in this section are the Johnson criteria, the ACQUIRE model, and its incorporation into a recent NVESD search model (FLIR92).

Across many current models, there are a few common underlying concepts. The instantiation of the concepts in the models, however, differs from model to model. Here, the concepts and basic instantiations of the concepts are discussed. The following five concepts have been identified as being basic to many models.

5.1 The Role of Contrast and Contrast Threshold

Central to all these ideas is that information used by the observer must be observable. That is, the information related to the target must have sufficient contrast, either between the target and the background or within the target, to allow the visual system to use it. The contrast threshold, C_T, is defined as the intensity of a stimulus required for it to be barely detectable with some reliability (usually 50% or 75%). It is typically described in terms of a lawful relation between the area of the target (or some other size-related quantity) and its intensity that holds at or near threshold called Ricco's Law.

Contrast and the Johnson criteria (see next section and appendix A) are intimately related. Johnson (1958) found that detection is typically afforded when a single cycle or less (a cycle being defined as a light and dark bar of a repeating bar pattern) on a target is visible. That the requirement for detection is near unity (see table 2) is consistent with the idea that the driving factor behind detection may be modeled by signal-to-noise ratio (SNR) or contrast. For near-threshold targets (e.g., targets with a small ΔT relative to their background support viewed through a FLIR sensor), the SNR is calculated in terms of a threshold SNR, below which a target is not visible (Howe, 1993; Johnson, 1958). For super-threshold targets (when SNR >> 1), the target contrast with respect to its immediate background is the crucial quantity.

Table 2. Resolvable cycles across critical dimension to perform 50% accurate acquisition (N50) at particular levels of target acquisition

Detection	Orientation (classification)	Recognition	Identification
1.0±0.25	1.4±0.35	4.0±0.8	6.4±1.5

Johnson also found that greater levels of target acquisition could be afforded when a greater number of cycles within a target are detectable. Once again, the concept of contrast comes into play in that these internal details must have sufficient contrast with their background to be detectable.

5.2 Johnson (1958) or Johnson-like Target Information Requirements for Levels of Target Acquisition Performance

Johnson (1958) found that ensemble target acquisition performance can be predicted by a determination of the number of resolvable bar cycles that can be perceived on a target (a quantity called N). (See appendix A for a detailed description of the method Johnson used.) Johnson found, not surprisingly, that the ability to perform increasing levels of target acquisition (i.e., detection → classification → recognition → identification) required that a greater number of bars be resolvable. The resulting "Johnson criteria," the amount of internal detail required to acquire a target, are widely cited and used in models of ensemble performance (e.g., ACQUIRE and FLIR92). Table 2 shows the findings from Johnson's original study.

The shape of the function describing the relationship between N and the probability of detection is, as one would expect, not a step function at or near 1.0 cycle. Rather, N50 describes the corresponding number of cycles for 50% ensemble performance on an ogive-shaped function called the target transform probability function (TTPF). The TTPF maps predicted probability of detection (P_d) for the ratio of N/N50. For detection, the TTPF can be described as follows:

$$P_d = \frac{(N/N50)^E}{1+(N/N50)^E}$$

in which N = number of cycles resolvable on the target,

N50 = number of cycles required for 50% of observers to detect the target, and

$E = 2.7 + 0.7(N/N50)$.

Note that the TTPF described performance at the ensemble level and is not intended to predict within-subject performance across trials[9].

Johnson's original idea has undergone few substantial changes since its first publication, although current so-called two-dimensional (2-D) extensions of the criteria take into account the height and width of the target rather than simply a "critical" dimension (e.g., the ACQUIRE model is based on such an approach).

That "information" resolvable about a target should drive performance as a unidimensional quantity is a powerful idea. Recent models have gone about determining the target-like information in the scene differently, but there remains a central requirement that a given amount of target information is needed for the average observer to acquire the target. (See the following section for a detailed description of these efforts.)

5.3 The "Classical Approach" to Modeling Search and Bailey's (1970) Separability of Time-Dependent and Time-Independent Search Processes

The so-called "classical approach" to search modeling was first put forth by Bailey (1970), in which probability of acquisition in search is a product of independently considered time-dependent and time-independent stages.

Bailey asserted that P_R, the probability of acquiring (recognizing or identifying) a target, is the product of P_1, the probability that a single glimpse will locate the target region of a scene, P_2, the probability that if the target is viewed foveally, it will be detected, and P_3, the probability that if the target is detected, it will be recognized or identified[10]:

---

[9]Such an analysis has been done, however, in order to evaluate the kinds of errors that such an ensemble predictor makes. For example, Valeton and Bijl (1995) looked at individual deviations from ensemble predictions in the evaluation of the Target Acquisition (TARGAC) model (which bases its predictions on a Johnson-type model), and Silk (1997) used such deviations to evaluate whether P_∞ is a biased estimator of ensemble performance.

[10]Indeed, this description of target acquisition is the same as was provided in the definition in section 2.

$$P_R = P_1 \times P_2 \times P_3$$

The first term, P_1, is time dependent in that it is assumed that during the search of a scene, a glimpse has a dwell time at a certain location and a certain amount of time between fixations for eye movements. Search progresses by the random selection of locations about the scene. The cumulative probability, $P_1(t)$, that a saccade will land sufficiently close to a target within time t, is described as the first arrival time of a Poisson process:

$$P_1(t) = 1 - e^{-t/\tau_{FOV}}$$

in which τ_{FOV} = the mean acquisition time, given that a target is fixated.

The second two terms are independent of time in that they are both conditional on the target having been fixated. Bailey (1970) derived separate terms for P_2 and P_3, which are of historical significance only (although Ryll [1962] incorporated the effect of scene clutter into the P_3 term). The most popular search models in use today incorporate a limiting term, P_∞, to denote that even after an infinite amount of time, some members of an ensemble of observers will be unable to acquire the target.

The current, widely accepted NVESD models ACQUIRE (Tomkinson, 1990) and FLIR92 (Scott & D'Angostino, 1992) instantiate this asymptotic term as the product of P_2 and P_3 and use the familiar TTFP as the limiting term:

$$P_2 \times P_3 \equiv P_\infty = \frac{(N/N50)^E}{1+(N/N50)^E}$$

in which N = number of cycles resolvable on the target,

N50 = N50 for detection, and

$E = 2.7 + 0.7(N/N50)$.

Thus, the entire ACQUIRE probability prediction equation can be expressed simply as a function of time and the number of resolvable cycles on target, which itself is a function of target area and contrast:

$$P(t) = P_\infty\left(1 - e^{-t/\tau_{FOV}}\right)$$

The average target detection rate, $1/\tau_{FOV}$, is related to target information available and required for 50% ensemble acquisition (Howe, 1993):

$$\frac{1}{\tau_{FOV}} = \frac{1}{6.8}\frac{N}{N50}$$

The theoretical and practical shortcomings of this model are discussed in various sections of this report.

5.4 Clutter and Its Impact on Performance

Counter to the assumption underlying the Johnson criteria, merely having a certain amount of target-related information available in the scene does not completely determine performance. The background in which the target is present must also be taken into account when one is making predictions of performance. The term "clutter" has no single agreed-upon definition. It has been described as scene complexity, number or density of target-like elements, number or density of objects, and overall scene "busyness" and has been quantified as any of several unitless metrics (e.g., signal-to-clutter ratio [SCR]). What can generally be agreed upon is that when certain kinds of terrain (such as desert) enable better target acquisition performance than others (such as partially wooded) when viewed optically, it is presumed that the driving force for this difference is that the former terrain is less cluttered (or has less clutter) than the latter. What exactly the clutter in the scenes *is* is not clear, although we can often determine it subjectively "just by looking" at the scene.

Clutter can be defined either locally or globally, depending on the metric enlisted to describe the scene. As stated before, certain kinds of terrain have more or less clutter, in general, than others. Likewise, some regions within a given scene may be more cluttered than other regions. This observation is obvious since terrain is rarely uniform and since some parts of a scene (such as an open field) can quickly be searched while rocks or trees surrounding the field provide for a more difficult search situation. Typically, local clutter metrics appear in models of time-dependent search, while global clutter metrics appear in models of pure acquisition (when eye movements are not needed because target location is known ahead of time)[11].

Clutter is known to adversely affect target acquisition performance at several levels. The impact of clutter on the Johnson criteria is to increase the number of resolvable cycles needed to acquire the target (e.g., Mazz, 1998). The effect on search is to decrease the size of saccadic eye movements between glimpses (meaning that the eccentricity from the fovea allowing for effective search decreases), and to increase the amount of time spent at each glimpse location (e.g., Akerman, 1992, 1993a). In addition and lending support to the definition of clutter as the number or density of target-like objects, local clutter affects where eye movements will occur. Fixations tend to be executed to "target-like" regions of the scene and not to locations at random. The presence of many target-like objects in the field is also known to increase the false detection probability compared to when there is relatively little clutter (Schmieder & Weathersby, 1983).

Clutter is discussed in more detail in a separate section of this report.

[11]This is not a "hard-and-fast" rule, of course.

5.5 Target Acquisition Models Based on the Decomposition of the Scene Into Oriented Spatial Frequency Channels

Models based on a spatial frequency analysis of a scene assume that visual perception is mediated by an array of spatially tuned pathways. Each pathway responds selectively to a band of spatial frequencies at a particular orientation and located at a particular position on the retina (i.e., corresponding to a particular position in the field of view). Information from these channels forms the building blocks of all visual percepts, including, of course, those of the target.

Justification for modeling the visual system with a set of oriented spatial frequency channels comes from a variety of sources. First, Hubel and Wiesel's Nobel prize-winning research (e.g., 1962, 1968) into the nature of cortical visual processing indicates that the receptive fields of neurons in early visual cortex (V1 and V2) seem to be sensitive to the presence of oriented line segments but largely insensitive to the presence of dots of light[12]. Second, the shape of the human contrast sensitivity function (the contrast threshold as a function of spatial frequency) and the selective adaptation of parts of the function can be explained elegantly by the summation of overlapping contrast sensitivities of a set of narrowly selective functions that varies over spatial frequency (Campbell & Robson, 1968).

In order to model a visual system based on selective sensitivity to spatial frequency, it is necessary to determine how many different frequency- and orientation-selective filters are required to define a wide variety of stimuli. The term "channel" is used to describe a mechanism that is maximally responsive to patterns of light of a certain spatial frequency and orientation.

Although there are theoretically 180 degrees of orientation and about three logarithm units of spatial frequency to which humans can respond within any orientation, a relatively small number of channels suffices to completely describe our percepts. Richards and Polit (1974) used a metameric texture-matching task to determine that one-dimensional textures can be described completely with only four channels. Metamers are two stimuli that differ physically but are perceived to be identical to each other. The existence of a metamer in a sensory modality implies that either the receptors in that modality cannot transduce the aspect of the stimuli that distinguish them or the nervous system cannot encode the stimuli as being different from each other. Richards and Polit found that any two textures that evoked the same responses along these four channels were perceived to be identical, regardless of their actual spatial frequency content. In two dimensions (expressed in polar coordinates), Wright and Jernigan (in Akerman, 1993a) used a similar method to determine that 42 channels (6 radial and 7 theta oriented) completely defined all the textures in their study. More pertinent to the modeling of the perception of objects by spatial frequencies, Vol, Pavlovskaja, and Bondarko (1990) found that objects with similar spatial frequency profiles tended to be more confusable than objects with disparate

[12]That the neurons in V1 and V2 are highly sensitive to sharp edges is not inconsistent with a spatial frequency interpretation of vision because such sharp edges approximate delta or step functions, which decompose into all wavelengths by Fourier transform. Thus such an edge should, in theory, affect *all* properly oriented spatial frequency channels.

spatial frequency profiles. This result indicates that, at some level, the visual system seems to compute the multi-dimensional distance between combinations of spatial frequencies to evaluate their similarity. In terms of target recognition, then, if the images of two targets do not differ greatly in their spatial frequency signatures (e.g., an M60 and a T-72 tank viewed at a distance), then they should be difficult to distinguish.

That a relatively small number of channels may completely determine a percept means that a model may be able to use these few channels as a set of feature detectors to extract perceptually important information from the scene. Operations can then be performed on the output of the channels in order to determine what the original image must have been to have precipitated the activations[13].

Two classes of models have used the Fourier decomposition of scenes in constituent spatial frequency information. One class of models performs the decomposition with the hope of finding information within the spatial frequency representation of the scene, which would come from the Fourier decomposition of a target. The assumption of these models is that a given target will have a spatial frequency profile that will stand out from that of the scene, and thus by monitoring particular channels, a model can detect the target. Additionally, because fine spatial detail resides at high spatial frequencies, the presence of such information may indicate that a higher level of target acquisition may be possible. These models assume that the human visual system itself may be monitoring spatial frequency channels when it searches for a target.

The second class of spatial frequency models is a subset of more general purpose human perception models that uses a Fourier decomposition of the scene as a "front end" for information feeding into the visual system. However, this second class of models then uses the information (in the form of channel strengths) as features, which are then combined into higher order percepts such as junctions, surfaces, and solids. This class of models tends to be more theoretically driven and typically comes from the realm of the perceptual psychology. Examples include Wolfe's Guided Search 3 (Wolfe & Gancarz, 1996) and Grossberg, Mingolla, and Ross's (1994) model of surfaces, edges, and attention.

6. Classic Modeling Concepts Revisited

Recent work in modeling has either augmented or attempted to replace the classic concepts. Efforts to incorporate new factors into old models, and challenges to the underpinnings of the old models are presented. The limitations of the classic concepts are discussed.

[13]An interesting perspective of what the visual system actually *does* comes from the fact that the brain's task is to try to determine the probability of an object being in the visual scene, given the stimulation along the visual pathway. This observation is often overlooked by scientists attempting to determine the visual system's response to a stimulus. The two things are the opposite conditional probabilities of each other (P(stimulus|response) versus P(response|stimulus)) and are in fact quite different (Reike et al., 1997).

6.1 Contrast Revisited

Contrast, like the various metrics proposed as alternatives to bar cycles on target from the Johnson (1958) criteria, is a one-dimensional quantity. It is typically assumed to vary according to the observer's contrast sensitivity function relating the required contrast between an object and its background (if both are uniform and untextured) in order to detect a target. Determining the contrast threshold, C_T, for real-world situations requires taking into account factors such as the reliability of detections (e.g., whether C_T is a 50% or 75% threshold), the retinal eccentricity of the target, the size of the target, its shape if it differs greatly from a 1:1 height-to-width ratio, its hue, and the observer's level of dark adaptation, to name a few.

The concept of contrast, as a single quantity indicating to a large degree the ease with which a target can be detected, has a number of problems. First, it fails to take into account various psychophysical findings that may be relevant to target acquisition performance in the field. For example, it is known that a non-uniform target against a uniform background is more detectable than a uniform target against a uniform background (Akerman, 1992).

Second, contrast is a local phenomenon and as such, cannot address issues related to the global scene such as clutter or highly salient events in other portions of the visual field. It is known, for example, that transient events in the periphery, even when known to be irrelevant, can render some objects difficult to detect (O'Regan, Rensink, & Clark, 1999). In these cases, the contrast of the target may far exceed what would be required for detection in the absence of the transient, yet it remains undetectable[14]. More details of this effect from perceptual psychology and its possible relevance to military target acquisition are discussed next.

Third, the flip side of irrelevant transients reducing the effective contrast of a target is the finding that a transient occurring at the target location or motion of the target can render the target *more* visible than it would otherwise be (Mazz, Kistner, & Pibil, 1998; Nakayama & Mackeben, 1989). Search models that incorporate motion tend not to adjust contrast threshold downward, however; they tend to change P_1 to make it more likely that a target is localized in a single glimpse[15]. This technique, of course, is empirically rather than theoretically motivated.

Fourth, contrast sensitivity is itself dependent on temporal aspects of the scene or display as well as light adaptation of the observer and retinal eccentricity, making its use as a single constant quantity related to a target somewhat questionable. Few Johnson criteria-based models incorporate this level of detail into their discussions of contrast. Models based on visual

[14]Studies in perceptual psychology that relate to transients and the transient capture of attention use the unidimensional term "salience" rather than contrast. In the luminance domain, it may be argued that the terms may be used interchangeably.

[15]In such models, motion may increase the size or characteristics of the hard or soft shell search lobe so that targets of greater eccentricity from fixation are detectable. Though such a change is consistent with an increase in target contrast, it is not specified as such in the models.

physiology and psychophysics, however, are more likely to include these details into the model front ends (see imminent section on psychophysical and physiological models).

Fifth, the contrast threshold below which a target cannot be acquired is not simply a function of the physical stimulus and adaptive state of the observer. Blackwell (1958 in Akerman, 1993a) lists several factors and how threshold contrast should be adjusted (always increased) to account for them. His results are summarized in table 3.

Table 3. The effect of various factors on target detection contrast threshold (C_T)

Factor	**Multiplier to C_T**
Uncertain frequency of occurrence (lack of vigilance)	1.19
Uncertain location	1.31
Uncertain occurrence	1.40
Uncertain size and occurrence	1.50
Uncertain occurrence and duration	1.60
Trained versus naive observers	1.90 - 2.00
Non-foveal target location	2.78

Note that all these factors, with the possible exception of the last one, are related to psychological variables. That such factors can so drastically change threshold contrast, yet are not included in models or are accounted for by appealing to a group of "trained military observers," indicates a lack of psychological sophistication and a clear case for the need to investigate how psychological factors influence performance.

6.2 Rethinking the Johnson Criteria

Although widely used and a good indicator of ensemble performance, the Johnson criteria are not without their problems. It is instructive to recall the kind of stimuli Johnson used in his initial study (see appendix A for details of his methods): bar patterns of uniform contrast against a uniform background. Such stimuli are obviously unrealistic, given that target and background characteristics vary greatly in the field. For example, using the results of Johnson's study to predict detectability of targets in a realistic setting requires N50s needed for various levels of acquisition to be increased, indicating that the criteria must be at least partly determined by particulars of the situation. Recall also that clutter is known to increase N50 across the board.

6.2.1 Other Issues Related to the Johnson Criteria

These issues hinge on a more realistic representation of real-world target acquisition situations.

6.2.1.1 Non-uniform Information in Targets (i.e., targets with large regions of little detail)

This problem arises from our attempting to apply the Johnson criteria to a wider variety of targets than were considered at the time of their inception. Regions of certain targets, such as ships, have relatively little detail and thus contribute little to our recognizing or identifying the target. Other regions of the same target contain the critical details. Since the Johnson criteria

depend on the area and the cycles across a critical target dimension, it seems obvious that area of the target alone is not a good indication of the information therein (Moser, 1972).

Moser proposed that instead of using area and resolvable cycles to determine the information in a target, the resolvable perimeter or the smallest resolvable perimeter element (i.e., a convex or concave region) would be a better indicator of performance. Work by Kennedy (1983) has led to the adoption of the square root of the area rather than simply the area when one is calculating N as a partial solution to the difficulties associated with using raw area. Overington (1982) suggested a similar approach to how recognition should be modeled. He proposed that detection performance (that is not biased by aspect ratio) can be predicted by an equivalent-size disk detection task, and that identification can be predicted by a disk discrimination task where the size of the disk in question was a fraction of the diameter of the target. Overington incorporated the psychophysical function relating disk discrimination and acquisition performance into an early version of the ORACLE model (see appendix A for details of the current ORACLE model).

6.2.1.2 Anisotropic Targets (i.e., targets that appear vastly different when viewed from different angles)

It is plainly apparent that most every target of interest is anisoptropic. Johnson and Lawson (1974) noted that many targets are more difficult to recognize from the front than from the side. (For example, envision an M-2 Bradley and an M1 tank from the front and the side. There is clearly more distinguishing detail available from a side view of the vehicles.) The authors found that N50 for recognition of ground vehicles increased by as much as 30% when viewed from the front. At intermediate aspects, however, performance remained relatively good as long as the details visible from a side view were still visible. This observation is very similar to how the RBC theory (Biederman, 1987) postulates that humans recognize objects. This theory is discussed shortly.

The effect of aspect has also been demonstrated to interact with the aspect ratio of the potential target. The increase in N50 as a function of aspect is even more pronounced for targets that have a large length-to-width ratio, such as a ship. In this situation, N50 increased by as much as 500% from the side to the front view (Johnson & Lawson, 1974; Ratches et al., 1973, in Howe, 1993). Thus, the Johnson criteria can no longer be considered a function of the level of target acquisition alone but must also incorporate target dimensions and aspect.

A different way to characterize target information within a Johnson-like framework (i.e., a set of criteria determining the amount of information required to acquire a target at different levels) is to use metrics other than N. Several such metrics have been defined and validated that do not depend explicitly on aspect. These metrics are said to have been validated in that they produce reliable criteria for each level of acquisition, similar to the Johnson criteria of a certain N for each level of acquisition.

As already mentioned, Moser has proposed that target information be a function of perimeter while Overington (1982) proposed that a detectable or discriminable disk size be used. Blumenthal and Campana (1981, 1983) proposed that image quality (operationally determined by the function of the inverse of the size of a barely detectable circle or square) be a metric for determining information about a target. Moser (1972) proposed an area-based metric (which he subsequently questioned) in which information is a function of the number of pixels on a target required for acquisition at various levels. Similarly, O'Neill (1974, in Howe, 1993) determined that Moser's number-of-pixels-on-target metric can be extended from silhouette images, used in Moser's study, to TV images.

A recent metric proposed by Bijl and Valeton (1998a) involves the contrast required to discriminate the orientation of an equilateral triangle. The underlying assumption of the triangle orientation discrimination (TOD) metric is that if a subject can reliably determine the orientation of a triangle of a dimension and contrast similar to a target, then he should also be able to discriminate the target. The critical dimension in the TOD metric is the square root of its area. That is, if a triangle and target have the same square root area, the probability of ensemble acquisition should vary together as a function of contrast.

Bijl and Valeton (1998b) validated the TOD metric against the cycles-on-target metric in the ACQUIRE model. ACQUIRE is used to predict the acquisition range for targets of a particular size and contrast. By comparing data about the discriminability of triangle orientations to data related to cycles on target and detection range, the authors found that (a) the TOD metric was a better predictor of acquisition range than ACQUIRE, and (b) the TOD metric is less susceptible to the aspect of targets, including ship targets known to have a large effect on N50.

6.2.1.3 The Reliance on a Single Quantity (e.g., cycles on target) to Determine Performance

One problem with the previously mentioned models that base performance predictions on the amount of information that can be derived from the target is the selection of a single aspect of the target that best captures the information content of the target. Area, resolvable cycles, perimeter, equivalent disk, square, and triangle size all capture some aspect of the target's information. However, it is likely a mistake to assume that all observers use the same source of target information. How then can the Johnson criteria be made to use more information?

As an example of a single metric that accounts for more than one aspect of a target, Akerman and Lucius (1990) defined the "useful area" as a function of both perimeter and area. Useful area is defined as the portion of the radius inward from the edge of an object's perimeter, which is to be used for assessing target acquisition performance. This metric has been incorporated into Akerman's visual observer model (VOM) (1992, 1993b). This technique of combining two largely independent features is a possible solution to the problem of selecting the one dimension most important for expressing the information content in a target. Physiological models and newer fuzzy logic models (discussed later) use this property.

6.2.1.4 The Relative Importance of Some Features Compared to Others

The concept that target information required for recognition or identification is related to the number of cycles resolvable on a target, and not what those cycles represent, is clearly a generalization. "More information" implies that some of it will likely be useful for discrimination performance, although the nature of that information is not clear. Johnson and Lawson's (1974) observation that N50 for anisotropic targets reaches a relatively stable minimum at aspects that include portions of the side view (e.g., a front left aspect angle) indicates that as soon as features of an object are visible (and themselves discriminable, of course) object recognition can proceed relatively independently of viewing angle. Thus, there may be critical details that, once visible, determine performance. This may be particularly true for targets that are easily confusable, such as a T-62 and T-72 tank. In a case such as this, the presence or absence of a single detail may be required for us to discriminate between the two. Should such a detail be small, the Johnson criteria for the discrimination would likely be quite large in that the size of a cycle on the target must be as small as the critical detail. The Johnson criteria, therefore, may be predictive but not very informative of the information that the observer uses to make a decision.

A popular model from perceptual psychology is Biederman's (1987) (see appendix A) RBC theory, which states that recognition of objects requires details (i.e., component geometric forms, called "geons" in the theory) of the object to be extractable from the image. If the aspect of the target is such that only a subset of the geons can be extracted (because others are not visible), then the object cannot be recognized definitively. In such cases, the observer uses the information available and performs the highest level acquisition decision possible—a classification or a recognition rather than an identification.

O'Kane, Biederman, Cooper, and Nystrom (1997) determined that the confusability between various military ground and air vehicles in a recognition task can be explained by an RBC-type model. The authors found that when particular features were obscured or not visible because of viewing angle, observers made errors in a manner consistent with their checking an internal representation based on the presence and configuration of basic geometric components of the objects.

Marr and Hildreth (1980) and Marr (1982) also modeled the process of recognition by asserting that objects in a scene are decomposed into a set of geometric primitives. Their approach was more computationally based than (and used a different mental representation of objects than) that of Biederman. However, a common fundamental aspect of the model is that it required the image to contain visual information sufficient to decompose it into its constituent primitives for recognition to take place.

Both RBC and Marr's theories differ from all the Johnson-like metrics and models in that the identity of constituent object components and not the quantity of information (however defined) determines identification performance.

6.3 The Bailey (1970), the Classical, and the Neoclassical Search Frameworks

All models of search must specify three aspects of the dynamic search process: search lobe type and size, fixation location selection, and whether over-searching is permitted. Search models all assume that a fixation must occur near a target in order for the target to be acquired. The distance required for acquisition need not define a hard "cut-off" between detectability and undetectablity, however. The visual lobe is defined as a set of probability contours that map the probability of acquiring the target at various eccentricities from the point of fixation. The shape of the function can be a step, indicating that no acquisition can occur after some eccentricity (and usually that there is equal probability of acquisition within that eccentricity) or a continuous, decreasing function of eccentricity. Models assuming the former are said to perform "hard shell" search; models assuming the latter are said to perform a "soft shell" search. There are also rare models (e.g., Georgia Tech Vision, discussed later) that require a target to be fixated directly before a detection can be made. In addition to how close to a target a fixation must fall, search models must also define how the locations of fixations are generated. Some models assume random selection with replacement, some assume random selection without replacement, and some assume guidance to target-like regions of the scene. Finally, models must also specify whether targets can be fixated more than once without being detected or eliminated from consideration.

In the instantiation of the Bailey framework, some assumptions must be made regarding how the time-dependent search operation is conducted. For example, selection of glimpse locations is typically considered to be random sampling with or without replacement. Also inherent in the selection of glimpse locations is the selection of the visual lobe. As discussed next, scenes will vary greatly as to the location of eye movements and distance moved in terms of the background and anticipated targets. Glimpse durations are usually assumed to be constant and independent of clutter, which is not necessarily the case. Clutter is known to increase dwell time (Akerman, 1992), indicating that P_1 may actually depend on processes involved in P_∞.

Two recent models that are based loosely on Bailey's logic but include more factors known to be involved with search performance are the Visual Detectability Model (VIDEM) (Akerman & Kinzly, 1979) and VOM (Akerman, 1992, 1993b). The most notable additions to the Bailey design are the effects of clutter (see appendix A and the section on clutter and conspicuity for more details) and the (optional) effect of display noise (VOM version 1.2, Akerman, 1993b). Display noise is represented by a final term, P_4, the probability of discriminating a target that has been fixated and detected, given the SNR inherent in the display upon which the target may be presented to the observer. Therefore, in the final model, $P = P_1 \times P_2 \times P_3 \times P_4$. It may be instructive to note that the independence assumption makes the combination of P_3 and P_4 possible and that no other model has separated this last term. The VOM is interesting in that although it uses a clutter metric (Waldman's SCR, discussed later) to alter glimpse time, it still uses a random selection of fixation locations.

The Bailey search step, as defined by P_1, the probability to fixate on the target in a single glimpse, assumes that the duration of the fixated eye movement is sufficiently long to allow for complete spatial sum of the stimulus. Spatial summation, which is only nearly complete for relatively small stimuli, requires between 50 and 200 milliseconds to take place (Howe, 1993). The time required to sum stimuli should certainly have an effect on the observer's decision as to the presence of a target, yet models tend to keep glimpse duration constant.

Self (1969, in Akerman, 1993a) summarized five aspects of eye movements in real-world visual search, which make their prediction problematic:

When a target is not found quickly, the observer tends to re-search areas of the scene he thinks are likely to contain the target while ignoring other areas of the scene which he thinks are unlikely to contain the target. Although knowledge of the target and where it is likely to appear could be helpful in many situations (and thus the justification for training the Soldier as to common concealment/placement methods), such dependence on where a target ought to appear could lead a Soldier to miss a target that is in an unexpected location.

This behavioral finding is in good agreement with a recent result by Chun and Wolfe (1996) that shows that subjects use different criteria for rendering a target present/absent judgment: when a target is located, search stops (as one may expect it to). When a target is not located, subjects employ a "conservative quitting criterion" and will over-search the scene until a more restrictive, task-dependent criterion for the target not being present is met.

The finding also indicates that cognitive processes related to knowledge of likely target characteristics and capabilities and possibly, familiarity with strategy and terrain types, has a strong influence on performance. Presumably, there should be a strong effect of training on this kind of behavior.

a. Most subjects first perform a cursory scan of the scene for the target before beginning any kind of systematic (trained or instructed) scan.

b. Targets closer to the center of the FOV tend to be detected more rapidly than those of the periphery.

This finding agrees with recent work in attention deployment in difficult (conjunction) search by Carrasco, Evert, Chang, and Katz (1995). The authors showed that, all things being equal, subject performance was faster and more accurate for detection of targets close to fixation. Given that a subject will likely begin perusal of a scene somewhere near the center, these results may be applicable to Self's observations.

a. Putting time pressure on the subject can lead to faster searching (i.e., shorter glimpse duration) without a loss in accuracy.

b. There are large, consistent individual differences between subjects related to performance. Some subjects are consistently faster and more facile at searching than others.

Although this point does not pose a specific problem for models based on Bailey, since these models predict ensemble performance, it means that less of the variance within a study will be captured by the situational variables of interest (e.g., N50).

In addition to Self's observations, other researchers have observed two additional aspects of eye movements that models must be able to address (e.g., Nicoll & Hsu's, 1995, analysis of field data from O'Kane, Walters, & D'Angostino, 1993):

c. Observers routinely visit the target many times before declaring a detection of the target.

d. Observers continue to visit non-targets and the target after detecting the target.

There exists substantial evidence that, as indicated by the observations by Self and Nicoll and Hsu, eye movements are anything but the random-selection-with-replacement phenomenon assumed by the Bailey model.

Eye movements in laboratory studies are a common means to determine if a model provides a good fit to empirical search data. A largely unaddressed problem for such a validation procedure is how to interpret brief glimpses of 100 to 200 milliseconds in duration. Such glimpses may be corrections for erroneous saccades or brief glimpses. At issue is what is considered a fixation (Karsh & Breitenbach, 1983). The neoclassical approach to search (discussed shortly) attempts to address this distinction in a theoretically meaningful way.

Eye movements are often considered nuisances in laboratory studies of perception because of their unpredictability unless intentionally recorded[16]. Some methodologies require subjects to perform a task without eye movements. However, more recent models from perceptual psychology have attempted to incorporate them, since there is now a fairly solid theoretical foundation for eye movement guidance based on the deployment of selective visual attention (Posner, Snyder, & Davidson, 1980; Schneider & Deubel, 1995; McPeek, Maljkovic, & Nakayama, 1999). What has become obvious to vision researchers, long after it was widely known to target acquisition modelers, is that eye movements do not agree with the randomness implicit in many basic models such as Bailey and ACQUIRE. In fact, some recent evidence shows that eye movements are not performed randomly without replacement but pseudo-randomly with replacement (Horowitz & Wolfe, 1998).

[16]Many such studies are interested in covert attentional shifts that do not require eye movements. Eye movements in these studies are considered unwanted noise.

6.4 Models of Visual Search

Because of the evidence for a link (perhaps even an obligatory one; see McPeek, et al., 1999) between focal attention and eye movements, it would be beneficial to briefly review some recent models of attention on visual search from the perceptual psychology literature. Models of interest include Wolfe and colleagues' Guided Search models (Wolfe, Cave, & Franzel, 1989; Wolfe, 1994; Wolfe & Gancarz, 1996), and Humphreys and Muller's SERR model (1993). All these models incorporate stimulus-driven and goal-directed selection of attention. That is, attention may be drawn to salient regions of the scene, or it may be directed overtly about the scene by the observer.

Common to the models is the notion of a pre-attentive stage of processing and an attentive stage. Pre-attentive processing is large capacity, parallel, and operates over much of the visual field. These mechanisms operate on the level of the features that constitute objects rather than objects themselves. Focal attentive processing is small capacity, serial or limited capacity parallel, and operates on objects in the field a few at a time. Focal attention, with or without overt eye movements to the region of the scene, is assumed to be required for the proper binding of features[17] into coherent objects (Treisman & Gelade, 1988) and for the conscious perception of objects (Rensink, O'Regan, & Clark, 1997).

The various versions of Guided Search all consist of two stages, a pre-attentive stage and an attentive stage (see appendix A). The pre-attentive stage extracts features from the scene along various feature dimensions separately (e.g., color opponency, orientation, luminance, motion). The attentive stage uses information about a known target (if one is available) to select from regions of the scene that weighed highly on relevant feature dimensions and then selects a single object to inspect. The interplay of top-down and bottom-up information is instantiated in the model by a master activation map. Search progresses in a time-limited serial self-terminating manner (i.e., one at a time until the target is found, all items have been searched, or a temporal cut-off has been met) from areas of high activation on the master map to areas of lower activation. The first two versions of Guided Search do not incorporate eye movements.

Wolfe and Gancarz (1996) have recently modeled visual search with eye movements but with fewer features than previous versions of the model. Guided Search 3.0 assumes that attention, both stimulus-driven and goal-directed, creates a spatiotopic saccadic activation map corresponding to the master activation map in earlier versions of Guided Search (see appendix A). Maxima in the map represent the input to the saccadic control system, which then causes an eye

[17]It is important to note that the stimuli used in most laboratory studies of visual perception consisted of such simple elements as oriented, colored line segments, rotated letters, and various shapes, usually presented on a blank background. Though such a methodology allows for a discussion of the basic features comprising a simple object, it may not immediately be generalized to studies of military target acquisition. Objects and scenes of military significance cannot be reduced to basic features, at least not features analogous to those discussed in the perceptual psychology literature. One could argue, though, that the various attempts to define metrics of target attractiveness, conspicuity, and distinctiveness are attempts to find such a set of basic features to describe the real world.

movement. Subsequent saccades to already searched locations are initially inhibited by inhibition of return (IOR). As IOR fades over several hundred milliseconds, the activation of the location can again increase until another saccade is produced. The model is quite simplistic to be sure (e.g., it is concerned only with luminance and orientation), but its input from the scene and the observer's intentions allows it to predict nearly all the performance characteristics mentioned by Self (1969).

Humphreys and Muller's (1993) SERR model focuses more on the stimulus-driven aspect of search than does Guided Search. The factors that drive the ease of search are based on target-target, target-non-target, and non-target-non-target similarity along any of several dimensions on which pre-attentive vision can operate, such as color, orientation, size, etc. (Duncan & Humphreys, 1989). Search is easy if targets are similar to each other, non-targets are similar to each other, and targets and non-targets are different from each other. As the degree of similarity *within* targets or non-targets decreases, or the similarity *between* targets and non-targets decreases, search becomes more difficult. The model progresses through search by rejecting regions of the scene recursively until it locates the target. Rejection is based on features dissimilar to the target and similar to each other; regions containing many such features are rejected *en masse*.

What is clear from both of these models and from other models that posit a pre-attentive feature extraction stage followed by an attentive selection stage (e.g., Feature Integration Theory by Treisman & Gelade, 1988, and Treisman & Sato, 1990), is that locations selected for attentional scrutiny are anything but random. As such, models that posit the random, independent selection of glimpse locations may be suspect since (a) that is not how search progresses, and (b) the probability that a target will be selected on a glimpse is a decreasing function of glimpse location rather than being constant (i.e., it is dependent rather than independent).

Some target acquisition models do indeed predict that glimpse locations are selected from regions of the image that are likely to be a target (i.e., that contain target-like information, however construed). For example, the GTV (Doll, McWhorter, Wasilewski, & Schmieder, 1998) model bases search on pre-attentively selected locations that have similar features as the (known) target. (GTV is described in more detail later and is detailed in appendix A.) Also, the evaluation of numerous local clutter, distinctness, and conspicuity metrics is based on the assumption that glimpses are directed to regions of the image that are relevant to the target. (These metrics comprise a major section of this report and are discussed at length shortly.)

Both models from perceptual psychology and most models of target acquisition assume that over-searching does not occur. Given the observations of Self (1969) and Nicoll and Hsu (1995), this assumption is obviously false. That is, there are cases when a target will fall within a prescribed search lobe, will be discarded as a non-target, and will be inspected later and at that time be judged a target. There are also cases when no target is present and the observer searches repeatedly over the scene before rendering a no-target judgment (e.g., Chun & Wolfe, 1996). Models of search, as mentioned before, typically assume that once a target falls within a search

lobe, it is either found or not. (Models that incorporate random glimpse location with replacement do not make this assumption; instead, however, they make another unrealistic assumption about how search progresses.)

6.4.1 The "Neoclassical" Approach to Search

Recall the assumptions of the classical search framework as described by Bailey and how they confront the reality of search behavior: the classical approach is serial and self-terminating, meaning that search progresses randomly one item at a time until the target is fixated at which time, it is either detected or not. If it is detected, search halts. Self (1969) pointed out that search does not progress in this orderly manner: objects are not selected at random, objects are searched more than once, and objects close to the center of the FOV tend to be searched first.

Though a pre-attentional saccadic guidance stage can alleviate some of these difficulties, such a remedy cannot address the fact that in the real world, observers search the same object more than once. The violation of this assumption draws into question the assumption that search can be described as a single Poisson process.

Nicoll and colleagues (e.g., Nicoll, 1994; Nicoll & Hsu, 1995; Cartier, Nicoll, & Hsu, 1998) have proposed a different way to model search and detection. The *neoclassical* framework is based on a different set of assumptions about how an observer actively goes about searching. The phenomenal underpinnings of the model are similar to Yarbus's (1967) description of eye movements: "the human eye can only be in one of two states: in a state of fixation or in a state of changing the point of fixation." When one is searching for a target, Yarbus's description can be described as having three states: (1) fixating on the target, (2) fixating on a non-target, and (3) changing the point of fixation. The modelers in the neoclassical framework describe the first two states as "examining points of interest (POIs)" and the third state as "wandering." Search can therefore be described by a Markov process containing these states and the rates of transitions between them.

Unlike the classical framework in which time to first fixation of a target (and thus detection) is an exponential function of total search time, the neoclassical framework assumes that detection of a target is an exponential function of time spent examining the target itself, not search overall. That is, a certain amount of time must be spent examining the target POI for it to be detected.

In more detail, a scene contains *i* POIs, with POI(0) being defined as the target and POI(1) through POI(i-1) defined as non-targets. Search can be in a state of examining any of these i POIs. In addition, search can be in an intermediate state in which it is wandering (without memory for where it has been) between POIs. This wandering state is referred to as W. The rate at which target information is accumulated is defined as $\alpha(0)$.

The rates describing the transitions between these states can be written as follows:

w = average rate of observer leaving a POI to wander

S_i = average rate of observer entering the ith POI from wandering

J_i = average rate of observer entering the ith POI from another POI

Markov processes are tractable mathematically because the solution for the behavior of the entire system is the linear combination of exponential terms for each state. Although tractable, such a solution may become very complex, since the number of potential POIs in a scene can be quite large. However, the solution may be simplified dramatically once the meanings of the transition rates is made clear. The rates of entering a POI, S_i and J_i, can be thought of as a function of the attractiveness of the POI. As mentioned before (and mentioned later in this report when clutter is discussed), there are a number of ways that the attractiveness of non-target regions can be modeled. The output of some pre-attentive mechanism, as mentioned before, seems to play a role in search performance. Various local metrics for conspicuity and clutter have also been proposed. All these metrics and processes are involved in the designation of local regions of the scene that contain information that is "target like."

In addition to points of *local* clutter in a scene, there is also good evidence that *global* measures and metrics of clutter influence performance (decrease P_d and/or increase response time) without appealing to the detailed spatial information in the scene. Such overall metrics of clutter or non-target scene attractiveness can be thought of as the rate at which any non-target POI is entered from wandering. This global attractiveness assumption allows us to simplify the model dramatically by lumping all the states wherein the eye is neither wandering nor examining the target as a single state: examining a non-target POI. The solution then becomes the linear combination of three states. The model can be further reduced into a two-state model if the target is not considered to have a different attractiveness than non-targets.

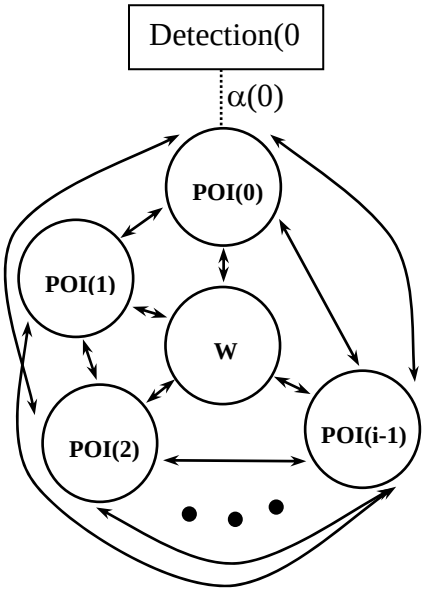


Figure 2. The complete state description diagram for the neoclassical search of a target, POI(0), among *i-1* distinct non-targets points of interest, POI(1) to POI(i-1).

The neoclassical framework has the advantage of making falsifiable predictions about search times, in that observer behavior should be the linear combination of exponential random variables. Nicoll and Hsu (1995) used eye tracker data from a NVESD study (O'Kane, Walters, & D'Angostino, 1995) to examine the specific predictions of the memory-less three-stage Markov search model. The predictions of the search portion of the model and the analysis of results are as follow:

1. *Targets are not always detected upon first visit. The probability of detection on a visit is independent of overall time spent searching.*

The first statement is obviously true. The second statement is not true; there is a weak correlation between time searching and P_d on a particular visit. The authors attribute this result to the non-exponential character of visit duration during detection visits (discussed next).

2. *A memory-less Markov process implies that the searcher will return to the target after detection (i.e., the process itself does not include a termination-upon-detection requirement as was assumed in Bailey [1970] and other classic framework models).*

Eye movement data clearly support this prediction.

3. *The duration of pre-detection visits to a target, during detection visits (when detection actually occurs), and post-detection visits should all be equivalent and should be exponentially distributed.*

The pre- and post-detection visit durations are exponential and essentially identical. However, the during detection visit durations tend to be longer (in the case of the test data, nearly twice as long) as pre- and post-detection durations, there were few very short-duration visits, and the distribution lacked a tail of long-duration visits. From these data, it seems that during detection visits are more normally than exponentially distributed. The authors posit that this delay may have been attributable to a motor response and some sort of inhibition in the eye movement system. As discussed in the next section of this report, it could also be that a different strategy was used for verification leading to a detection rather than checking when no detection decision was made.

4. *The distribution of the time to the first target visit is described by one or two exponentials (depending on whether all POIs are equivalent or target and non-target POIs are different).*

The distribution of first visit times is actually close to an exponential but only after a delay. This result is consistent with observations in the scene perception literature, indicating that observers do not begin immediately searching the scene when it appears. When an observer is confronted by a new scene, he first spends a few hundred milliseconds glancing around at it to "get his bearings" and extract the spatial layout or "gist" of the scene (Intraub, 1981)[18].

[18]Upon reflection, this observation is obvious even within the logic of the neoclassical framework. Some visual and possibly cognitive process has to extract scene information sufficient to delineate points of interest before the search process as described by the model can begin.

5. *The distribution between gaps (times between visits to the target) is described by one or two exponentials.*

The data examined indicate that a two-exponent process provides good agreement with the data.

6. *A memory-less Markov process implies that the gaps before and after detection will be distributed in the same way.*

After detection, the gaps are not distributed exponentially. The search process returns to the target too soon after detection for it not to have learned (i.e., search is not a memory-less process).

The detection process (i.e., the assumption that detection is based on time exploring the target and not search time overall) makes two additional predictions within the framework of the Markov process:

7. *The probability of detection is exponential in the time on target.*

This basic premise of the detection process is supported by the data.

8. *The distribution of the number of targets detected (across all trials in the data set) is described by two or three exponentials.*

This hypothesis, too, is supported by the data when the search time is shifted to account for the delay in first visit (see [4]), though a few finely grained anomalies remain. For targets with high P_∞, a one-exponential model and the classical framework both do well; for targets with low P_∞, a two-exponential model can account suitably while the classic model's predictions are too low by a nearly constant amount.

There are several strengths in the neoclassical approach to search and detection. First, it provides theoretical rationale for known eye movement phenomenology such as searching the target more than once and continuing to search after detection of a target. Second, the assumption that detection depends on time spent examining the target is more likely an accurate description than the assumption in the classical framework (that detection depends on time spent searching in general). Third, the notion that the attractiveness of POIs determines the rates at which their states are entered provides a way to insert conspicuity, attractiveness, or clutter, at a global or local level, into a theoretical framework. If the neoclassical framework proves to be a better predictor of overall behavior than the classical framework, then the assignment of rates by attractiveness may permit the objective analysis of such metrics.

Nicoll (1994) extended the basic model to include field of regard searches, multi-target searches, searches when a particular state is assumed to begin the process (in accordance with the observation by Carrasco et al. [1995] that targets near the center of the FOV tend to be examined first), and time-limited searches. Not only can the framework accommodate such concepts, but it still provides testable predictions.

A disadvantage is that the neoclassical model does not completely account for the data set examined in the Nicoll and Hsu study. The time constant that must be added to first target visit times, the fact that there is evidence for memory of detection (by the post-detection visit gaps), and the non-exponential distribution of during detection target visit durations all provide evidence that the Markov process model cannot account for search without additional mechanisms.

Perhaps the most glaring shortcoming of the model is its assumption of memory-less search. Such an assumption negates the possibility of cognitive search strategies (e.g., systematic search of the scene or deciding not to revisit a previously searched region), when it is obvious that observers use such strategies to search! Of course, Markov model predictions are based on distributions across trials, so unless subjects used similar, consistent search strategies, the model would be unable to determine if its assumptions were incorrect. That is, if subjects used an evenly distributed (in space) variety of search strategies, then the data would still, by chance alone, show an exponential distribution of detection numbers, gap times, etc. Presumably, though, the data would not fit as tightly around an exponential curve. Once again, individual differences are relegated to the error term.

6.4.2 What is Happening During Detection?

Nicoll and Hsu's (1995) finding that distributions of target visit durations are longer and less exponential when a detection is made than before or after a detection is made indicates that some other process is involved in detection. What is that process? It may be instructive to be more clear what the authors meant by a "visit" to a POI. Eye movements do not simply go to a potential target, sit there, then fly to another point. (If that were the case, then no "wander" points could be empirically determined.) Rather, eye movements tended to be of two types: sequences of short (in distance) saccades around a small region, and one or two long saccades between these sequences. The inflection points between two long saccades were defined as "wander" points (they typically lasted only around 100 ms, a period likely too short to extract much information [Cartier et al., 1998]). The sequences of short saccades around a region were defined as "examination" points around a single POI. In other words, detection was based on the accumulation of time spent making saccades and extracting information from a region, not fixating directly at a target.

This distinction gives rise to the possibility that the process of detection of a target may actually be a discrimination process in which the target must be discriminated from a non-specific "non-target" class of scene elements[19]. Since the assumption of all these models is that the observer is aware of what a target looks like (how else could the non-target POIs have been selected?), perhaps the time examining the potential target POI is actually spent by a discrimination process.

[19]This redefinition of detection is, of course, a tautology. However, it may be meaningful in the context of a difficult search.

Stark and colleagues (e.g., Noton & Stark, 1991; Hacisalihzad, Stark, & Allen, 1992; Stark, 1993) proposed the *scan path* theory positing that observer eye movements examine a potential target for known features and then recognize or reject the object based on the concordance of observed and expected features. The examination of a target for discrimination requires a sequence of anticipatory saccades toward known points (corners) of a target. Unfortunately, the scan path theory's limitation to large, clearly defined, familiar objects in a particular orientation makes it unsuitable for target acquisition modeling. A more complete description of a number of Stark's models is presented in Lind (1995).

What then is going on during detection that slows the search process? Some sort of feature-matching process may be in play. Also, as Nicoll and Hsu (1995) postulated, there could be a motoric delay that slows search while a detection decision is physically rendered (though why that should change the distribution from an exponential is unclear). It could also be that as more information accumulates about the target, the more processing time is required for the addition of information and for evaluations of that information. (The actual process of detection of a target is not specified in this model, only its temporal character.)

6.5 Clutter and Its Effects on Performance

It may be worth mentioning at the beginning of this section that the term "clutter" has no analog in perceptual psychology. Perceptual psychology tends to view a scene as a collection of features (e.g., Wolfe, 1998), surfaces (Nakayama & He, 1994), or oriented visual primitives extracted by early cortical mechanisms (such as line segments, e.g., Grossberg, 1997). However, one of the most consistent findings in the visual search literature is that response time increases with display size (number of non-target distractors). As mentioned earlier, a non-target is only considered to be a hindrance in search (i.e., is only considered to be clutter or to be a distractor in the literal sense) if it cannot readily be eliminated from consideration because it is similar to the target (Egeth, Virzi, & Garbart, 1984; Duncan & Humphreys, 1989). Just what it is about a non-target that is important (e.g., the color, size, shape, orientation, proximity, depth, etc.) is unclear.

It is also known that the homogeneity and distribution of non-targets influence search difficulty. Duncan and Humphreys (1989) found that search performance suffered when (a) non-targets were similar to targets, and (b) when the non-targets were dissimilar to each other. Nothdurft (1991) found that targets were easy to see if they differed from their neighbors in a single feature, but the same targets embedded near similar features were quite difficult to see. Wolfe et al. (1989, 1994, 1998) have modeled the selection of locations for the deployment of focal attention and eye movements as a function of similarity as well as distance between non-targets and targets, and Humphreys and Muller (1993) have modeled the elimination of non-targets based on these feature-based similarities.

Taken together, perceptual psychology has a relatively simplified conceptualization of what might be considered clutter. As such, the quest to find a single explanation of clutter and a single

numerical metric for its magnitude comes largely from work in the target acquisition and ATR modeling communities[20].

In this section, several metrics for clutter, conspicuity, and distinctness are discussed in terms of what they measure, why or how they are purported to work, and how well they have fared at predicting target acquisition performance. The terms clutter, conspicuity, and distinctness, plus the term "attractiveness," are all attempts to define what it means for a target to be easy or difficult to acquire. Clutter may be considered the inverse of the other three terms, all of which (for the purposes of this report) are used interchangeably.

Metrics for clutter can be local, semi-local, or global. Local metrics refer to parts of a scene that are confusable with the target; semi-local metrics refer to the amount of clutter in particular regions of a scene; global metrics refer to the overall measure of scene clutter without any specific information about regions or locations within the scene.

6.5.1 Early Clutter Models/Metrics

Clutter and conspicuity have long been included in models of target acquisition. As mentioned earlier, clutter can affect search processes (by slowing search, shrinking a hard shell lobe, and influencing eye movements) and detection and discrimination processes (by increasing the amount of information required from the target in order to acquire it). Different metrics and models of clutter have therefore been inserted into models at different stages of processing.

By far, the most common way that clutter is modeled is in its effect on detection. It is important to note that a model predicting only P_d for an ensemble cannot determine in what stage of target acquisition (search, detection, recognition) clutter has its effect. However, if a local clutter metric can predict eye movements (e.g., Rotman, Kowalczyk, & George, 1994; Engel, 1977), then such a distinction may be made, even when P_d is the only performance measure. For example, if eye movements reveal that high-clutter scenes contain few fixations near the target, then the effect of clutter was on the search process; otherwise, it was on the detection process.

Ryll (1962) modeled the effect of clutter in terms of the probability of recognition within a fixation:

$$P_3 = \frac{1}{1+\left(\frac{M}{0.29t^{0.93}}\right)^{1.29}}$$

in which M = the number of "confusable forms" in the fixation and

t = the single glimpse time.

[20]It could be that the very term "clutter" with its negative connotation as a collection of undesirable things may be traced to the fact that in target acquisition, clutter is defined to be negative, a non-target.

As M increases, recognition performance drops. However, Ryll's instantiation of clutter may be incompatible (by itself) with metrics that model clutter's effect by increasing the average glimpse time because as search slows, performance improves[21]. A question is, of course, what factors determine whether an object in the scene is deemed confusable. In the original studies, observers "eye-balled" the scene to make this determination. In a recent model (VOM, Akerman, 1992, 1993b) the Ryll metric is incorporated with the number of confusable forms determined empirically by means of Waldman's clutter metric C_N.

Bailey (1970) instantiated the effect of clutter into the search portion of his model. Clutter (defined as a "scene congestion factor" ranging from 1 to 10) influences the probability that a target will be located during a glimpse:

$$P(t)_1 = 1 - \frac{1}{e^{-[\frac{700\, a_T}{G\ A_s}]t}}$$

in which a_T = target size,

A_s = search area,

G = scene congestion factor {1..10}, and

t = search time.

6.5.2 Conspicuity, Distinctness, and Attractiveness

Williams (1966) was the first to insert a metric for target conspicuity into a target acquisition model. His metric relates to clutter's effect on detection probability over time:

$$Pd = 1 - e^{-K_p t / A_d}$$

in which K_p = target conspicuity,

t = search time, and

A_d = display area.

Williams' K_p concept is a way of modeling the specific effect that clutter has on the number of fixations required to locate the target. Given an infinite amount of time, however, target performance will be perfect. Williams recognized that many factors would contribute to a single measure of the conspicuity of a target, but at the time, only psychophysical data and sophisticated metrics existed to describe luminance contrast.

Similarly to Bailey's (1970) instantiation of clutter, Williams' conspicuity metric slows search but does not determine the probability of eventually detecting the target. Such an instantiation of

[21]An example of a model that includes clutter at several points in processing is the VIDEM model (Akerman & Kinzly, 1979). Clutter was in so many places that Akerman removed some of its effects from his later VOM (Akerman, 1993b). See appendix A for details of VIDEM and VOM.

clutter requires an account for the known effects of clutter on detection and discrimination performance elsewhere in the model.

Pratt (1991) described several first order metrics of target distinctiveness. These metrics are based on various first order statistics of the gray-level representation of the scene. The metrics are based on the mean and standard deviations of gray levels across the target and the target's local background. Note that the various metrics, depending on how the background is defined, may be considered local, semi-local, or global (see appendix B for expressions and details of the metrics).

- Absolute average intensity difference,
- Root mean square (rms) intensity and target variance difference,
- Adjusted rms intensity and target variance difference,
- Absolute mean intensity plus absolute mean standard deviation (SD),
- Absolute mean intensity plus target SD,
- The Doyle metric (Copeland, Trivedi, & McManamey, 1996),
- The $\text{Doyle}_{\text{mod}}$ metric (Copeland et al., 1996),
- The *nrms* metric (Moulden, Kingdom, & Gatley, 1990; Kosnik, 1995).

First order metrics do not relate pixels to one another but are descriptors of the regions of the image in which the target and background exist. They lack any information about where different levels of luminance are with respect to each other. An additional class of first order metrics is the histogram and histogram intersection metrics. They are discussed later.

In addition to the previously mentioned first order metrics, there are metrics that take into account the spatial structure of the gray-level images rather than simply the distributions across a target or background area. These metrics are referred to as second order metrics. Metrics that take into account structure can begin to address issues related to specific information within a target, which, if present in a background, will lead to a decrease in conspicuity (and thus an increase in clutter).

One such metric that has been used in clutter and conspicuity metrics (e.g., Waldman, Wooton, Hobson, & Leutkemeyer, 1988; Rotman, Tidhar, & Kowalczyk, 1994; Tidhar et al., 1994; Rotman, Kowalczyk, & George, 1994; Copeland & Trivedi, 1996, 1998) is the gray-level co-occurrency matrix. This matrix represents, within an area of a pixilated image, the frequency of one gray-level occurring in a specified linear spatial relationship with another gray-level. The co-occurrency matrix, $P_{\Delta}(i,j)$, is a GxG dimension matrix in which G is the number of gray-scale levels in the image. It is defined by

$$P_{\Delta}(i,j) = \frac{1}{N}\sum_{k=1}^{N} f(x_k = i, x_{k+\Delta})$$

in which $(x_k, x_{k+\Delta})$ = a pair of pixels with gray-levels *i* and *j*;

i and j = gray-level values from 0 to a maximum, G, separated by

Δ = a displacement vector, which is a function of the distance, *s*, between the pixels and the angle θ between them.

f = {1 if x_k=i and $x_{k+\Delta}$=j, or 0 otherwise};

N = number of pixels in the area of the image.

Waldman et al. (1988) used the co-occurrency matrix to calculate a normalized clutter metric, C_N, which has been used in Akerman's VIDEM (Akerman & Kinzly, 1979) and VOM (Akerman, 1992, 1993b). C_N represents the degree to which the background texture is similar to the target in shape, size, and orientation. (See appendix B for the calculation of C_N.)

The normalized clutter measure is computationally demanding and makes some assumptions that may not be realistic when one is dealing with naturalistic images. It is symmetrical in orientation and size and assumes that as similarity between target and background texture elements decreases, clutter decreases uniformly. That is, texture elements different in size from the target by some amount will produce the same clutter (all other things being equal) regardless of whether the target or texture element is larger. The same assumptions are made for orientation; there is no absolute difference in orientation. These results contradict a phenomenon from perceptual psychology known as *search asymmetry* (Wolfe, Cave, & Franzel, 1989; Wolfe, 1994). Search asymmetry occurs when the reversal of target and non-target features results in drastically easier or more difficult searches. (For example, searching for a vertically oriented target among oblique oriented non-targets is much more difficult than searching for an oblique oriented target among vertical non-targets.)

Also, the C_N metric yields zero clutter if the background is uniform, regardless of the structure of the target. Such a result is obviously overly simplistic and points to a limiting case to which the metric may or may not decay gracefully as background uniformity increases. No literature regarding whether such gradual decay actually occurs has been found.

Similar to the normalized clutter metric is another metric based on the gray-level co-occurrency matrix: the texture-based image clutter (TIC) (Shirvaikar & Trivedi, 1992; see appendix B for details). Like C_N, the TIC metric depends on the size of target and background elements. However, unlike the linear weight given to transitions between gray levels as a function of the magnitude of their difference in C_N, TIC squares the difference, thereby giving more emphasis to larger disparities in luminance. According to the authors, TIC is only marginally better than C_N at extracting the meaningful structural information from the co-occurrence matrix.

Co-occurrency matrices are calculated one per displacement vector, Δ. That is, an image has as many co-occurrency matrices as there are positions between the target and background blocks. In order to overcome this inherent specificity, Copeland and Trivedi (1996, 1998) created a metric of target distinctness based on the average co-occurrency matrix (ACE) (see appendix B for details). This matrix is used in determining the distinctness of two patches of texture of a particular size. It is based on all possible displacement vectors in the texture model. In psychophysical tests involving the detectability of low-contrast geometric targets embedded in texture noise, the ACE metric was judged more accurate than either a first order Doyle metric or the target complexity metric, described next (Copeland & Trivedi, 1998).

Schmieder and Weathersby (1983) attempted to quantify the global clutter in an image by using a measure of statistical variance, SV (see appendix B for details). From the global SV, an SCR is calculated on the basis of absolute target contrast. SCR is then used rather than SNR as a predictor of detection in a cluttered scene.

The premise underlying the SV metric is the notion that the visual system is interested in areas of the scene with high gray-level variability. Unlike the second order metrics based on the gray-level co-occurrence matrix, SV is not concerned with the *structure* of the target or the background but only with its variance. As such, two perceptually different patterns could produce identical SVs. The theoretical justification for using the variance of the gray levels rather than a structure-based metric such as the co-occurrence matrix may have arisen as much from the lack of computing power in the early 1980s as anything else.

Schmieder and Weathersby (1983) found an orderly relationship between N50 for detection and the SCR,

$$N50 \cong \frac{1}{\sqrt{SCR}}$$

which was integrated into the Night Vision Model by Nichols and Paik (1993). The resulting increase in correlation between predicted and recorded detection performance as a function of clutter (from r^2=0.04 to r^2=0.64) was significant.

In evaluating SV and SCR in an urban environment, Cathcart, Doll, and Schmieder (1989) found that the metric underestimated performance compared to “rural” clutter. Such a result indicates that factors such as expectations and other sources of contextual scene information may be as important as image variance in determining performance in some situations. Birkmire, Karsh, Barnette, and Pillalamarri (1992) found that global SV was a poor predictor of overall search time. When SV was calculated for blocks of a display, Rotman, Kowalczyk, and George (1994) found that SV did not correlate highly with eye movements (i.e., fixations to regions of high clutter) in search.

Based on two assertions (that most targets tend to be more symmetric than non-targets and the visual system is able to efficiently detect regions of high local symmetry), Reisfeld, Wolfson,

and Yeshurun (1995) proposed a semi-local or global metric for eight-axis (circular) symmetry, CS_8 (see appendix B for details). Although the authors reported that the metric did a reasonable job of predicting near-target fixations for aerial views of symmetric ground targets, Rotman et al. (1994a) found that the circular symmetry did not perform well at predicting general human fixation behavior in a naturalistic scene. (That the model arose from the discipline of machine vision may indicate that it is better suited for locating man-made objects in general than for predicting human search performance.)

Tidhar et al.'s (1994) probability of edge (POE) clutter metric is founded on the idea that high spatial frequency edge information is important for the detection of targets. Related to this idea is the finding that the visual system seems to perform edge extraction early in visual processing, thereby creating a representation of the scene from which objects and surfaces can be readily extracted (Marr & Hilldreth, 1980; Marr, 1982; Nakayama & He, 1994; Biederman, 1987). Rather than extracting complete edges and treating them as elementary features, however, the POE metric (see appendix B for details) quantifies clutter by counting the number of edge pixels in sub-regions of the scene. Unlike the SV metric, in which sharp edges (i.e., regions with high luminance gradients) lead to a higher SV magnitude, POE merely counts the pixels. Like SV, however, POE relates only the amount of something rather than the structure of the image.

Unfortunately, also like SV, the POE metric fails to accurately predict response time (Birkmire et al., 1992) and fixation location during search (Rotman et al., 1994a). Presumably, a problem with the metric is that although edges of objects lead to edge-defined pixels, edge-defined pixels do not necessarily indicate the edges of real objects. Rotman, Hsu, Cohen, Shamay, and Kowalczyk (1996) evaluated a co-occurrency matrix-based clutter metric and the POE metric. The authors determined that the co-occurrency-based metric outperformed the POE metric in predicting observer false alarm responses. The stimuli used in the Rotman et al. (1996) study may have been biased more toward the co-occurrence matrix since they were "large targets, possibly camouflaged, where the texture of the target region is of crucial importance" (p. 673). As such, there may simply have been less information in an edge representation of the targets than in their internal texture-like detail.

Rotman, Tidhar, and Kowalczyk (1994) introduced the peak signal (ΔT) metric to describe the difference between average "temperatures" across clusters of pixels (though any intensity measure such as luminance will also work). (See appendix B for details of how the metric is calculated.) In averaging across the gray-scale image in order to form clusters, we must realize that all fine structural detail in the scene will be smoothed. (One input to the calculation is the minimum cluster size, and no group of pixels smaller than that size is permitted in the cluster representation.) An interesting aspect of this metric compared to other second order metrics is that it does not require knowledge of the target's structure; it concerns only the gray-level map of the image.

The authors found that the metric was a good indicator of human fixation performance in naturalistic scenes. No other evaluations of the metric were found. Toet (1996) has called the model too computationally expensive to be of practical use in his comparison of clutter and conspicuity metrics.

Another metric that purports to extract meaningful information about a target from a description of edge-based information is the target complexity (TC) metric of Tidhar et al. (1994). The metric is similar to the POE, but it adds the assumption that target objects will have more pronounced edges than interior details. (Defeating such a real-world property of objects is one of the goals of cryptic coloration in animals and camouflage patterns on targets, so the metric has a degree of face validity.) The metric is based on the cumulative distribution of difference of offset Gaussians (DOOG)-extracted edge points on the target and its immediate surround. (See appendix B for the rather complex description of this metric.)

Although Tidhar et al. (1994) determined that the metric did a reasonable job of predicting overall detection RT, the fact that the metric is only defined for a target and its immediate surroundings (usually taken to be twice the height and width of the target) leads to problems. For example, a target with a uniform local background will result in a measure of TC indicating a very simple search, even though the scene may contain much complexity that would cause performance to be quite poor. Grossman, Hadar, Rehavi, and Rotman (1995) used TC as a basis for calculating a signal-to-noise metric (analogous to the calculation of Schmieder & Weathersby's SCR) in order to model false alarms in cluttered environments. The authors found that the metric was as effective as either POE or SCR at predicting the trade-off between P(FA) and P_d. (That is, that they all made similar predictions for how subjects change their thresholds as clutter increases to produce more false alarms.)

A second order metric that incorporates both the concept of contrast and its ability to drive search performance and the fact that contrast as defined by a first order metric does not take into account the contrast variations along the boundary of the target, is the complex contrast metric, K (Lillesæter, 1993). Instead of modeling contrast as a function of maximum or average absolute difference between target and background regions, K includes a term for the integrated point-by-point contrast around the perimeter of the target. (The metric is defined in appendix B.) The U.S. Army Night Lab Static Performance Model for Thermal Viewing Systems (Skjervold, 1995) has incorporated this metric. Although the metric does not take into account target structure, that omission may not be important for its inclusion in a detection model.

The last class of non-empirical conspicuity metrics to be discussed is based on how the human visual system analyzes the scene with and without the target. These metrics will produce estimates of how distinct an observer will perceive the target to be within the context of the scene; they do not estimate the conspicuity of the target alone. The basic rationale for these models is that although the visual system may seem to pay attention to such things as complex contrast, the probability of edges within a region, the distribution of light and dark pixels, etc.,

visual processing does not occur on a pixel-by-pixel basis. Visual processing begins with an analysis of the scene akin to a Fourier analysis. As such, metrics should work from that point onward.

Information about portions of a scene can also be characterized in terms of their gray-level histograms (i.e., the rank-ordered gray value distribution of pixels in the portion of the image). Such histograms can be normalized by the division of the level of a gray-level bin by the fraction of pixels that have that value. Image regions that appear visually similar should have similar normalized histograms. Since the normalized histogram is a first order metric and conveys no information about the internal structure of a region of the image, the converse is not necessarily true; regions that have identical histograms may have dramatically different appearances. Also not necessarily true, though usually the case in reality, is that two image regions appearing visually different (e.g., containing a target and not containing a target) will have different normalized histograms. Conspicuity metrics based on the normalized histograms of images determine the degree of histogram overlap by a logical intersection of target and background histograms. A greater degree of overlap indicates less conspicuity (see appendix B).

The Camaeleon model (Hecker, 1992; see appendix B) calculates normalized histograms not on the raw gray-level representations of images but on images convolved with band-pass filters. Regions of the scene are designated target and background, and after band-pass filtering, normalized histograms are created for the local energy (based on chromatic or achromatic contrast), spatial frequency, and orientation of each region. The degree of camouflage (analogous to magnitude of clutter, or the inverse of conspicuity, but bound on [0,1]) is defined as the product of the intersections of all target and background histograms. The main shortcoming of this metric, of course, is the fact that it is uninterested in structural details of the target, and thus may judge a target to be well camouflaged when it is not!

Another detectability metric based on neurophysiology is Watson's (1987) Cortex Transform (see appendix B for details). This metric is based on a multi-channel-oriented spatial frequency analysis of an image adjusted by a contrast sensitivity function. It is called the cortex transform because it mimics the oriented edge detection of area 18 (V1) of visual cortex. Two images, one of a scene containing the target and one without, are first converted to luminance contrast images and then subjected to the cortex transform. The result of the transform is a four-dimensional representation of the scene, with each of 20 or 24 channels (five or six frequencies at four orientations each) weighed at every point (i.e., the four dimensions are x, y, frequency, and orientation). The difference between the strengths of the target and no-target components is the component's contribution to the overall distinctness. Masking is implemented in the metric when the distinctness component is reduced by a factor related to the component's background signal strength. The distinctness of the scenes is determined by the Minkowski sum of the coefficients.

The cortex transform, when masking is implemented, has been shown to be a good predictor of human detection performance for low-contrast scenes (Ahumada & Beard, 1996; Rohaly, Ahumada, & Watson, 1997). Without the masking term, performance tends to be overpredicted. The cortex transform is an elegant implementation of known early visual physiology and psychophysics in that it integrates inter-channel masking and known contrast sensitivity functions and is based on human and animal physiology. However, it is a predictor of pure detection in static, achromatic scenes, so its current usefulness is limited.

Also relying on the assumption that differences in individual oriented spatial frequency channels constitutes a distinctness metric from which the detectability of a target can be determined is the Perceptual Distortion distinctness metric of Martinez-Baena et al. (Martinez-Baena, Fdez-Valdivia, Garcia, & Fdez-Vidal, 1998; Martinez-Baena, Toet, Fdez-Vidal, Garrido, & Rodriguez-Sanchez, 1998). Like the cortex transform, the metric involves a spatial frequency decomposition. However, the distinctness metric is based on changes registered only in the few channels that provide the principal structural components of the image.

The image is first decomposed into radial spatial frequencies representing distinct structural components of the image. The relative contributions of each band (wavelength and orientation) to the overall image structure are computed, and the principal components are identified. Then a set of oriented Gabor filters is applied to the image, based on the principal components. Finally, a difference metric is created on the basis of a combination of the differences of filter output on the images containing and not containing a target.

The metric was evaluated against a set of field images taken during the DISTAF (distributed interactive simulation, search and target acquisition fidelity) field test at Ft. Hunter Liggett, California, in 1995 (Toet, Bijl, Kooi, & Valeton, 1997; see reference for information about acquiring image set) in which nine vehicles were deployed at various locations. Scenes were digitized still photos. To evaluate the model, the authors digitally removed the target from each scene and applied the metric to the images with and without the target. The resulting distinctness metric was then compared to an empirical metric of distinctness by Toet and colleagues (described next). The empirical and calculated distinctness correlated highly ($r = 0.81$). The calculated metric also correlated highly with response time to detect the target in the scenes ($r = 0.82$). These results indicate that the distortion-based metric may be a good overall indicator of what subjects use to guide their search for a target in a static scene. Like many of the metrics in this section, the distortion-based distinctness metric is achromatic and concerns only static scenes.

6.5.3 An Empirical Measure of Conspicuity

Toet (1996) and Toet, Kooi, Bijl, and Valeton (1998) described an empirical method for determining the conspicuity of a target in a scene. They used Engel's (1977) operational definition of conspicuity as being the peripheral area around the center of the visual field from which specific target information can be extracted in a single glimpse. This definition is obviously similar to the concept of a visual lobe. Toet and colleagues define detection conspicuity and identification conspicuity as the maximum distance between the target and fixation that permits the respective level of acquisition.

Toet and colleagues assert that it requires only a small number of subjects to perform a psychophysical experiment on a scene and its target in order to determine conspicuities consistent across a large group of observers. The results of Toet et al. (1998) indicate that two experienced subjects are able to determine conspicuity measures that accurately predict overall search performance (response time to detect a target) for a group of observers viewing the same stimuli. The agreement between conspicuity and response time is a good indication that the measure may serve as an efficient and effective means of determining conspicuity.

Such an empirical method may be of more use in future laboratory-based investigations of conspicuity than in the prediction of performance for scenes encountered in real time. The authors have in no way determined lawful or predictive relationships between characteristics of the scene and the target and conspicuity as empirically measured. On the other hand, their relatively simple empirical method allows accurate measures of conspicuity to be extracted quickly, thus making a factorial investigation of scene features and their role in conspicuity feasible.

6.5.4 Other Clutter Issues

Related to the idea that discrimination may require the extraction of specific target features is the possibility that clutter is perceptually masking such target features. Legge and Foley (1980) and Tolhurst and Barfield (1978) demonstrated the contrast necessary for the detection of a sine wave grating when it was accompanied by a nearby masking grating of a similar frequency and orientation. Given that high spatial frequencies contain information about edges and fine detail, background elements of similar frequency and orientation to target features may make them less visible. Masking is difficult to measure since its 2-D characteristics are as yet unknown (see Olacsi & Beaton, 1998). However, implementing masking into a spatial frequency-based model of vision or target acquisition has been accomplished successfully in the cortex transform.

Although clutter can dramatically influence performance, there are some visual events that are known to "cut through" the clutter: visual transients and motion. These visual events have a temporal character that is absent from static visual clutter. As discussed in another section of this report, motion has long been known as a feature to which the human visual system can readily attend. Kosnik (1995), in particular, found that search for a moving target was nearly as easy when the target is viewed on naturalistic terrain as a uniform background. Likewise, transient visual events as used in laboratory studies are not only easy to see but may also be effectively impossible to ignore (e.g., O'Regan, Rensink, & Clark, 1999). If a target is known to be associated with such a visual event, clutter will not play nearly so vital a role in acquisition.

6.6 Models and Metrics Based on Human Visual Physiology/Psychophysics

Models based on human visual physiology and psychophysics focus their attention on how the human visual system processes actual scene information, rather than on how overt performance may be related to scene variables such as clutter. These models are of interest because their goal is to predict human performance for any situation in which an observer attempts to acquire a visual target. As such, a model should inherently be able to address such factors as sensor type,

number of targets, moving or stationary target, presence of obscurants, level of clutter, etc. These factors are not of separate interest since a model should be able to compensate for them by virtue of the fact that it is an accurate depiction of human visual processing and decision making.

For this report, the broad class of these models can be described as lying along a continuum from psychophysical but non-physiological all the way to highly physiological and predictive of psychophysics. All the models attempt to model early human visual properties. However, some go about it more by processing information in stages related to closed form expressions of psychophysical performance or by appealing to psychometric functions to determine human perception of stimuli. Others approach it by processing information based on stages corresponding to the transformations that information in the visual system undergoes during vision. Neither style is necessarily better or worse than the other, so long as (a) the physiology agrees with the psychophysics, and (b) the physiology and/or psychophysics are well understood enough that a broad class of phenomena can be modeled. This discussion will begin with highly psychophysical models and move to more physiological models.

6.6.1 The British Aerospace ORACLE Model

The ORACLE model from British Aerospace (Overington, Brown, & Clare, 1977; Cooke, Stanley, & Hinton, 1995) attempts to model search, detection, and discrimination performance for a human observer. (See appendix A for details of the model's operation.) The model is based more on known psychophysics than on the physiology underlying the psychophysics. An important note about ORACLE is that it is modular and proprietary, and no full implementation of all the modules is known by the author of this report to exist outside British Aerospace. The documentation available for this report concerns search, detection, discrimination, and clutter in an achromatic image only.

ORACLE bases its predictions on the retinal image of the target and how the visual system responds to the retinal image. The primary assumptions behind the model are (a) the edges of a target are more important than the target's total energy in determining detectability, (b) discrimination is a function of the visual system's ability to distinguish between two adjacent features of a target, each of which is approximated to be half the target size, (c) signal strength must exceed a noise strength in order for a detection or discrimination to be made.

Much of the model's detail is concerned with how the non-linearity of eye optics and the modulation transfer function of the cornea determine the point spread function of the eye. Images of known resolution, contrast, and sharpness are then subjected to this function and retinal images are produced. The sum of the activity and the gradient of the responses of adjacent photoreceptors constitute the basic signal of the target.

ORACLE models search as a soft-shell process. Fixation locations are selected at random with replacement. Glimpse time is a constant 1/3 second. If the target lies within the soft shell lobe, then acquisition can occur. The lobe size is modeled as a distribution of hard shells and may

change throughout a trial. The effect of clutter in the model is to influence the distribution of lobe sizes to favor the selection of smaller shells. That is, clutter makes search less efficient because less of the image area can be searched at a time.

An important aspect of ORACLE is its ability to equate its distinction definition to the Johnson (1958) criteria embodied in so many other models. It does so by Fourier decomposition of a Johnson-like bar pattern into a fundamental and several odd sinusoids and determines whether ORACLE can distinguish between the component spatial frequencies at a given resolution, contrast, and size.

Although the model available to this reviewer did not incorporate color, Cooke et al. (1995) mentioned that such a version of ORACLE does exist. Its implementation is based on color opponency between R and G cones only. Although it is unclear how such an implementation of color processing can be a reasonable facsimile of the human visual system, the model seemed to do well at a laboratory color distinctness task. The visibility (signal strength relative to clutter strength) of colored shapes on a colored background was judged by the model to correspond highly with human judgment of the conspicuity of the same colored stimuli. Insufficient detail of the study and the implementation of the model were provided to evaluate this claim, however.

Though the model's various steps in processing the image from the outside world (e.g., display or sensor or optical device) through optics, photoreceptor anatomy and physiology, adaptation and luminance effects, and contrast sensitivity functions are all based on well-documented psychophysics, the model as a whole has not been evaluated against what Cooke et al. consider a set of images sufficient to test it *in toto*. Some caution is urged before such an evaluation, especially at the limits of the known psychophysics. Models such as this likely become less accurate as the stimuli on which they are based approach the limits of the psychophysical measurements used to develop the models. Overington (1982) pointed out that models based on psychophysics have specific "envelopes of usage" where their predictions are accurate. Outside such envelopes, error propagates from step to step in calculation, resulting in a potentially dramatic degradation in overall performance.

A more serious shortcoming of the model is that its firm foundation in psychophysics has made the integration of top-down (i.e., observer) factors extremely difficult. Currently, there are no such factors in the model, probably because the psychophysics behind the effects of training, attention, stress, etc., often involve setting a decision criterion or a processing speed rather than changing the shape of a psychometric function. Since there is no single objective set of data relating observer variables to psychophysics, the authors have taken the conservative route and omitted it entirely.

A related shortcoming is the fact that the model processes information in a single stream from image to retina to signal to response. There is no operation that takes into account goal-directed (top-down) or stimulus-driven (pre-attentive or bottom-up) information. A manifestation of this shortcoming is in the assumption that fixation location is random, as opposed to guided by interactions of low- and high-level processes (e.g., Wolfe, Cave, & Franzel, 1989; Wolfe, 1994;

Wolfe & Gancarz, 1996; Doll et al., 1998). The authors readily admit that this assumption is unrealistic and that eye movements tend toward target-like portions of the scene, but they argue that "...the effort in modeling an equivalent level of detail is far greater than the reward for many situations" (Cooke et al., 1995, p. 167).

Motion is incorporated into ORACLE only in terms of looming motion (i.e., motion directly toward the observer). Such motion is modeled as an increase in target contrast and size, from which an increased signal will occur. However, such a gradual increase in signal strength may not account for the particular salience characteristic of such stimuli.

6.6.2 The Georgia Tech Vision (GTV) Model

The GTV model by Doll, McWhorter, Schmieder, and Wasilewski (1995), Doll, McWhorter, Wasilewski, and Schmieder (1998) and its military counterpart, visual/electro-optical (VISEO) by Doll et al. (1997) are general purpose models of human vision that can be used to model search and detection in dynamic, cluttered scenes. Because the models are intended to be true to the human visual system, they are based more on human visual physiology than on ORACLE. The optics of the eye, as well as retinal and cortical areas V1 (edge detection), V4 (color processing), and MT (motion processing) are integrated into the model's processing. The physiology must, of course, produce the same psychophysical functions that underpin ORACLE. However, the authors chose to be more general in order to handle situations that do not agree closely with existing psychophysical findings. (The model is detailed in appendix A.) Much detail is provided in the text of this report because GTV comes closest (in this author's opinion) to integrating what is known about the spatial frequency aspect of early vision with what is known about the phenomenology of visual search and attention.

GTV is quite complex and incorporates many aspects of visual processing. The primary processes of interest include a multi-channel-oriented SF model of feature extraction, texture-based scene segregation into object-like "blobs," and parallel pre-attentive and attentive modules to calculate two probabilities for locations in the image: the probability that a blob will be the target of fixation (P_{fix}) and the probability that, once fixated, the blob will be detected ($P_{yes|fix}$). A neural network learning algorithm determines the features that are to be stressed in determining these probabilities. Signal detection theory is then used to determine whether a blob will be determined to be a target. Search proceeds by the selection of locations that have high P_{fix} without replacement and determining if a decision is to be made based on $P_{yes|fix}$. Outcome measures of the model are P_d, P(FA), d′, and RT.

In more detail, GTV consists of five modules: (a) a front end, (b) a pre-attentive module, (c) an attentive module, (d) a selection and training module, and (e) a performance module. GTV is similar to Wolfe et al.'s Guided Search model in that parallel pre-attentive processes extract peripheral feature information that is used for eye movement guidance. Concurrent with pre-attentive processing, an attentive process extracts foveal feature information that is used for discriminating between clutter and a target.

- Front end processing in GTV concerns retinal factors such as pigment bleaching, pupil size, flicker, and transient luminance changes. Color information is converted from responses of the three photoreceptors to responses on two (R/G and B/Y) opponent process pairs and an average achromatic cone luminance signal.
- The pre-attentive and attentive processing modules in GTV use sets of filters tuned for peripheral and central color, temporal, and spatial sensitivities to extract features (e.g., motion, orientation, and spatial frequency) from the image. Motion information in the image (sampled at 30 Hz) is filtered to produce a scalar local motion signal and integrated to add blur to the image. Each module has a pattern perception unit that decomposes the temporally integrated spatial information into 24 frequency and orientation selective channels. More spatial information comes from the cone luminance than the color opponency, in agreement with psychophysics. Interactions between the channels are simulated to incorporate spatial masking. Finally, a second order texture metric is calculated and blobs (regions of different textures, corresponding presumably to object-like regions of the image) are segregated from the background. Features in SF and orientation domain are assigned to the blobs for their region.
- The selection/learning module takes the feature loadings on the blobs from both the pre-attention and attention blob map and assigns weights to them, based on the state of a neural network that has been trained (or not) to look for a specific target. This module is intended to mimic the ability of a human to improve in performance of a task that is initially quite difficult (i.e., to switch from controlled, conscious processing of sensory information to automatic processing [Schneider, Dumais, & Shiffrin, 1984]).
- The performance module determines blob P_{fix} and $P_{yes|fix}$, and simulates a search process to determine P_d, P(FA), d′, and RT for a trial. P_{fix} for each blob is based on a noisy decision process that takes into account the weights on the relevant features of blobs as well as noise (quantum and neural for near-threshold stimuli), clutter (defined as “the extent to which another blob’s luminance, texture, chromatic information, and temporal contrast match the current blob”), and the spacing of other blobs nearby.

We determine $P_{yes|fix}$ and RT by first calculating the SCR for each blob in the image. The SCR is taken to be equivalent to an effective d′, which in turn determines $P_{yes|fix}$ for a blob. Assuming that search progresses without replacement from highest P_{fix} to lowest and that search occurs at a constant rate, then search for a trial can be modeled. The RT to a decision (either a false alarm or a detection) is determined by the number of blobs that will be encountered before a decision is made.

6.6.2.1 Comments

The model is interesting in that it incorporates many human physiology and perceptual psychological principles. However, there are serious issues related to learning and to motion

processing. Learning is assumed to be the selection of features and combinations of features indicating the possible presence of a target among clutter. The processing, especially in the pre-attentive module, is meant to mimic the function of learning a task so well that it can be done "without thinking" (i.e., automatically [Schneider et al., 1984]). After sufficient training, GTV can perform even quite difficult searches with ease. The problem with the implementation of learning is that *any* combination of features can be learned pre-attentively—a phenomenon that cannot occur in humans. (For example, performance in a rotated T/L discrimination task will never become automatic even after tens of thousands of trials [Wolfe, 1998].) Some features cannot be processed in parallel pre-attentively but require focal attention (Wolfe & Bennett, 1997; Rensink et al., 1997). The authors acknowledge that after training, noise needed to be added to the input images in order for the model not to outperform humans (Doll et al., 1998).

Motion is included in the model. However, the temporal filtering only adds a scalar motion feature to blobs in the image. Because motion information is scalar (only related to speed, not direction), the model's attention mechanism has no direction selectivity as the human visual system has. Therefore, GTV can only distinguish between speeds. This does not allow the system to extract information about motion parallax (e.g., how a moving target's violation of parallax may be plainly visible).

Other, more minor issues relate to assumptions made about when the model calculates some quantities and how it operates to make a decision. The calculation of all foveal features at the same time (by attention module) is not physiologically realistic. The model would be more realistic and behave identically if it were to calculate the foveal features only after a blob is selected by the performance module. (This behavior takes into account the unbound feature aspect of pre-attentive vision by Wolfe & Bennett, 1997.) Also, foveation of a target is required for a "yes" decision to be made. Even though the model is ostensibly based on the conspicuity of targets, highly conspicuous targets must still be fixated for the model to produce a "yes" response. This result is inconsistent with "pop-out" (i.e., rapid search largely insensitive to the number of distracting elements).

6.6.3 The Wilson (1991) Spatial Vision Model

The basic assumption underlying Wilson's (1991) model is that at the detection and identification threshold, information from only a small number of spatial channels that are most sensitive to the target determines performance. This assumption makes intuitive sense since a signal in the visual system from the target will naturally be carried by those channels most responsive to the target. The interesting aspect of the theory comes from the idea that *decisions are based* on the output of these few most active channels. The model is based on results from human and non-human primate psychophysical and physiological experiments, indicating that spatial tuning of six mechanisms comprises the behavior of the primate retino-geniculate-cortex (V1) pathway.

The six mechanisms correspond to different spatial frequencies. Lower frequency mechanisms corresponding to coarser grain details are selective to fewer orientations; higher spatial frequency

mechanisms are sensitive to a greater number of mechanisms. The locations on the retina that correspond to the different mechanisms also differ, with higher frequency mechanisms at smaller eccentricities than lower frequency mechanisms. In addition, the filters have different contrast sensitivities, consistent with the contrast sensitivity functions of humans. (See appendix A for a table describing the filters comprising each purported mechanism.)

Although the Wilson model of spatial vision is general purpose in nature, it does have implications for the thresholds required for the detection and identification of targets in real-world scenes. The model assumes that the degree to which a target can be acquired depends on the response of the six spatial mechanisms to the target image. More to the point, the model assumes that a few highly selective filters are the ones that determine the detectability of the target. If two different targets stimulate these basis channels identically, then they will be identified as the same target, and discrimination between them will not be possible. In fact, additional information at other spatial frequencies will not permit discrimination because the information is not present in the filter responses that go into the decision.

In order to test Wilson's spatial model, Thomas and Barsalou (1995) determined whether a target with sufficient contrast to be barely detectable or identifiable will be perceived differently if information is added to non-basis filter channels. The authors used images of B-1B bombers and analyzed them with a set of filters corresponding to Wilson's model. The three most active channels were identified and a new image consisting only of information on these channels was created. Subjects judged the two images as identical, indicating that the decisions seemed to be based on these channels alone.

MIRAGE (Watt & Morgan, 1985) and MIDAAS (Kingdom & Moulden, 1992) are not models of target acquisition per se but are models of how physiological processes can extract meaningful feature information from a scene. Both models concern one-dimensional stimuli only. The image is sampled at all locations at four different spatial scales. The output of the filters at the different scales can only be interpreted as an edge or a bar. The central difference between the two models lies in how the information from the different spatial scales is combined. In MIRAGE, the responses of all the filters are combined before they are interpreted; in MIDAAS, the filters are first interpreted, and then their interpretations are combined across scales. The scale dependence of MIDAAS is viewed by the authors as an asset since it provides for more than one possible interpretation of the scene.

6.6.4 The Limits of Direct Access Spatial Frequency Models

Models such as Wilson's (1991) Spatial Vision Model assume that detection and discrimination decisions are based on output from a single set of tuned pathways. In such models, the only difference between detection and discrimination arises from how information from those pathways is used. Models based on this assumption (rather than an assumption that different basic operations provide information to detection and discrimination stages) are referred to as

"direct access multi-channel models" (Olzak & Thomas, 1992). The authors examined four assumptions inherent in this class of models in terms of discrimination performance:

1. The observer has direct access to the output of the channels.
2. The observer can selectively attend to a subset of these channels.
3. The pathways are independent of one another. (Mathematically, they are independent Fourier components.)
4. Information from the pathways is integrated probabilistically in order to determine the presence or absence of information in the image based on the channel activations.

Unfortunately, these assumptions do not withstand scrutiny well. Olzak and Thomas (1992) demonstrated that the channels were not independent by cueing one channel and measuring effects in other channels. Verghese and Pelli (1994) and Lamb and Yund (1996) found that observers are quite poor at selecting a scale bandwidth to attend to and search, indicating that at least consciously, selection of individual channels is limited. There is some evidence that lateral masking of spatial frequencies can occur and that they are not restricted to within-channel frequencies (Ackerman, 1993a). Finally, Thomas and Olzak (1990) found that integration of disparate bandwidths was worse than integration of similar bandwidths.

Similar to the Wilson (1991) model is the physiological saliency-based models of Itti and Koch (2000). The underlying premise of the model is that an observer directs his or her gaze at the most visually salient location in the currently visible retinal image. Performance in the model is based on eye movements to successive points of high salience in a scene, with this saliency represented as a spatiotopic map of the scene.

The Itti and Koch (2000) model determines the saliency of locations of the retinal image through a multi-feature, multiple scale scheme based on known visual psychophysiology and psychophysics. The extraction of early visual features takes place at nine spatial scales for each of three features: luminance intensity, color, and orientation at four orientations: 0, 45, 90, and 135 degrees. Extraction at each location is performed by simulated center-surround excitation-inhibition regions akin to the known physiology of early cortical visual processing. Each set of multiple scale feature maps creates one feature conspicuity map by means of competition between areas of high activation within each feature. This competition takes the form of large spatial scale inhibition corresponding to the behavior of so-called non-classical receptive fields present in visual cortex (Gilbert et al., 1996). The three conspicuity maps are then combined into a single saliency map by means of linear combinations, the relative weights of which are determined empirically, based on model performance, and then fixed as constant.

The model posits a "winner-take-all" process so that the next fixation location is determined by the location of highest activation in the saliency map. After simulated saccade selection takes place, the area of highest saliency is temporarily inhibited (for approximately 500 to 900 ms) so

that it is not immediately selected as the next fixation location. This inhibition instantiates the previously mentioned IOR effect widely demonstrated in perceptual psychology.

Although the Itti and Koch (2000) model is explicitly "bottom up" in nature[22], the authors assert that with proper selection of weights, the model can be applied "hands off" to a variety of visual search situations. These weights include the relative weight given to specific values of features (e.g., to a particular orientation or a particular color), the relative strength of features in the calculation of the salience map, and the temporal characteristics of the simulated search (e.g., dwell time, frequency of saccades, duration of IOR). During evaluation of the model (described next), the authors found a single such set of these characteristics and ran the model on a variety of scenes ranging from simple and conjunctive visual search tasks to search for military vehicles in the DISTAF image set.[23]

Overall, the authors report that the model showed "reasonable results" (Itti & Koch, 2000, p. 1497) across a variety of scenes ranging from simple search to artistic paintings to outdoor scenes. Although it is notoriously difficult to empirically evaluate a set of saccades, the time and number of saccades required for the model to generate a fixation close enough to a target to acquire it may be objectively compared to human search for targets in the same or similar situations. The model was successfully able to produce pop-out effects for simple feature searched and slower search (with number of saccades increasing as a linear function of number of distracting elements) for conjunctive search. Thus, for these simplified scenes, an entirely bottom-up search strategy may be sufficient to explain human behavior.

The model fared less well when compared to human performance searching for military targets in the DISTAF image set. After some changes in the temporal dynamics of search to better match average human characteristics such as saccade frequency and latency (recall that the Toet et al., 1997, human performance data set did not contain information about eye movements but only response times to locate the target), the model was able to detect the targets adequately and in far fewer saccades than would be required if fixations occurred at random locations. However, both the overall response time required to locate the targets and the pattern of scene difficulty as determined by human response time rankings were quite different between the model and the human data. Specifically, although scenes that required more time for humans to detect the target also required more time for the model to detect the target, the correlation is extremely weak (it was not mentioned in Itti & Koch, 2000). In addition, there was significantly more variability in human response time across scenes than there was in model response times,

[22]The authors write, "Our model is limited to the bottom-up control of attention, i.e., to the control of selective attention by the properties of the visual stimulus. It does not incorporate any top-down volitional component" (Itti & Koch, 2000, p. 1492).

[23]Note that when this author refers to "search" for a target, it is not intended to imply that the model actually had a goal of finding a particular target. Rather, performance was judged on the basis of overall pattern of simulated saccades which, eventually, fell close enough to the target for it to be acquired.

and in 35 of the 44 scenes evaluated, the model was able to detect the target in many fewer fixations than humans could.

In order to account for this lack of agreement between human and model performance, Itti and Koch (2000) noted the differences between the task set for human participants in the Toet et al. (1997) study and those set for the model. Specifically, the participants were trained in the appearance (from three vantage points) of all possible military targets before they viewed the DISTAF images. Itti and Koch (2000) assert that, given the difficulty of many of the searches[24], the goal-directed knowledge of possible target identity possessed by human participants may have biased them toward poorer performance by continually drawing their attention to areas of the scene "in inappropriate ways" (Itti & Koch, 2000, p. 1502).

Parkhurst, Law, and Niebur (2002) modified the Itti and Koch (2000) model to add a more realistic decrease in peripheral contrast sensitivity. More importantly, their study included the collection of eye movements for human observers viewing the same scenes to which the model was subjected. Similar to Itti and Koch (2000), the tasks in the current study did not include visual search for a target. Rather, participants were told to "look around at the image" for the 5 seconds of each trial (Parkhurst et al., 2002, p. 112). The model was evaluated in terms of how well its predictions of locations of high scene salience correlated with observer fixation locations.

Results indicated that stimulus-based saliency predicted a significant proportion of variance in fixation location variance, with strongest correlation occurring early during scene presentation. That is, when scenes were first presented to observers, the early fixations were better predicted by the model than were later fixations. Nevertheless, the saliency-based model continued to produce significant correlations between predicted and observed fixations throughout the trial. These findings are consistent with the notion of gist extraction (Intraub, 1981), as described earlier during discussion of the Nicoll and Hsu (1995) results. Specifically, in search tasks, the first few hundred milliseconds of viewing a scene may be consumed by the extraction of overall spatial layout and schematic information from the scene (not by the active search for a target). The saccades required to extract this information, which take place by definition before there is any high-level cognitive representation of scene content, are likely guided by local scene salience. Only later do top-town aspects of gaze selection come into play. Since the Parkhurst et al. (2002) tasks did not involve search, this initial stimulus-based guidance of eye movements may have been extended.[25]

---

[24]Itti and Koch (2000) omitted the eight most difficult of the 52 DISTAF images because the model or the human participants were unable to detect reliably within a 10-second window.

[25]Note that Parkhurst et al. (2002) also found that observers showed a bias toward fixations near the center of the scene, particularly in early fixations. This finding may correspond to "orienting" in the scene before saccades that support gist extraction.

Contrary to the Parkhurst et al. (2002) findings of significant stimulus-based influences throughout all fixations in a trial, Turano, Geruschat, and Baker (2003) showed that the Itti and Koch (2000) model failed to predict fixation location above chance levels in a specific goal-directed task unless goal-directed information was inserted into the model. Participants in the study were asked to navigate an unfamiliar hallway and to "walk through the third door on the left" while wearing a head and eye tracker. Recorded fixations were compared to those predicted by (a) an unmodified Itti and Koch (2000) model, (b) an Itti and Koch (2000) model weighted toward target features (vertical orientation and large spatial scale), (c) an Itti and Koch (2000) model weighted toward target location (the model was restricted to making fixations only on the left side of fixation), and (d) an Itti and Koch (2000) model weighted for target location and features.

Analysis of model predictions and observed fixation locations was different from that done by Parkhurst et al. (2002) in that fixations were not assigned (x, y) coordinates but were assigned to regions of the scene, based on contiguous surfaces or objects. Fixations were thus turned into a series of categories visited by observer and model predictions. These sequences of categories formed the data to be correlated.

Results indicated that the unmodified Itti and Koch (2000) model and the model weighted for target features performed no better than chance at predicting the regions of the display fixated. The model weighted for target location, however, performed better and predicted 35% of fixation regions. The model incorporating both location and feature weighting fared best, predicting nearly 48% of fixation regions. Together, these results show that (at least for simple goal-directed behaviors such as walking toward a target) bottom-up and top-down information is required for a model to be able to predict human fixation performance.

---

7. Other Topics of Interest, Not Previously Addressed

7.1 Perceptual Psychology

In considering what would make a good model of target acquisition, one can take the point of view that it would be an application of a model of basic vision or basic visual performance to a situation in which the observer seeks a target. The perceptual psychology community has long been interested in these basic models and in the basic properties and processes underlying human vision. It is this author's opinion that target acquisition models should attempt to incorporate as many of these basic principles as possible in order to be flexible and robust. As such, this section of the report discusses aspects of vision and visual perception gleaned from the perceptual psychology literature, which have bearing on target acquisition. The section includes discussions of color vision, motion perception, and the effects of visual transients.

7.1.1 Color Perception

Color perception is a key aspect of human vision. In order to account for how humans perceive color, any model must incorporate the following factors: luminance, eccentricity, and target-background luminance contrast. Color perception is a function of the cone-type photoreceptors, which are sensitive to light only in the photopic range of luminance. During low-light conditions, the cones do not respond and all vision is achromatic. Cones are concentrated at the macula (the center 1 degree of the retina) and decrease in density quickly with eccentricity; thus, good color vision is afforded only for foveated targets. (These two factors interact in that during low-light conditions, the poorest acuity will be at fixation.) If the target and its background support have the same luminance and differ only by color, the target will not stand out clearly, and its motion (if it is moving) will be difficult to perceive. In addition, a considerable fraction of the male population suffers from one kind or another of congenital color blindness, indicating that consideration of an impaired population may be justified in considering a general purpose model.

Color is processed in the human visual system by three types of photoreceptors, each receptive to a broad range of wavelengths. These three photoreceptors are interconnected in the retina by bipolar and horizontal cells and innervated ganglion cells representing combinations of excitatory and inhibitory center-surround pairs of red-green and blue-yellow sensitivity (Zeki, 1993). Substantial differences in ganglion cell anatomy and physiology between color-sensitive and luminance-only sensitive neurons result in psychophysical differences between human color vision and non-color vision. (See Zeki, 1993, for a very readable overview of visual neurophysiology in general and color vision in particular.)

Most models of target acquisition tend not to address color as a driving factor in performance. (Models of low observable [LO] targets and camouflage, such as CAMELEON[26], do, but they are the exceptions.) This lack of consideration in the modeling literature likely arises from two basic facts: (a) the enemy would be foolish to send an oddly colored target into a battle since an object's color is relatively simple to change to fit an environment, and (b) electro-optical sensors such as I^2, synthetic aperture radar (SAR), and FLIR have historically used non-color displays, so any color in visible light would be lost. However, with the advent of fused sensor systems, full-color I^2 devices (image intensifiers that use more than single-wavelength phosphor), and false color FLIR, it would seem that color may become more important in the future of target acquisition modeling.

When one is considering color in perceptual psychology, there are some issues relevant to target acquisition modeling efforts: (a) detectability under equiluminance or near-equiluminance, (b) how color space is to be represented, (c) what levels of target acquisition are aided by the presence of color information, and (d) how color contrast or salience can be defined. These topics are interrelated to a certain degree.

[26]CAMELEON stands for camouflage assessment by evaluation of local energy, spatial frequency, and orientation (Hecker, 1992).

Vision for colors displayed at equiluminance (i.e., figures differing from their background by hue alone) is known to be quite poor. Theeuwes (1995) showed that search for a newly displayed stationary target whose color differs from its background is very slow unless the target also differs from its background in luminance. When luminance differences are sufficiently large, the target will become readily apparent, even though its color may not be known in advance, thus demonstrating that color differences can drive attentional selection[27]. However, search for a target of unknown color, even when its luminance is substantially different from the background, may be quite difficult if other elements of the scene are also uniquely colored. That is to say, if the observer is looking for a target on the basis of color, he would have more difficulty finding it if it is not the only object with a unique color in the scene (Bacon & Egeth, 1994; Theeuwes & Burger, 1998; Duncan & Humphreys, 1989).

In certain circumstances, color may be able to reduce clutter effectively. Clutter, as defined by the number or density of confusing non-target objects within a scene, may be reduced if non-targets are known to be of a different color than the target. Egeth, Virzi, and Garbart (1984) demonstrated that non-targets of a particular color do not influence search response time if they are of a color that is sufficiently different from that of a known target. Humphreys and Muller (1993) incorporated this factor into their SERR model of search, as discussed earlier. This conceptualization of clutter has a certain circularity about it since if an object in the scene is confusable with the target, it is clutter; if it is not, then it is not clutter. If the observer is aware of the color of the target ahead of time, then it may be argued that the differently colored non-targets do not represent clutter. Clutter metrics that do not take chromaticity into account would not be able to incorporate this ability of the visual system.

Motion detection is also quite poor during conditions of equiluminance. Cavanagh and Anstis (1991), Kooi and deValois (1992), Ramachandran and Gregory (1978), and others have demonstrated that objects defined only by color are difficult to detect. Kooi and deValois argue that the neurophysiology of the parvocellular ganglion cells that carry color signals from the retina to the cortex, as well as the cortical projections themselves, account for this lack of color-based motion perception. Color information is sent through different parts of cortical areas V1 and V2 to V4, where largely motion-insensitive color processing occurs. Non-chromatic motion information, on the other hand, is processed in the medial temporal (MT)[28] area. Near equiluminance, however, motion perception quickly recovers as the luminance difference between target and background increases. Since most target acquisition situations involve targets that would be close to equiluminance and similar in hue to their backgrounds (presuming

[27]The uniquely colored item may have *lower* luminance contrast than the non-targets. It need only be different in some way for its color to become important.

[28]The underlying logic of this segregation was postulated by Mishkin, Ungerleider, and Macko (1983) as the separate processing of "what" information (related to form and identity) includes color and "where" information (related to where the object is and where it is going) that does not.

that appropriate CCD measures are in use), their motion signals may be attenuated compared to the motion of a more obvious target.

Research in experimental psychology typically uses (x,y) CIE (Commission Internationale de l'Eclairage) color space plus luminance to describe colored stimuli. Other choices are (u',v') color space, red, blue, green; hue, saturation, brightness; or cyan-yellow-magenta-black coordinates, and weights of R/G and B/Y opponent pairs. Models of target acquisition tend to use (x,y) space or opponent pairs (e.g., ORACLE and GTV). Although photoreceptor responses are well characterized, there is some disagreement about the behavior of the color opponent cells. The basis of this disagreement comes from the fact that within a population of, say, R+/G- center-surround cells, there is much variability in the receptive field characteristics and the response magnitudes near and above thresholds, indicating that current physiological understanding may be inadequate to model the system effectively.

Recent research by Olds, Cowan, and Jolicoeur (1999) indicates that mapping stimuli into 3-D color space allows predictions to be made about their salience and distinctiveness. The authors found that targets were readily detectable in a background of differently colored non-targets if the coordinates of the colors of the targets and non-targets were planar separable[29]. Eastman (1968) similarly used distance between points in (u', v', w' [luminance]) space as a definition of color contrast. Color contrast has also been modeled by Frome, Buck, and Boynton (1981) as an equivalence term for luminance contrast. That is, the overall contrast of a target is modeled as a linear combination of its achromatic, R, G, and B color dimensions.

Thus far, color has been mentioned only in its role in detection of a static or moving target. Research from perceptual psychology indicates that, inasmuch as real-world objects are concerned, color plays little role in recognition or identification[30]. That is to say, the addition of target color information when sufficient information already exists in the image for us to recognize the target does not aid recognition performance. By far, the best example of this is the work by Biederman and Ju (1988), which demonstrates that in agreement with RBC theory, the surface characteristic of color does not improve response time to name a common object. Even an object readily associated with a color, such as a banana, is as quickly recognized in a line drawing as a full color picture.

Biederman and Ju's finding is not surprising, given that first, objects tend not to be defined by color alone and second, much of what the color vision system does is provide color constancy, whereby a colored object will appear to be the same, regardless of the source of illumination (Zeki, 1993). This so-called "discounting of the illuminant" means that the physically measured color spectrum of a surface does *not* correspond one to one with an observer's perception of the

[29]The "points" in color space actually correspond to Gaussians with steep sides; therefore, linear separability of the peaks does not ensure visual distinctness of the objects.

[30]Most research uses line drawings for object recognition. However, more powerful PC-based rendering software is making the use of solid models more common.

color. Implications of this finding may be important for observers who view potential targets in two very different lighting conditions. Models would have to compensate for the illuminant in order to accurately predict performance in both conditions.

7.1.2 Motion

Most models of target acquisition focus on static images and the static characteristics known to affect performance (e.g., clutter, contrast, size, resolution, range, atmospheric interference). Electro-optical models or front ends to models (such as TARGAC) may include the temporal response characteristics of the sensor or display, but the treatment of time dependence in such models typically relates to how the sensor addresses changes in the scene over time rather than targets that may be moving.

The two visual effects of a target moving relative to its background are a motion signal arising from the target itself, and the flicker-like changes in contrast around the borders of the target (e.g., if a light target moves over terrain that is alternatively dark and light, it may appear to flicker with respect to its background). The first effect is well studied and has been instantiated into several models; the latter has not been modeled successfully.

The human perception literature is useful when we consider how motion can influence performance in target acquisition. Motion and objects defined by motion (such as a cryptically colored animal that suddenly moves) are known to be especially good at directing visual attention (Hillstrom & Yantis, 1994; Yantis & Egeth, 1999; Wolfe, 1994). That is, a moving stimulus only needs to be a fraction of the physical intensity (e.g., luminance, size) of a static stimulus in order to immediately become visible. Some models (discussed next) take advantage of this fact by using motion to adjust the effective contrast of the moving target. That said, however, it is important to note that the effect of motion on detectability is not constant; instead, it interacts with contrast. If the contrast of a moving target is very low, it will remain quite difficult to see, regardless of its speed (Mazz, Kistner, & Pibil, 1998; Meitzler, Kistner, et al., 1998).

Of particular importance to the human visual system is the appearance of a “looming” stimulus whose motion is toward the observer (Schmidt, 1997; Yantis & Hillstrom, 1994). The only way that current models of target acquisition have incorporated looming motion has been to address the resultant increase in size and contrast of the object. However, the size increase necessary for looming objects to capture attention is quite small, so a simple contrast increase might not be adequate to account for the phenomenon. Therefore, looming should be treated as a special case of motion.

In addition to looming stimuli, two important aspects of motion perception are the ability of the visual system to segregate objects that are not moving from those that are (Watson & Humphreys, 1999), and the detection of objects that are moving in a way inconsistent with a moving observer viewing a field of stationary objects (Kaiser & Montegut, 1995). That is, if

there are objects at a variety of ranges from the observer, their motion with respect to the observer (actually, about his point of fixation) will be determined by (a) the range from the observer, (b) the speed of the observer, and (c) whether the objects are stationary or moving. Kaiser and Montegut determined that humans are particularly sensitive to objects whose motion is inconsistent with the expected motion parallax at their position. In other words, humans are good at spotting moving objects when they themselves are moving.

This ability comes into play in target acquisition when the observer himself is not always stationary; rather, the observer may be moving. Both of these aspects of perception relate to situations when an object in a scene is moving differently from other objects, indicating that its retinal velocity is different from what a stationary object at that location in space should be. Therefore, the object is self-propelled and is likely of military interest. No models thus far encountered have taken relative motion signals into account as they relate to such implied depth-related motion parallax, although simple relative motion signals should be able to be modeled when the frame of reference of the scene is changed from stationary to moving.

Models that account for motion tend to relate it to the probability of detection instead of discrimination since, if anything, the structural detail of a moving object will *decrease* because of the loss of high spatial frequencies (blur). Electro-optical systems are particular susceptible to blur, depending on the integration time of the sensor and the sampling and display rates. The effects of blur induced by motion are considered in some models (e.g., GTV).

7.1.2.1 Early Models of Motion

An early, empirical, and somewhat cognitive inclusion of motion into search performance was in Bishop and Stollmack's (1968) DYNTACS model. DYNTACS incorporated the effect of motion as an increase in the probability of detection within a time window, Δt. The model parameters are in terms of range and linear velocity, and the model included a term for "terrain complexity" which corresponds to possible paths in the scene along which a moving target may travel. As mentioned earlier, the TARGAC model is based more than most models on atheoretical data fits, indicating that its results may not be generalizable to other studies or situations.

Rogers (1972) found that the (luminance) contrast threshold of a moving object remains relatively constant as the retinal eccentricity increased to around 55 degrees. In order for a stationary target to remain barely visible as its eccentricity increases over the same level requires a five-fold increase in contrast. (Note that for a small target, the change in receptive field size with eccentricity cannot account for this finding.) Peterson and Dugas (1972) modified the search term (P_1) in Bailey's (1970) model to account for motion by increasing the size of the hard shell lobe as a function of angular velocity:

$$A_g = A_{g_0} C\left(1 + 0.45\omega^2\right)$$

in which A_{go} = typical glimpse aperture (hard shell lobe diameter),

C = contrast of target with background, and

ω = angular velocity of target with respect to the observer.

(Note that this instantiation includes the contrast-motion interaction mentioned earlier when the adjustment in the glimpse aperture weights contrasts very heavily by angular velocity, but when contrast is near zero, the magnitude of the effect of velocity will be negligible.) Presumably, such a simple modification could be made in a soft shell visual lobe calculation, possibly by the probability-to-detect drop-off becoming much shallower with eccentricity. Indeed, an increase in soft shell lobe is exactly what Rogers' result seemed to indicate.

7.1.2.2 More Recent Approaches to Modeling Motion

Meitzler, Kistner, et al. (1998) and Mazz, Kistner, and Pibil (1998) investigated the effects of motion on target detection in controlled laboratory experiments. Findings from both studies indicated that angular velocity was as important a factor as, or perhaps even more important a factor than, range (which determines target size) or contrast alone in the detection of a target. However, the effect of velocity was not independent of other factors in the study. Mazz et al. found that velocity interacted significantly with range and with range and contrast. Meitzler et al. found that velocity interacted significantly with range and with the background used in the studies (backgrounds were digitized images of different clutter levels). Thus, it was clear that an isolated velocity term would be insufficient for a model to account for the effects of target motion.

NVESD's ACQUIRE model (Tomkinson, 1990) was modified by Meitzler, Kistner, et al. (1998) to include a parameter for target velocity by making the probability of detection a function of target image size and the target image size necessary for 50% ensemble detection of the target:

$$P_d = \frac{(A/A_c)^E}{1+(A/A_c)^E}$$

in which A = target angular extent,

A_c = target angular extent necessary for 50% ensemble detection, and

$E = 2.7 + 0.7(A/A_c)$.

This function is purposefully similar to the TTPF used in other NVESD models. However, the angular extent necessary for 50% performance (A50) is itself considered a function of target size, contrast, and angular velocity:

$$A_c = aT_e + bC + cV_a + d$$

in which T_e = target angular extent,

C = target contrast[31], and

V_a = target angular velocity.

Although all three terms (and their interactions) are known to affect performance, the conditions in which one factor may be more important than or may interact with others are not clear. Therefore, the authors did not attempt to fit the constants. Instead, Meitzler et al. used a fuzzy logic approach (Zadeh, 1965) to derive fuzzy rules governing the influence of these factors in different conditions. One half of the data derived from a laboratory study in which target size, contrast, and angular velocity were controlled was used as input into fuzzy inference neural network (The MathWorks, 1995) to derive rules against which the other half of the data was tested. The authors report that the correlation between derived fuzzy rules and the test data was 0.95.

Another way that motion has been incorporated into models of target acquisition has been to include human visual physiology, as related to motion perception, into models based on early visual processes. How humans process motion information has been the subject of active research in the vision literature for decades. Studies that may have some bearing on models of target acquisition include those focused on the detection of motion signals among noise (e.g., Snowden & Braddick, 1991; Verghese & Stone, 1995; Verghese, Watamaniuk, McKee, & Grzywacz, 1999), those attempting to derive the basic motion features to which the visual system is sensitive (e.g., Adelson & Bergen, 1985), and those testing motion processing as related to known visual psychophysics and physiology (e.g., Grossberg & Rudd, 1991).

7.1.3 Transient Visual Events

Soldiers in the field routinely encounter situations in which events occur that are only visible for a brief time, such as the glint off a sight, a muzzle flash, the momentary appearance of an object from behind an occluder, or an explosion. The presence of such transient visual events can aid or hinder search for a target.

Before we discuss the specifics of how transients can affect search and target acquisition, it is necessary to understand how the visual system responds to such stimuli during search. It is obvious that before an observer can acquire a target, some representation of the target must exist in the observer's visual system. At issue is the amount of information accumulated over time as the observer views a scene. It has long been argued that the representation is an "integrative visual buffer" that collects information and becomes progressively more detailed over time (Rayner, McConkie, & Ehrlich, 1978). There actually is no such buffer and very little information about objects in a scene remains when the scene disappears or when we look away (see Vaughan, 1998, however, for evidence that some information does persist). This effect can be seen in almost any situation: close your eyes, turn around, and open your eyes for 1 or 2 seconds. Then close your eyes and describe as much of the scene as you can. You will probably only be able to recall details of a handful of objects.

[31]The contrast term does not account for the flicker-like effect of rapid changes in target contrast as it moves across terrain.

The reason why so little information persists is that our mental representation of the scene is actually very sparse, consisting of only four or five objects at a time (Rensink, 1996). The mechanism that selects objects from the scene, binds their features properly, and inserts them into this representation is selective attention. Objects in the scene that are clearly visible, yet unattended, are not perceived consciously or acted upon consciously (Mack & Rock, 1998). O'Regan (1992), Rensink (1997), and Minsky (1985) have argued that observers are not consciously aware of the sparseness of their mental representation because the scene itself serves as an external memory. In order to acquire information about a scene, the observer must focus his attention on a part of the scene, and that part is then encoded into the mental representation.

The role that selective attention plays in conscious perception is the key to understanding how transient visual events affect target acquisition. Search for a target includes a series of eye movements to locations in the scene similar to a target along some dimension or to locations as determined by a top-down scan path. It is in the first case that transients have their effect, since attention is presumed to precede eye movements to a location in the scene that is of interest. This "spotlight" of attention can readily be deployed to salient or conspicuous regions of the scene (Yantis & Egeth, 1999); thus, target conspicuity may determine the probability that the target will be attended and fixated. Transient visual events have the ability in certain circumstances to disrupt this salience-based attentional deployment system (e.g., Yantis, 1996; O'Regan, Rensink, & Clark, 1999) and involuntarily summon or "capture" attention to their locations.

If the transient event occurs at the location of the target (such as a glint or muzzle flash), then such a transient will increase the probability that the target will be fixated. In addition to a sudden increase in luminance or contrast, the appearance of a new perceptual object (e.g., when an object suddenly becomes visible as it appears from behind an occluder) is also known to capture attention (Hillstrom & Yantis, 1994; Yantis & Hillstrom, 1994). Attention may be captured even if the contrast of the new target is not sufficiently high to be judged as salient or conspicuous if it had not just appeared. An interesting aspect of attentional capture is that it can occur even if the scene is highly congested. Therefore, any model that incorporates clutter into dynamic scenes must treat visual transients as a special case in which the effects of clutter are strongly attenuated.

Attention is not always captured by transient events, however. As is the case with moving objects, transients will only capture attention when the increase in luminance or contrast or when the contrast of the new object is sufficiently high[32]. Enns and Austen (1999) found that low contrast targets failed to capture attention in such circumstances, but moderate contrast targets did so quite effectively. Valeton and Bijl (1995), in evaluating the TARGAC model on data from the Battlefield Emissive Sources Trials (BEST) under the European Theater Weather and

[32]Note that the increase in luminance or contrast associated with the transience will render it far more likely to capture attention than a static object with the same high luminance or contrast (Yantis, 1996).

Obscurants (TWO); NATO, 1990) studies, found that targets that appeared suddenly were particularly difficult to see. The targets tended to be small and of low contrast. The reason why observers in the BEST TWO study found these targets particularly difficult is that in the low contrast conditions, they were no more salient than other targets and were available in the scene for less time.

In addition to transient events failing to alert the observer to a potential target, they may also hinder search. If a transient event occurs at a non-target location or at an already acquired target, attention may be captured, thereby disrupting a salience- or conspicuity-driven search of the scene. O'Regan, Rensink, and Clark (1999) demonstrated that when a "mud splash" (a convex gray region) was repeatedly added to and taken away from a scene, the time required to search for a target increased dramatically[33]. Even when the target itself represented a transient event (such as a sudden movement or color change or a sudden appearance of a new object), the more salient mud splash disrupted search.

In target acquisition situations, the effect of irrelevant transients is likely to be manifested by a change in search strategy from an efficient one to an inefficient one. Target conspicuity has been the result of much study because it is a good predictor of search performance. The reason why conspicuity drives search performance is that the attention system is able to quickly select conspicuous regions in the scene during search. Other, less conspicuous regions are not searched because targets are deemed less likely to be in them. When search is difficult, a more systematic, top-down search strategy is employed that involves a conscious pattern of searching the scene, often including parts of the scene where no target is likely to be. In the presence of transients, an analogous strategy is employed. O'Regan et al. found that repeated mud splashes forced subjects to abandon a conspicuity-driven search and adopt a slower, systematic search. Search models that include transients may benefit from the addition of a systematic search that occurs when such repeated transients occur.

7.2 Multiple Targets

War game simulation as well as real-world combat situations are not exclusively single-target scenarios. Often, a Soldier is confronted by several targets, all of which may be obscured, camouflaged, or otherwise difficult to acquire. The issues of how to model target acquisition in such an environment are complicated because additional assumptions must be made regarding what the observer's task is, how search progresses, and how limiting conditions arise. In addition, models must be altered differently, depending on whether they predict individual or ensemble performance.

Looking first at the task, a model may predict the probability of first acquisition (i.e., the probability that any target will be acquired) or the probability that multiple targets are acquired. The simplest solution to the problem of multiple targets would be to work within the framework

[33]This research was funded by Nissan Motor Corporation. The researchers were interested in the effects of material splashed onto car windshields on a driver's ability to spot important changes in the scene, such as a person stepping into the roadway.

of individual performance prediction. In such a framework, the only additional assumption needed would be a specific statement of search-quitting criterion. For example, first-detection performance could be modeled with no changes in a single-target model except that the probability of a target within a glimpse would increase. (Of course, depending on the model, the presence of multiple targets may also affect factors such as fixation selection, decision criterion, etc.) Predicting multiple acquisition performance requires the simulated observer to know how many targets there are and to stop after they are all acquired or to place a time limit on the search process and let it continue until the time limit. In either case, the model must keep track of targets that have already been acquired so that a single target will only be acquired once. (This addition of a memory component to search is built into some models, such as GTV, but is lacking from others, such as Nicoll & Hsu's [1995] model.)

Search models from perceptual psychology rarely use multiple targets except as a test of the serial or parallel nature of a purported search process by examining a phenomenon called redundancy gain (e.g., Egeth & Mordkoff, 1991)[34]. The lack of interest in multiple target search may also stem from the fact that these models are all based on individual performance and, as mentioned before, are easily extendible to multiple target situations.

Predicting individual performance in a static detection (i.e., non-search) task requires the targets in question to be within the observer's search lobe. That condition being met, assumptions must be made regarding limits to an observer's performance in the task. Decisions made on the basis of signal detection theory (e.g., based on SNR or SCR) may proceed in one of two ways in multi-target scenarios. First, the task may be redefined as several independent decisions (one for each target) with a logical OR determining the probability of first acquisition. Second, the signal and noise terms must be redefined to take into account contributions from all the targets; then a single global decision must be made to judge if the signal arose from a target (or targets) or noise.

Predicting ensemble performance in a static search is considerably more daunting. The difficulty arises from how asymptotic performance terms such as P_∞ are conceptualized. That is, depending on what it actually *means* that a particular target in a particular scene will be acquired by P_∞ of an ensemble of observers, the predictions for how P_∞ changes with the number of targets will be different. Six possible meanings of P_∞ are discussed.

Rotman, Gordon, and Kowalczyk (1989) considered three possible reasons why ensemble detection performance is imperfect, given infinite time.

1. The ensemble of observers is strictly ordered in terms of target acquisition competence. That is, some of them are simply better at detecting targets than others. These observers

[34]The issue of whether visual search progresses in a serial or parallel manner has long been a contentious issue in perceptual psychology (e.g., Palmer & McLean, 1995; Townsend, 1971, 1990) since the processes underlying parallel and serial search differ dramatically. Redundancy gain is but one technique for obtaining data that may be able to tease apart the serial/parallel distinction. In terms of target acquisition modeling, the difference is not as important because the very fact that serial and parallel processes can mimic each other in RT or accuracy measures indicates that neither type of model is likely "better" at predicting relevant performance.

will be consistently better, regardless of the target or background (i.e., they will require fewer cycles on target to acquire the target). Silk (1997) refers to this explanation of P_∞ as the "observer-only" account.

2. Observers are equivalent in a statistical sense, but the responses to any given target will be stochastic (within the bounds that an ensemble must perform at a level of P_∞). Some observers will confuse the target with background clutter and will not evaluate it any further while other observers will not make this confusion. In this case, observers who cannot detect a target in one situation may be able to do so in another. Silk (1997) refers to this explanation of P_∞ as the "observer-target" account[35].

3. Observer performance will decrease over time because of mental weariness. Some observers are able to acquire the target within a critical period and some are not.

Rotman et al. (1989) derived predictions on the basis of these three assumptions and compared them to data based on images from Hughes Aerospace (Scanlan & Agin, 1978). The proportion of the population of observers who were able to detect targets of varying degrees of difficulty strongly favored either explanation 2 or 3 over explanation 1. The authors point out, however, that the number of observers in the study reduced the statistical power of their tests to the point that no explanation could be eliminated definitively.

In addition to the three explanations mentioned, common sense tells us that a combination of these factors is probably occurring: Some observers *are* better than others, and some targets (for reasons unknown) *will* be more difficult than others, regardless of how facile a target acquirer any given individual is. Whether mental weariness comes into play is unclear. Likely, in the case of testing the explanations with empirical data, weariness would not be a factor, given the controlled situations in which the data were collected. Combinations of these explanations have been termed "hybrid" models by Silk (1997).

Silk (1997) analyzed a data set from O'Kane, Walters, and D'Angostino (1993) to determine whether a "hybrid" explanation of P_∞ could be based on the deterministic observer-only and observer-target stochastic processes. He found that those two factors, plus a degree of uncertainly that exists as a result of uncertainty in target signature computation[36], completely defined observer performance. In other words, given the inherent uncertainty in determining target characteristics, observer performance can be described as a combination of observer-only and observer-target explanations.

In addition to Rotman et al.'s (1994a) explanations, Nicoll (1994) put forth three additional possibilities for P_∞ that have bearing within a neoclassical search framework.

---

[35]The Army combat model JANUS (not an acronym) assumes that P_∞ is purely observer-target based.

[36]Silk (1995) demonstrated that the modeling uncertainty in Johnson-like models is statistically unbiased. That is, the uncertainty in predictions of target detectability is independent of the actual detectability.

4. There is another state in the search process called "quit," and the corresponding rate, Q, at which this state is entered from any other state. The probability of quitting as a function of time is then very much like the probability of fixating a target as a function of time, except that it is the linear combination of three rather than two exponentials. The number of targets detected before quitting (rather, the distribution of such trials) determines P_∞.
5. The number of visits to a target may be restricted (possibly because of a temporal cut-off or a moving FOR).
6. Assume that the Markov process is not memoryless but that the amount of information accumulated during visits to the target decreases over time. If the asymptotic amount of information obtainable about the target is below that required to detect it, then detection cannot occur.

Nicoll (1994) has not offered any data to support any explanation over the others but presented them as examples of the flexibility of the neoclassical framework.

Other issues related to the multi-target scenario relate to the expectations of the observer and the difference between the targets. For example, if there are two very different targets in the scene, the observer must know that there are two (according to many models that base their predictions on a known target representation) or must base his search on a general metric or search strategy that makes no assumptions about the appearance of the targets. Also, if the subject is expecting to see or has been trained in a target-rich scenario, his performance in a low-contrast multi-target scenario will be different from someone trained in a different scenario (e.g., Doll & Schmieder, 1993). Specifically, the former observer will be more likely to hazard many false alarms whereas the latter will be more conservative. More is said about dependent measures other than Pd in a later section.

Classic studies in perceptual psychology have shown that if the various targets are similar to each other in appearance and are different from non-targets in appearance, then little training will be required for search performance in the multi-target situation to be as good as in the single-target situation (Schneider, Dumais, & Shiffrin, 1984). However, as targets become different from each other and more similar to non-targets, training will take much longer to achieve the same level of performance (Schneider et al., 1984; Duncan & Humphreys, 1989).

7.3 Blur, Noise, and Obscurants

Different factors limit human target acquisition performance in threshold and super-threshold situations. At or near threshold, human performance is noise limited; above threshold, human performance is contrast limited (Lloyd & Sendall, 1970). The role of noise in target acquisition is not limited to the threshold of our sensory system, however. The same limits to detection of visible form apply when noise is relative to signal strength. That is, when noise is high, human perception is noise limited; when noise is low, human performance is contrast limited.

Noise in target acquisition comes from a variety of sources: the absolute threshold of vision for a dark-adapted observer is determined partly by quantum noise (probabilistic absorption of photons by photochemical molecules) and neural noise (photochemical breakdowns and firing of neurons in the visual system). For the military observer, visual noise of interest typically comes from the display or the sensor on and through which he is viewing an image of the scene.

Noise varies, depending on the type of sensor. FLIR sensors are susceptible to noise from IR atmospheric emissions and scatter, thermal noise within the sensor, and scintillation noise because of turbulence of the air along the line of sight of the sensor. The latter noise can take the form of blur (the loss of high spatial frequency information) if the integration time of the sensor is long or motion artifacts (small moving images that do not correspond to objects moving in the field) if the integration time of the sensor is brief and its spatial resolution is high. Image intensifiers do not suffer from so many sources of noise because the wavelengths of light intensified by the sensor do not interact so readily with particulate matter in the air column[37].

The effects of atmospheric noise in FLIR sensors are well understood and modeled quite effectively (e.g., the TARGAC front end for NVESD static detection models). However, the effect of noise on human decision making is not as clear.

Blur, the loss of fine spatial detail (i.e., an attenuation of high spatial frequency information), is well understood, in theory at least. An across-the-board degradation in performance is expected for all levels of target acquisition because the loss of detail is akin to a reduction of contrast of targets to the point that the modulation of their fine details falls below threshold. Blur can be instantiated in a model with a digital blur operation on an input image (such as Gaussian blur) or by modulation of the Fourier components of an image with a high-frequency-attenuated modulation transfer function. The resulting decrease in effective contrast can be traced along a TTPF to determine the concomitant loss in performance.

Aleva and Kuperman (1997) evaluated the effects of various kinds of scene degradation on the detection and recognition of a variety of Army vehicles at various ranges using a signal detection paradigm[38]. The authors manipulated scenes by increasing scene modulation (reduction in contrast), blur, and white noise. The authors noted two effects of significance: first, modulation and blur interacted (as one would expect). Second, the effect of blur and modulation was manifested as a decrease in hit rate only; false alarm rate remained constant. From these results, it was concluded that the sensitivity of the observer was decreasing as a result of the image degradations. Such a result is consistent with the loss of information or in signal detection terms, the decrease in SNR in conditions of blur and modulation.

---

[37]Stereoscopic image intensifiers, with an intensifier tube for each eye, are even less susceptible to noise. The scintillation noise in the tubes is uncorrelated, and the visual system has little trouble discounting it from the otherwise stereoscopic image of the scene. The per-item cost of such systems remains prohibitive, however.

[38]The manuscript by Aleva and Kuperman serves as an excellent review of basic visual psychophysics and of signal detection theory.

Obscurants make target acquisition difficult by blocking the electromagnetic radiation reflected or emitted by the target so that it is never detected by a sensor. Unlike noise or blur, however, obscurants have a temporal character since the consistency, density, and amount of obscurant between the sensor and the target are not uniform over time. Rotman, Gordon, and Kowalczyk (1991) extended the NVESD static detection model to account for time-varying obscurant smoke by assuming that as the smoke obscures more target information, the proportion of observers who will be able to detect the target will decrease. The authors modeled the performance for an ensemble by estimating a mix of performance for an unobscured target and a steady state obscured target. The main predictions of the model are that time-varying obscurant performance will reach an asymptote at the level for an unobscured target. The model has been applied to engineer specifications of several fielded FLIR sensor systems, but it has not yet been validated by human data from field tests.

7.4 Measures of Performance Other Than P_d

The most influential modeling concept in the past 40 years has been the Johnson criteria and the corresponding TTPF. The resulting static target discrimination model incorporated into several NVESD models is considered one of the most common (and most effective) models for ensemble performance. However, the model only makes predictions of a single variable in a single type of situation: P_d, the probability of detecting a target when one is present. Other performance measures can be inferred if one makes assumptions about how a decision is made (e.g., Rotman et al., 1991), but the Johnson criteria are by themselves limited in how they inform us about the process of target acquisition.

Over the years, different tasks and different analyses have led to several ways of characterizing observer performance in target acquisition tasks. This section discusses a number of them: Schmieder and Weathersby's (1983) P_{acq} measure, the false detection percentage (FDP), response time, the determinants of performance according to signal detection theory (P_{hit}, P_{FA} (FAR), d', A', and β), and real-time eye movement data.

Although popular models such as ACQUIRE predict search performance over time, they do so by predicting P_d as a function of time only. Such a measure is useful, especially for war game simulation in which it is important to predict the detectability of a target when only a certain amount of time is available to scrutinize the scene. However, this measure in and of itself is limited in how well it predicts overall observer behavior. The primary shortcoming of the measure is that it does not address observer false alarms (i.e., reporting a target when none was present). False alarms, also called false detections in some analyses, can be further subdivided into cases when no target was present and cases when a target was present but the observer mistakenly reported that a non-target element was the target. Such a distinction can be made when an observer is forced to localize a target in a scene in addition to reporting merely its presence, or it can be inferred from observer response and eye movement data. (The potential value of eye movement information is discussed shortly.)

In their analysis of observer performance in cluttered environments, Schmieder and Weathersby (1983) determined that P_d might not always be a meaningful measure since a high rate of false alarms is typically observed in conditions of high clutter. The authors proposed instead the measure P_{acq}, the probability of acquisition, defined as the probability that an observer can correctly acquire a target after n+1 investigations in which n false targets were first correctly rejected:

$$P_{acq} = \sum_{i=1}^{n} \frac{P_d[1-P(FA)]^i}{n+1}$$

in which P(FA) = fixed probability of false alarm (based on clutter level),

P_d = probability of detection, and

n = the number of objects investigated when a target is located.

This measure of performance assumes that clutter attracts eye movements in a discrete manner. It also presumes that clutter's effect is on the probability of false alarms and to a lesser extent, on the probability of detection, as determined by the SCR.

A further problem with using response time and a single accuracy measure (e.g., P_d) is that it ignores how an observer makes a decision. For example, a speed-accuracy trade-off may occur. Speed-accuracy trade-offs result when an observer with a lax criterion for deciding that a target is present responds faster and makes more errors than an observer with a more stringent criterion, who responds more slowly and makes fewer errors. This pattern of errors and response times may occur even if the observers are equally good at detecting the target. The difference in decision criterion not only varies between observers (see, e.g., Rotman, Gordan, & Kowalczyk, 1989, for an analysis of performance based on this assumption) but within observers as a function of training, stress, fatigue, expectation, the costs and benefits ("payoffs") of rendering a decision, and concurrent task load. As such, it is impossible to determine how sensitive an observer is to the presence of a target by looking solely at RT and P_d.

The method used most often to separate the contributions of observer sensitivity and criterion in making a decision is called Signal Detection Theory (SDT) (Green & Swets, 1966). Briefly, signal detection theory asserts that the detection of a signal requires an observer to be able to distinguish between noise inherent in the sensory system and a signal added to that noise. Signal and noise distributions are assumed to be normal and have equal variance. An observer bases his decisions on sensitivity (his visual system's ability to distinguish between the noise and the signal-plus-noise distributions) and the criterion that he sets for determining if a given sensory signal arose from the signal or noise distribution. A sensory signal whose strength is above the criterion will be reported as a signal; one whose strength falls below the criterion will be reported the absence of a signal. See MacMillan and Creelman (1991) for an excellent introduction to SDT.

The assumption that signal and noise distributions are normally distributed (with the same standard deviation) allows the decision criterion, β, to be separated from the observer sensitivity, d′[39]. In order to perform an SDT analysis, the hit rate, defined as P(target reported|target present), which is the same as P_d for a single subject) and the false alarm rate (FAR), defined as P(target reported|target absent), is needed. Further assumptions may be needed for us to perform SDT analysis on data in which false alarms include non-targets misidentified as targets when targets were present elsewhere.

SDT has been used extensively to examine target acquisition. (For a more thorough review of SDT analysis as it applies specifically to target acquisition, see Wilson, 1992.) Of particular interest is how false alarms are affected by various factors. As mentioned earlier, Aleva and Kuperman (1997) used SDT to evaluate the effects of modulation, blur, and noise on target acquisition performance. Their results showed a decrease in hit rate but no change in FAR as scene quality decreased, indicating that subjects in the study did not shift their criteria but were becoming less sensitive to the targets.

Doll and Schmieder (1993) were the first study to look at the effects of clutter, as measured by a quantitative metric, on false alarm rate[40]. The authors used a measure of clutter called the SCR, which is related to the gray-level statistical variance metric (see the section of this report on clutter and conspicuity for details). The authors looked at overall probabilities of detection and FAR and found that as SCR decreases (i.e., as clutter increases), observers shift their criterion to produce more "target present" responses, thus increasing the FAR.

Grossman, Hadar, Rehavi, and Rotman (1995) also used SDT to investigate how clutter affects FAR. The authors defined noise to be the strength of a clutter metric (the probability of edge metric or Schmieder & Weathersby's [1993] SCR metric) and modeled search performance over time as a function of per-glimpse SCR. Glimpses were assumed to be independent and attracted to regions of high clutter. Their results indicated that the average accumulated number of false alarms increased as a linear function of clutter, the slope of which was determined by the time permitted for search. That is, the false alarm *rate* within each glimpse was constant. The difference between their results and those of Doll and Schmieder were likely attributable to assumptions made by Doll and Schmieder to predict overall FAR rather than examining FAR as a function of search time. The hit rate (P_d) decreased as a function of clutter, indicating that

[39]If normality is known to be violated or cannot be evaluated directly by normalized receiver operating characteristic curves (see MacMillan & Creelman, 1991), then a non-parametric measure of sensitivity, A', may be calculated (Pollack & Norman, 1964):

$$A' = \frac{\frac{1}{2} + (P_{hit} - P_{FA})(1 - P_{hit} - P_{FA})}{4P_{hit}(1 - P_{FA})}$$

[40]The Doll and Schmieder (1993) paper contains a good introduction to SDT and how it applies to target acquisition in cluttered environments. The paper also addresses the effects of display resolution and its interactions with clutter.

while subjects kept their criteria relatively constant, they were less sensitive to targets that appeared in cluttered scenes. The authors also argue that time reduces the decision criterion.

Silk (1995a) argues that scene-based characteristics such as blur and clutter are not the only determinants of observer decision threshold. In a study involving detection of altered IR target signatures (i.e., digitally modified to reduce the signature), observers were more likely to generate false alarms in a test situation if they had been trained in a target-rich environment. The hit rate of the observers was the same across training situation, indicating that observers shifted their decision criteria downward when they thought more objects in the scene were likely to be targets.

In addition to being able to disentangle the effects of sensitivity and decision criterion, the use of methods amenable to signal detection analysis has advantages of its own. First, such methods are likely to be standardized across studies, so researchers may be better able to relate their theories and analyses to existing data rather than having to run additional studies. Also, forcing observers to perform a two-alternative forced choice (2AFC) or a detection-plus-confidence task rather than simple go/no-go detection task or deliberately manipulating pay-offs for the different types of errors (misses and false alarms) gives the experimenter additional information about the nature of the discrimination. Valeton and Bijl (1995) found, for example, that subject performance was better in a 2AFC task (picking which of two trials contained a target) than a go/no-go task (only reporting if a target is seen).

A FAR-like measure, the FDP, has been used successfully within the framework of the ACQUIRE model to explain the variability in N50 with scene clutter. FDP is defined as

$$FDP = \frac{\#\text{"}present\text{"}\ responses \mid no\ target}{total\ \#\text{"}present\text{"}\ responses} \times 100\%$$

Mazz (1998) noted that much of the variability noted in empirical N50 resulting from different levels of clutter can be accounted for if one also accounts for the false detection percentage. FDP is analogous to and largely independent from N50. That is, FDP and N50 can vary freely within a study, indicating that both quantities should be taken into account when one is performing an analysis of the effect of clutter[41].

Eye movement data from search tasks are an often-overlooked source of information for how subjects perform target acquisition experiments. Eye movements during search can provide insight into (a) the evaluation of local metrics of clutter, conspicuity, distinctness, and attractiveness; (b) evaluating model parameters such as glimpse aperture and glimpse duration; (c) determining whether the classical or neoclassical search framework provides a better fit to overt behavior. As has been mentioned elsewhere in this report, eye movements during search tend toward regions of the scene that are "target like." Several metrics have been proposed to

[41]Though this result is not surprising given the independence of hit rate and false alarm rate in SDT, it is interesting that such a result holds in the case of N50 and FDP in that both are ensemble performance measures.

determine the attractiveness (or distinctiveness, or conspicuity) of various regions of the scene. The evaluation of these metrics is almost always performed by an examination of the degree to which eye movements about the scene tended to land on regions that score highly on a metric (e.g., Tidhar et al., 1994; Rotman, Kowalski, & George, 1994; Toet, 1996; Cartier, Nicoll, & Hsu, 1998). Also, the evaluation of search models that posit a fixation guidance mechanism based on regions of the scene that are likely to contain the target (e.g., Doll et al., 1998) can be aided by an evaluation of eye movements. By examining the spacing of eye movements and how that spacing changes as a function of clutter, we can obtain information about the size of a glimpse aperture, whether soft- or hard-shell search is occurring, and any effects of clutter on glimpse parameters. Finally, looking at the degree to which fixations return to previously visited regions of the scene and when during searching a decision is made can corroborate or disprove predictions of the classical and neoclassical search models.

7.5 Validation Issues

O'Kane has written an excellent overview of the process of target acquisition model development and validation (1995). The author specifies and gives concrete examples of three different methods and the roles they play in the process of model development: (a) perceptual experiments using hybrid imagery, (b) perceptual experiments using calibrated field imagery, and (c) field trials controlled and documented as well as possible. The discussion herein focuses on the second of these three steps, as the models considered in this report were arguably past the point of using hybrid imagery to test their underlying theories. At the same time, though, the authors (wisely) chose not to put forth the risk and expense required for field trials. If the field imagery is calibrated sufficiently and all relevant observer, task, and dependent variables are recorded in detail, then much can be learned about target acquisition without our leaving the laboratory. (Of course, field experiments will be required to validate major models, especially if the models predict an effect of a variable, such as observer stress, that cannot be readily manipulated in a laboratory setting.)

As alluded to earlier, evaluation of scene metrics and models of target acquisition performance depends on the existence of a standardized data set of images, tasks, observer variables, and performance measures. The generalizability of models is determined by the underlying psychophysical data upon which the models are based. The validation of parts of models such as ORACLE or GTV depend on a database of psychophysical results. Models of vision are currently constrained by the lack of a readily available database of stimuli, methods, and threshold[42].

A useful data set for target acquisition development and validation must contain four things: (a) standardized, calibrated stimuli with complete descriptions of the scene geometry,

[42]A special interest group at the 1999 Annual Meeting of the Association for Research in Vision and Ophthalmology (ARVO) called for the creation of such a database of thresholds and called for its availability on the internet.

atmospheric conditions, and scene manipulations, (b) information about the task that observers must perform, (c) observer variables and how they were measured, and (d) the performance measures used and subject performance data.

Although it is a relatively simple task for basic vision science, since the optical spectrum, the unaided eye, and established psychophysical measures are of interest, creating such a database for use in military target acquisition research represents a more daunting challenge. One reason for the difficulty (and the need) is that different sensors have different specifications, any of which may be important in the determination of observer behavior. In addition, observer tasks, levels of target acquisition desired, and dependent measures (e.g., RT, P_d, FAR, eye movement) will differ greatly. In order for us to grasp observer variables, much data about subject training, levels of fatigue, concurrent task load, etc., must also be collected. The performance measures should include ensemble and individual data, preferably with sufficient detail that different analyses can be performed on the same data set (e.g., SDT analysis can be performed on data from an ensemble-performance study). Individual data in ensemble studies are of particular interest because, as pointed out by Rotman et al. (1989), the reason why ensemble performance predictors such as P_∞ have the values they do remains unknown.

8. Prognostication: The Future State-of-the-Art Target Acquisition Model

This section describes the current state of the art and where modeling is headed. This final section discusses the author's thinking in terms of the most profitable avenues to be pursued in target acquisition modeling.

There is no clear state-of-the-art target acquisition model. Some models do a good job of predicting performance in general but do not incorporate many factors known to influence performance (e.g., ACQUIRE, FLIR92). Other models incorporate many such factors but have so many degrees of freedom that their applicability to a given situation may be questionable (e.g., ORACLE, GTV). Although there is little benefit to having a single model that accounts for everything as opposed to several models that each account for a piece of the target acquisition pie, there is undoubtedly a benefit to models that take more than a single factor into account.

The need for a multi-factor approach to target acquisition modeling comes from various lines of evidence. First, studies by Mazz, Kistner, and Pibil (1998), and Meitzler, Kistner et al. (1998) demonstrated that the effects of variables such as scene clutter and target velocity, range, and contrast had effects on performance independently and as interactions. Second, many commonly studied and validated metrics of clutter and conspicuity are based on the co-occurrency matrix, which incorporates structure as well as contrast in determining what parts of a scene are target like or are based on measures that take into account more than one scene factor at a time (e.g., CAMELEON). Third, from the perceptual psychology literature, it is known that contrast alone

does not determine salience or attentional capture. Rather, it interacts with factors such as motion, transient visual events, and color.

The trick will be to incorporate the various factors in a way that makes sense and provides a good analog to the cues that the human visual system used to perform target acquisition. Meitzler, Kistner, et al. (1998) and Meitzler, Singh, et al. (1998) used a fuzzy logic approach to incorporate several factors into the ACQUIRE model. (Recall that this model is based on the Johnson criteria.) The result of the study were sets of fuzzy rules, gleaned from half of a human performance data set and applied to the other half, which predicted more than 90% of the variance in performance. This result raises two questions: Can the rules from one such study can be applied more generally to other studies? Why did the rules arise the way they did? The first question is a practical matter since it applies only to models within the ACQUIRE framework. The second question is more interesting. What is it about the target acquisition situations in the study that prompted observers to use some factors in one case and other factors in another?

This reviewer is convinced that a theoretically driven research program into how human observers use information in the scene will allow general rules to be derived for integrating multiple factors in future models. The starting place for such a program should be an aspect of visual perception that is well understood in theory and has been shown to have an impact on search and detection. One possibility would be to investigate the role played by selective attention in real-world target acquisition and the observer and scene-based factors that influence the deployment of attention. A team at ARL's Human Research and Engineering Directorate is endeavoring to study attention in just such a way. With a principled understanding of the role of attention and the influences on attention, models may be modified or developed to include known effects of measurable factors.

Current models best able to accommodate the effects of selective attention are models of individual rather than ensemble performance. GTV, in particular, already contains modules to prioritize and guide eye movements based on attention and to include training. Incorporating attention into a neoclassical framework model would require a non-random search step that dramatically complicates calculations. Fitting attention into a Johnson criteria-based model also presents somewhat of a challenge, since there are so few free parameters to work with. (Presumably, N50 or the shape of the TTPF may be modulated by attentional parameters.)

It is readily acknowledged that regardless of the emphasis placed on multi-factor approaches to target acquisition modeling, the Johnson criteria and models based on it will not go away. It is therefore important to determine the extent to which models based on the criteria can be extended to include additional factors. NVESD's static performance models have undergone such scrutiny in an attempt to see if they can accommodate multiple observers, multiple targets, clutter, false detection predictions, and the presence of scene obscurants. Analyses such as the one by Silk (1995b, 1997) should be emphasized before we attempt to encompass additional variables in such models.

9. References

Adelson, E. H.; Bergen, J. R. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A* **1985**, *2*, 284-299.

Ahumada, A. J.; Beard, B. L. Object detection in a noise scene. In B. Rogowitz & J. Allebach, Eds., *HumanVision, Visual Processing, and Digital Display VII, SPIE Proceedings, 2657*, SPIE: Bellingham, WA, 1996.

Akerman, A. *User Manual for the Visual Observer Model, Version 1.2*, I-MATH Associates, Inc., (Report 921R32, 31), 1992.

Akerman, A. Visual signature management target acquisition model improvement program for requirements specification, design analysis, and concept evaluation, Volume II (interim draft). Prepared for US Army Tank-Automotive Command, Warren, MI, by OptiMetrics, Inc., Ann Arbor, MI, 1993a.

Akerman, A. The Visual Observer Model for CCD analysis. In (A. M. L. LaHaie, Ed.). *Proceedings of the 3rd Annual Ground Target Modeling and Validation Conference*, pp. 324-330, Ann Arbor, MI, August, 1992, 1993b.

Akerman, A.; Kinzly, R. E. Predicting aircraft visibility. *Human Factors* **1979**, *21*, 277-291.

Aleva, D. L.; Kuperman, G. G. Effects of scene modulation image blur and noise upon human target acquisition performance (interim report AFRL-HE-WP-TR-1998-0012). Air Force Research Laboratory, Crew Systems Interface Division, Human Effectiveness Directorate, Wright-Patterson AFB, Ohio, 1997.

Bacon, W. F.; Egeth, H. E. Overriding stimulus-driven attentional capture. *Perception & Psychophysics* **1994**, *55*, 485-496.

Bailey, H. H. Target detection through visual recognition: A quantitative model. Santa Monica, CA: Rand Corporation, AD 721446. (Rand report RM-6158/1-PR), 1970.

Biederman, I. Recognition-by-components: A theory of human image understanding. *Psychological Review* **1987**, *94*, 115-117.

Biederman, I.; Ju, G. Surfaces. edge-based determinants of visual recognition. *Cognitive Psychology* **1988**, *20*, 38-64.

Bijl, P.; Valeton, J. M. Triangle orientation discrimination: The alternative to minimum resolvable temperature difference and minimum resolvable contrast. *Optical Engineering* **1998a**, *37*, 1976-1983.

Bijl, P.; Valeton, J. M. Validation of the new triangle orientation discrimination method and ACQUIRE model predictions using observer performance data. *Optical Engineering* **1998b**, *37*, 1984-1994.

Birkmire, D. P.; Karsh, R.; Barnette, S. D.; Pillalamarri, R. Target acquisition in cluttered environments. *Proceedings of the 36th Annual Human Factors Society*, 1425-1429, 1992.

Bishop, A. B.; Stollmack, S. The tank weapons system: DYNTACS model. Ohio State University. (DTIC report AD 850367), 1968.

Blackwell, H. R. The effects of certain psychological variables on target detection. University of Michigan. (ERI Report 2455-12-F, June, 1958), 1958.

Bliss, W. D. *In Air-to-ground target acquisition source book: a review of the literature*, (D. B. Jones, et al., Eds.), Office of Naval Research, Arlington, VA. (DTIC report ADA 015079), 1974.

Blumenthal, A. H.; Campana, S. B. An improved electro-optical image quality summary measure. *Proceedings of the SPIE* **1981**, *310*, 43-52.

Blumenthal, A. H.; Campana, S. B. Development of an image quality model for object discrimination. *Proceedings of the SPIE* **1983**, *467*, 24-32.

Campbel, F. W.; Robson, J. G. Application of Fourier analysis to the visibility of gratings. *Journal of Physiology* **1968**, *197*, 551-566.

Carrasco, M.; Evert, D. L.; Chang, I.; Katz, S. M. The eccentricity effect: Target eccentricity affects performance on conjunction searches. *Perception and Psychophysics* **1995**, *57*, 1241-1261.

Cathcart, J. M.; Doll, T. J.; Schmieder, D. E. Observer detection performance in urban clutter. Air Force Wright Aeronautical Laboratories, Avionics Laboratory. (DTIC report AD B122931, LIMITED), 1988.

Cathcart, J. M.; Doll, T. J.; Schmieder, D. E. Target detection in urban clutter. *IEEE Transactions on Systems, Man, and Cybernetics* **1989**, *SMC-19*, 1242-1240.

Cartier, J. F.; Nicoll, J. F.; Hsu, D. H. Target attractiveness model for field-of-view search. *Optical Engineering* **1998**, *37*, 1923-1936.

Cavanagh, P.; Anstis, S. The contribution of color to motion in normal and color-deficient observers. *Vision Research* **1991**, *31*, 2109-2148.

Chun, M. M.; Wolfe, J. M. Just say no: How are visual searches terminated when there is no target present? *Cognitive Psychology* **1996**, *30*, 39-78.

Cooke, K. J.; Stanley, P. A.; Hinton, J. L. The ORACLE approach to target acquisition and search modeling. In (E. Peli, Ed.), *Vision models for target detection and recognition: in memory of Arthur Menendez*, pp. 135-171, River Edge, NJ: World Scientific, 1995.

Copeland, A. C.; Trivedi, M. M. Texture perception in humans and computers: Models and psychophysical experiments. In (W. R. Watkins & D. Clement, Eds.), *Targets and Backgrounds: Characterization and Representation II*, *Proceedings of the SPIE* **1996**, *2742*, 436-446.

Copeland, A. C.; Trivedi, M. M. Signature strength metrics for camouflaged target corresponding to human perceptual cues. *Optical Engineering* **1998**, *37*, 582-591.

Copeland, A. C.; Trivedi, M. M.; McManamey, J. R. Evaluation of image metrics for target discrimination using psychophysical experiments. *Optical Engineering* **1996**, *35*, 1714-1722.

Doll, T. J.; McWhorter, S. C.; Hetzler, M. C.; Stewart, J. M.; Wasilewski, A. A.; Schmieder, D. E.; Owens, W. R.; Shedder, A. D.; Galloway, G. L.; Harbert, S. L. Visual/Electro-optical (VISEO) detection analysis system final report. Prepared under contract DAALJ02-92-C-044 with the Army Aviation and Troop Command, Aviation Applied Technology Directorate, Georgia Tech Research Institute, Atlanta, 1997.

Doll, T. J.; McWhorter, S. W.; Schmieder, D. E.; Wasilewski, A. A. Simulation of selective attention and training effects in visual search and detection. In (E. Peli, Ed.), *Vision models for target detection and recognition : in memory of Arthur Menendez*, pp. 396-418, River Edge, NJ: World Scientific, 1995.

Doll, T. J.; McWhorter, S. W.; Wasilewski, A. A.; Schmieder, D. E. Robust, sensor-independent target detection and recognition based on computational models of human vision. *Optical Engineering* **1998**, *37*, 2006-2021.

Doll, T. J.; Schmieder, D. E. Observer false alarm effects on detection in clutter. *Optical Engineering* **1993**, *32*, 1675-1684.

Dudgeon, D. E. ATR Performance Modeling and Estimation. Massachusetts Institute of Technology, Lexington, Lincoln Lab. 7 Dec 1998. (Report: ESCTR-97-137), 1998.

Duncan, J. Selective attention and the origin of visual information. *Journal of Experimental Psychology: General* **1984**, *113*, 501-517.

Duncan, J.; Humphreys, G. W. Visual search and stimulus similarity. *Psychological Review* **1989**, *96*, 433-458.

Eastman, A. A. Color contrast vs. luminance contrast. *Illuminating Engineering* **1968**, *63*, 316-319.

Egeth, H. E.; Mordkoff, J. T. Redundancy gain revisited: Evidence for parallel processing of separable dimensions. In (Lockhead, G. R., & Pomerantz, J. R., et al., Eds.), *The perception of structure: Essays in honor of Wendell R. Garner*, (pp. 131-143). Washington, DC: American Psychological Association, 1991.

Egeth, H. E.; Virzi, R. A.; Garbart, H. Searching for conjunctively defined targets. *Journal of Experimental Psychology: Human Perception and Performance* **1984**, *10,* 32-39.

Engel, F. L. Visual conspicuity, visual search, and fixation dependencies of the eye. *Vision Research* **1977**, *18,* 95-100.

Enns, J. T.; Austen, E. L. The role of visibility in attentional capture. Paper presented at the 40th Annual Meeting of the Association for Research in Vision and Ophthalmology, Ft. Lauderdale, Florida, May 1999.

Frome, F. S.; Buck, S. L.; Boynton, R. M. Visibility of borders: Separate and combined effects of color differences, luminance contrast, and luminance level. *Journal of the Optical Society of America* **1981**, *71,* 145-150.

Gilbert, C. D.; Das, A.; Ito, M.; Kapadia, M.; Westheimer, G. Spatial integration and cortical dynamics. *Proceedings of the National Academy of Sciences of the United States of America, 93,* 615-622, 1996.

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

Greening, C. P. Alternative approaches to modeling visual target acquisition. Rockwell International for the Naval Weapons Center, China Lake, CA, 1974.

Greening, C. P. Mathematical modeling of air-to-ground target acquisition. *Human Factors* **1976**, *18,* 111-148.

Grossberg, S. Cortical dynamics of three-dimensional figure-ground perception of two-dimensional pictures. *Psychological Review* **1997**, *104,* 618-658.

Grossberg, S.; Mingolla, E.; Ross, W. D. A neural theory of attentive visual search: Interactions of boundary, surface, spatial, and object representations. *Psychological Review* **1994**, *101,* 470-489.

Grossberg, S.; Rudd, M. E. A neural theory of visual motion perception. In (B. Blum, Ed.), *Channels in the visual nervous system: Neurophysiology, psychophysics and models,* (pp. 151-194). London, England: Freund Publishing House, Ltd., 1991.

Grossman, S.; Hadar, Y.; Rehavi, A.; Rotman, S. R. Target acquisition and false alarms in clutter. *Optical Engineering* **1995**, *34,* 2487-2494.

Hacisalihzad, S. S.; Stark, L. W.; Allen, J. S. Visual perception and sequences of eye movement fixations: A stochastic modeling approach. *IEEE Transactions on Systems, Man, and Cybernetics* **1992**, *22*, 496.

Hayward, W. G.; Tarr, M. J. Testing conditions for viewpoint invariance in object recognition. *Journal of Experimental Psychology: Human Perception and Performance* **1997**, *23*, 1511-1521.

Hecker, R. CAMELEON – Camouflage assessment by evaluation of local energy, spatial frequency, and orientation. *In Proceedings of the SPIE Conference on Characterization, Propagation, and Simulation of Sources and Backgrounds II*, *SPIE, 1687,* (pp. 342-349), Bellingham, WA: SPIE, 1992.

Hillstrom, A. P.; Yantis, S. Visual motion and attentional capture. *Perception & Psychophysics* **1994**, *55*, 399-411.

Horowitz, T. S.; Wolfe, J. M. Visual search has no memory. *Nature* **1998**, *357*, 575-577.

Howe, J. D. Electro-optical imaging systems performance prediction. In (J. S. Accetta, D. L. Shumaker, Eds.), *The Infrared and electro-optical systems handbook*, Infrared Information Analysis Center: Bellingham, Wash.: SPIE Optical Engineering Press, 1993.

Hubel, D. H. *Eye, Brain, and Vision*. New York: Scientific American Library: distributed by W. H. Freeman and Company, 1988.

Hubel, D. H.; Wiesel, T. N. Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology* **1962**, *155*, 385-398.

Hubel, D. H.; Wiesel, T. N. Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology* **1968**, *195*, 215-243.

Hummel, J. E.; Biederman, I. Dynamic binding in a neural network for shape recognition. *Psychological Review* **1992**, *99*, 480-517.

Humphreys, G. W.; Muller, H. J. Search via recursive rejection (SERR) – A connectionist model of search. *Cognitive Psychology* **1993**, *25*, 43-110.

Intraub, H. Rapid conceptual identification of sequentially presented pictures. *Journal of Experimental Psychology: Human Perception and Performance* **1981**, *3*, 604-610.

Itti, L.; Koch, C. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research* **2000**, *40*, 1489-1506.

Johnson, D. F. Suprathreshold conspicuity of coloured stimuli in high ambient lighting. (British Aerospace Report JS11511), 1990.

Johnson, J. Analysis of image forming systems. *Proceedings of the Image Intensifier Symposium*, pp. 249-273, U.S. Army Engineer Research and Development Lab, Ft. Belvoir, VA. (DTIC report AD 220 160), Oct. 6-7, 1958.

Johnson, J.; Lawson, W. R. Performance modeling methods and problems. *Proceedings of the IRIS Imaging Symposium*, pp. 105-123, Infrared Information Analysis Center, ERIM, Ann Arbor, MI, 1974.

Kaiser, M. K.; Montegut, M. J. Rotational and translational components of motion parallax: Observers' sensitivity and implications for three-dimensional computer graphics. *Journal of Experimental Psychology: Applied* **1995**, *4*, 321-331.

Karsh, R.; Breitenbach, F. W. Looking at looking: The amorphous fixation measure. In (R. Groner, C. Metz, D. F. Fisher, and R. A. Monty, Eds.), *Eye Movements and Psychological Functions: International Views*, pp. 53-72, Hillsdale, NJ: Lawrence Erlbaum Associates, 1983.

Kennedy, H. V. Two dimensional modeling of FLIR systems. *Proceedings of 1983 Meeting of IRIS Specialty Group on Infrared Imaging*, Infrared Information Analysis Center, ERIM, Anne Arbor, MI, 1983.

Kingdom, F.; Moulden, B. A multi-channel approach to brightness coding. *Vision Research* **1992**, *32*, 1565-1582.

Kooi, F. L.; deValois, K. K. The role of color in the motion system. *Vision Research* **1992**, *32*, 657-668.

Kosnik, W. Quantifying target contrast in target acquisition research. In (E. Peli, Ed.), *Vision models for target detection and recognition: in memory of Arthur Menendez*, pp. 380-395, River Edge, NJ: World Scientific, 1995.

Lamb, M. R.; Yund, E. W. Spatial frequency and attention: Effects of level-, target-, and location-repetition on the processing of global and local forms. *Perception & Psychophysics* **1996**, *58*, 363-373.5.

Legg, G.E.; Foley, J.M. Contrast Masking in Human Vision. *Journal of the Optical Society of America* **1980**, *70*, 1458-1471.

Lillesæter, O. Complex contrast, a definition for structured targets and backgrounds. *Journal of the Optical Society of America, A* **1993**, *10*, 2453-2457.

Lind, J. H. Searching and scanning: A review of Lawrence W. Stark's vision models. Navy Postgraduate School, Monterey, CA. (DTIC report ADA 302569), 1995.

Livingstone, M. S.; Hubel, D. H. Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *The Journal of Neuroscience* **1987**, *7*, 3416-3468.

Lloyd, J. M.; Sendall, R. L. Improved specifications for infrared imaging systems. *Proceedings of IRIS Imaging Symposium*, pp. 109-129, Infrared Information Analysis center, ERIM, Ann Arbor, MI, 1970.

Mack, A.; Rock, I. *Inattentional Blindness*. Cambridge, MA: MIT Press, 1998.

MacMillan, N. A.; Creelman, C. D. *Detection theory: A user's guide.* Cambridge, MA: Cambridge University Press, 1991.

Marr, D. *Vision*. New York: W. H. Freeman and Company, 1982.

Marr, D.; Hildreth, E. Theory of edge detection. *Proceedings of the Royal Society of London Series B – Biological Sciences, 207*, 187-217, 1980.

Martinez-Baena, J.; Fdez-Valdivia, J.; Garcia, J. A.; Fdez-Vidal, X. R. A new image distortion measure based on multisensor organization. *Pattern Recognition* **1998**, *31*, 1099-1116.

Martinez-Baena, J.; Toet, A.; Fdez-Vidal, X. R.; Garrido, A.; Rodriguez-Sanchez, R. Computational visual distinctness metric. *Optical Engineering* **1998**, *37*, 1995-2005.

Mazz, J. ACQUIRE model: Variability in N50 analysis. (USAMSAA technical report TR-631), 1998.

Mazz, J. P.; Kistner, R. W.; Pibil, W. T. Detection of low-contrast moving targets. (USAMSAA draft document), 1998.

McPeek, R. M.; Maljkovic, V.; Nakayama, K. Saccades require focal attention and are facilitated by a short-term memory system. *Vision Research* **1999**, *39*, 1555-1566.

Meitzler, T. J.; Kistner, R. W.; Pibil, W. T.; Sohn, E.; Bryk, D.; Bernarz, D. Computing the probability of target detection in dynamic scenes containing clutter using fuzzy logic approach. *Optical Engineering* **1998**, *37*, 1951-1959.

Meitzler, T. J.; Singh, H.; Arafeh, H.; Sohn, E.; Gerhart, G, R. Predicting the probability of target detection in static infrared and visual scenes using the fuzzy logic approach. *Optical Engineering* **1998**, *37*, 10-17.

Minsky, M. *The Society of Mind*. New York: Simon and Shuster, 1985.

Miskhin, M.; Ungerleider, L. G.; Macko, K. A. Object vision and spatial vision – 2 cortical pathways. *Trends in Neuroscience* **1983**, *6*, 414-417.

Moser, P. M. Mathematical modeling of FLIR performance. NAVAIRDEVCEN Technical Memorandum NACD-20203: PMM. (DTIC report ADA 045247), 1972.

Moulden, B.; Kingdom, F.; Gatley, L. F. The standard deviation of luminance as a metric for contrast in random-dot images. *Perception* **1990**, *19*, 79-101.

Muller, M. J. Texture boundaries: important cues for human texture discrimination. *In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 464-468, 1986.

Nachman, M. The influence of size and shape on the visual threshold of the detectability of targets. Boston University, Optical Research Laboratory. (Technical Note 109, December 1953), 1953.

Nachmias, J. On the psychometric function for contrast detection. *Vision Research* **1981**, *21*, 215-223.

Nakayama, K.; He, Z.J.J. Visual surface representation – an intermediate state between early filtering and object recognition. *Investigative Ophthalmology and Visual Science* **1994**, *35*, 1477-1477.

Nakayama, K.; Mackeben, M. Sustained and transient components of focal visual attention. *Vision Research* **1989**, *29*, 1631-1647.

NATO. Battlefield Emissive Sources Trials under the European Theater Weather and Obscurants. Organized by the North Atlantic Treaty Organization, AC243/Panel4/RSG.15, 1990.

Nichols, W.; Paik, H. A methodology for evaluating clutter effects on observer detection performance. (A.M.L. LaHaie, Ed.). *Proceedings of the 3rd Annual Ground Target Modeling and Validation Conference*, pp. 214-221, Ann Arbor, MI, 1992. (DTIC report AD-B171 616, LIMITED), 1993.

Nicoll, J. F. A mathematical framework for an improved search model. Alexandria, VA: Institute for Defense Analysis. (Institute for Defense Analysis paper P-2901. DTIC report ADA 288858), 1994.

Nicoll, J. F.; Hsu, D. H. A search for understanding. Analysis of human performance on target acquisition and search tasks using eyetracker data. Alexandria, VA: Institute for Defense Analysis. (Institute for Defense Analysis paper P-3036. DTIC report ADA 297602), 1995.

Nothdurft, H. C. Texture segmentation and pop-out from orientation contrast. *Vision Research* **1991**, *31*, 1073-1078.

Noton, D.; Stark, L. W. Eye movements and visual search. *Scientific American* **1991**, *224*, 34-43.

O'Kane, B. L. Validation of prediction models for target acquisition with electro-optical sensors. In (E. Peli, Ed.), *Vision models for target detection and recognition: in memory of Arthur Menendez*, pp. 192-218, River Edge, NJ: World Scientific, 1995.

O'Kane, B. L.; Biederman, I.; Cooper, E. E.; Nystrom, B. An account of object identification confusions. *Journal of Experimental Psychology: Applied* **1997**, *3*, 21-41.

O'Kane, B. L.; Walters, C. P.; D'Angostino, J. Report on experiments in support of thermal TAMIP. Night Vision and Electronic Sensors Directorate, Fort Belvoir, VA, 1993.

O'Neill, G. J. The quantification of image detail as a function of irradiance by empirical tests. (NAVAIRDEVCEN Technical Memorandum NADC-202139:GJO), 1974.

Olacsi, G.S.; Beaton, R.J. *Examining Visual Masking Effects on Target Acquisition Using Two-Dimensional Fourier Analysis Techniques*; unpublished manuscript; Virginia Polytechnic Institute: Blacksburg, VA, 1998.

Olds, E. S.; Cowan, W. B.; Jolicoeur, P. Spatial organization of distracters in visual search. *Canadian Journal of Experimental Psychology* **1999**, *53*, 150-159.

Olzak, L. A.; Thomas, J. P. Configural effects constrain Fourier models of pattern-discrimination. *Vision Research* **1992**, *32*, 1885-1898.

O'Regan, J. K. Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology* **1992**, *46*, 461-488.

O'Regan J. K.; Rensink R. A.; Clark J. J. Change-blindness as a result of 'mudsplashes.' *Nature* **1999**, *398*, 34-34.

Overington, I. Towards a complete model of photopic visual threshold performance. *Optical Engineering* **1982**, *21*, 2-13.

Overington, I.; Brown, M. B.; Clare, J. N. (B.A.C. (GW) Report ST15153), 1977.

Palmer, J.; McLean, J. Imperfect, unlimited-capacity, parallel search yields large set-size effects. Paper presented at the Society for Mathematical Psychology, Irvine, CA, 1995.

Parkhurst, D.; Law, K.; Niebur, E. Modeling the role of salience in the allocation of overt visual attention. *Vision Research* **2002**, *42*, 107-123.

Peterson, H. E.; Dugas, D. J. The relative importance of contrast and motion in visual detection. *Human Factors* **1972**, *14*, 207-216.

Pollack, I.; Norman, D. A. A non-parametric analysis of recognition experiments. *Psychonomic Science*, 1964.

Posner, M. I.; Snyder, C. R.; Davidson, B. J. Attention and the detection of signals. *Journal of Experimental Psychology: General* **1980**, *109*, 160-174.

Pratt, W. K. *Digital Image Processing, 2nd Ed.*, New York: Wiley, 1991.

Quadripartite Working Group on Army Operational Research. *Search and target acquisition nomenclature,* Quadripartite Advisory Publication, Special Working Part on the Modeling of Target Acquisition, May 1990, 1990.

Ramachandran, V. S.; Gregory, R. L. Does colour provide an input to human motion perception? *Nature* **1978**, *275*, 55-56.

Ratches, J. A.; Lawson, W.R.; Ober, L.P.; Bergemann, R.J.; Cassidy, T.W.; Swenson, J.M. *Night Vision Laboratory Static Performance Model for Thermal Viewing Systems*; ECOM-7043; U.S. Electronics Command, April 1973.

Rayner, K.; McConkie, G. W.; Ehrlich, S. Eye-movements and integrating information across fixations. *Journal of Experimental Psychology: Human Perception and Performance* **1978**, *4*, 529-544.

Reike, F.; Warland, D.; deRuyter van Steveninck, R.; Bailek, W. *Spikes: Exploring the Neural Code*. Cambridge, MA: MIT Press, 1997.

Reisfeld, D.; Wolfson, H.; Yeshurun, Y. Context-Free Attentional Operators: The Generalized Symmetry Transform. *International Journal of Computer Vision* **1995**, *14*, 119-130.

Rensink, R. A. The attentional capacity of visual search under flicker conditions. *Perception (Suppl.)* **1996**, *25*, 2.

Rensink, R. A. How much of a scene is seen? The role of attention in scene perception. Paper presented at the 38th Annual Meeting of the Association for Research in Vision and Ophthalmology, Ft. Lauderdale, Florida, 1997, May.

Rensink, R. A.; O'Regan, J. K.; Clark, J. J. To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science* **1997**, *8*, 368-373.

Richards, W.; Polit, A. Texture matching. *Kyberkenetik* **1974**, *16*, 155-162.

Rogers, J. G. Peripheral contrast thresholds for moving images. *Human Factors* **1972**, *14*, 199-205.

Rohaly, A. M.; Ahumada, A. J.; Watson, A. B. Object detection in natural backgrounds as predicted by discrimination performance and models. *Vision Research* **1997**, *37*, 3225-3235.

Rosenfeld, D.; Wolfson, H.; Yeshurun, Y. Context-free attentional operators: the generalized symmetry transform. *International Journal of Computer Vision* **1995**, 14, 119-130.

Rotman, S. R. Modeling human search and target acquisition performance: II. Simulating multiple observers in dynamic scenarios. *Optical Engineering* **1989**, *28*, 1223-1226.

Rotman, S. R.; Gordon, E. S.; Kowalczyk, M. L. Modeling human search and target acquisition performance: I. First detection probability in a realistic multi-target scenario. *Optical Engineering* **1989**, *28,* 1216-1222.

Rotman, S. R.; Hsu, D.; Cohen, A.; Shamay, D.; Kowalczyk, M. L. Textural metrics for clutter affecting human target acquisition. *Infrared Physics & Technology* **1996**, *37,* 667-674.

Rotman, S. R.; Kowalczyk, M. L.; George, V. Modeling human visual search and target acquisition performance: Fixation-point analysis. *Optical Engineering* **1994**, *33,* 3803-3809.

Rotman, S. R.; Tidhar, G.; Kowalczyk, M. Clutter metrics for target detection systems. *IEEE Transactions on Aerospace Electronics Systems* **1994**, *30,* 81-90.

Ryll, E. Subject effectiveness and map-of-the-earth: Final report of project TRACE. Buffalo, NY: Cornell Aeronautical Laboratory. (Report VE-1519-G-1), 1962.

Scanlan, L. A.; Agin, A. K. A behavioral model of target acquisition in realistic terrain. Man-Machine Systems Section, Display Systems Lab, Engineering Division and Advanced Programs Lab, Tactical Systems Division, Hughes Aircraft Co., Culver City, California. (Hughes Technical Report P78-70R, pp. 1-95), 1978.

Schmidt, W. C. Artificial looming yields improved performance over lateral motion: Implications for stereoscopic display techniques. *Human Factors* **1997**, *39,* 352-358.

Schmieder, D. E.; Weathersby, M. R. Detection performance in clutter with variable resolution. *IEEE Transactions on Aerospace Electronics. Systems* **1983**, *19*, 622-630.

Schneider, W. X.; Deubel, H. Visual attention and saccadic eye movements: Evidence for obligatory and selective spatial coupling. In (J. M. Findlay, R. Walker, et al., Eds.), *Eye movement research: Mechanisms, processes and applications. Studies in visual information processing, 6,* pp. 317-324, Amsterdam, Netherlands: Elsevier Science, 1995.

Schneider, W.; Dumais, S. T.; Shiffrin, R. M. Automatic and controlled processing and attention. In (R. Parasuraman & D. R. Davies, Eds.), *Varieties of Attention,* pp. 1-27. Orlando, FL: Academic Press, 1984.

Scott, L.; D'Angostino, J. *FLIR92 Thermal Imaging Systems Performance Model.* NVESD, Ft. Belvoir, VA, 1992.

Self, H. Image evaluation for the prediction of the performance of a human observer. NATO Symposium on Image Evaluation, 1969.

Shirvaikar, M. V.; Trivedi, M. M. Developing texture-based image clutter measures for object detection. *Optical Engineering* **1992**, *31*, 2628-2639.

Silk, J. D. A model of false alarms in target acquisition by human observers. Alexandria, VA: Institute for Defense Analysis. (Institute for Defense Analysis paper P-3076. DTIC report ADA 301181), 1995a.

Silk, J. D. Statistical and modeling uncertainties in the Thermal Target Acquisition Model Improvement Program (TAMIP) predictions. Alexandria, VA: Institute for Defense Analysis. (Institute for Defense Analysis paper P-3078. DTIC report ADA 301182), 1995b.

Silk, J. D. Modeling the observer in target acquisition. Alexandria, VA: Institute for Defense Analysis. (Institute for Defense Analysis paper P-3102. DTIC report ADA 326220), 1997.

Skjervold, J. Extensions of the US Night Vision Laboratory model for thermal viewing systems on structural targets and backgrounds in cluttered scenes. *Proceedings of the SPIE Conference on Targets and Backgrounds: Characterization and Representation,* SPIE vol. 2469, pp. 568-575, Bellingham, WA: SPIE, 1995.

Snowden, R. J.; Braddick, O. J. The temporal integration and resolution of velocity signals. *Vision Research* **1991**, *31,* 907-914.

Stark, L. W. New quantitative evidence for the scanpath theory: Top-down visionl in humans and robots. *In Proceedings of the First Meeting of the International Society of Theoretical Neurobiology*, Milano, Italy, 1993.

Tarr, M. J.; Bulthoff, H. H. Is human object recognition better described by geon structural descriptions or by multiple views – Comment on Biedermand and Gerhardstein (1993). *Journal of Experimental Psychology: Human Perception and Performance* **1995**, *21,* 1494-1505.

Theeuwes, J. Abrupt luminance change pops out; abrupt color change does not. *Perception & Psychophysics* **1995**, *57,* 637-644.

Theeuwes, J.; Burger, R. Attentional control during visual search: The effect of irrelevant singletons. *Journal of Experimental Psychology: Human Perception and Performance* **1998**, *24,* 1342-1353.

The MathWorks, Inc. *Fuzzy Logic Toolbox*, for use with MATLAB, 1995.

Thomas, J. P.; Olzak, L. A. Cue summation in spatial discrimination. *Vision Research* **1990**, *30,* 1865-1875.

Thomas, S. R.; Barsalou, N. Applying human spatial vision models to real-world target detection and identification: A test of the Wilson model. In (E. Peli, Ed.), *Vision models for target detection and recognition: in memory of Arthur Menendez*, pp. 219-244, River Edge, NJ: World Scientific, 1995.

Tidhar, G.; Reiter, G.; Avital, Z.; Hadar, Y.; Rotman, S. R.; George, V.; Kowalczyk, M. L. Modeling human search and target acquisition performance: IV, detection probability in the cluttered environment. *Optical Engineering* **1994**, *33*, 801-808.

Toet, A. Target acquisition in complex scenes, part A: Search and conspicuity models. TNO Human Factors Institute, Soesterberg, Netherlands, ADA 332390, 1996.

Toet, A.; Bijl, P.; Kooi, F. L.; Valeton, J. M. Image data set for testing search and detection models. TNO Human Factors Research Institute, Soesterberg, the Netherlands. (TNO Report TNI-TM 1997 A-036), 1997.

Tolhurst, D.J.; Barfield, L.P. Interactions Between Spatial Frequency Channels. *Vision Research* **1978**, *18*, 951-958.

Tomkinson, D. *ACQUIRE*. Night Vision and Electronic Sensors Directorate, Ft. Belvoir, VA, 1990.

Townsend, J. T. A note on the identification of parallel and serial processes. *Perception & Psychophysics* **1971**, *10*, 161-163.

Townsend, J. T. Serial and parallel processing: Sometimes they look like Tweedledum and Tweedledee but they can (and should) be distinguished. *Psychological Science* **1990**, *1*, 46-54.

Treisman, A.; Gelade, G. A feature-integration theory of attention. *Cognitive Psychology* **1980**, 12, 97-136.

Treisman, A.; Sato, S. Conjunction search revisited. *Journal of Experimental Psychology: Human Perception and Performance* **1990**, *16*, 459-478.

Turano, K. A.; Geruschat, D. R.; Baker, F. H. Oculomotor strategies for the direction of gaze tested with a real-world activity. *Vision Research* **2003**, *43*, 333-346.

Valeton, J. M.; Bijl, P. Target acquisition: Human observer performance studies and TARGAC model validation. (DTIC report ADA 297133),1995.

Vaughan, B. D. Gone but not forgotten: Characteristics of the occluded scene. *Dissertation Abstracts International*, 1998.

Vecera, S. P.; Farah, M. J. Does visual attention select objects or locations? *Journal of Experimental Psychology: General* **1994**, *123*, 146-160.

Verghese, P.; Pelli, D. G. The scale bandwidth of visual search. *Vision Research* **1994**, *34*, 955-962.

Verghese, P.; Stone, L. S. Combining speed information across space. *Vision Research* **1995**, *20*, 2811-2823.

Verghese, P.; Watamaniuk, S.N.J.; McKee, S. P.; Gryzwacz, N. M. Local motion detectors cannot account for the detectability of an extended trajectory in noise. *Vision Research* **1999**, *39,* 19-30.

Vol, I. A.; Pavlovskaja, M. B.; Bondarko, V. M. Similarity between Fourier-transforms of objects predicts their experimental confusions. *Perception & Psychophysics* **1990**, *47,* 12-21.

Waldman, G.; Wooton, J.; Hobson, G.; Leutkemeyer, K. A normalized clutter metric for images. *Computer Vision, Graphics, and Image Processing* **1988**, *42,* 137-156.

Watson, A. B. The cortex transform: Rapid computation of simulated neural images. *Computer Vision, Graphics, and Image Processing* **1987**, *39,* 311-327.

Watson D. G.; Humphreys G. W. Segmentation on the basis of linear and local rotational motion: Motion grouping in visual search. *Journal of Experimental Psychology: Human Perception and Performance* **1999**, *25,* 70-82.

Watt, R. J.; Morgan, M. J. A theory of the primitive spatial code in human vision. *Vision Research* **1985**, *25,* 1661-1674.

Williams, L. G. Target conspicuity and visual search. *Human Factors* **1966**, *8,* 80-92.

Wilson, H. R. Psychophysical models of spatial vision and hyperacuity. In (D. Regan, Ed.), *Vision and Visual Dysfunction: Spatial Vision*, *10,* CRC Press: Boca Raton, FL, pp. 64-68, 1991.

Wolfe, J. M. Visual search in continuous, naturalistic stimuli. *Vision Research* **1994a**, *34,* 1187-1195.

Wolfe, J. M. Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review* **1994b**, *1,* 202-238.

Wolfe, J. M. What can 1 million trials tell us about visual search? *Psychological Science* **1998**, *9,* 33-39.

Wolfe, J. M.; Bennett, S. C. Preattentive object files: Shapeless bundles of basic features. *Vision Research* **1997**, *37,* 25-43.

Wolfe, J. M.; Cave, K. R.; Franzel, S. L. Guided Search: An alternative to the Feature Integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance* **1989**, *15,* 419-433.

Wolfe, J. M.; Gancarz, G. Guided search 3.0: A Model of Visual Search Catches Up With Jay Enoch 40 Years Later. In (V. Lakshminarayanan, Ed.), *Basic and Clinical Applications of Vision Science*, p189-192, Dordrecht, Netherlands: Kluwer Academic, 1996.

Yantis, S. Attentional capture in vision. In (A. F. Kramer, M. G. H. Coles, & G. D. Logan, Eds.), *Converging operations in the study of visual selective attention*. pp. 45-76, Washington, DC: American Psychological Association (1996)..

Yantis S.; Egeth H. E. On the distinction between visual salience and stimulus-driven attentional capture. *Journal of Experimental Psychology: Human Perception and Performance* **1999**, *25*, 661-676.

Yantis, S.; Hillstrom, A. P. Stimulus-driven attentional capture: Evidence from equiluminant visual objects. *Journal of Experimental Psychology: Human Perception & Performance* **1994**, *20*, 95-107.

Yarbus, A. L. *Eye Movements and Vision*. New York: Plenum Press, 1967.

Zadeh, L. Fuzzy sets. *Information Control* **1965**, *8*, 338-353.

Zeki, S. *A Vision of the Brain*, Oxford, England: Blackwell Scientific Publications, Inc,1993.

10. Bibliography

Avital, Z.; George, V.; Hadar, Y.; Kowalczyk, M. L.; Reiter, G.; Rotman, S. R.; Tidhar, G. Modeling human search and target acquisition performance: Detection probability in the cluttered environment. In (A. M. L. LaHaie, Ed.). *Proceedings of the 3rd Annual Ground Target Modeling and Validation Conference*, pp. 53-68, U.S. Army Tank-Automotive Command, Warren, MI, 1993.

Blackwell, H. R.; McCready, D. W. Foveal detection thresholds for various durations of target presentations. *Minutes and Proceedings of the NAS-NRC Vision Committee*, p. 249 AGSIL/53/4405, 1952.

Doll, T. J.; Schmeider, D. E.; McWhorter, S. W. Simulation of human visual search and detection. In (A. M. L. LaHaie, Ed.). *Proceedings of the 3rd Annual Ground Target Modeling and Validation Conference*, pp. 247-255, Ann Arbor, MI, 1992. (DTIC report AD-B171 616, LIMITED), 1992.

Driggers, R. G.; Cox, P. G.; Leachtenauer, J.; Vollmerhausen, R.; Scribner, D. A. Target and intelligence electro-optical recognition modeling: A juxtaposition of the probabilities of discrimination and the general image quality equation, 1998.

Lawson, W. R.; Cassidy, T. W.; Ratches, J. A. A search prediction model. *Proceedings of IRIS Specialty Group on Imaging*, Infrared Information Analysis Center, ERIM, Ann Arbor, MI (June 1978), 1978.

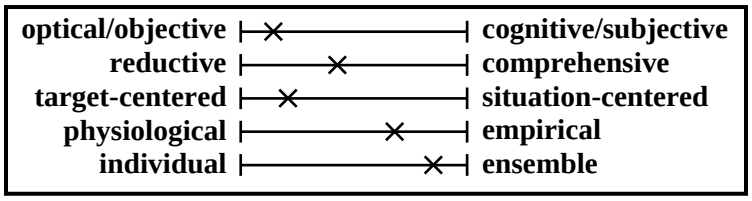
van Meeteren, A. Characterization of task performance with viewing instruments. *Journal of the Optical Society of America A* **1990**, *7*, 2016-2023.

Vos, J. J.; van Meeteren, A. PHIND: An analytic model to predict target acquisition distance with image intensifiers. *Applied Optics* **1991**, *30*, 958-966.

Appendix A. Models and Modeling Concepts of Interest

This appendix contains descriptions of models that have been influential or discussed in detail in the report. Models are discussed in terms of where they fall on the five classification axes, how they function, what they predict, their relations to the topics of interest, and a critique of their strengths and weaknesses.

British Aerospace ORACLE Model (Overington, Brown, & Clare, 1977; Cooke, Stanley, & Hinton, 1995)



CAUTIONARY NOTE: The ORACLE model is proprietary to British Aerospace and (so far as this reviewer can determine) has never been published *in toto*. The model consists of several modules for performing specific visual tasks such as motion, color, depth, etc. The following description is for the general ORACLE framework and its application in search and discrimination of achromatic, static, luminance-defined targets.

Basic operating principles:

- Focus is on the known physiology/anatomy of visual system, primarily the optics of the eye and the anatomy of the retina.
- The model bases its predictions on retinal image of elements of the scene.
- Assumption: Edges of a target rather than the total energy within it are significant.
    - Threshold detection is therefore based on strength of signal arising from luminance gradients across adjacent retinal receptors.
- Signal strength must exceed a noise term for a decision to be made.

Flow of processing:

- Mean scene luminance is used to determine the level of adaptation of the visual system.
- Mean scene luminance and field of view determine pupil diameter.
- Pupil size and non-linear optical properties of eye structures determine point spread function and modulation transfer function of the eye's optics.
- The point spread function determines how a target image of a particular size and luminance contrast (including edge gradient or sharpness) is represented as an image on the retina.
- The sum of the activity of photoreceptors around the edge of the target constitutes the signal.

- All of the above processes are based on known anatomical and psychophysical properties of the visual system and include such factors as eccentricity, photopic and scotopic acuity, the distribution of retinal receptors, vertical/horizontal asymmetries in acuity.

How search is characterized:

- Glimpse duration is constant (1/3 of a second).
- Glimpse locations are independent. (I.e., random sampling with replacement.)
- Search progresses in soft-shell manner.
- Soft shell characteristics are modeled as a distribution of population (i.e., known) hard shell sizes.
- Clutter causes soft shell distribution to lean more towards smaller shells.

How detection is characterized:

- Detection is based on Ricco's law (i.e., that threshold contrast of a target is proportional to its area).

How recognition is characterized:

- Based on ability to resolve detail within the target signature (i.e., detectable changes in the perimeter of the target).
- Two adjacent features must be resolvable for discrimination to be possible.

How color is characterized:

- NOTE: The available documentation on the model did not go into detail although the authors do acknowledge that ORACLE's predictions related to color conspicuity are accurate (see below).
- Color in ORACLE is based on R and G cones only, using cone response sensitivity data.

ORACLE framework can be used to calculate response to Johnson-like bar patterns by determining point spread of constituent Fourier components (odd sinusoids) of the bar pattern's square waveform.

- Four spatial scales (analogous to responses from sets of single, 3, 9, and 27 adjacent photoreceptors) are incorporated into the model in order to accommodate psychophysical results related to the overall contrast sensitivity function of the eye.

NOTES:

- Motion not included in model.
- Model has not been validated in general – only piecewise agreement with psychophysics.
    - There is no sufficient database with which to validate the model (authors).
- More interested in optics of the eye than other models.

Targets are assumed to be larger than 9 arc min in diameter.

- The model assumes that edges are important, yet it is able to model Ricco's Law for small targets. This is only possible for targets that are not elongated; otherwis,e the edge-based

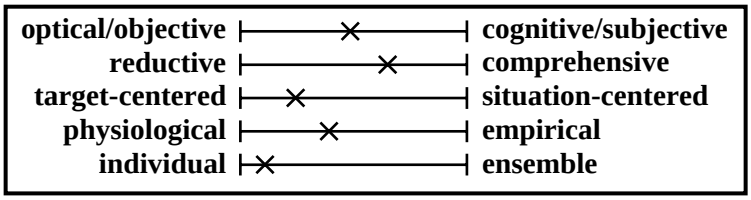
signal might be stronger than the area-based signal. Possibly small targets are blurred so that this is not a problem?

- ORACLE's visibility (target SNR) was compared to human ratings of conspicuity of colored targets against colored backgrounds (Johnson, 1990). The model agreed well with human data. However, no specifics were given as to how color was included in the model tested.

CRITIQUE:

- Model suffers from lack of fixation guidance mechanism. Effects of clutter only alter attributes of soft shell lobe.
- ** Top-down characteristics are not implemented into model because they govern how search will progress through a particular scene. The authors argue that "…the effort in modeling at an equivalent level of detail is far greater than the reward for many practical situations." As a result, they select lobe sizes that will produce experimentally measured cumulative search distributions over time.
- Models such as this one are likely less accurate since stimuli upon which they operate approach the limits of any psychophysical measurements upon which the model is based. Overington (1982) pointed out that that models based on psychophysics have specific "envelopes of usage" where their predictions are accurate. Outside such envelopes, error propagates from step to step in calculation, resulting in degradation in overall performance.
- Looming targets (targets that approach the observer along their line of sight) are modeled as an increase in size and apparent contrast only. Such a characterization is inadequate to model the phenomenon of looming.
- Perceptual learning not included in model, so performance cannot improve with practice.
- It is unclear how practice effects could be included since the model's psychophysical basis does not include data for trained versus untrained observers.

Georgia Tech Vision (GTV/VISEO) Model (Doll, McWhorter, Schmieder, & Wasilewski, 1995; Doll, McWhorter, Wasilewski, & Schmieder, 1998)



GTV is the general purpose vision model produced by Georgia Tech. The military target acquisition model VISEO (Doll et al., 1997) incorporates GTV into a number of processing modules.

Basic operating principles:

- Focus is on psychophysics and multi-channel SF modeling.
- Decisions are based on the object-by-object point probability of being fixated and that a fixated object will be judged a target.
- SNR and clutter are incorporated.
- Conspicuities of objects in the scene determine the probability that they will be fixated.
    - Clutter and training are involved in determining these.
- Pre-attentive scene segregation into "blobs" (i.e., target-like regions) is based on texture segmentation.
- GTV models visual system as output of many (56) oriented spatial frequency-selective channels, the output of which undergoes various operations. The goal of the model is to intelligently combine information from channel output of early vision so that targets can be distinguished from clutter.
    - GTV includes a simulation of optics of the eye as well as retinal and cortical V1, V4 (color), and MT (motion) visual processing area.
- GTV consists of a pre-processor that takes a scene or display and converts it into a map of one rod plus three cone output, followed by five processing stages: a "front end," pre-attentive and attentive modules that run in parallel, a selective attention/training module, and a performance module.

Each stage, in more detail:

Stage 1 – Front End

- Luminance: concerns receptor pigment bleaching, pupil dilation, receptor thresholds/ bleaching, flicker, and transient luminance changes
- Color: converts the pre-processed image from short, medium, long wavelength receptor activity to R/G and B/Y color opponent pairs and cone luminance signal
- → output to pre-attention and attention stages in parallel

Stage 2 – Pre-attention module (search information)

- Performs calculations of conspicuity for peripheral vision.
- Motion: temporal filtering extracts local motion signals; temporal integration adds blur to high spatial frequencies of the image

- Filtered separately from spatial information, similarly to human V4/MT processing (Livingstone & Hubel, 1987)
- Motion processing based only on scotopic (rod) and mean photopic (cone luminance) information, not chromatic information.
- Pattern perception unit in module decomposes image into oriented SF channels
- Number of channels depends on source (though always four orientations): two from rods, 12 from each color opponency, 24 from cone luminance.
- Interactions between rectified, filtered channels are simulated.
- Texture information extracted.
- → output to selective attention module

Stage 3 – Attentional module (detection discrimination information)

- Performs similar calculations to stage 2, only now for foveal feature extraction, i.e., different acuities, color and motion sensitivities, etc.
- → output to selective attention module

Stage 4 – Selective attention module (assignment of P_{fix}, $P_{yes|fix}$; training)

- Uses weighted pre-attentive output to segment scene into objects.
- Uses neural network to set weights. Weights for discriminant function attempt to distinguish between target and background pixels. The neural network uses training to set up this discriminant function.
- →output of pre-attentional operations is a set of blobs representing potential targets. Uses weighted attentive output to segment foveal scene into objects.
- → output of same processes as on pre-attentive information (only now using filters tuned for the fovea) is a map containing target-like foveal objects.
- → output to Performance Module
- NOTE: The neural network that sets the weights of pre-attentive and attentive representation features that are to be stressed must be trained before GTV is run.

Stage 5 – Performance module

- Performance module computes measures of search and discrimination performance based on output of selective attention module:
- Calculation of P_d, P(FA), d', and RT:
- The model simulates an observer selecting fixation locations by means of a noisy decision process-based conspicuity. Conspicuity is a function of the pre-attentional P_{fix} calculation, noise, clutter, and the spacing of objects:
- quantum noise, neural noise, and clutter (defined as "extent to which a clutter blob's luminance, texture, chromatic information, and temporal contrast match the target")
- Spacing of target blob with respect to clutter blobs also influences conspicuity (consistent with Duncan & Humphreys, 1989).
- At each location, the signal-to-clutter ratio is calculated:
- The fixated object signal is based on the pooled attentive output summed over the blob area.
- The SCR is = (signal – average clutter blob signal)/standard deviation of

clutter blob signals

- Appealing to signal detection theory, the SCR is equated to d'. Thus, $P_{yes|fix}$ and P(FA) can be calculated once a decision criterion has been assumed or measured. $P_d = P_{fix} \times P_{yes|fix}$
- Once $P_{yes|fix}$ is known, the model can determine how many glimpses are required before a decision is rendered. (The model assumes that fixations are selected from high P_{fix} locations without replacement.) Given a constant glimpse duration, RT can be calculated.

Model predictions:

- Probability that a "blob" is fixated on glimpse i: $P_{fix}(i)$
- Probability that a blob, once fixated, is determined to be a target: $P_{yes|fix}(i)$
- P_d (given a criterion for decision making according to SNR)
- RT, based on number of glimpses required to make judgment.

NOTES:

- Motion contributes to conspicuity and causes blur before SF analysis.
- Masking (interactions between SF channels) is implemented in the model.
  - Channels are therefore not independent (see Olzak & Thomas, 1992, for a discussion of such models)
- Glimpse duration assumed to be a *constant* 1/3 second. That is, all glimpses during search are exactly 333 ms.
- Motion can but does not necessarily increase the conspicuity of a moving object.

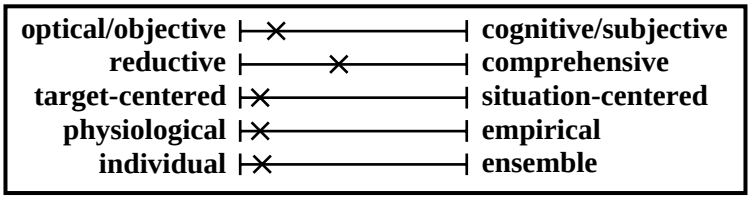
CRITIQUE:

- The calculation of all foveal features (by attention module) at the same time is not physiologically realistic. The model would be more realistic and behave identically if it were to calculate the foveal features only after a blob has been selected by the performance module. (This behavior takes into account the unbound feature nature of pre-attentive and post-attentive vision by Wolfe & Bennett, 1997.)
- Incorporation of pre-attentive stage to drive eye movements is a good idea.
- Training at both pre- and attentive levels is also a good idea.
- Eye movement assumptions (i.e., selection without replacement) are unrealistic (e.g., Horowitz & Wolfe, 1998; Nicoll & Hsu, 1995).
- Although the model can in theory handle target-absent trials (i.e., no response is made if every pre-attentive blob is investigated and none has sufficient signal strength to trigger a "yes" response), it can only do so if serial self-terminating search processes are assumed. Such an assumption, that after each potential target is investigated once only an absent judgment is made, does not appear to be the case (Chun & Wolfe, 1996). Observers tend to over-search and are hesitant to report the absence of a target.
- Attention can only refer to the *presence* of features, not their absence. As such, a less green object among more green objects should be quite inconspicuous, though in reality it may be quite conspicuous (although the search asymmetry literature indicates that it would not be *as* conspicuous as the obverse e.g., Wolfe, 1994b]).
- Training issues:
  - Training's effect is entirely based on automaticity (Schneider, Dumais, &

Shiffrin, 1984). That is, with training, any combination of features can cause an increase in conspicuity. In reality, some conjunctions of features cannot be learned by humans (such as orientation-color combinations). This lack of constraint on what can be learned manifests itself as the model's out-performance of humans and the need to add noise to make it behave more like a human observer (Doll et al., 1998).

- It is unclear to what degree the training is generalizable to slightly different targets or to what degree more than a single target can be trained at a time (as are many neural net-based representations). These possibilities were addressed in neither paper.
- Problem with motion implementation:
  - Because motion information is scalar (only related to speed, not direction), the model's attention mechanism has no direction selectivity, which the human visual system does.
  - Therefore, GTV can only distinguish between speeds. This does not allow the system to extract information about motion parallax and how a moving target's violation of parallax is plainly visible.
- Foveation is required for detection! Even though the model is ostensibly based on the conspicuity of targets, highly conspicuous targets must still be fixated for the model to produce a "yes" response. This result is inconsistent with pop-out (e.g., Yantis & Egeth, 1999).

Itti and Koch (2000) Saliency-Based Attention and Fixation Selection Model

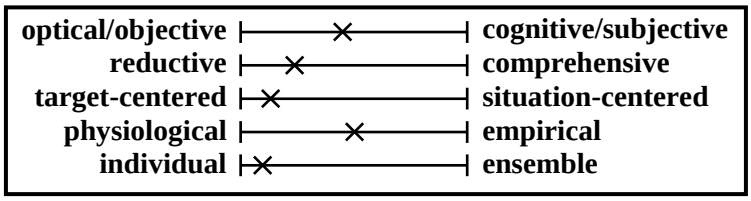


The Itti and Koch (2000) model is purely bottom-up in nature, though it was modified with limited success by Turano et al. (2003) to incorporate crude representations of target features and target location. The model is based heavily on known aspects of human, primate, and mammalian (feline, primarily) visual psychphysics, neuroanatomy, and electrophysiology. The model encodes the visual scene along three feature dimensions (luminance intensity, orientation, and opponent-pair color contrast) at multiple scales. Activation within each feature dimension is used to create a conspicuity map for that feature. These three conspicuity maps are then combined into a single saliency map. The model defines the next fixation location as that corresponding to the point of maximum activation in the saliency map. Inhibition of return is invoked as a temporary inhibition of this location in the saliency map to prevent immediate re-fixation of the same location in the scene.

CRITIQUE:

- → The model does not incorporate transient visual events (flashed, motion, etc.) into its calculation of saliency, even though those events have been demonstrated to capture visual attention (e.g., Yantis, 1996).
- → The model's performance for simple stimuli such as oriented and colored line segments is a good match for human performance. However, though the model's behavior in real-world scenes seems subjectively to be reasonable and actually located targets far faster than would a random fixation generator, results from Itti and Koch (2000) indicate that it does a poor job of predicting the response times for human observers to detect targets. Further results from Turano et al. (2003), though coded differently for fixation location, indicate that the Itti and Koch (2000) model performed at chance levels during a real-world mobility task.

Guided Search (Wolfe, Cave, & Franzel, 1989; Wolfe, 1994b; Wolfe & Gancarz, 1996)



Wolfe and colleagues' Guided Search models integrate stimulus-driven (bottom-up) and goal-directed (top-down) mechanisms in the deployment of attention (and eye movements in Wolfe & Gancarz, 1996) about a scene. That is, the models all incorporate observer knowledge of target attributes and guide attention to objects in the scene that have those attributes. (Note that these models are based on simple objects such as oriented, colored line segments with simple, separable features. They were not intended to be applied in their present form to real-world target acquisition situations. Wolfe [1994a] did, however, apply the guided search framework to "naturalistic" stimuli with some success.)

Pre-attentive system attributes

- Operates in parallel across visual scene.
- Creates a map of features present at various locations in the scene
    - Features: orientation, size, color, luminance, motion, depth
    - One spatial map per feature, with activation level indicating feature presence.
    - Activation level a function of feature and both difference within feature dimension from neighbors (dissimilar neighbors → higher activation than similar neighbors) and distance between items (close → higher activation than far). Therefore, pre-attentive system calculates feature loadings of items and also *distinctness* of items.
        - (This incorporates findings of Duncan & Humphreys, 1989, and Nothdurft, 1991.)
- Noise is added to feature map locations.

Attentional system attributes

- Top-down feature maps contain information about features present in the target.
- Features that are unique are given highest weight.

Combination of bottom-up and top-down activations:

- A master activation map is created. High activations result from locations that weigh heavily on several feature maps from top-down and bottom-up processing.

Search and detection:

- Search progresses in order from highest activation location to lowest.
- IOR is implemented so that once a location is searched, it is not searched again. (This is an unrealistic assumption. See Nicoll & Hsu, 1994, for data contradicting this.)
- Search progresses until (1) a target is found, (2) a specific period of time has passes without finding a target, or (3) activations are judged by the observer to be too low to be targets.

- Target detection is based on SDT: If activation of bottom-up maps is greater than a decision criterion, a detection is rendered.
- → Situation (1) is a hit or a false alarm; (2) or (3) may create misses or correct rejections.
- False alarms are a result of the decision criterion being shifted downward after a miss (assuming subjects are given feedback). False alarms, in turn, shift the criterion upward.

CRITIQUE:

- → Model makes specific, testable predictions about search performance in simple tasks.
- Guided Search is useful in that it assumes (probably correctly) that covert shifts of attention (i.e., attentional movement without subsequent eye movement) and eye movements are determined in large part by a parallel pre-attentional system.
- Features are weighed so that search asymmetry results, pop-out, and top-down attentional control settings are accounted for.
- Incorporation of several results from search literature (e.g., pop-out for feature singletons, similarity and proximity effects).
- The assumption of serial self-terminating search is almost certainly incorrect.
- The generation of errors in the model is problematic and seems almost atheoretical.
- Inclusion of the mechanism to generate errors does create a reasonable looking speed-accuracy trade-off.
- Cannot be extended to situations in which features are not clearly delineated.
- Does not work for continuous, naturalistic stimuli such as textures.
- Does not have a mechanism to perform a difficult detection or any kind of discrimination task.
- Inclusion of IOR is interesting, though it is unclear what role IOR actually plays in search. See section of report on assumptions of neoclassical search framework. Memoryless search does not permit IOR to occur.

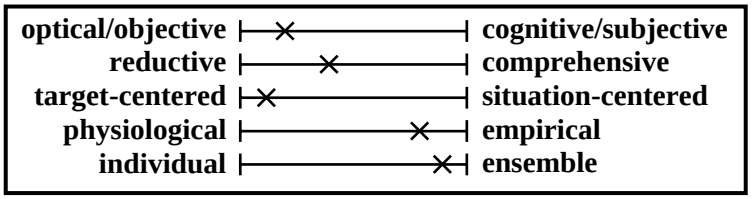
Human Spatial Vision Model (Wilson, 1991) → Filters for spatial vision

The filters that determine spatial vision:

Mechanism	Basis Frequency	Number of Orientations	Weighted Locations	Number of filters	Filter Contrast Sensitivity
A	0.9 cpd	6	6	36	30.0
B	1.7 cpd	7	36	252	70.0
C	2.8 cpd	8	49	392	140
D	4.0 cpd	9	100	900	150
E	8.0 cpd	11	256	2816	76.7
F	16 cpd	12	961	10,532	18.4
Total number of filters: 15,928					

cpd = cycles per degree

Johnson's (1958) bar pattern equivalence study and the Johnson Criteria



Basics:

- Static performance model
- Stationary targets
- Achromatic
- Uniform background

Johnson attempted to establish a relationship between the number of lines resolvable on a target through an imaging device and the degree to which that target could be acquired. Subjects viewed scale models of eight vehicles and a Soldier through an I^2 device and were asked to (1) detect, (2) determine the orientation of, (3) recognize, or (4) identify the target. (The level of discrimination in task (2) is referred to as classification.)

Bar charts of the same contrast and scale as the target models were also displayed to subjects. At each scale and contrast, Johnson desired to know how many cycles were resolvable. The maximum number of resolvable bar cycles across the target's critical dimension was determined for each task:

$$N = H_{targ} \cdot f_x$$

in which N = number of cycles resolvable across target critical dimension,
H_{targ} = critical dimension of the target,
f_x = highest bar pattern spatial frequency (fundamental frequency of bar).

Johnson found that so long as the contrasts of the bar (light versus dark bands) and target (target versus background) were equal, the number of cycles on target was found to be independent of both target contrast and scene luminance. In other words, the ability of an observer to perform a discrimination task was related solely to their ability to resolve bar patterns. The following table lists the average number of cycles required for an ensemble of observers to acquire various military targets with 50% accuracy, defined as Pd = P(detect | present):

Resolution across critical dimension to perform 50% accurate acquisition at a level of:			
Detection	Orientation (classification)	Recognition	Identification
1.0±0.25	1.4±0.35	4.0±0.8	6.4±1.5

This number of cycles for 50% accurate ensemble performance is referred to as N50.

The shape of the psychometric function relating ensemble accuracy, P_d, to the number of

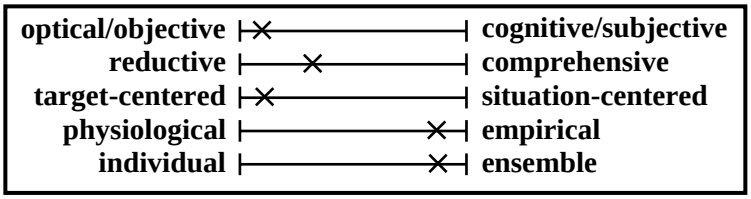
resolvable cycles on target, N, is known as the target transform probability function (TTPF), and has been empirically determined to be:

$$P_d = \frac{(N/N50)^E}{1+(N/N50)^E}$$

in which

$$E = 2.7 + 0.7(N/N50)$$

NVESD models: ACQUIRE (Tomkinson, 1990) and FLIR92 (Scott & D'Angostino, 1992)



FLIR92 is based on the Bailey (1970) framework of separable detection and discrimination stages in search. ACQUIRE represents the non-time-dependent discrimination stage of FLIR92 and does not incorporate search. Both models use the Johnson (1958) bar pattern equivalence metric for target discriminability using electro-optical devices. ACQUIRE in particular is designed to predict the range at which a known target can be acquired by an ensemble of observers.

Pure detection in ACQUIRE proceeds as follows:

1. The area of a rectangle with the same width and height as the target is calculated. Call it A.

2. The mean temperature difference between the target and its immediate background, Δ T, is calculated.

3. The number of resolvable cycles on the target, N, is calculated as

$$N = \frac{\sqrt{A}}{R} \times f_r$$

in which R = the target range,

f_r = the maximum spatial frequency resolvable from the minimum resolvable temperature difference (MRTD) curve defined for the sensor and atmosphere

4. The ensemble probability of acquisition is then calculated as :

$$P_d = \frac{(N/N50)^E}{1+(N/N50)^E}$$

in which $E = 2.7 + 0.7(N/N50)$

N50 = number of cycles resolvable for 50% ensemble acquisition at the desired level of acquisition (i.e., detection, recognition, identification)

NOTES:

- ACQUIRE is sensitive to clutter in that N50 increases as level of clutter increases
- ACQUIRE is not able to handle motion effectively, though attempts to do so are under way (e.g., Mazz, Kistner, & Pibil, 1998)

FLIR92 adds a front end search process before the ACQUIRE discrimination stage. As mentioned before, the model takes advantage of the following assumptions:

- Glimpse duration is constant (at around 0.3 second).
- Glimpse location is random with replacement.
- Each glimpse has an equal probability of locating the target.
- Asymptotic performance (given infinite time) will not be perfect; rather, it will converge on the predictions of ACQUIRE.

The model uses these assumptions to achieve the following performance prediction as a function of time:

$$P_d(t) = P_d^{\infty}(1 - e^{-t/\tau_{FOV}})$$

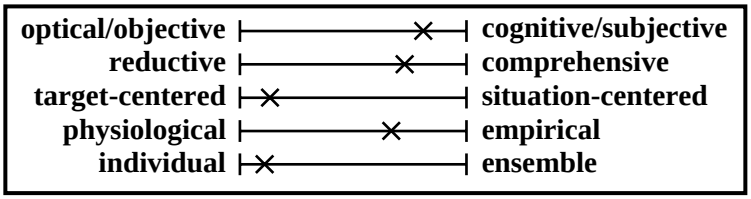
in which $P_d^{\infty} = P_d$, above (asymptotic performance given an infinite search time)

τ_{FOV} = the mean time to detect the target (equals average glimpse time divided by the probability of locating the target in a single glimpse)

…the average target detection rate, $1/\tau_{FOV}$, is related to target detail available and required for acquisition within a field-of-view search:

$$\frac{1}{\tau_{FOV}} = \frac{1}{6.8}\frac{N}{N50}$$

Recognition by Components (RBC) Theory (Biederman, 1987)



The basic idea behind RBC theory is that through the extraction of so-called "non-accidental" properties of a 2-D retinal image, a 3-D representation of the object can be formed. The representation consists of a selection of basic geometric forms called "geons." The representation of the object is then compared to internal representations of known objects. Recognition occurs when the geon representations match. The internal representations, according to the model, are scale and viewpoint invariant in that, so long as an object can be broken into sufficient geons in a well-defined spatial relationship with one another to produce a representation matching an internal representation, the location in space, size, and viewing angle of the object are unimportant.

The primary non-accidental properties of an image are based on regions of deep concavity, which correspond highly with Marr and Hildreth's (1980) concept of zero crossings in a difference-of-Gaussian-processed image. (In a neural network model instantiating RBC theory, Hummel & Biederman, 1992, used a DOG or DOOG (difference of offset Gaussians) operator to extract edges at an early stage of processing.) The edges, however extracted, hint at 3-D surface characteristics by means of Gestalt-like principles such as grouping, symmetry, and similarity, and by the interpretation of T- and L-junctions. These principles and properties are used to generate inferences about the underlying geon structure of the object that precipitated the retinal image. A key feature of the theory is the idea that not all parts of an edge drawing of an object are necessary for the extraction of the object's shape. Rather, it is the "cusps" or junctions that are crucial.

Theoretical problems with the model:

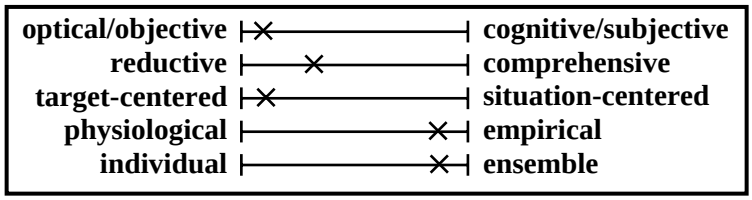
- The notion of invariance has not withstood uniformly empirical examination (e.g., Hayward & Tarr, 1997), indicating that the internal representation of objects may not be as simple as RBC holds.
- Surface characteristics may have an effect on extraction of a geon-based representation of objects (Hayward & Tarr, 1997).
- Tarr and Bulthoff (1995) argue that a geon-based structural description is inadequate for the recognition of category-level objects.
- (However, good agreement with field testing of sub-category object recognition and the errors made in such recognition seems to lend support to the generalizability of some aspects of RBC theory [O'Kane, Biederman, Cooper, & Nystrom, 1997].)
- As the aspect ratios of geons is not hypothesized in RBC theory to be included in the internal representation of objects, some discriminations cannot be performed. For example, the

only distinction between a Boeing 747-400 and a 747-ST is that the latter is shorter. Both objects are composed of the same components, however, so that RBC-based recognition cannot distinguish between them.

Modeling target acquisition with RBC theory:

- Since RBC relies on the extraction of object primitives that are based on a line drawing of the object in 2-D space, models that use a DOG or DOOG operation to extract edges from an image may be particularly suitable. Such an edge extraction technique would have to have some way of eliminating the edge artifacts of surrounding clutter and shadows.
- The fact that RBC ignores surface characteristics such as texture and color indicates that during certain circumstances, it may be inapplicable.
- A key difficulty for RBC theory as a general purpose object recognition explanation is the fact that it can distinguish only between basic categories of objects, such as tanks and jeeps. Because geons do not have extent, internal object representations may not be able to distinguish between members of the category, that is, identification discrimination.

Rand/Bailey's (1970) classical model of search



Primary contributions of Bailey models:

- Target acquisition is considered to consist of three distinct steps: time-dependent search, time-independent detection, and time-independent discrimination. Each step is considered independent of the others, although all depend on the same information in the scene (obviously), and each subsequent step presupposes that the previous step has occurred. The information is treated separately, though.
- The independence assumption allows the probability of discrimination to be the product of the probabilities of each stage succeeding:

$$P = P_1 \times P_2 \times P_3$$

in which P_1 = probability of locating target in a single glimpse,
P_2 = probability of detecting a located target,
P_3 = probability of discriminating a detected target

- P_1 is a hard-shell search with a fixed glimpse aperture A_g.
- P_2 is contrast-based, assumes SNR >> 1, based on observed target size and contrast, though contrast for targets specifically modeled (ground targets as seen from the air) are rarely of absolute contrast >1.
- P_3 is based loosely on the Johnson criteria in that it is based on the number of resolvable "cells" across the smallest target dimension. The model attempts to fit the asymptotic probability of discrimination to the number of cycles (i.e., the TTPF) with an inverse exponential cut-off at 0 probability of discrimination when cycles < 2.

Particulars

- Validated against Blackwell data.
- Not a near-threshold model. Targets were detected, based on contrast rather than SNR.
- Location stage is a non-guided, deliberate search.
- Detection stage is dependent on unconscious visual detection of contrast.
- Discrimination stage is conscious, effortful process.
- Glimpse rate and duration are constant (0.3 second).
- Eye movements not selected at random or completely systematically:
  - Distance of search saccade should be affected by A_g, the effective glimpse aperture over which foveal search can occur. A_g is influenced by size of known target with respect to FOV size.
  - Bailey models probability of glimpse landing on target as function of glimpses (or time) as 1 minus an inverse exponential, that is, as the distribution of first arrival times of a Poisson process.

- Given sufficient time, the search portion of the model will eventually fall on target.

How clutter is handled:

- Clutter is modeled as a scene congestion parameter, G, which varies from 1 to 10 and indicates the density of target-like scene elements. G's primary influence in P1 is to reduce the size of the glimpse aperture. That is, more clutter causes smaller glimpses and shorter saccades (which is, in fact, the case).

$$P(t)_1 = 1 - \frac{1}{e^{-[\frac{700}{G}\frac{a_T}{A_s}]t}}$$

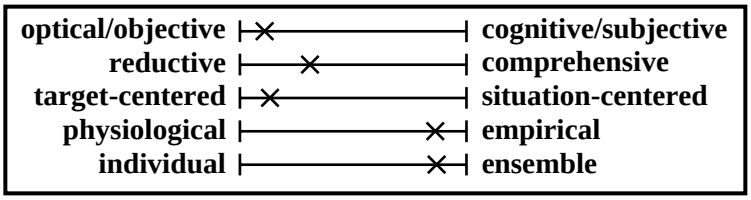
in which a_T = target size,
A_s = search area,
G = scene congestion {1..10}, and
t = search time

- Clutter only plays a role in location, not detection or discrimination.

CRITIQUE:

- No guidance of search process aside from knowledge of target size that drives saccade size (A_g).
- Search process cannot terminate on "target not found" decision.
- Contrast modeling for required detection in P2 may be too specific for general use.

VIDEM (Akerman & Kinzly, 1979)



Particulars:

- validated by Blackwell data (in terms of contrast threshold)
- Search type: soft shell
- Background: cluttered
- Targets
    - stationary, single targets, non-chromatic, equivalence to circles of a certain diameter
- Search location selection: random
- Bailey (1970) search framework

Detection stage:

- target contrast is modeled to be that of a disk of diameter equivalent to the target's critical dimension
- driven by contrast threshold
    - contrast threshold is a function of target size (equivalent disk diameter) and retinal eccentricity:

$$C_T = 0.0352\theta^{0.24} + 0.584\theta^{1.6}/\alpha^2$$

Discrimination stage:

- driven primarily by clutter (see below)

Clutter inclusion:

- clutter increases glimpse duration, decreases distance of search saccades, increases eye response time, and increases contrast threshold
- clutter is assumed to be a GLOBAL metric
    - Mean scene clutter, M-bar, is calculated by Waldman, et al.'s (1988) gray-level co-occurrency metric, which bases clutter on similarity of background and target structure.
    - Target must be known in order to calculate M-bar
- instantiated in similar manner to Greening's (1976) MARSAM model:

$$P_3 = [1 + M/29t_g^{0.93}]^{-1.29}$$

in which M = number of confusable objects (from Waldman, et al., 1988),
t_g = average glimpse time

- effect on t_g:

$$t_g = (0.5782 + M)\theta_s^{-0.2132}$$

in which θ_s is circular search field size (equivalent to saccade distance)

- effect on saccade distance:

$$\theta_s = 0.152 t_g^{-3.127}$$

- effect on contrast threshold (instantiated as a multiplier to contrast threshold):

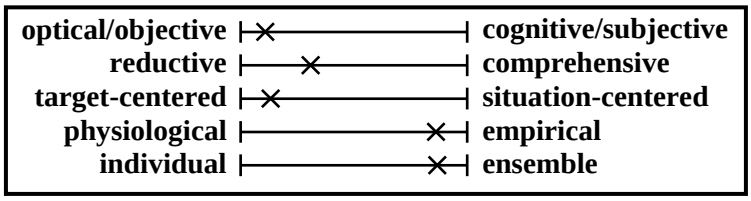
$$F_H = \exp(-5.46 M^{2.37})$$

CRITIQUE:
VIDEM does a good job of representing the effects that clutter is known to have on search. However, the numerous effects of clutter (and the number of parameters that must be fit to a validation data set) yield a model where it may be difficult to weigh the effects on a single process. Also, treating targets as equivalent disks will be problematic for targets that are known to have a high degree of anisotropy, a length-to-width ratio vastly different from 1:1. VOM attempts to address these two shortcomings.

Also:

- Random saccade locations are unrealistic, given that Waldman's clutter metric has been used by itself to predict fixations in a cluttered scene. That is to say, the co-occurrence metric used to calculate a local clutter metric yields regions of the scene that are highly similar to a target. As such, attentional guidance to regions of similarity to the known target (recall that the target must be known in detail to calculate M) will cause saccades to known target locations.
- The fact that VIDEM uses only a global clutter metric is the basis for their assumption of random saccade location selection.

Visual Observer Model (VOM) – Akerman (1992, 1993b)



Particulars:

- An extension of the VIDEM model (Akerman & Kinzly, 1979)
- Excludes some effects of clutter
- Targets may be represented differently: by a function of their "useful Area, A_u" (rather than an equivalent disk area) equal to a portion of the area projected inward from the perimeter of the target

Detection stage:

- Contrast threshold can now be calculated by the VOM criteria (excluding the clutter multiplier, F_H) or by Nachman (1953) criteria:

$$C_T = K_1 p k_2 / A_u$$

in which K_1 and K_2 are constants, empirically derived, based on the adaptation luminance,
p = target perimeter, and
A_u = useful area measured inwards from perimeter (in angular distance).

Differences from VIDEM:

- Clutter is not assumed to have an effect on contrast threshold in search stage.
- The notion of the eye's response time (an additive factor to glimpse duration that depends on clutter) is eliminated.

CRITIQUE:

The same criticisms based on saccade location selection still hold for VOM as they did for VIDEM.

The modification of target "size" to include useful area is potentially quite useful. The useful area notion introduces more observer-based knowledge to the target acquisition situation since the area of a target that is deemed "useful" will depend on its structure, which the observer is also presumably looking for. Given that the gray-level co-occurrency matrix upon which the clutter metric is based concerns target and background structure, it may be argued that a detection stage that keys onto useful area is also incorporating some knowledge of structure and thus may give better agreement with the clutter metric. Eye movements may therefore be better accounted for, albeit not directly.

Appendix B. Proposed Metrics for Motion, Clutter, Conspicuity, and Distinctness

This appendix contains details of the various metrics discussed throughout the review.

Number of confusing forms clutter metric (M) – Ryll (1962)

$$P_3 = \frac{1}{1+\left(\frac{M}{0.29t^{0.93}}\right)^{1.29}}$$

in which M = number of confusable forms visible within a glimpse and
t = glimpse duration.

NOTES:

- Global metric.
- Only affects recognition.
- Used in VIDEM and VOM models with M calculated by means of Waldman's co-occurrence clutter metric, C_N.

Scene congestion (G) metric (Bailey, 1970)

$$P(t)_1 = 1 - \frac{1}{e^{-[\frac{700\,a_T}{G\ A_s}]t}}$$

in which a_T = target size,
A_s = search area,
G = scene congestion factor {1..10}, and
t = search time.

NOTES:

- Clutter is modeled as a scene congestion parameter, G, which varies from 1 to 10 and indicates the density of target-like scene elements. G's primary influence in P1 is to reduce the size of the glimpse aperture. That is, more clutter causes smaller glimpses and shorter saccades (which is actually the case [e.g., Akerman, 1992]).
- Clutter only plays a role in location, not detection or discrimination.
- Global metric.

Target conspicuity (Kp) (Williams, 1966)

$$Pd = 1 - e^{-K_p t / A_d}$$

in which Kp = target conspicuity,
t = search time, and
Ad = display area.

NOTES:

- First mention of how conspicuity can be instantiated into a model.
- Author realized that of the many possible factors incorporated in conspicuity, only luminance contrast was well defined (at the time).
- Clutter affects detection probability (P_1) only.
- K_p empirically determined.

Simple First Order scene metrics (Pratt, 1991, for overview)

- Absolute average intensity difference: $|\mu_T - \mu_B|$
- RMS intensity and target variance difference: $\sqrt{(\mu_T - \mu_B)^2 + \sigma_T^2}$
- Adjusted RMS intensity and target variance difference: $\sqrt{(\mu_T - \mu_B)^2 + 4\sigma_T^2}$
- Absolute mean intensity difference plus absolute standard deviation difference: $|\mu_T - \mu_B| + |\sigma_T - \sigma_B|$
- Absolute mean intensity difference plus target standard deviation: $|\mu_T - \mu_B| + \sigma_T$
- The Doyle metric (Copeland, Trivedi, & McNamey, 1996): $\sqrt{(\mu_T - \mu_B)^2 + (\sigma_T - \sigma_B)^2}$
- The $\text{Doyle}_{\text{mod}}$ metric (Copeland, et al., 1996): $\sqrt{(\mu_T - \mu_B)^2 + k(\sigma_T - \sigma_B)^2}$
- The *nrms* metric (Kosnik, 1995): $\frac{\sigma_{T+B}}{\mu_{T+B}}$

NOTE: μ_T = mean of gray-level distribution over the target area
μ_B = mean of gray-level distribution over background support (typically area immediately around target area)

σ_T = standard deviation of gray-level distribution over target area

σ_B = standard deviation of gray-level distribution over background support

k = modulation factor for variance difference

NOTES:

- First order metrics or any combination of them lack structural information about the target or background support and thus cannot be used for feature extraction.
- A more complex class of first order metrics is based on normalized histograms, described next.

The gray-level co-occurrency matrix

This matrix represents, within an area of a pixilated image, the frequency of one gray level occurring in a specified linear spatial relationship with another gray level. The co-occurrency matrix, $P\Delta(i,j)$, is a GxG dimension matrix in which G is the number of gray scale levels in the image. It is defined by

$$P_{\Delta}(i,j) = \frac{1}{N}\sum_{k=1}^{N} f(x_k = i, x_{k+\Delta})$$

in which $(x_k, x_{k+\Delta})$ = a pair of pixels with gray levels *i* and *j*,
i and j = gray-level values from 0 to a maximum, G, separated by
Δ = a displacement vector which is a function of the distance, *s*, between the pixels and the angle θ between them.
f = {1 if x_k=i and $x_{k+\Delta}$=j, or 0 otherwise}
N = number of pixels in the area of the image.

Normalized Clutter metric (C_N) (Waldman, Wooton, Hobson, & Luetkemeyer, 1988)

(Note: the gray-level co-occurrency matrix $P_{\Delta}(i,j)$ is detailed in the text.)

To Calculate:

- The amount of clutter C is calculated as the mean of the product of the relative texture size and the distance-weighted transition probability (i.e., the probability of transitions between gray levels in the co-occurrence matrix):

$$C = \frac{s}{T}B(\Delta)$$

in which s = average texture element size,
T = average target size,
Δ = polar displacement (see text), and

$$B(\Delta) = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} |i-j| P_{\Delta}(i,j)$$

- The normalized clutter is defined as either C/B_E or 1, whichever is smaller.
  - B_E is the expected value of B.

NOTES:

- works for uniform textures only
- Has overly-simplistic mathematical properties:
  - It is symmetric with respect to target size and background texture size; ignores search asymmetry literature.

Texture-based clutter (TIC) metric (Shirvaikar & Trivedi, 1992)

To Calculate:

- This metric is also based on the gray-level co-occurrence matrix. It is similar to normalized clutter, except that it puts quadratic instead of linear weight on differences in gray level.
- First, calculate the "inertia" of the co-occurrence matrix, ΔI:

$$I(\Delta) = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} (i-j)^2 P_{\Delta}(i,j)$$

- We calculate the global TIC by dividing the inertia by the target size, Δ:

$$TIC = \frac{I(\Delta)}{\Delta}$$

NOTES:

- global metric
- depends on target size
- Performs marginally better than SV.
- The authors recognize that the metric alone, because it fails to capture internal target structure, may not capture perceptually meaningful information and should be used in addition to such measures (Shirvaikar & Trivedi, 1992).

Average Co-occurrence Error (ACE) metric (Copeland & Trivedi, 1996, 1998)

To Calculate:

- Define a target and background region
- Define the "texture model" as the number of pixels away from each other; two pixels within each region are then compared (typically eight pixels are considered)
- ACE is the absolute difference between corresponding elements of target and background co-occurrence matrices, summed over all possible displacement vectors of the length specified within the texture model (see Copeland & Trivedi, 1996, for more details):

$$ACE = \frac{1}{\Theta_{NGLC}} \sum_{\Delta \in \Phi} \sum_{i=0}^{G} \sum_{j=0}^{G} \left| P_T(i,j \mid \Delta) - P_B(i,j \mid \Delta) \right|$$

in which Θ_{NGLC} = total number of displacement vectors in the set Φ of vectors in the texture model

G = number of gray scale levels

$P_T(i,j \mid \Delta)$ = joint probability of a pixel of gray level i and gray level j given the displacement vector Δ for the target pattern

$P_B(i,j \mid \Delta)$ = corresponding joint probability for the background pattern

- The total of displacement vectors of separation eight pixels is 144 displacements.

- To simplify calculation, the number of gray levels is typically reduced to eight, since computation becomes very laborious with 144 256x246 matrix operations to calculate the ACE.

NOTES:

- Authors used this metric to predict human judgments of texture differences.
- Metric outperformed both Doyle metric and a model based on boundary strength (Muller, 1986).
- Local clutter metric.

Circular Symmetry (CS_8) clutter metric (Reisfeld, Wolfson, & Yeshurun, 1995)

To calculate:

- Take each pixel P and calculate a set of values based on the local gradient in the area and the symmetry of the point in eight radial directions about points in the area: $S_8(i, P)$, where i is the direction of symmetry.
- The symmetry, $CS_8(P)$, for each point is the product of S8's for all eight directions:

$$CS_8(P) = \prod_{i=1}^{8} [1 + S_8(i, P)]$$

- A *global* symmetry metric is calculated thus:
  - Divide the scene into k rectangular blocks.
  - Calculate the sum of symmetry values within each block:

$$CS_{8,k} = \sum_{P \in k} CS_8(P)$$

- The global metric is the root mean square of the block-wise symmetries:

$$CS_8 = \left(\frac{1}{N} \sum_{j=1}^{N} CS_{8,j}^2 \right)^{1/2}$$

NOTES:

- Assumptions behind metric are that (1) man-made objects are more likely than natural scene elements to have a high degree of symmetry, and (2) visual system is able to readily locate regions of high local symmetry in a scene.

Statistical Variance (SV) clutter metric and the SCR (Schmieder & Weathersby, 1983)

To calculate:

- Divide scene into N blocks, each twice the height and width of a known target.
- Calculate gray-level variance of pixels within each block i.
- SV is the root mean square of the block variance:

$$SV = \left(\frac{1}{N} \sum_{i=1}^{N} \sigma_i^2 \right)^{1/2}$$

We calculate SCR by dividing the target contrast with its immediate background by SV:

$$SCR = \frac{\left| maximum\ target\ value - maximum\ background\ value \right|}{SV}$$

NOTES:

- Based on idea that the visual system is interested in areas of the scene with high gray-level variability. Consistent with the notion of "confusing forms" in that targets are presumed to have high gray-level variability, although SV does not take into account actual target structure. (Instead, it uses the variance of targets as a generalization of target-like structure.)
- SV is a global measure, though σ_i^2 represents a local metric for clutter.
- This metric *underestimated* performance for urban clutter, indicating that variance alone does not completely instantiate clutter (Cathcart, Doll, & Schmieder, 1989).

Probability-of-Edge (POE) clutter metric (Tidhar et al., 1994)

To calculate:

- Convert gray-scale image into edge map.
- Image is divided into regions. Regions are assigned a value, depending on the fraction of pixels within it that are edges, POE_i
- Overall probability of edge for an image is the rms of local POE_i's:

$$POE = \left(\frac{1}{N} \sum_{i=1}^{N} POE_{i,T}^2 \right)^{1/2}$$

in which $POE_{i,T}$ = probability of edge in region i with DOOG filter threshold T

NOTES:

- Based on idea that early visual processing is involved in edge detection and extraction (e.g., Marr & Hilldreth, 1980).
- Edge detection performed with a DOOG whose output over the scene is thresheld to a level T to yield a yes/no pixel-by-pixel edge map of the scene.
- Local or global metric of clutter, depending on whether POE_i or POE is of interest.

Peak-Signal (ΔT_{PS}) clutter metric (Rotman, Kowalczyk, & George, 1994)

To calculate:

- Set a tolerance ΔT and a minimum cluster size.
- Start with a pixel at a corner and compare it to its neighbor. If the intensity difference is within ΔT, average the two intensities and join them into a cluster.
- If the difference is greater than ΔT, then the new pixel is assigned to a new cluster.
- Repeat this for all pixels, then for all existing clusters until clusters are at least as large as the minimum cluster size.
- The peak-signal ΔT_{PS} is calculated as:

$$\Delta T_{PS} = \frac{T_{\max} - T_{\min}}{1 + \frac{|A(T_{\max}) - A(T_{\min})|}{A(T_{\max}) + A(T_{\min})}}$$

in which T_{max} and T_{min} = intensities of the most and least intense clusters and $A(T_{max})$ and $A(T_{min})$ = areas of the most and least intense clusters.

- We may calculate block-wise $\Delta T_{PS,i}$ by computing ΔT_{PS} for arbitrary blocks of the scene
- This step may be useful for eye movement validation of the metric, but otherwise, it runs the risk of cutting clusters down the middle.

NOTES:

- Based on the contrast between local extrema and their background.
- Divides scene into clusters by grouping pixels of the image together into regions of high and low intensity (the T in the metric is short for temperature) based on the contrast of the pixel with its neighbor. Groups of pixels are likewise grouped together with their neighbors until clusters of the minimum size defined by the user are achieved.
- A global metric for clutter.

Target Complexity (TC) metric (Tidhar et al., 1994)

To Calculate:

- Calculate a histogram of edge intensities by means of a DOOG filter over the target area and its immediate surround. Let the histogram be:

$$\{N_i\}_{i=0}^{G-1}$$

in which 0 and G-1 are the minimum and maximum histogram values
N_i = the number of pixels in the bin at level i

- The corresponding cumulative distribution of edge intensity levels is:

$$S_N(i) = \frac{1}{M^2} \sum_{j=0}^{i} iN_j$$

in which M = the number of points in the histogram.

- SN has the following properties:
  $S_N(i) = 0$ when $i<0$
  $S_N(i) = 1$ when $i>G-1$
  $S_N(i) < S_N(i+1)$

- Target detectability is proportional to the absolute mean distance between cumulative edge histograms of the observed target section (S_N) and the situation when all pixels have the same value (P(i)):

$$TC = \frac{1}{G}\sum_{i=0}^{G} | S_N(i) - P(i) |$$

in which $P(i) = \frac{i}{G}$ (uniform distribution)

NOTES:

- local metric
- Reasonable correlation to overall search RT (authors).
- The size of the surroundings taken with the target seems to be crucial, as a uniform local surround yields a prediction of zero clutter even when the overall scene may be very complex.

Complex Clutter metric (K) (Lillesæter, 1993)

To Calculate:

- An image with a visible target-background border must be selected.
- Let Z be the entire length of the target contour.

$$K = a \mid \overline{\ln G_T} - \overline{\ln G_B} \mid + \frac{b}{Z}\oint_Z \left| \ln\left(\frac{G_T}{G_B}\right) \right| dz$$

in which G_T = pixel gray value distribution of the target area
G_B = pixel gray value distribution of the background support area
a, b = weight factors that sum to unity (usually assumed to be 0.5).

NOTES:

- Incorporates variable target-background contrast around border with a first order metric of contrast.
- The first term is the mean area contrast.
- The second term corresponds to the contrast around the entire target-background boundary.
- The amount of the background to incorporate into the contrast calculation is arbitrary.
- Does not take into account structure of target or length of perimeter (which, in extreme circumstances, may inflate the metric).
- Local metric

Normalized Histogram Intersection and CAMELEON camouflage strength (C) (Hecker, 1992)

Normalized gray-level histogram calculation (for *n*-bit gray level):

- Let h(v) denote the histogram entry for value v, and let the image represent a function with 2^n levels on a rectangular array of width w and height h:

$$\sum_v h(v) = w \times h$$

with $v \in [0, 2^n]$

- If n_a is the number of pixels in the area over which the histogram is computed, then the normalized histogram H(v) is defined as:

$$H(v) = \frac{h(v)}{n_a}$$

(Note: The area over which the histogram is calculated does not need to be rectangular. However, it is assumed to be a rectangular region around the target for the calculation of this and most other metrics used in target acquisition models.)

Histogram Intersection calculation:

- Let H_T be the normalized histogram containing the target and H_B the normalized background histogram.
- The intersection of the matrices is defined as the cumulative sum of the pairwise minimum of corresponding histogram bin heights:

$$H_T \cap H_B \equiv \sum_{v-1}^{n} \min\{H_T(v), H_B(v)\}$$

in which n = number of bins

- The value of the intersection will be between 0 and 1: 0 = no overlap, 1 = complete overlap.

Camaeleon calculation:

- Start with a gray level or color image input image.
- Specify a region in the image as target and a region as background (need not be same size).
- Camaeleon convolves image with set of quadriture band-pass filters to derive pixel-by-pixel representations for the target and background regions:
  - Local energy based on sum over all bands of the energies of individual filters
  - Local spatial frequency based on vector sum of complex frequency averaged over all filter bands
  - Local orientation is computed as vector sum of directions over all filter bands
- Normalized histograms are then calculated for target and background pixels in energy (HE_T and HE_B), frequency (HF_T and HF_B), and orientation (HO_T and HO_B)
- Camouflage strength, C, is calculated as the product of the histogram intersections:

$$C = (HE_T \cap HE_B) \cdot (HO_T \cap HO_B) \cdot (HF_T \cap HF_B)$$

NOTES:

- C is inverse of conspicuity (C may be thought of as measure of local clutter) but is only defined on [0,1]. The rank order of targets with different values of C will reflect the rank order of their clutter.
- Assumption is made that orientation, frequency, and energy are all equally important to estimates of camouflage.
- Boundaries between target and its background are not necessarily taken into account.
- Being based on first order metrics (i.e., histograms) of individual features, the spatial configuration of the features is not specified in the metric.

Cortex Transform-based distinctness (d) (Watson, 1987; Ahumada & Beard, 1996; Rohaly, Ahumada, & Watson, 1997)

To Calculate:

- Take image with target, I_1, and image without target, I_0.
- Convert images to luminance contrast by subtracting and then dividing by the mean background image luminance:

$$I_j \leftarrow (I_j - \bar{I}_0)/\bar{I}_0$$

- We then applied contrast sensitivity filter S to I_1 by multiplying its Fourier components by the magnitude of S's component wavelengths and then recombining the components with the inverse Fourier transform:

$$I_j \leftarrow F^{-1}[SF[I_j]]$$

- Next, the Cortex transform is applied to the image. The cortex transform corresponds to a set of 20 filters: five spatial frequencies with four orientations each, applied to every point (x,y) in the image. The resulting coefficients, corresponding to the signal strength of the channel, for image I_j are $c_{j,k}$, where k ranges over four dimensions: orientation, frequency, x, and y.
- We compute the detectability of each coefficient (d_k) by taking the absolute difference between image and background coefficients:

$$d_k = \left| c_{1,k} - c_{0,k} \right|$$

- We implemented masking for super-threshold channels by decreasing d_k by a factor related to the background channel signal when the background channel exceeds detection threshold:

$$d_k = \frac{\left| c_{1,k} - c_{0,k} \right|}{\max\left(1, |\, c_{o,k}\,|^{0.7}\right)}$$

- The overall distinctness metric, d, is calculated as the Minkowski sum of the individual coefficients:

$$d = \left(\sum_k d_k^{\beta} \right)^{\frac{1}{\beta}}$$

NOTES:

- Based on psychophysics and physiology
  - Contrast sensitivity function and SF decomposition of image are part of calculation.
- Comparison between two images (one with target and one without) to determine detectability of the target.
- Incorporates masking (was determined to over-predict performance without it [Rohaly et al., 1997])
- Only for achromatic images.

NO. OF COPIES	ORGANIZATION
1 (PDF ONLY)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA 8725 JOHN J KINGMAN RD STE 0944 FORT BELVOIR VA 22060-6218
1	US ARMY RSRCH DEV & ENGRG CMD SYSTEMS OF SYSTEMS INTEGRATION AMSRD SS T 6000 6TH ST STE 100 FORT BELVOIR VA 22060-5608
1	INST FOR ADVNCD TCHNLGY THE UNIV OF TEXAS AT AUSTIN 3925 W BRAKER LN STE 400 AUSTIN TX 78759-5316
1	DIRECTOR US ARMY RESEARCH LAB IMNE ALC IMS 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	DIRECTOR US ARMY RESEARCH LAB AMSRD ARL CI OK TL 2800 POWDER MILL RD ADELPHI MD 20783-1197
2	DIRECTOR US ARMY RESEARCH LAB AMSRD ARL CS OK T 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	ARMY RSCH LABORATORY - HRED ATTN AMSRD ARL HR M DR M STRUB 6359 WALKER LANE SUITE 100 ALEXANDRIA VA 22310
1	ARMY RSCH LABORATORY - HRED ATTN AMSRD ARL HR MA J MARTIN MYER CENTER RM 2D311 FT MONMOUTH NJ 07703-5630
1	ARMY RSCH LABORATORY - HRED ATTN AMSRD ARL HR MC A DAVISON 320 MANSCEN LOOP STE 166 FT LEONARD WOOD MO 65473-8929
1	ARMY RSCH LABORATORY - HRED ATTN AMSRD ARL HR MD T COOK BLDG 5400 RM C242 REDSTONE ARSENAL AL 35898-7290
1	COMMANDANT USAADASCH ATTN ATSA CD ATTN AMSRD ARL HR ME MS A MARES 5800 CARTER RD FT BLISS TX 79916-3802
1	ARMY RSCH LABORATORY - HRED ATTN AMSRD ARL HR MO J MINNINGER BLDG 5400 RM C242 REDSTONE ARSENAL AL 35898-7290
1	ARMY RSCH LABORATORY - HRED ATTN AMSRD ARL HR MM DR V RICE BLDG 4011 RM 217 1750 GREELEY RD FT SAM HOUSTON TX 78234-5094
1	ARMY RSCH LABORATORY - HRED ATTN AMSRD ARL HR MG R SPINE BUILDING 333 PICATINNY ARSENAL NJ 07806-5000
1	ARL HRED ARMC FLD ELMT ATTN AMSRD ARL HR MH C BURNS BLDG 1467B ROOM 336 THIRD AVENUE FT KNOX KY 40121
1	ARMY RSCH LABORATORY - HRED AVNC FIELD ELEMENT ATTN AMSRD ARL HR MJ D DURBIN BLDG 4506 (DCD) RM 107 FT RUCKER AL 36362-5000
1	ARMY RSCH LABORATORY - HRED ATTN AMSRD ARL HR MK MR J REINHART 10125 KINGMAN RD FT BELVOIR VA 22060-5828
1	ARMY RSCH LABORATORY - HRED ATTN AMSRD ARL HR MV HQ USAOTC S MIDDLEBROOKS 91012 STATION AVE ROOM 111 FT HOOD TX 76544-5073

NO. OF COPIES ORGANIZATION

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MY M BARNES
2520 HEALY AVE STE 1172 BLDG 51005
FT HUACHUCA AZ 85613-7069

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MP D UNGVARSKY
BATTLE CMD BATTLE LAB
415 SHERMAN AVE UNIT 3
FT LEAVENWORTH KS 66027-2326

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MJK J HANSBERGER
JFCOM JOINT EXPERIMENTATION J9
JOINT FUTURES LAB
115 LAKEVIEW PARKWAY SUITE B
SUFFOLK VA 23435

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MQ M R FLETCHER
US ARMY SBCCOM NATICK SOLDIER CTR
AMSRD NSC SS E BLDG 3 RM 341
NATICK MA 01760-5020

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MY DR J CHEN
12423 RESEARCH PARKWAY
ORLANDO FL 32826

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MS MR C MANASCO
SIGNAL TOWERS RM 303A
FORT GORDON GA 30905-5233

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MU M SINGAPORE
6501 E 11 MILE RD MAIL STOP 284
BLDG 200A 2ND FL RM 2104
WARREN MI 48397-5000

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MF MR C HERNANDEZ
BLDG 3040 RM 220
FORT SILL OK 73503-5600

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MW E REDDEN
BLDG 4 ROOM 332
FT BENNING GA 31905-5400

NO. OF COPIES ORGANIZATION

1 ARMY RSCH LABORATORY - HRED
ATTN AMSRD ARL HR MN R SPENCER
DCSFDI HF
HQ USASOC BLDG E2929
FORT BRAGG NC 28310-5000

1 ARMY G1
ATTN DAPE MR B KNAPP
300 ARMY PENTAGON ROOM 2C489
WASHINGTON DC 20310-0300

1 US ARMY NATICK SOLDIER CTR
FUTURE FORCE WARRIOR
ATTN AMSRB NSC W C BLACKWELL
NATICK MA 01760-5020

1 UNIV OF CENTRAL FLORIDA
DEPT OF PSYCHOLOGY R GILSON
4000 CENTRAL FLORIDA BLVD
ORLANDO FL 32816-1390

ABERDEEN PROVING GROUND

1 DIRECTOR
US ARMY RSCH LABORATORY
ATTN AMSRD ARL CI OK (TECH LIB)
BLDG 4600

1 DIRECTOR
US ARMY RSCH LABORATORY
ATTN AMSRD ARL CI OK TP S FOPPIANO
BLDG 459

1 DIRECTOR
US ARMY RSCH LABORATORY
ATTN AMSRD ARL HR MR F PARAGALLO
BLDG 459

1 DIRECTOR
US ARMY RSCH LABORATORY
ATTN AMSRD ARL HR S L PIERCE
BLDG 459

3 DIRECTOR
US ARMY RSCH LABORATORY
ATTN AMSRD ARL HR SD B VAUGHAN
BLDG 459