(2)
B.S.

(6)

TARGET ACQUISITION THROUGH VISUAL RECOGNITION: AN EARLY MODEL

(10) H. H. Bailey

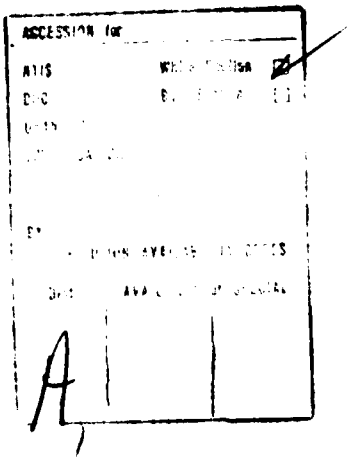
(11) Sep[illegible] 1972

(12) 18 p.

(14) P-4918

This Paper is to be presented as an invited paper at the Target Acquisition Symposium, Orlando, Florida, 14-16 November 1972.

CONTENTS



TARGET ACQUISITION THROUGH VISUAL RECOGNITION: AN EARLY MODEL

H. H. Bailey
The Rand Corporation, Santa Monica, California

I. INTRODUCTION

Human vision is achieved by a remarkable electro-optical system that has evolved naturally to a form that is far more versatile than anything we have been able to build ourselves. Everyone is aware of the two classes of primary receptors, the efficient distribution of the high-resolution components, the huge range of luminance levels over which the system can function, and so forth.

There is an additional feature of the dynamic range capability that has been brought to light in a little-known fundamental paper by Ory*. Using modern statistical decision theory, he was able to obtain an excellent fit to the 6-sec/8-position Tiffany data by making the following assumptions: primarily, the decision criterion is assumed to be adaptable, or programmed, and to depend on both the luminance (total number of neural excitations per unit time) and its fluctuations (square root of number of excitations per unit time); in addition, both the background and the target luminance and fluctuations are to be included, and the image extension due to both optical aberrations and the retinal neural organization must be accounted for. With these assumptions, all of the experimental data on the threshold luminance ($\Delta L\theta^2\tau$) can be plotted on two curves, one for foveal-photopic vision and one for extrafoveal-scotopic vision. Each curve is a simple quadratic in the total fluctuation rate, and one of them has a small constant term.

The interesting point here is that these curves define two regions, depending on whether the fluctuation or the luminance term is dominant; and these correspond to the DeVries-Rose and the Weber-Fechner

*Ory, H.A., *Statistical Detection Theory of Threshold Visual Performance*, RM-5992-PR, The Rand Corporation, September 1969.

regimes of the literature. These regimes are interpreted as follows. At low light levels, discriminable increments are limited by the noise level, $\Delta L \sim \sqrt{L}$, and system performance is determined by the signal-to-noise ratio, $\Delta L/\sqrt{L}$. As the luminance is increased, the number of discriminable levels rises as the square root, until another limit sets in -- the number, n, of levels that our visual equipment can process simultaneously. From this point on, the discriminable increments are just the fraction 1/n of the maximum luminance, $\Delta L \sim L$, and system performance is limited by the contrast, $\Delta L/L$.

Now, with modern man-made electro-optical systems, despite the fact that the display is bright enough for the eye to operate photopically, there is so much noise in any practical system that the overall performance is again noise-limited. The number of distinguishable grey levels on some of the old radar scopes was 3 or 4 at most, and now we struggle to get 7 or 8; but these are still well below the number (probably around 20) where the human processing system imposes its contrast limitation.

The conclusion is that signal-to-noise is the appropriate parameter for analyzing the capabilities of an observer viewing an electro-optical display; but it is still appropriate to use contrast when you look out the window.

Now I will go back and describe my old observer model. This was primarily intended for, and is best applied to, the contrast regimes. A description of this model was first published in 1970*; however, it was developed during the mid-sixties and was used internally at Rand prior to its publication. This paper gives an abridged description of the model, emphasizing the concepts involved, together with two extensions which have been added since that time.

*Bailey, H. H., *Target Detection Through Visual Recognition - A Quantitative Model*, RM-6158-PR, The Rand Corporation, February 1970; also appeared with classified applications in Journal of Defense Research, Vol. 3B, p. 54, 1971.

II. GENERAL FORM OF THE MODEL

The performance of a human observer is often a very complicated function of many interacting variables. In order to simplify this difficult situation and yet stay reasonably close to reality, we consider explicitly the task of finding known and fixed objects in a complex field in a short time. This process, even when so restricted, is still complex, but it can be considered to consist of the following three distinct steps: deliberate search over a fairly well-defined area; detection of contrasts (a subconscious retino-neural process); and recognition of shapes outlined by the contrast contours (a conscious decision based on comparison with memory).

On the basis of assorted experimental data, three formulas can be devised for the probabilities of completing each of the three steps separately. It is postulated that the overall target recognition probability can be expressed by the product of these three terms. Accordingly, we establish the following definitions:

- P_1 is the probability that an observer, searching an area that is known to contain a target, looks for a specified glimpse time (viz., 1/3 sec) in the direction of the target with his foveal vision. P_1 is a function of the ratio of an acceptable search rate to that demanded in a given situation; the loosely defined concept of foveal vision is replaced by that of an effective glimpse aperture.

- P_2 is the probability that if a target is viewed foveally for one glimpse period it will, in the absence of noise (i.e., $S/N \gg 1$), be detected. P_2 is determined by psychophysical limits operating on the observed target size and contrast.

- P_3 is the probability that if a target is detected it will be recognized (again during a single glimpse and in the absence of noise). Recognition is usually (but not necessarily) accomplished on the basis of intrinsic shape without reliance on context.

We then write for P_R, the probability of target recognition,

$$P_R = P_1 \times P_2 \times P_3 . \quad (1)$$

Inasmuch as the three steps described above are independent events, P_2 and P_3 (as defined) each representing a conditional probability under the one preceding it, the product formulation of Eq. (1) is obvious and rigorously correct. This is so despite the fact that the individual terms are not completely independent in the sense that they may be functions of some of the same variables (contrast, for example). This and certain other subtle interactions are discussed briefly in a later section. In the following paragraphs the nature of each of the three terms is examined in some detail, and a specific analytical expression is developed for each one.

The Search Term

The first term, P_1, describes the search limitations. When searching from the air for a terrestrial object whose location is known only approximately, it becomes both possible and necessary to utilize foveal vision and to search fairly systematically. The maximum acuity of foveal vision is a necessity, since there is always a need to find targets at the earliest possible moment during approach, and at long range either the apparent size or the available contrast or both may be marginal; few military targets really stand out. Foveal vision is also usually feasible, since only a limited area needs to be covered. The required area may be as much as the whole of an electro-optical display, but more commonly it is an area set by navigation errors and target location uncertainties, centered on a predicted or expected target location. Even under these conditions, however, search rates are extremely variable and almost intractable for the fundamental reason that pieces of terrain (not to mention possible targets) differ widely and almost defy quantification. Nevertheless, some bounds can be set.

It is well known that the eye moves in discrete steps, ordinarily with about three stops, called fixations, per second. (Actually an

observer occasionally takes longer to examine certain points, but this does not affect very much the average search rates described below.) Our approach, therefore, is to postulate that an experienced observer searches by moving an apparent aperture (essentially his foveal vision) in some fairly regular pattern over the area of interest, and furthermore that he adjusts his average interfixation distance, and hence the effective size of his scanning aperture and his overall search rate, in accordance with his *a priori* information on the size and contrast of the target or its image. Intuitively, one recognizes that an observer will scan the floor around him differently if he is looking for a pencil or an ant. Stated more formally, the observer estimates how far off his visual axis he will still have an adequate probability of detecting the expected image, and he automatically adjusts his search rate accordingly. A key concept, therefore, is the size of the effective scanning aperture -- here called a glimpse aperture, A_g. This is a quantity that commonly ranges from 10 to 100 times the area of the target, a_T, but can sometimes vary between 1 and 1000 times a_T.

The reason for this huge spread is not just the observer's inability to predict the nature of the image or his own detection probabilities. It lies also in a second important factor -- the structure, complexity, or "congestion" of the surrounding scene. The search for an ant mentioned above will also be quite different depending on whether the floor is covered with a nearly featureless linoleum or a textured and patterned rug. However, this "congestion" cannot be described solely by the two-dimensional spatial-frequency content in a scene. What really matters is the density of contrast points -- the natural fixation centers for the eye -- or other "confusion objects" that are present in the scene. The writer once experienced a striking example of many such false targets (natural decoys, as it were) while flying over the notorious Coso Range in California. This region contains scattered trees and bushes which appear very dark against the background of sandy soil or dried grass, as do the vehicles and "bridges" which were placed in the area as "targets". Almost every tree had to be examined to see whether or

not it had straight sides before the true targets could be found. Indeed, tests there have produced some of the lowest target acquisition probabilities ever measured.

The kind of adaptive search rate described here, in which the observer automatically reacts to both the character of the scene and the (anticipated) nature of the target imbedded in that scene, has been advocated informally by this writer for several years. The only independent reference to such a concept found in the literature is by Williams*. He talks about target "conspicuity," which is measured by the rate at which a particular target can be successfully searched for in a particular field, and he points out that the commonly observed lack of dependence of target acquisition on display scale factor (within limits, and assuming no change in information content on the display) is another manifestation of observer adaptation.

A heuristic derivation of an expression for P_1 follows. If an area A_s is to be searched, the number of glimpses (each of area $A_g = ka_T$) required to cover the area is A_s/A_g. The number of glimpses that are available in t sec, at 1/3 sec per glimpse, is 3t. With perfectly systematic search, the probability of "looking at" the target (i.e., including it within a glimpse aperture) would be just the ratio of the available glimpses to the total number required, or $3t/(A_s/A_g)$. This would give P_1 the form of a linear ramp function with time. Real search is probably something between perfectly systematic and purely random, so that P_1 should have a form that lies between the ramp and an exponential rise. We conservatively adopt the latter and postulate

$$P_1 = 1 - e^{-K \times 3t/(A_s/ka_T)},$$

where K is a constant and k is a parameter related to scene congestion.

---

*Williams, L. G., "Target Conspicuity and Visual Search", Human Factors, Vol. 8, February 1966, p. 80.

The supporting experimental evidence and the evaluation of the constants can be found in the original reference. However, one more idea must be introduced. Since the value of k, the number of target areas in a glimpse aperture, is usually found to lie in the range 10 to 100, we substitute 100/G for k, where G is a measure of the congestion in a scene. Formally, it is the number of fixation centers within the nominal glimpse aperture of 100 a_T; in practice, it is nothing more than an estimate by the observer of the congestion of a particular scene relative to his experience, and he is asked to assign a number which usually falls between 1 and 10. Although this estimate may be quite crude, such as no better than within a factor of two, it is believed to be far better than ignoring the problem altogether. Further studies are needed to refine this troublesome point.

Accordingly, we propose the following expression for P_1:

$$P_1 = 1 - e^{-[(700/G)(a_T/A_s)t]} . \qquad (2)$$

The Contrast Term

The second term, P_2, has to do with the basic process of contrast detection by the human visual system. Blackwell's classical experiments provide the fundamental data here, yielding curves of threshold contrast (50 percent detection probability) versus size of circular discs under various levels of ambient illumination. These are commonly called "demand" contrast functions. However, there is a good deal of evidence that the best (i.e., lowest) threshold values obtained by Blackwell must be adjusted upward substantially for application to the practical situations discussed in this paper. Ignoring some of the details, which can be found in the original reference, it is proposed that the shape of the "demand" curve of threshold contrast C_T versus angular subtense α in minutes of arc (min) be taken from the average of the two best-known sources of 1/3-sec data, and that this curve be adjusted upward by a factor of about 5.5 in contrast -- or that 0.75 be added to log contrast. Since, in the absence of bright lights or specular glint, target

contrasts (using the absolute value definition) greater than unity are rarely observed through the real atmosphere, and even less frequently on military targets, the resulting curve on log-log paper can be approximated by the hyperbola

$$(\log C_T + a)(\log \alpha + 0.5) = 1. \tag{3}$$

This simplification is often convenient and usually adequate, but, whenever contrasts greater than unity are important (for example, on certain electro-optical displays), a more accurate curve with an asymptotic slope of -1/2 should be used.

The probability of detection, P_2, at the threshold contrast is, by definition, 50 percent. The probability of detection for other values of observed contrast, C, has been shown to depend only on the ratio C/C_T and to have the form of the cumulative normal distribution with $P_2 = 0.9$ for $C/C_T = 1.5$. This is equivalent to setting the value of the Gaussian standard deviation equal to 0.39, and it indicates that on the average Blackwell's subjects chose to operate at a false-alarm rate of about 1/200, corresponding to an S/N of roughly 2.6:1. Accordingly, we write

$$P_2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\{[(C/C_T)-1]/0.39\}} e^{-u^2/2}\, du. \tag{4}$$

The Resolution Term

The third term, P_3, has to do with the more subjective act of deciding what particular image forms represent in the real world. But, since we are primarily concerned with shape recognition of known or briefed objects as distinct from the interpretation of unfamiliar imagery, the problem can be reduced to the visibility -- or detectability in the sense of the previous section -- of sufficient geometrical detail for shapes to be compared with memory and thereby recognized. The concept of "sufficient" detail might lead one into the morass of "critical details" -- those unique features that permit various classes of objects to be distinguished from one another.

However, when all portions of an image are equally detectable so that the whole shape is either visible or not, Johnson of NVL has demonstrated the remarkable fact that, for a variety of military objects,* a single parameter -- namely N_r, the number of resolution cells contained in the shortest dimension across a target -- is all that is required to describe what constitutes "sufficient" detail for detection or for recognition. He found values of N_r between 3.3 and 4.8, or 4.0 $\pm$ 20 percent, for high-confidence recognition. Other experimenters have confirmed this simplification and derived values for N_r close to Johnson's or slightly larger. We adopt a conservative value and write

$$\begin{aligned} P_3 &= 1 - e^{-[(N_r/2)-1]^2} && N_r \geq 2 \\ &= 0 && N_r < 2 \end{aligned} \tag{5}$$

which makes $P_3 \approx 0.9$ when $N_r = 5$.

It is important to emphasize the meaning of N_r. As previously defined, it is the number of resolution cells contained in the minimum dimension (e.g., width or height) of the projected image of an object to be recognized. In the present context, "resolution cells" means independently detectable spots -- the subject of the previous discussion. The proper procedure is to first calculate, from Eq. (3), the size of the smallest spot that can be seen -- at the contrast level with which the target is presented to the observer. Then, the number of these detectable spots contained in the shortest dimension of the target image gives the value of N_r.

It was implied above that recognition in unfamiliar situations may be much more complicated, and far more difficult to predict, than the mere detection of shape details. An extreme example might be the classical one of the photo-interpreters searching for completely unknown elements of the Peenemunde launching areas during

---

*One should probably add "in a military context." The amount of detail required to distinguish a truck from an oxcart is far less than that required to discriminate between various truck models; but the simple separation of objects into classes is usually sufficient for designating targets.

World War II. No attempt is made to extend this model to cover such cases. It should also be mentioned, however, that under certain other circumstances recognition may be very much easier than this model would predict. Consider the approach of unauthorized aircraft, or the presence of vehicles along a road in enemy territory. Both are cases in which the mere detection of objects might be sufficient to justify the decision, "There is a target!". These cases can be handled by assigning artificially high values to P_3 (when the prior information so justifies), thus effectively equating detection as given by P_2 to recognition. This point is discussed further below. Our model of P_3 covers the more common intermediate cases in which shape provides the primary criterion for recognition.

Discussion

Equations (2) to (5), combined as indicated by Eq. (1), constitute the proposed model of a human observer. The fundamental concepts and the basic product formulation are explained at the beginning of this paper. The result, after analyzing each of the three terms, is an expression for the probability of recognition as a function of several observable quantities -- the apparent size and contrast of the target, and the required search rate and the false-target density in the scene.

An important and useful property of the model is the separation of variables that has been achieved. Each of the terms is expressed as a function of a rather small number of input parameters, and target size is the only parameter which appears in more than one term. This rather significant simplification arises from a careful consideration of the consequences of the product formulation and a detailed evaluation of each of the terms over only the ranges of the input variables for which that term is controlling or otherwise of interest.

For example, the model is not applicable to a target that is so isolated or whose contrast is so high (relative to the background

clutter) that it can easily be seen with peripheral vision,* since in that case the search rate can be very much faster than postulated in Eq. (2) and P_1 will still be very high. But then P_2 will also be very high (essentially unity), and the problem is almost trivial. The search model assumes only that target contrast is *not* that high, so that fairly systematic and fine-grained search must be carried out. In fact, the actual search rate employed by an observer is determined by some sort of average false-target density over the scene. If the actual contrast of a specific target against its contiguous background turns out to be less than sufficient for recognition to take place during a single properly-directed glimpse, this fact will show up in P_2 and P_3, which will correctly reduce the value of P_R.

The relationship between the conditional probabilities, P_2 and P_3, can be discussed in a similar way. While it is clear that they are intimately related, they are separated, with further separation of variables, for several reasons. Ordinarily, detection not only precedes but also dominates shape recognition. That is, unless P_2 is rather high, there is probably no point in even calculating P_3, since P_R will be too low to justify the sortie. When P_2 is high, then P_3 controls. On the other hand, as has been mentioned, there are cases for which *a priori* or contextual information may suffice to obviate the need for shape recognition per se. In such cases -- boats on a river or trucks on a road, for example -- P_3 can be ignored (i.e., set to unity without regard for Eq. (5)) and P_2 will control. By keeping the two terms separate, model flexibility is preserved. Further arguments for this separation revolve around the role of resolution. First, as a practical matter, most man-made sensor systems are resolution-limited, since resolution always costs something. (This is true at least of systems whose displays are properly designed.) Accordingly, the sometimes-difficult calculation of system MTF need be applied only once (namely, when it is most

*This is essentially what is achieved by multispectral cueing or by MTI radar, as indicated on page 8 of the original reference.

critical) in the shape-recognition term. More importantly, there are many cases with multiscaled or zoom-capable systems in which the combination of *a priori* information and required search area may make a two-step identification of the target desirable. In such cases an initial and tentative detection on a wide field of view is followed and confirmed (or denied) by shape recognition on a magnified image. At the first step P_2 controls, but P_R is incomplete; at the second step P_3 controls.

The product of the three terms provides a viable model for a wide variety of circumstances; it can be used in predicting the capabilities of a broad class of manned systems, since it deals only with the observer and the information presented to him, whether this be directly to his unaided eyes or through optical aids, or with sophisticated artificial sensors provided that the signal-to-noise ratio is high.

III. APPLICATION TO FLIGHT TOWARD A TARGET

The foregoing basic model really applies only to static situations, as when looking from a hovering helicopter. In normal flight approaching a target, the time constraints enter in a somewhat different manner.

Firstly, it is observed that the product P_2P_3 gives a single-glimpse probability of recognition as a function of the distance from the target. The distance, together with the flight path and altitude of course, determine the proper projection, the angular subtense, and the apparent contrast (through the atmosphere) of the target. Secondly, repeated glimpses in the same direction (i.e., fixed on the same contrast point) are not independent in the usual probabilistic sense and do not, per se, increase P_R significantly. What does change with repeated looks in normal flight is that the observer is getting closer to the target. Thus the P_R vs D curve with $P_1 = 1$, which may look like a cumulative probability curve, really is not that; it is simply a plot of the increasing probability of recognition as one approaches a target -- on the assumption that no search at all is required.

The need to search for a target can only delay its acquisition, or, saying it differently, a given probability of recognition will be realized at some shorter range. To determine the amount the original curve must be shifted toward later (i.e., smaller) values of ranges, one must utilize the concept of searching over a specific area, A_s -- such as a SAM-site pattern, or the area within a revetment, or the area enclosed by a reticle that is adjusted to indicate the navigational error, to mention a few examples. If the search over this area is thoroughly systematic, which it might be since the search is completely structured, then the maximum delay, τ, would be the time taken to search the indicated area (given by $A_s/3A_g = A_s 3ka_T = A_s G/300a_T$). On the average the delay would be close to half that quantity, and the corresponding range shift would then be $V\tau/2$, where V is the approach velocity. (When the search area is

not well defined, as in locating the first and grossest checkpoint, an arbitrary expression of the form τ (sec) = 1 + V(kn)/200 has been used.)

If the search were indeed ideally systematic, the time delays would lie between zero and twice the average value; and indeed this kind of spread in field test data must be anticipated. With a search pattern that is less than ideal, the P_R vs D curve will be both displaced and flattened, with the result that the probabilities calculated by the method described will be a little optimistic at the high side (i.e., close to the target).

IV. APPLICATION TO MOVING TARGETS

In 1970, Rand conducted some simple simulation experiments* on the detection of stationary and moving targets. The data consisted of the times taken by several observers to search a TV screen and find a simple, electronically generated, rectangular target against a background that was variously a piece of felt, an artificial grid pattern, or scenes from aerial photography. The exponential time-dependence was confirmed, and the effect of the target motion could be well accounted for by introducing into the exponent of the P_1 term a factor $(1 + 0.45\ V^2)$, where V is the target velocity in degrees/second subtended at the observer's eye. There was some indication that part of the effect in moving objects being "easier" to detect is due to the changing contrast as the object passes over a complex background -- analogous to the detection of flashing lights. However, this effect could not be quantitatively separated out in these experiments.

*Dugas, D. J. and H. E. Petersen, *An Experimental Investigation of the Effect of Target Motion on Visual Detection*, R-614-PR, The Rand Corporation, February 1971; and Peterson, H.E. and D. J. Dugas, *The Relative Importance of Contrast and Motion in Visual Target Detection*, R-688-PR, The Rand Corporation, March 1971.

V. CONCLUDING REMARKS CONCERNING NOISE

The foregoing description has ignored noise altogether. Historically, I used to think that this model would be equally applicable to an observer viewing an E-O display, except that his performance would necessarily be degraded by the insertion of electrical noise. I therefore multiplied the $P_1P_2P_3$ product by a fourth term which was given by

$$\eta = 1 - e^{-[(S/N)-1]} \qquad S/N \geq 1$$

$$= 0 \qquad S/N < 1 .$$

This form has been used at Rand, and by other people too (I'm told), to make performance predictions that seem to be reasonable. It is certainly better than nothing; but it is now clear that the concept of "perceived signal-to-noise" described here this morning, and the procedures that were given for applying it, constitute a much better approach to understanding recognition through noise. The procedures may appear to some to be a little arbitrary or *ad hoc*-ish, but they do work. Perhaps a more fundamental (or at least a more satisfying) theoretical approach can be evolved building on Ory's work, mentioned earlier in this paper. Perhaps someone in this audience will do just that in the very near future.