Sperling, G., Reeves, A., Blaser, E., Lu, Z-L., and Weichselgartner, E.
Two Computational Models of Attention.
In J. Braun, C. Koch and J. L. Davis (Eds.),
Visual attention and cortical circuits.
Cambridge, MA: MIT Press (2001). Pp. 177-214. + four color plates.

10 Two Computational Models of Attention

George Sperling, Adam Reeves, Erik Blaser, Zhong-Lin Lu, and Erich Weichselgartner

10.1 Introduction

10.1.1 Attention Models for the Twenty-first Century

Models of attention phenomena should (1) explain or account for significant phenomena and (2) be physiologically plausible. There are other good properties models might have, such as simplicity, parsimony, efficiency, application to naturalistic situations, and so on. Here, we concentrate on two models of visual attention that account for the overall behavior of an observer in psychophysical tasks. Among the many models that have been proposed for attention phenomena, we offer a particular reason for giving especially serious consideration to the ones proposed here. They do not merely predict that an observer should do better in one situation than another, or that an interaction between two variables should be observed under certain circumstances. They account for relatively large amounts of data quite efficiently. "Large amounts" of data means (in 2000) minimally dozens, preferably hundreds, and sometimes more than a thousand data points obtained from each observer in the experiments with, preferably, an average of a hundred or so observations for each of the hundreds of data points. As the number of data points becomes large, the data increasingly constrain possible models. By "account for," we mean that a model accounts for more than 80%, and preferably more than 90%, of the variance in the data.

When a model efficiently accounts for a large amount of data, the concepts embodied in the model, such as an attention window or attention-switching time or attentional amplification, achieve face validity, like the concepts of an electron or of electron spin in physicists' models. The large-scale quantification is essential to make the attention processes analogous to twentieth-century physical concepts. Without such quantification, attention theories are underconstrained, and correspond to speculative theories about the nature of matter that characterized earlier stages of physics.

10.1.2 Two Attention Models: Overview

Models will be considered for two phenomena: the time course of attention windows and attention amplification involved in selective attention. Following these two quite well-determined models, a speculative proposal is offered for the overall functional architecture in which these models are embedded.

The first model derives the form of an attention window from psychophysical experiments. Here, we concentrate primarily on the derivation—how a sufficiently detailed data set implies the shape of an attention window and the properties of certain related processes,

such as cue interpretation and the storage of attended items in the visual short-term memory. Elsewhere (Sperling and Weichselgartner, 1995), this model has been applied to make accurate, quantitative predictions of the data from the paradigms that have been most widely used to measure shifts of visual attention. In particular, it makes predictions of the pattern of speeded reaction times in response to valid attentional cues (Posner's cost/benefits paradigm) with traditional go/no-go responses and also with choice reaction-time responses. It predicts the pattern of more accurate responses at locations that have been validly cued, and also has other applications (Sperling and Weichselgartner, 1995). This is the model to consider when there are spatial or temporal attention cues.

The second model describes the processes involved in the attentional amplification of attended features. It shows how the relative importance (salience) of features is determined by bottom-up processes and altered by top-down processes of selective attention, providing a precise description of these processes in terms of the attentional amplification of selected inputs to a salience map. This model provides a general theory, derived from detailed psychophysical data, for how bottom-up and top-down attentional influences combine. It is especially applicable to studies of visual search, in terms of providing a mechanism for so-called guided search. (Note: The ideal attention experiment presents the same stimuli, and records the same responses, in two conditions that differ only in attentional instructions and in the payoff matrices. Typically, search experiments are not formally attention experiments, although they often are considered together with that category. See Sperling and Dosher, 1986 for a detailed review.) The second model also applies to figure–ground segmentation, and to a host of selective-attention paradigms.

The two models use quite different mathematical structures because of the psychophysical phenomena from which they are derived. However, these are merely different aspects of the same underlying control structure, a salience map and associated processes, that is developed in the more speculative overall formulation, and that provides a general framework for combining bottom-up and top-down attentional processes. Together, the two models encompass most of the paradigms that have been used to study attention.

10.2 Determining the Time Course and Structure of Attention Windows

10.2.1 Measuring Attention Reaction Times

Indirect Measures of Motor Reaction Time: The Grabbing Response The procedure for measuring the reaction time of a shift of visual attention can be best understood by an analogy with an unusual way of measuring a motor reaction time. Imagine an observer, as shown in figure 10.1, seated at a conveyer belt on which balls, about the size of billiard balls, pass by. The speed of the belt has been calibrated so that ten balls pass per second.

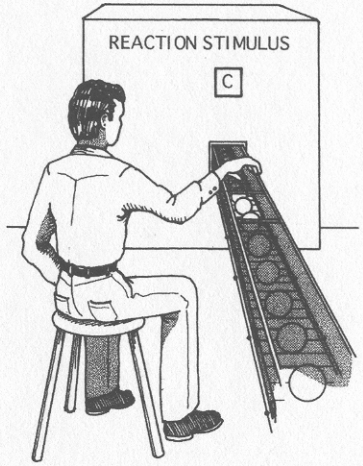


Figure 10.1
The grabbing response: an indirect measurement of individual reaction times. Balls are placed on the moving conveyer belt so that a new ball passes the opening every 0.1 s. When a critical "reaction stimulus" appears above the conveyor belt (say, the letter C), the observer reaches into the opening and grabs the first ball possible. A code number inside the ball indicates its place in the sequence, and hence the moment in time at which it passed the opening. This grabbing response is analogous to the procedure in which items from the "next-to-be-attended" stream are admitted to short-term memory because a cue has triggered a shift of attention to that stream.

There is a small opening through which the observer can reach to grab a ball. His task is to monitor a screen until a critical character (a target) appears. In this example, the target is the letter C. As soon as the observer detects a target, his task is to reach into the opening and grab the first ball that he can. Once he has grabbed a ball, he opens it and reads the number painted on the inside.

Suppose the experimenter has arranged the situation so that the number of the ball that is simultaneous with the target is 0, the ball that passes one tenth of a second later is 1, two tenths of a second later is 2, and so on. From the number that is reported by the observer, the experimenter can infer the reaction time of the observer's "grabbing" response on that trial to an accuracy of 1/10 s. From a long series of trials, the experimenter can observe the entire distribution of reaction times for a particular target in a particular environment of nontargets (distractors).

There is a minor problem with this reaction procedure. Suppose that the observer has consistently been grabbing balls numbered 3 and 4. Then, in one trial in which he was not quite prepared, the observer grabs a ball with the number 9. The observer knows he

was slow, and might improve his response by calling out a lower number than 9, especially if there were a reward for quick reactions. To eliminate the possibility of cheating, the balls are assigned arbitrary numbers. The experimenter knows the number of the ball that passed the opening at each instant, but the observer does not. The identification number of the ball is of interest only insofar as it indicates the instant at which the ball passed the opening.

Grabbing Items for Short-term Memory The procedure for measuring the grabbing response would be an indirect and unnecessarily complicated procedure for measuring a motor reaction time because direct measures of motor reaction times are easily obtainable. However, there is no direct measurement of an attention reaction time. But the indirect grabbing procedure is easily applied to measuring attention reaction times in a paradigm in which the attention response is to grab one item from a rapidly passing stream of items and enter it into short-term memory. The procedure is as follows.

The observer views two adjacent streams of items, a stream of letters on the left and a stream of numerals on the right. (A stream of items is a spatial location where consecutive frames containing new visual items fall one on top of the other.) Initially, the observer's attention is focused on the stream containing the target, the search stream. When the target is detected, the observer's task is to shift attention to the numeral stream (the measurement stream, the next-to-be-attended stream) and to report the earliest possible numeral. To eliminate eye movements, experience has shown that it is best if the observer maintains fixation on the next-to-be-attended stream throughout the trial. Thereby, when the time to shift attention arrives, there is no urge to move the eyes because they already are fixated on their destination. In the early experiments described here, however, the observer maintained fixation between the two streams, and the streams were centered 1.87° apart (figure 10.2).

The (target-containing) search stream consisted of a sequence of thirty randomly chosen letters of the alphabet. The letters B, I, O, Q, S, and Z were omitted because of their similarity to the numbers 8, 1, 0, 5, and 2. A target letter was embedded at a random position in the middle of the stream. In different blocks of trials, the target was either the letter C or the letter W, or simply an outline square with no letter in the middle. The rate of target stream presentation was 4.6 letters per second (218 ms between consecutive onsets). This rate was chosen to make the target detection task sufficiently difficult that observers had to devote all their attention to it. The rate of the next-to-be-attended stream differed between blocks: 4.6, 6.9, 9.2, or 13.4 numerals per second. (For additional details, see Sperling and Reeves, 1980).

The observers' task was to detect the target letter and then to report the first numeral they could from the numeral stream (i.e., to grab the earliest possible numeral). Addition-

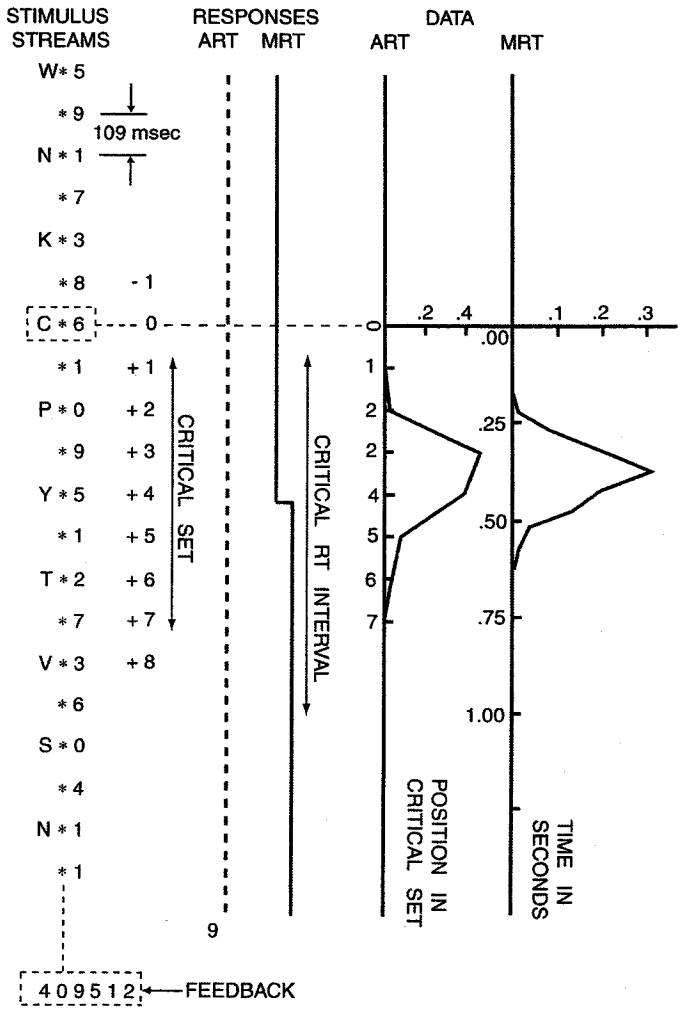


Figure 10.2
Attention and motor reaction times (ARTs and MRTs). The subject fixates the central * and attends the letter-containing stream (left) until a target letter (C) is detected, then shifts attention (but not his eyes) to the next-to-be-attended numeral-containing stream, to "grab" and report the first possible numeral. The critical set is a sequence of all-different numerals in the to-be-attended stream, centered on the time when the response is expected. The ART graph shows the histogram of temporal positions from which numerals were reported (middle). In addition to reporting numerals, the subject also made a rapid finger response upon detecting the target letter. The MRT graph shows the histogram of these motor-reaction times (right). Although the abscissa is the same in both ART and MRT graphs, the units of the MRT graph give the actual time in seconds, whereas those of the ART graph indicate the onset times of critical-set items.

ally, the observers were required to make a motor reaction-time response, lifting a finger from a response key. After considerable practice, performance on both these tasks was almost independent (i.e., differed little from control conditions in which either task was performed alone).

Attention Reaction Times (ARTs) Figure 10.2 shows the data for an observer with the target letter C, and the next-to-be-attended stream rate of 9.6 numerals/s. The data show that the observer nearly always reported the numerals occurring 3 or 4 positions after simultaneity, that is, numerals that occurred 327 or 436 ms after target onset. By analogy to the grabbing response, this is a distribution of attention reaction times (ARTs). For comparison, the histogram of motor reaction times (MRTs) shown in figure 10.2 is remarkably similar. Figure 10.2 illustrates that it is possible to obtain as good information about the implicit, unobservable reaction time of an attention-grabbing response as it is about a motor reaction time.

Reeves (1977) obtained 17 pairs of motor and attention reaction times in a variety of conditions. The ART and MRT distributions are not always quite as similar as in figure 10.2, although they are highly correlated. An increase in difficulty of target detection causes a somewhat greater increase in mean ART than in mean MRT. This implies that the target is processed somewhat more fully before an ART (as opposed to an MRT) is initiated.

Reporting Four Numerals To obtain more information about the attention microprocesses that underlie ART performance, it is useful to gather more extensive data than are illustrated in figure 10.2. The "grabbing" procedure described above was elaborated to require the observer to report not merely the first numeral that he could from the numeral stream, but the earliest four. In all other respects the procedure was identical. The observer merely had to, after reporting one numeral from the numeral stream as before, now report three more numerals. Control experiments showed that when the observer was reporting four numerals, the first-reported numeral had the same statistical properties as the only reported numeral when the observer was reporting just one. Thus, the three additional numeral reports are obtained at no cost.

Single-Item Data Figure 10.3 shows a complete set of data for one observer and one target. It shows the four reported numerals at each of the four numeral rates. There are eight temporal positions at which numeral reports are recorded, 4 such eight-point curves per figure panel, and four figure panels, for a total of 128 data points.

Data from Pairs of Items Figure 10.4 shows a different aspect of the data from the same experiment and same conditions as in figure 10.3. These data are used to test a strength model of visual short-term memory, and are based on the method of paired

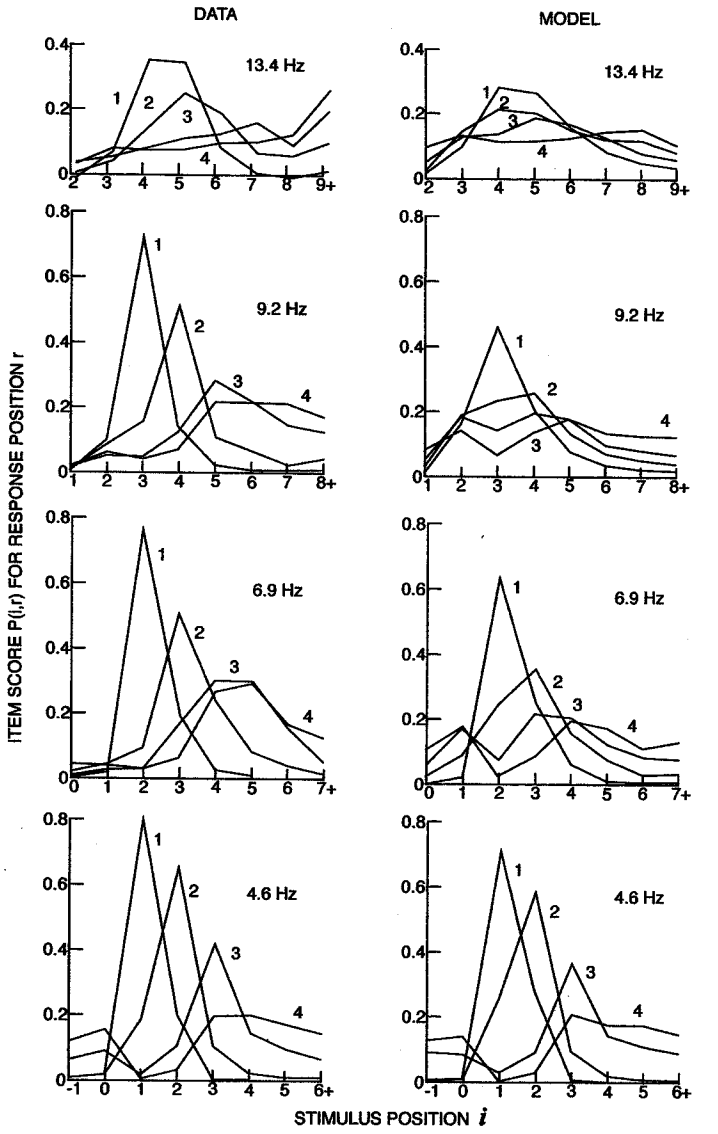


Figure 10.3
Data and model predictions for a modified attention-gating experiment ("Report four numerals!"). The abscissa is the position i of the reported items from the to-be-attended stream. The ordinate $P_i(r)$ is the estimated probability of reporting the item from stimulus position i in response position r. The curves labeled 1, 2, 3, 4 represent r, the first-, second-, third-, and fourth-reported items within a response. In all graphs, there is a progression from left to right of the response items: earlier response items tend to come from earlier stimulus positions. The speed of the to-be-reported stream is indicated in terms of the number of items per second (13.4 to 4.6). The left column shows the data for one subject; the right column shows the model fit to these data. (See text for details.)

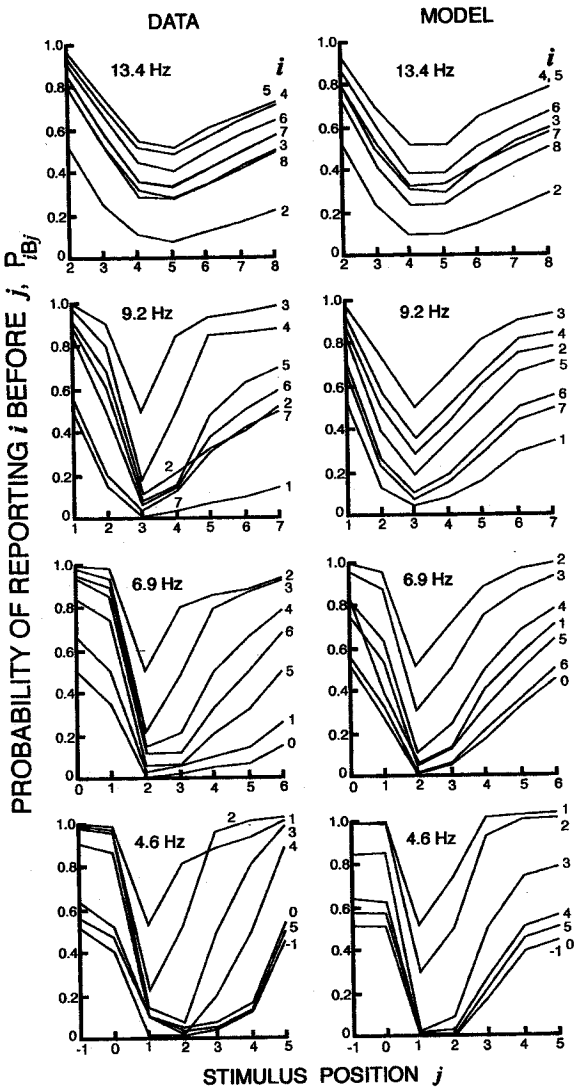


Figure 10.4
The probability P_{iBj} of reporting an item from stimulus position i earlier in the response than an item from stimulus position j as a function of j. Target and stimulus speeds are as in figure 10.3. The left column shows data for one subject; the right column shows the model fit to these data. Each curve represents a particular stimulus position i (indicated at extreme right, adjacent to the curve). Model curves are perfectly laminar (do not cross); their relative heights therefore precisely represent the relative strengths of the memory representation of the indicated stimulus positions.

comparisons. Suppose two numerals, *i, j* occur in the same response. If numeral *i* is reported *before* numeral *j*, we write *iBj;* otherwise we write *jBi.* We regard being reported first as "winning" or achieving primacy in short-term memory. Each trial is analogous to a sports or chess tournament in which we are given the order of the four best competitors (the four reported positions).

There is an extensive mathematical development that deals with precisely this situation: determining the relative strength of different players, even when they may not have played against each other, by determining how they fare against common opponents. The relevant data are *paired comparisons:* the collection of available *iBj* pairs.[1]

Figure 10.4 shows P_{iBj}, the observed probability of reporting position i before position j, as a function of j. Each curve is for a different i. In order to display continuous curves, we arbitrarily (but logically) define $P_{iBi} = 0.5$. There are seven critical positions for which data were collected, and this results in twenty-one independent P_{iBj} values in each panel, yielding eighty-four data points in the four panels. The top panel of figure 4 (numeral rate 13.4/s) shows that position 5 is the strongest: 90% of the time it is reported before position 2, 80% of the time before position 3, and never less than 50% of the time before any other position. However, position 5 is in a virtual tie with position 4, which is second strongest. Third strongest is position 6, followed by positions 7, 3, 8, and 2.

Laminarity and Folding The data have two interesting properties: laminarity and folding. Laminarity means that the curves do not cross. A failure of laminarity means a circle: *iBj* and *jBk,* but *kBi* (instead of the expected *iBk*). The data predicted by the model are perfectly laminar (righthand panels, figure 10.4). The real data have 5% crossings, and statistical analysis shows that this number, although very small, is slightly higher than the number (2–3%) that would would be expected by chance. To a very good approximation, however, laminarity holds for the real data. This means that, to a very good approximation, the data can be described by a strength model.[2]

A strength model means that the order of reporting an item from a position, except for random variation, is determined entirely by the memory strength of the position. The strongest position occurs roughly 300–400 ms after the target. Item strength is roughly symmetric around the strongest position. Items from weaker positions before and after the strongest position alternate in the response. This property is *folding.*

10.2.2 Model for a Temporal Attention Window: The Engine

The properties of laminarity and folding in the item pairs of the panels of figure 10.4, and the progression of the individual item reports from chaotic to orderly in the panels of figure 10.3, can be nicely encapsulated in an attention-gating model. This is a model of an attention window that gates the flow of information from the input to short-term

memory. The engine of the model is illustrated in figure 10.5, which shows the time course of the attention window.

The strength of an item in memory is determined by the height of the window function during the time the item is visually available, which is the time from its initial exposure until it is overwritten by the next item. (A more elaborate model would assume that an item is stored in *sensory memory* and that its availability decays exponentially. For the short time intervals under consideration here, this is an unnecessary complication.) The integral of visual availability over time determines total attention strength. The laminarity property then implies that items are reported in order of their strength, independent of when they might have occurred within the attention window.

The attention window principle is much like the old Compur-Rapid camera shutter that opened a diaphragm embedded within the lens to expose the whole image and then closed it. This type of shutter opens over a period of tens of milliseconds and then closes with a similar time course. If such a shutter had photographed the stimulus array, and the observer then reported items from the photograph simply in order of their clarity in the final image, it would correspond exactly to the presumed process here.

10.2.3 Model for a Temporal Attention Window: The Full Model

Just as a car needs more than an engine to make it useful, so an attention model needs more than an attention window. To generate data, a representation of both input streams is needed, as well as an explicit response-generating mechanism. The target-detecting and -interpreting mechanism is represented in the model[3] by a simple delay τ.

The Temporal Attention Window (Attention-Gating Function) The attention window must be represented by a causal function. It cannot be a normal density function, because this begins at $-\infty$ and therefore is not causal. The simplest causal function is an exponential decay function, that is, a one-stage RC circuit (first-order Gamma function). The absolutely instantaneous onset from zero to maximum value makes this function unrealistic. The next simplest function is two successive, identical RC stages (a second-order Gamma), and that is what was chosen to represent the shape of the attention window. The Gamma function controls the gate to short-term memory.

The Next-to-Be-Attended Pathway The next-to-be-attended stream from which the response items will be chosen is represented in the top row of figure 10.5. Items are assumed to be visually available until they are overwritten. Their access to memory is determined by the attention gate, which at each instant of time multiplies the next-to-be-attended-item by the height of the attention window. The integrated product determines strength in memory. Item strength is subject to random variation, represented as added noise. Items are output in order of their net strength.

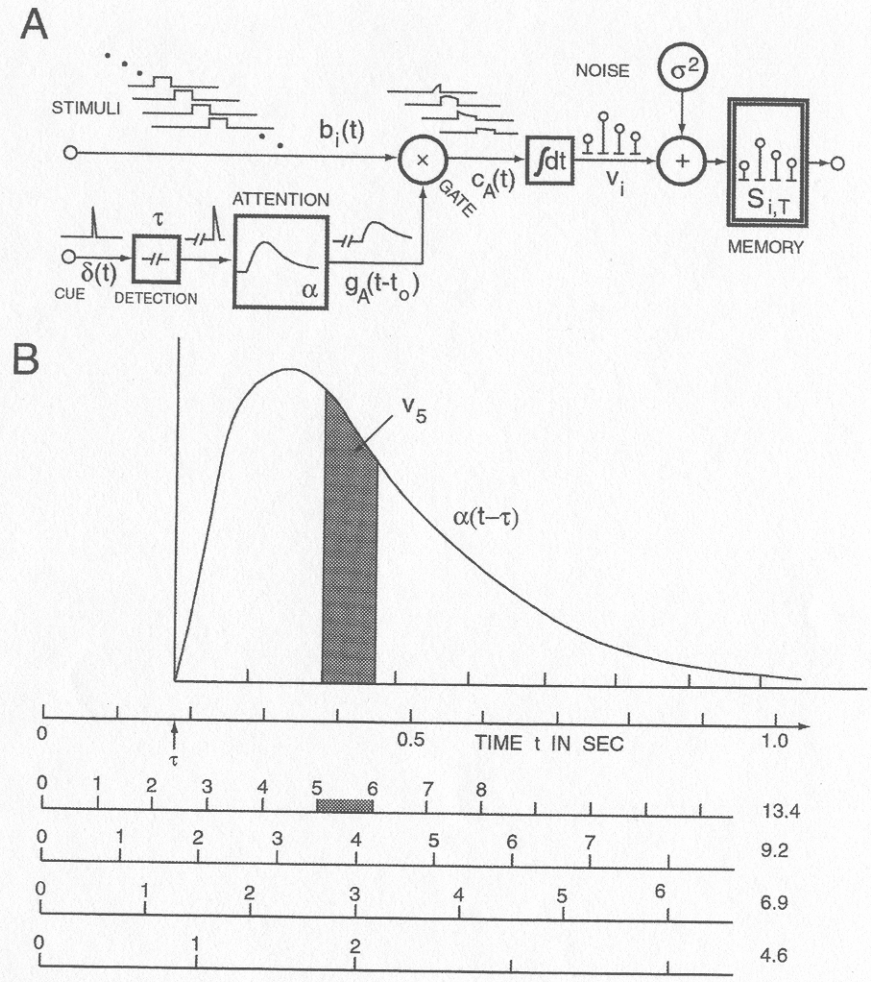


Figure 10.5
Model for the attention gating experiment. (*A*) Block diagram of the model. There are two input streams: the upper one receives the stream of to-be-attended items, $I(t)$; the lower one receives the target—the cue to switch attention, $\delta(t)$. Detection of the target occurs after a delay τ, at which time an attention window is generated, as represented by the box α. The attention window is produced by two consecutive RC stages, each with time constant α. Although items of the to-be-attended stream are presented instantaneously, they are visually available until the arrival of a subsequent item, as indicated by $b_i(t)$. The attention gate, $\times$, multiplies the visual information $b_i(t)$ by the attention window to produce $c_A(t)$, the temporal function that describes the instantaneous availability of the ith item. The integral $v_i = \int c_A(t)dt$ gives the strength of item i. On a particular trial T, strength is perturbed by random Gaussian noise with variance σ^2 to produce the net strength of item $S_{i,T}$ in short-term memory. Response items are output in order of their net strength. (*B*) Detailed illustration of the attention window. The curve $\alpha(t - \tau)$ describes the time course of an attention window. The strength of a particular item (here, item 5) is given by the area v_5 under the window during the time that item 5 is visually available. The example is for a presentation rate of 13.4 items per second. Slower rates would produce bigger areas under the attention window.

Efficiency The model has only three estimated parameters: τ, the time needed to detect and interpret the target, which in this instance is also the cue to shift attention; α (the effective width of the attention window is $\alpha\sqrt{2}$); and σ, the standard deviation of the memory noise. In effect, σ scales the memory strength, because it determines by how much two positions, *i* and *j*, must differ in strength in order for position *i* to be reported before position *j* with a probability *P*. Without noise, the order of report would be completely deterministic.

The parameter τ has nothing to do with the attention mechanism *per se;* it reflects the processes that detect and interpret the cue. Thus only two attention parameters need to be estimated from the data: the width of the attention window and the power of the memory noise. The model with one detection and two attention parameters generates the 212 predictions shown in figures 10.3 and 10.4. These predictions account for 0.85 to 0.90 of the variance of the data, depending on the observer and the condition. For example, changing to another target generates a new set of 212 points but requires only one new parameter (τ, which characterizes the speed of target detection). Accounting for 0.85 to 0.90 of the variance is not perfect prediction, but it is impressively efficient. Three targets were investigated for each observer so, with five estimated parameters, the model accounts for 636 data points per observer.

10.2.4 Extended Attention Models

One Attentional Episode The attention-gating model implies that, for items accumulated within a single attention window, observers have no intrinsic information about the temporal order in which items were entered into memory. In the absence of information provided by lower-level processes such as apparent motion, and in the absence of correlations between successive items (as might occur with meaningful words), the attribute used to order memory items is their memory strength. In this theory, discriminating the temporal order of two successive events requires two successive attention windows.[4]

Two Consecutive Attentional Episodes When two successive attention episodes occur, such as detecting a target (and remembering it) and then switching attention to a next-to-be-attended stream (and remembering items from that stream), observers can discriminate memory items that belong to the target stream from items that belong to the next-to-be-attended stream. They can discriminate these two episodes even when the target is embedded in the next-to-be-attended stream and does not itself differ from other items, as illustrated below.

Figure 10.6 illustrates two successive attention episodes. The task of the observer was to attend a single stream of characters until a target was detected, and then to report that target and the next three characters. The target was one of the characters in the stream

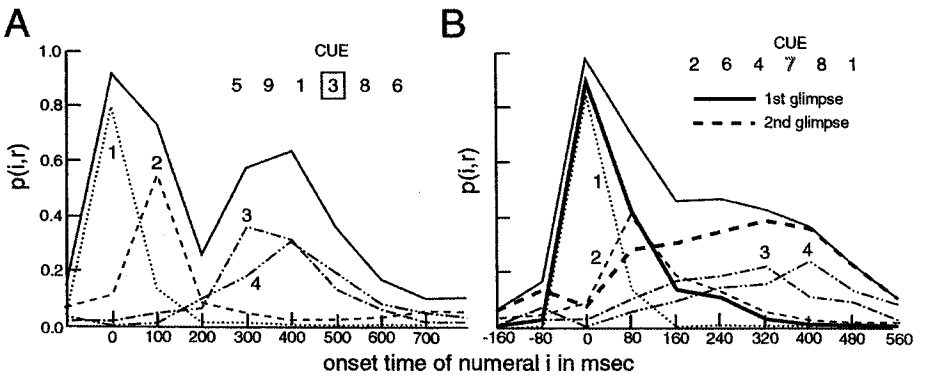


Figure 10.6
Attention windows generated by two successive attention episodes. The subject (EW) monitors a stream of items until a cued item occurs. He then attempts to report the cued item and the subsequent three items. (*A*) The probability of reporting items from a particular stream position when the cue is an outline square around the target item (as shown). The envelope curve indicates the cumulative probability of reporting an item from a particular stream position in any of the four responses. The four curves under the envelope, ordered from left to right, indicate the probability of reporting an item from position *i* in the first, second, third, and fourth response positions, respectively. (*B*) Here the target item is more intense than other items. In addition to reporting each item, the subject indicated whether it was in the first glimpse (thick solid curve) associated with the target or in the second glimpse (thick dashed curve) associated with a subsequent voluntary shift of attention. The form of the second glimpse coincides exactly with subject EW's results in the ART experiment (i.e., when reporting items from a next-to-be-attended stream with no requirement to report the cue in the attended stream; see figure 10.2).

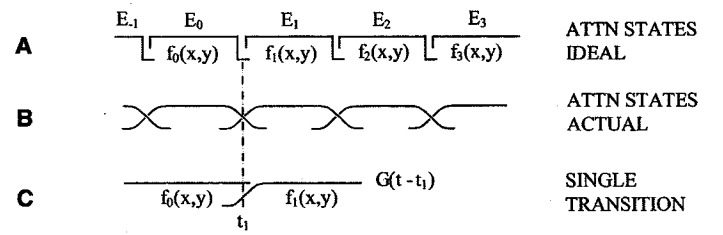
that either had greater luminous intensity than the other characters or was surrounded by an outline square. In addition to reporting four characters, the observer reported whether each character was associated with the target (described by the observers as the first glimpse) or with a second group of characters (second glimpse). Observers were able to report the target almost 100% of the time, and frequently the next occurring item. These constituted the first glimpse. The second glimpse had precisely the same distribution of items as the items in an attention shift from one location to another (as described above).

In terms of mechanisms, the two successive glimpses described by the observers correspond to two consecutive memory episodes. The distinction between the two episodes (glimpses) is quite clear. For both episodes, the memory structure maintains successive items simultaneously—unlike visual sensory memory (iconic memory), in which the contents are overwritten by succeeding items. Accessing the contents of such a memory requires a memory access code, usually called a retrieval cue. For the target item, the

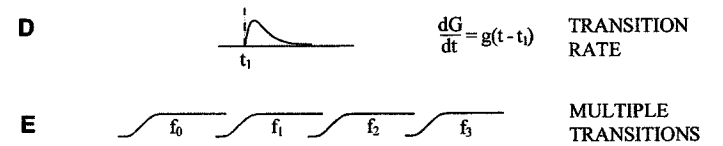
retrieval cue is simple: it is "all the items that are stored in association with an outline square (or with a sudden intensity increase)." For the second episode, the access code is an internally generated code: "all the items that are associated with the attention window created in such-and-such circumstances and at such-and-such a time." (The observer does not have direct access to the attention window itself, only to its contents and to their context.) It is not surprising that items associated with a brief visual retrieval cue are much more tightly grouped in time and more reliably reported than items associated with an internally generated retrieval cue.

Multiple Attention Episodes: Discrete Spotlight Model Visual attention can be well represented by a spotlight model that is actually used in many theaters. In the model or in the theater, there is a collection of available spotlights. For convenience, they are numbered in the order of their use, so that the same physical spotlight may have many numbers. A spotlight i illuminates some portion of the stage; its spatial distribution of illumination is given by $f_i(x,y)$. Only one spotlight is turned on at a time. The lighting program is a sequence of immediately consecutive events designated as episodes E_i, each characterized by a starting time, an ending time, and a spatial distribution of light, as illustrated in figure 10.7A. When, at time t_i, power is switched from spotlight $i - 1$ to spotlight i, the transfer of power takes a nonnegligible amount of time. Thereby, there is a certain amount of unavoidable overlap in the light from adjacent successive episodes during the transfer period (figure 10.7B). The time course of the transfer of power from one spotlight to another is described by a temporal function $G(t - t_i)$. This function is a cumulative probability distribution function that increases monotonically from zero to unity as t increases (figure 10.7A). For example, in switching from the initial spotlight with light distribution $f_0(x,y,t)$ to spotlight $f_1(x,y,t)$, the amount of light on the stage, $A(x,y,t)$, is given by equation 1 in figure 10.7. In the more general case, there is a very large number of successive episodes that could extend (for mathematical simplicity) from $-\infty$ to $+\infty$, as formalized by equation 2 in figure 10.7. Because different power transfers may have different transition functions, the temporal function G_i in equation 2 is subscripted with episode i.

The extension of the stage illumination model to visual attention is quite straightforward. Illumination in the theater model is analogous to attention in an attention model. In the theater, information is available primarily from illuminated portions of the stage. In visual tasks, information is available primarily from areas of the visual field in which there are significantly nonzero values of attention $f_i(x,y,t)$. In the theater, the positions of the spotlights are fixed during rehearsals. Actors learn to move to where the lights will appear or the actors will find themselves in the dark. Similarly, in attention experiments, observers learn the typical sequence of events during a practice period that is often quite



Eq. 1 $$A(x,y,t) = E_0 + E_1 = f_0(x,y)\,[1-G_1(t-t_1)] + f_1(x,y)\,G_1(t-t_1)$$



Eq. 2 $$A(x,y,t) = \sum E_i = \sum f_i(x,y)\,[G_i(t-t_i) - G_{i+1}(t-t_{i+1})]$$

Figure 10.7
Attention as a sequence of space–time separable episodes. (*A*) A sequence of ideal attention states (episodes), E_0, E_1, E_2, E_3, . . . Each episode E_i is characterized by an onset time t_i, an offset time t_{i+1}, and a function $f(x,y)$ that describes the spatial distribution of attention (salience) during E_i. (*B*) Actual attention episodes start and turn off gradually (not instantaneously). (*C*) An isolated, single attention transition from E_0 to E_1 that occurs with temporal transition function $G(t - t_1)$. For the example in (*C*), attention $A(x,y,t)$ is the sum of E_0 and E_1 (equation 1). (*D*) The rate of an attention transition is a probability density function. (*E*) A sequence of attention episodes showing the spatial attention distribution functions that are in effect during each episode. Equation 2 is the general formulation of attention as the sum of a sequence of episodes.

extended, so that during the experiment proper, the observer's performance is highly reproducible and stereotypic.

One intrinsic property of the discrete spotlight stage model is that switching time does not depend on where spotlights happen to be pointing. More specifically, switching time is independent of distance. That attention switches should be independent of the distance of the attention shift is counterintuitive, but it has been verified in different laboratories (Cheal and Lyon, 1989; Sperling and Weichselgartner, 1995).

The shape of the attention window (as in figure 10.5) comes about from three successive episodes: (1) wait for and detect the cue to switch attention; (2) switch attention to the next-to-be-attended stream and admit items to memory; (3) close the attention window to avoid memory overflow. The net outcome of these processes is illustrated in figure 10.7.

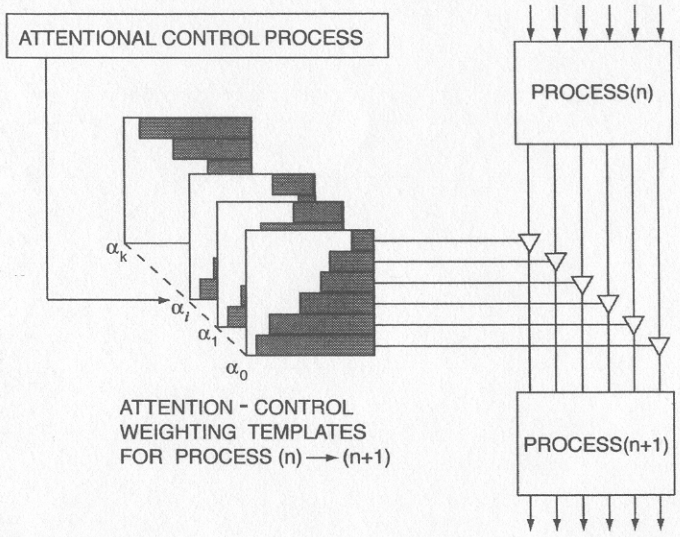


Figure 10.8
Attention in a neural network. Attention modifies the passage of signals from a neural process n to $n + 1$. In well-practiced experimental subjects, a cue to shift attention causes a previously learned template of attentional weights α_0 to be quickly put into place. This may occur simultaneously at several different levels n.

Neural Implementation of Attention Neurally, attention is implemented as a control process that modulates the passage of information between neural processes (n) and ($n + 1$), as illustrated in figure 10.8. In the experimental situations in which attention is measured, there are typically thousands of trials, so performance becomes both optimal and, concurrently, quite stereotypic. The attention templates of weights (e.g., the $f_i(x, y, t)$) are well learned and quickly instantiated. Indeed, attention acts not only at the gateway to memory but also concurrently, at many levels. Attention determines, in perceptual stages, what information is passed on to pattern recognition processes or to memory; in decision stages it determines bias and sensitivity parameters; and in response selection and response execution stages, it determines the speed and accuracy with which particular responses are executed.

The model illustrated in figure 10.7 has been applied to four of the most widely used attention paradigms, and it quantitatively accounts for the results of quite diverse experiments. For more detail, the reader should consult Sperling and Weichselgartner (1995); the remainder of this chapter describes methods for examining the microstructure of attention processes.

10.3 The Salience Map: An Implementation of Attention

The logic behind this section is that apparent motion can be used as a delicate assay of attention. In particular, it is possible to construct third-order motion stimuli in which the direction of apparent motion is determined by attention. The fact that attention influences the direction of motion is itself diagnostic, and gives important insights into the mechanisms of attention. It is used here to develop a computational model of how attention to a feature, such as "red," is implemented via a salience map. To proceed, we need first to clarify what motion systems are, and in particular what a third-order motion system might be. This, in turn, requires the concepts of figure–ground segmentation and of a salience map.

10.3.1 Motion Systems, Flow Fields, Attention-Driven Apparent Motion

First-Order Motion A motion system is a neurophysiological concept derived from psychophysical experiments; the essential ingredient is a flow field computation. To illustrate this, we consider the input to the first-order motion system, namely, the dynamic sequence of images that is formed on the retina and transformed by the early processing stages of the visual system. Processing by the retina removes the mean stimulus luminance from the signal (for nearly all neurons), so that only contrast signals (i.e., deviations from mean luminance) are transmitted to the lateral geniculate nucleus and cortex.

Let the stimulus luminance at a point with spatial coordinates x,y at time t be $l(x,y,t)$. Then the point contrast $c(x,y,t)$ is the normalized amount by which the luminance $l(x,y,t)$ differs from the mean luminance l_0:[5]

$$c(x,y,t) = (l(x,y,t) - l_0)/l_0 \tag{3}$$

Positive values of $c(x,y,t)$ are carried by retinal ganglion cells and lateral geniculate cells with ON-center receptive fields, and negative values by OFF-center cells (Kuffler, 1953). The first-order motion system takes point contrast as its input and produces the first-order flow field as its output. The flow field $F_1(x,y,t)$ is a vector function that indicates the direction and velocity of motion in the neighborhood of location x,y, at time t.

A flow field does not directly indicate what may have caused the motion; it represents only the motion itself. And though we do not know exactly how the brain computes velocity, we do know that the first-order motion flow field is used to compute 3D structure from 2D motion (kinetic depth effect), and that it contributes to the control of locomotion, balance, orientation, and all the other functions usually attributed to motion perception (Dosher et al., 1989). Subsequent processing stages combine the information from a motion flow field with contour, color, texture, and other features to serve object and scene perception.[6]

Second-Order Motion The second-order motion system computes a flow field analogously to the first-order system except that it discards the sign of $c(x,y,t)$ before the flow field computation, that is, it rectifies the point contrast and uses the absolute value (or squared value) instead of the point contrast directly. In neural terms, the outputs of ON-center and OFF-center cells are treated identically instead of oppositely.

As with first-order motion, there are complications. In second-order motion processing, rectification is preceded by spatiotemporal filtering, a combination that has been called texture grabbing (Chubb and Sperling, 1988). Spatial filtering followed by rectification means that the second-order system is sensitive to the amount of texture in each neighborhood of the stimulus, which is closely related to the luminance *variance* within the neighborhood. Whereas while the first-order motion system reports on the movement of areas that have fewer or more photons than their surround, the second-order system reports the movement of areas that have fewer or more texture features than their surround.

Salience Map, Third-Order Motion The third-order system generates its flow field from figure–ground information. Most visual images can be segmented by the perceptual system into figure (the important parts that are designated for further processing) and ground (the remainder). According to Lu and Sperling (1995a, 1995b), Sperling and Lu (1998), and Blaser et al. (1999), the results of the figure–ground computation are stored in a salience map where figure is represented, for example, by 1 and ground by 0.

Not every point in every image can be unambiguously classified as figure or ground. Therefore, it is useful to define a real-valued variable, *salience,* to indicate the relative importance (or ''figureness'') of each image point in space and time. The instantaneous

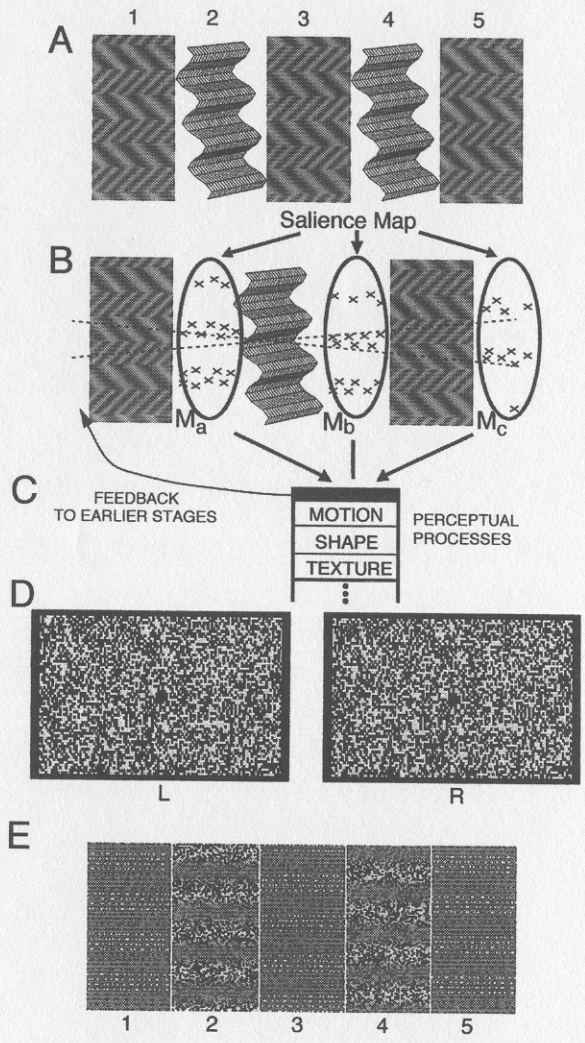


Figure 10.9
How attention influences ambiguous motion displays via a salience map. (*A*) Five frames of an ambiguous motion display (1, 2, 3, 4, 5) with alternating features: odd frames modulate texture, even frames modulate binocular depth. Consecutive frames are shifted in phase by 90°, so that a motion signal arises only from the combination of odd and even frames (i.e., no motion within only the odd, or only the even, frames). (*B*) Three frames of the display (*a, b, c*) with their salience map representations, M_a, M_b, M_c, (*ellipses*) immediately to the right of each frame. When a subject attends the coarse texture patches, these patches acquire a higher salience value, as indicated by the X marks in the salience map (M_a, M_c). Areas with less binocular depth are automatically perceived as foreground, as indicated by the X marks (M_b). The dotted lines indicate the two possible directions of apparent motion (downward when attention selects the coarse texture, upward when it selects the fine texture). (*C*) Outputs of the salience map go to subsequent processes that compute motion, shape, and texture. (*D*) Left-eye and right-eye images (L, R) of one frame of the dynamic random-dot stereogram used to create a translating corrugated surface in depth (as shown schematically in *A* and *B*). (*E*) Five frames of an ambiguous third-order motion display. In the texture frames (2,3), salience is unambiguosly high in the high-contrast regions. In the other frames (1,3,5), the black-spot and white-spot regions have equal salience when there are no attentional instructions. Thus, no motion is seen without such instructions. Attention to white spots produces upward apparent motion; attention to black spots produces downward apparent motion.

values of salience at each point of the visual field constitute a *salience map* of the visual field. The third-order motion system uses the time-varying salience map as its input, and computes a flow field that gives the direction and the magnitude of salience movement at each point as a function of time.

The third-order motion system computes the motion of those parts of the visual field which are designated as "figure." This can be demonstrated by producing a succession of images in which the distinguishing features of the "figure" change from image to image. Figure may be defined by stereo depth in one image, by an area of greater texture contrast in the next image, and so on. If the areas defined as figure are displaced in a consistent direction from image to image, then observers perceive motion in that direction. It is worth noting that observers do not discriminate between motion that is produced by first-, second-, or third-order computations: they merely report "apparent motion."

The term "salience map" was first popularized by Koch and Ullman (1985), who used the concept to describe a winner-take-all network that determines a region in space from which information from various topographic feature maps is combined and directed to a central processor. Related concepts have emerged independently as an attention map (Mozer, 1991), a priority map (Ahmad and Omohundro, 1991), a selective tuning mechanism (Tsotsos et al., 1995; Tsotsos et al., chapter 14 in this volume), a hierarchical pruning mechanism (Burt, 1988), and under other names, with different authors giving somewhat different interpretations to these concepts.

Attention to Feature A remarkable aspect of third-order motion is that attention can strongly influence the direction of motion perception (Lu and Sperling, 1995a), but this is not true for first- or second-order motion (Solomon and Sperling, 1994). Lu and Sperling arranged ambiguous motion displays (figure 10.9) so that when observers attended to one feature (e.g., coarse stripes [figure 10.9A] or white spots [figure 10.9E]), the display appeared to move in one direction; when they attended to the other feature (fine stripes or black spots), the display appeared to move in the opposite direction. Attention to a feature determined the direction of apparent motion even when the sequence of displays occurred so rapidly (five displays in 333 ms) that observers were unable to track any specific elements.[7] That attention can determine motion direction even when feature tracking is impossible implies that there must be another mechanism by which attention operates in these displays. Lu and Sperling (1995a) proposed that attention enhances the attended features at a level prior to conscious perception, and that these enhanced features are recorded as figure (versus ground) in the salience map. By influencing the input to the salience map, attention can determine third-order motion.

10.3.2 Selective Attention to Color

To investigate the proposed role of attention in increasing the salience of features, Blaser et al. (1999) used a third-order motion display involving attention to a color, red or green. Their experiment was designed to answer the following question: To what extent is selective attention to red (or green) equivalent to increasing the redness (or greenness) of a motion stimulus? Ultimately, this enabled them to measure the amount by which attention to a feature amplifies its salience.

Stimulus Sequence The procedure used by Blaser and colleagues (1999) is shown in figure 10.10 (see also plate 6). A motion sequence consisted of five consecutive frames. In figure 10.10, the even frames (numbers 0, 2, 4 . . .) contain a contrast-modulated texture grating, and the odd frames (numbers 1, 3, 5 . . .) contain an isoluminant red–green grating. There is a 90° phase shift between consecutive frames. The phase shift between two color frames is 180°, so there is no directional motion signal within the color frames. Similarly, the phase shift between consecutive texture frames also is 180°, so there is no directional motion signal within the texture frames, either. To perceive a direction of motion, information from the color and texture frames must be combined.

The luminance is the same (average luminance in the case of texture areas) in all parts of all frames, both color and texture, so there is no usable first-order motion signal. This was verified by a sensitive calibration procedure (Anstis and Cavanagh, 1983; Lu and Sperling, 1999). Similarly, there is no significant texture in the isoluminant grating to stimulate the second-order motion system. Indeed, without attention instructions, observers usually do not report motion from this stimulus sequence.

To create the isoluminant color grating, the red gun of the display monitor was set to maximum intensity, and the green gun was adjusted to be of equal luminance. When the red and green stimulus colors were mixed 50/50, the result was a yellow that was equal in luminance to both the red and the green. The background was formed of this yellow. To create desaturated stripes of a color, say red, between 0 and 50% of the red was exchanged for green.

Varying Salience: Red Advantage In the isoluminant color grating, the salience of a stripe (red or green) is assumed to be monotonically related to the amount by which it differs from the background (Lu et al., 1999). Blaser and colleagues (1999) called this the "chromaticity difference," which here is defined as follows: Let r and g represent the intensities of the red and green guns, respectively ($r,g \leq 1$). To maintain isoluminance, $r + g = 1$ at every location and point in time. The chromaticity difference $|R|$ of a red stripe from a yellow background is $|R| = r - g (r > g)$; the chromaticity difference of a green stripe is $|G| = g - r (g > r)$.

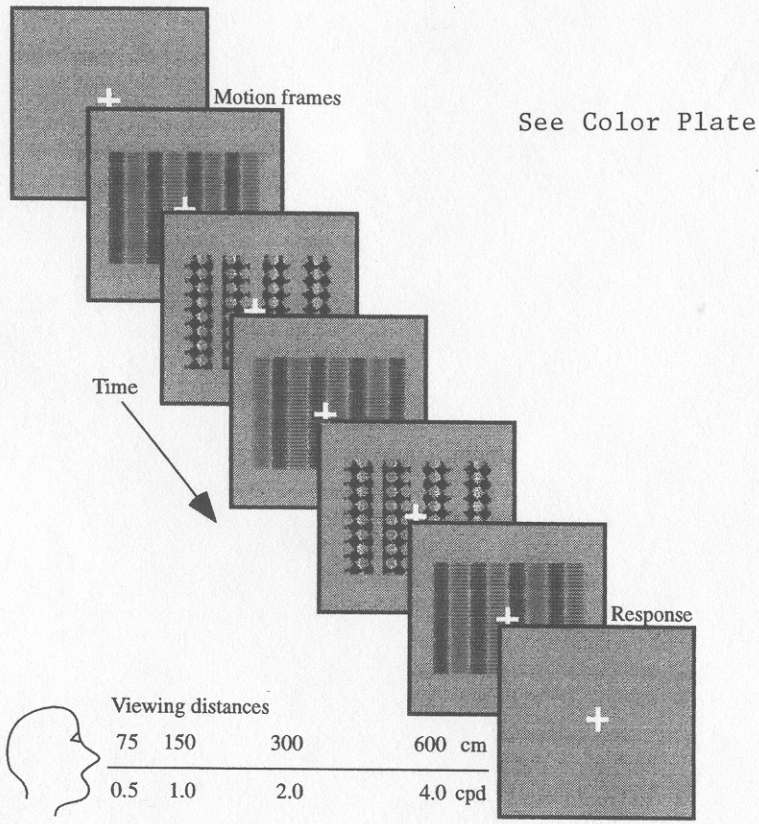


Figure 10.10
Procedure using amplification principle in third-order motion to measure the attentional amplification of salience. Even frames are texture-contrast gratings, with unambiguously high salience in the high-contrast texture bands. Odd frames are red–green color gratings, characterized by separate red-saturation and green-saturation values. Motion strength (e.g., as measured by Reichardt and motion energy detectors) is determined by the product of the modulation amplitudes in even and odd frames. When the texture modulation in the even frames is far above threshold, even weak salience modulations in the odd frames can produce apparent motion. Different viewing distances determine the spatial frequency (cycles per degree) of the gratings on the retina. (See plate 6 for color version.)

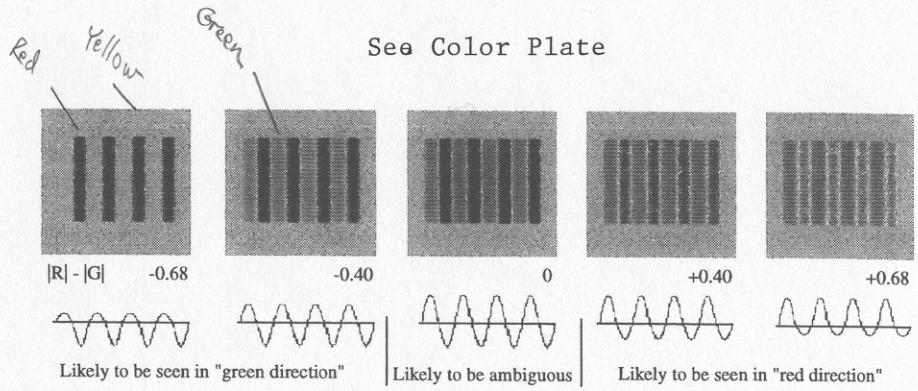


Figure 10.11
Five stimuli with different red advantages. From ~~top~~ L to ~~bottom~~ Rt: −0.68, −0.32, 0, +0.32, +0.68. For the displays of figure 10.11, attending to green (or red) produces apparent motion equivalent to a stimulus approximately one level higher (or lower). (See plate 7 for color version.)

It is critical whether red or green differs more from the background, because the stimulus will appear to move in one direction (the red direction) when red differs more, and in the other direction (the green direction) when green differs more. This aspect of the stimulus is characterized by a quantity called *red advantage,* which is simply $|R| - |G|$. For example, a stimulus that has only red stripes on a neutrally yellow background without green stripes (i.e., $|R| = 1$ and $|G| = 0$) would have a red advantage of 1. A stimulus that has only green stripes on a neutrally yellow background without red stripes (i.e., $|G| = 1$ and $|R| = 0$) would have a red advantage of -1. Finally, a stimulus with $|R| = |G|$ has a red advantage of 0. Stimuli actually used in the experiment had red advantages of -0.68, -0.4, 0, $+0.4$, and $+0.68$ (see figure 10.11, plate 7).

Experimental Procedure In all sessions, a trial consisted of 0.5 s of a blank frame with a fixation point, followed by a five-frame stimulus at 100 ms/frame, and the observers simply judged the direction of movement. The stimulus grating was four cycles wide, and it was embedded in a much larger yellow background. There were many possible stimuli: five different chromatic gratings with various degrees of red advantage, randomized spatial phase, and randomly chosen direction of movement. The assignment of color gratings to odd frames and texture to even frames was reversed randomly from trial to trial. There were four different viewing distances, blocked by session, to produce four different stimulus spatial frequencies. Initially, observers were not given any attention instructions and ran through the whole sequence of trials. Subsequently, they were told to attend to the red (green) stimulus, and the entire procedure was repeated. Then the observers were told to attend to the previously unattended color and the entire procedure was repeated again.

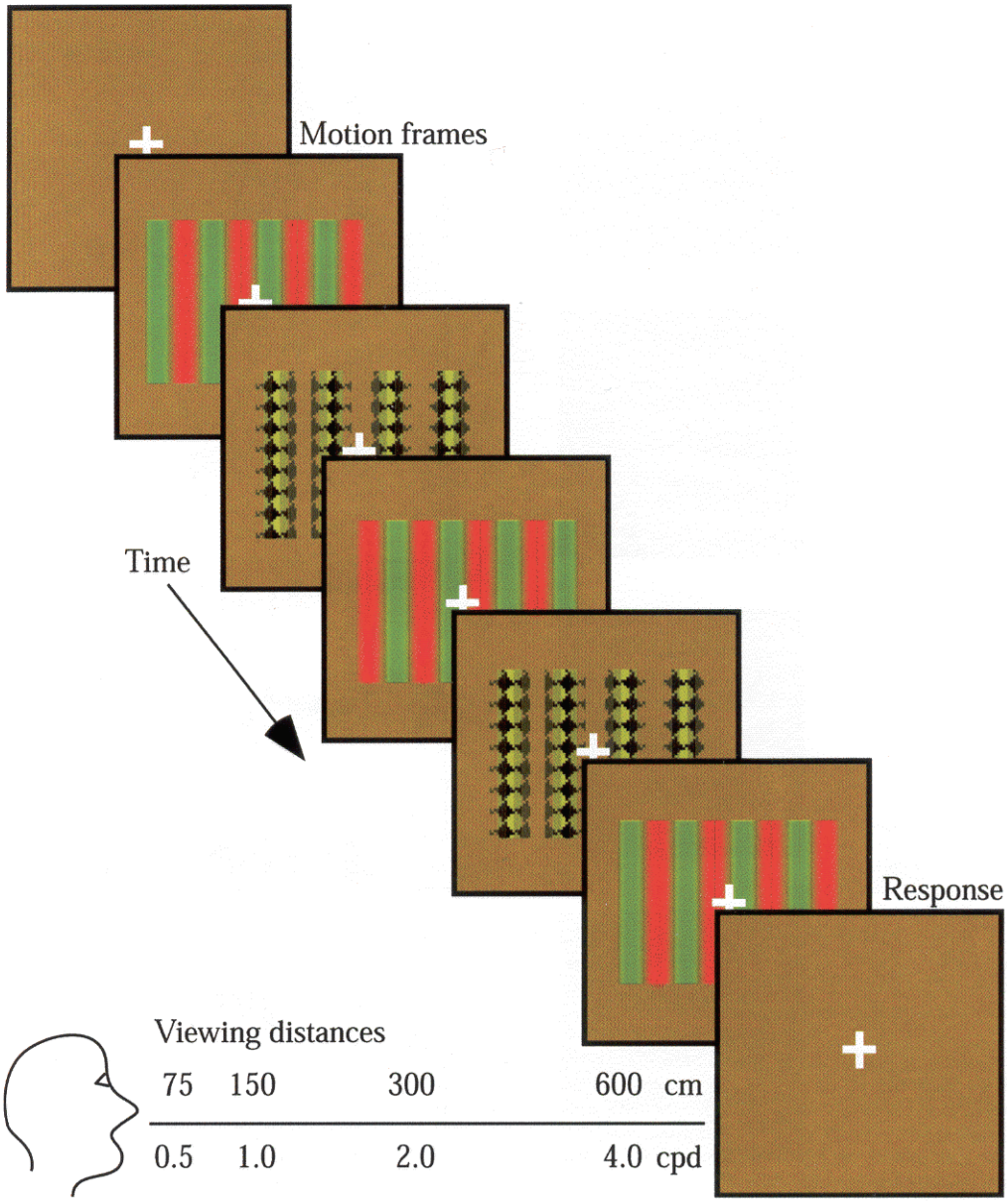


Plate 6 Procedure using amplification principle in third-order motion to measure the attentional amplification of salience. Even frames are texture-contrast gratings, with unambiguously high salience in the high-contrast texture bands. Odd frames are red–green color gratings, characterized by separate red-saturation and green-saturation values. Motion strength (e.g., as measured by Reichardt and motion energy detectors) is determined by the product of the modulation amplitudes in even and odd frames. When the texture modulation in the even frames is far above threshold, even weak salience modulations in the odd frames can produce apparent motion. Different viewing distances determine the spatial frequency (cycles per degree) of the gratings on the retina. See chapter 10.

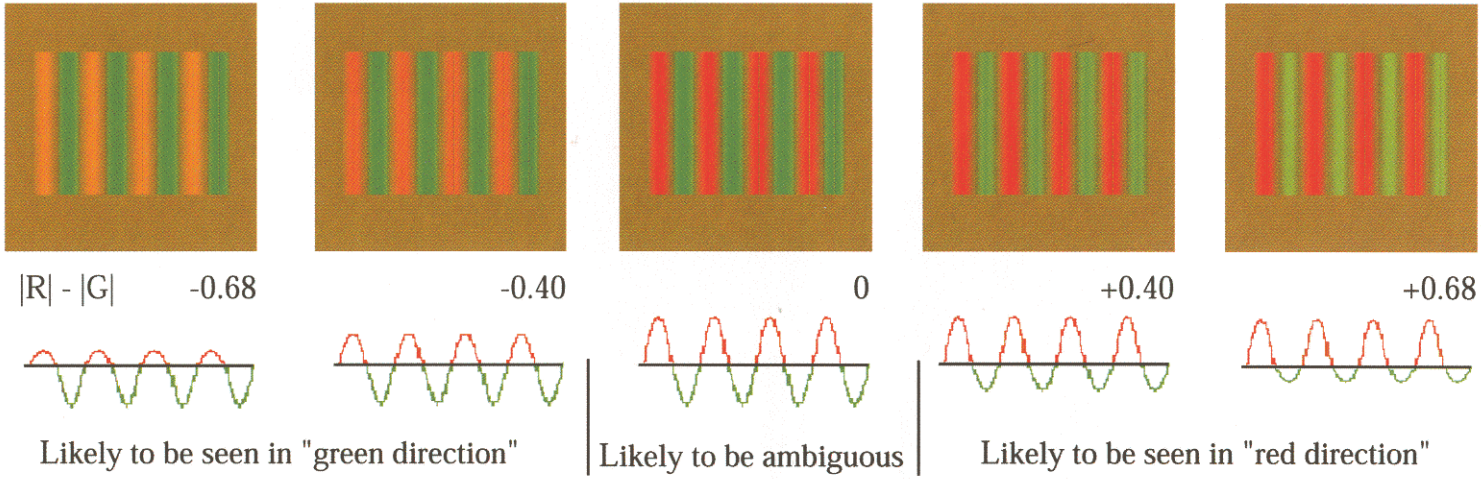


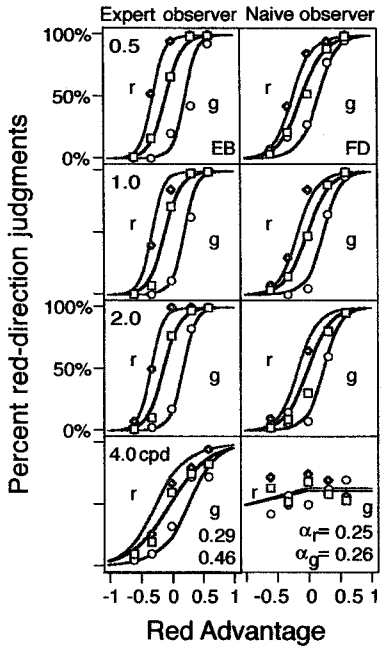
Plate 7 Five stimuli with different red advantages. From top to bottom: −0.68, −0.32, 0, +0.32, +0.68. For the displays of plate 6, attending to green (or red) produces apparent motion equivalent to a stimulus approximately one level higher (or lower). See chapter 10.

Results and Discussion Results are shown here for two observers, both practiced psychophysical observers. One was naive about the purpose of this experiment; the other was one of the experimenters. In the neutral condition (without attention instructions), when the red and green stripes both had maximum chromatic difference from the background ($|R| = |G|$), motion responses were random for one observer and showed a slight bias in favor of the red direction for the other. However, when there was a large red advantage ($|R| = 1.0$, $|G| = 0.32$, so $|R| - |G| = +0.68$), the direction of perceived motion was almost 100% in the red direction. For the same stimulus sequences, but with green advantage ($|R| = 0.32$, $|G| = 1.0$), the perceived motion direction was almost 100% in the opposite direction. For a spatial frequency of 4 cycles per degree (cpd), the resolution of the salience system for these stimuli is exceeded for one observer and his motion direction responses are almost random. The other observer's performance is impaired but remains far above chance.

The psychometric functions for the three resolvable gratings, in neutral attention conditions, go from 0 to 100% of apparent movement in the red direction as a function of red advantage (top three rows, figure 10.12, plate 8). This reflects the bottom-up control of salience. The greater the difference of a color stripe from the background, the greater its salience. In this kind of display sequence, when red is more salient, motion is in the red direction; when green is more salient, motion is in the green (opposite) direction; and when red and green are equally salient, there is no consistent apparent motion.

When observers pay selective attention to red, the psychometric functions appear to be shifted to the right; and when the observers attend to green, the opposite shift occurs. For example, under attention to red, direction judgments to the stimulus, $|R| = |G|$, are approximately the same as under neutral attention to a stimulus with a red advantage of +0.3, i.e., ($|R| - |G| = 0.3$).

Attention Does Not Change Appearance An interesting, informal observation is that attending to red or to green does not make a stimulus sequence look different than in the neutral attention condition. Certainly, selective attention to color in a static display does not produce any noticeable change in the appearance of the static display. This is entirely consistent with previous observations that attention reduces the variance of various psychological judgments but does not alter the appearance of simple features (Prinzmetal et al., 1998). Indeed, one would expect selective attention to make judgments of a feature more accurate, not to bias the judgments in a particular direction. The difference in appearance between stimuli having red advantages of 0 and of 0.3 is very obvious, and if attention produced even 1/10 of this difference in appearance, it would be quite noticeable.



See Color Plate

Figure 10.12
Results of the attention-amplification experiment. The percent of red-consistent motion judgments versus the red stimulus advantage, $|R| - |G|$ which is the difference $|R|$ of red from background yellow minus the difference $|G|$ of green from background yellow. As red advantage increases, the probability of perceiving motion in the red-consistent direction increases. Five data points are shown for each of four spatial frequencies (rows), three attentional conditions, and 2 observers (columns). Solid curves are model fits (see figure 10.13A). Middle curves indicate the baseline condition (no attention instructions); curves on right (g) and on left (r) are model fits for the attend-red and attend-green conditions, respectively. The estimated model parameters for the additional $|R|$ and $|G|$ amplification due to attention, α_r and α_g, are indicated in the bottom panel for each observer. (See plate 8 for color version.)

10.3.3 A Dynamical Systems Model of Salience and Related Processes

The continuous curves drawn through the data in figure 10.12 account for 99% of the variance of the data for these observers. These curves are generated by the model of figure 10.13A. The "reduced" model of figure 10.13A includes just those components needed to generate the particular predictions in figure 10.12. Figure 10.13B shows these same components embedded in a larger system that illustrates how they relate more generally to attentional and perceptual processes.

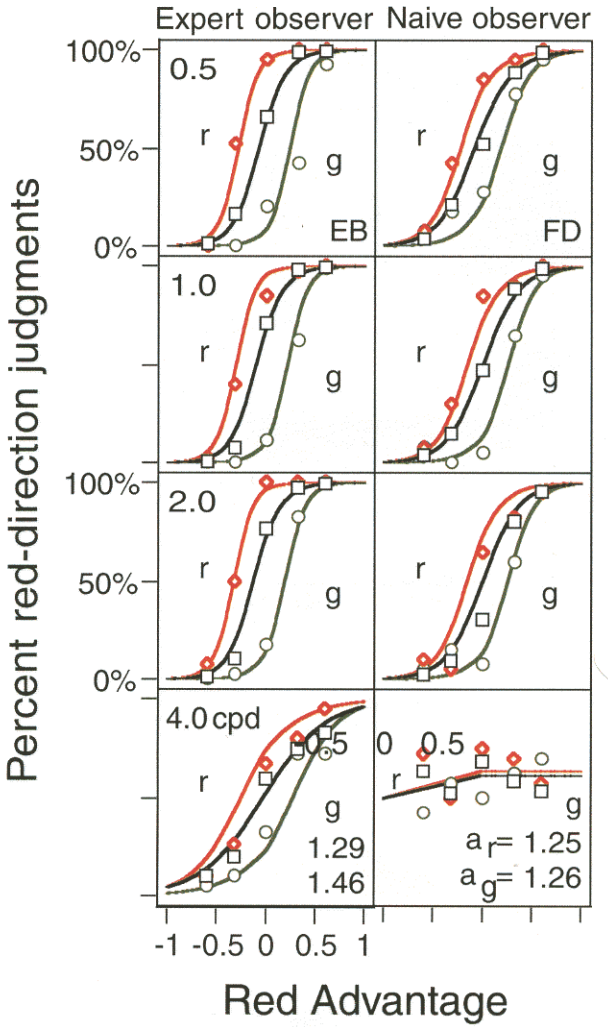
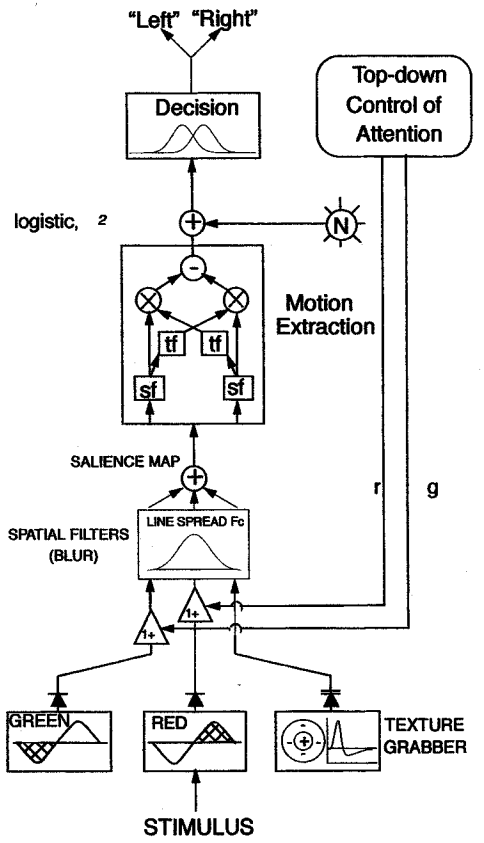
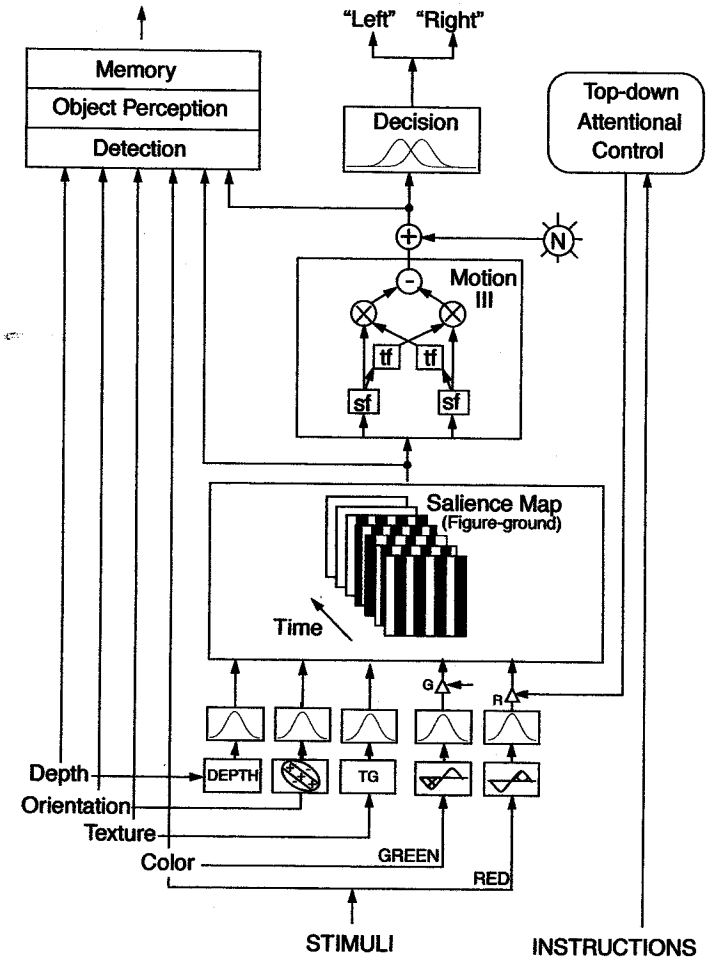


Plate 8 Results of the attention-amplification experiment. The percent of red-consistent motion judgments versus the red stimulus advantage, $|R| - |G|$, which is the difference $|R|$ of red from background yellow minus the difference $|G|$ of green from background yellow. As red advantage increases, the probability of perceiving motion in the red-consistent direction increases. Five data points are shown for each of four spatial frequencies (rows), three attentional conditions, and two observers (columns). Solid curves are model fits (see figure 10.13a). Middle curves indicate the baseline condition (no attention instructions); curves on right (r) and on left (g) are model fits for the attend-red and attend-green conditions, respectively. The estimated model parameters for the $|R|$ and $|G|$ amplification due to attention, α_r and α_g, are indicated in the bottom panel for each observer. See chapter 10.



A

Figure 10.13
(*A*) A computational model of attention processes in third-order motion. The inputs are stimuli and attention instructions; the output is a direction-of-motion judgment. Stimuli are analyzed along the dimensions of texture and color. Instructions to attend to a color (red or green) are assumed to increase the gain of the attended color channel in the salience pathway by a factor of α to $1 + \alpha_r$ or $1 + \alpha_g$ depending on which color is attended. The texture channel produces output proportional to the amount of local texture. The salience map is the sum of all the stimulus inputs in the salience pathway; its output goes to the Motion III (third-order) computation. The third-order motion computation is represented as a Reichardt model (Reichardt, 1961; van Santen and Sperling, 1984); it produces a real-valued output that indicates a direction of motion and is perturbed by additive noise *N*. A decision processes outputs a response "right" if its input is greater than a criterion, and "left" otherwise. (*B*) A more comprehensive model of visual processing that shows sensory inputs bypassing the salience computation en route to subsequent processing. Although high salience does not seem to perturb the appearance of objects, it does eventually determine which signals are analyzed and remembered. The third-order motion signal is also available to subsequent perceptual processes, as indicated.



B

10.3.4 Components of the Computational Model

Texture Grabber In the experiments, there are two kinds of inputs: visual stimuli and attention instructions. The stimuli are texture gratings and color gratings. To extract texture from the texture gratings requires a texture grabber (Chubb and Sperling, 1988, 1989a; Werkhoven et al., 1993). A texture grabber is composed of a linear bandpass filter (center-surround receptive field) that is most sensitive to spatial frequencies in a particular frequency range, a temporal filter, and a rectifier (figure 10.13). Because the output of a filter may be positive or negative, the filter output is rectified (absolute value or square) so that it represents the total quantity of texture. The texture grabbers of the second-order motion system are isotropic (circularly symmetric; see Werkhoven et al., 1993), but those of the third-order system are sensitive to orientation (Chubb and Sperling, 1991; Werkhoven et al., 1994). Although the filter could, in principle, process any texture, texture was not varied in this experiment. Therefore, for the present experiment, it is sufficient to assume that the output of the texture grabber is 1.0 in regions of maximum texture contrast and that it is 0 in regions where there is no texture.

Color Grabber Extracting an arbitrary color that differs from an arbitrary background is a complex problem. For the present experiment, it is sufficient to extract red or green from a yellow background, and this is simple. In direct analogy to a texture grabber, a color grabber can be constructed from a wavelength-sensitive filter responding positively to red and negatively to green (or vice versa) followed by a rectifier. It is assumed that when red or green areas of the stimulus are at maximum intensity, the output of the color grabber is 1, whereas the output is 0 for a yellow stimulus, and is between 0 and 1 for intermediate stimuli.

In the visual system, positive and negative signals are carried by separate neurons (e.g., ON-center and OFF-center neurons). The red (positive) and green (negative) outputs are assumed to be carried by separate neurons. This is critical for the attention amplification, which acts separately on the red and green outputs.

Attention Amplification Attention amplification is determined by instructions that have to be interpreted (a high-level cognitive process) and implemented (at a lower level). Under instructions to attend to red, the red amplifier is turned on and the output of the red channel is amplified, that is, multiplied by a factor of $1 + \alpha_r$, $\alpha_r > 0$, while $\alpha_g = 0$. Under instructions to attend to green, the green channel is amplified by $1 + \alpha_g$, $\alpha_g > 0$, while $\alpha_r = 0$. It is important to note that attentional amplification is independent of the stimuli being presented. An attention state is described by parameters $[\alpha_k]$ that represent the amplification of the various inputs. Once an attention state has been established, it determines the (altered) response to whatever stimuli may be presented.

Spatial Filter The experiments do not distinguish the spatial resolution of the color and the texture systems, so limited spatial resolution arises from the same spatial filter for all inputs. Because the data are essentially the same for spatial frequencies of 0.5, 1.0, and 2.0 cpd, with a severe decline in performance only at 4 cpd, the spatial filter need have only a single parameter, F_c, the corner frequency at which resolution declines. For greater accuracy, spatial resolution could be modeled perfectly with three parameters. This would ensure that estimates of attention components are not contaminated by errors in estimating spatial resolution.

Salience Map We assumed that in the brain, inputs from various sources sum at the salience map, and a complex figure–ground computation is performed. For the present experiment, it is sufficient to consider just the summing aspect of the computation, so in the block diagram (figure 10.13A) the salience map is represented by simple summation.

Standard Motion Analysis For humans, the extraction of the direction of movement from dynamic (first-order and second-order) stimuli is very well modeled by a Reichardt detector (van Santen and Sperling, 1984). Other theories—based on Fourier motion energy (Adelson and Bergen, 1985), on Hilbert detectors (Watson and Ahumada, 1985), and on spatiotemporal gradients (Adelson and Bergen, 1986)—have been shown to be similar or indistinguishable (in terms of their overall computation) from an elaborated Reichardt detector (van Santen and Sperling, 1985). This overall computation has been called "standard motion analysis" by Chubb and Sperling (1989b), and it applies to both first- and second-order motion.

The third-order motion computation clearly is different from standard motion analysis because it fails the pedestal test (Lu and Sperling, 1995b) and because it seems to be more sensitive to displacement than to motion energy (Krauskopf et al., 1999). Whether this is due to an intrinsically different motion computation or to the preprocessing of the input (so that amplitude is only very coarsely quantized) has not been resolved. So, the motion component is represented simply as standard motion analysis. It produces a positive output for motion in one direction and a negative output for motion in the opposite direction.

Noise and the Decision Process Psychophysical data are not deterministic; the same stimulus evokes different responses on repeated presentations. This is taken into account by adding Gaussian noise to the output of the motion detector. The variance of this noise determines the slope of the psychometric functions in figure 10.12. The decision process simply determines whether the net output is greater or less than zero, which represent the two permissible directions of motion in the experiment.

Parameters and Efficiency of Prediction There are four data panels representing the four stimulus sizes (spatial frequencies), and each panel has five data points for each of three attention conditions: sixty data points per observer. The model has two attention parameters, α_r and α_g, and one noise parameter, σ, that determines the slope of the psychometric function. Just one parameter is needed to describe the spatial filter for observer FD, but three are needed for the other observer. The four- and six-parameter predictions account for 99% of the variance of the data. This is efficient prediction.

The actual values of the attention amplification for the two observers in figure 10.12 are the following: observer EB: $\alpha_r = 0.29$, $\alpha_g = 0.46$; observer FD: $\alpha_r = 0.25$, $\alpha_g = 0.26$. The average value of 0.32 represents an attentional amplification of over 30%, which is quite significant.[8]

10.3.4 The "Full" Model

The computational model described above sufficed to fit the motion-direction data of the experiments. However, it deals neither with the observation that attention does not seem to change the appearance of stimuli, nor with the issue of how the processes described above relate to the more general functioning of a salience map. The full model links salience-related attentional processes to more general attentional processes.

Three Pathways Modeling the attention-motion experiment requires three conceptual pathways. The first is a pathway for the instructions to attend to a color. These instructions are interpreted at a high cognitive level. In the model, these high level processes then send a control signal that modulates the inputs to the salience map, that is, it controls amplification prior to the salience computation. The second pathway conveys the stimulus to the salience map. The third pathway conveys the stimulus directly to other perceptual processes, such as motion perception, shape recognition, object perception, memory, and subsequent cognitive processes. The direct pathway is suggested by the informal observations that attention has no effect on appearance even though it produces a large effect on salience as determined by the direction of apparent motion of ambiguous stimuli.

In addition to color and texture, which were investigated in the experiment described here, previous studies of third-order motion showed sensitivity to depth and to texture orientation, so these inputs to the salience map are also represented in figure 10.13B. And surely there will be others as well.

The salience map has outputs that control detection, object perception, access to memory, and other perceptual and cognitive processes. For example, the salience map is assumed to generate the temporal attention window, described in the first part of this chapter, that controls access to short-term visual memory.

Two Kinds of Amplification in Attention Processing One kind of amplification is the modulation of inputs to the salience map. The other kind is the actual implementation of salience. For example, in controlling memory access, the salience map may determine what input information is to be stored; the actual control of access is a different process, and it is useful to maintain the distinction.

Salience Theories In the present model, the salience map has a privileged place in the processing hierarchy. A relatively small difference in the input to the salience map determines the figure–ground relations in the map, and these are assumed to ultimately determine the flow of information to other perceptual processes. In the neural network model of Koch and Ullman (1985), the salience map determined in which parts of the visual field features from other stimulus maps (e.g., color and shape maps) could be combined, and it embodied some of the computations envisioned here. At that time, it had not occurred to the authors that one might make direct measurements of salience.

In the neural network model of Tsotsos and colleagues (chapter 14 in this volume), control inputs modulate perceptual processing at various levels of a visual hierarchy. Once a particular area of the visual field is selected for further processing, the entire cone of information in the visual hierarchy that derives from the selected area is amplified relative to everything else. In this scheme, there is no privileged salience map per se; rather, salience is distributed throughout the visual hierarchy. Both of these models, as well as others that have been proposed, could be elaborated to take into account the experimental evidence and theoretical considerations reported here.

10.3.5 Salience Map: Applications to Other Paradigms

Third-Order Motion The basis of the experiments at issue is that the output of the salience map can serve as an input to a third-order motion flow field. One of the useful features of the apparent motion paradigm is that it takes advantage of an amplification principle: the strength of apparent motion in a sequence of frames in which there is a spatial 90° phase shift from frame to frame is proportional to the *product* of the modulation amplitudes in each frame (van Santen and Sperling, 1984). Introducing high-contrast-texture stripes in a background of zero texture renders the textured regions highly salient. Introducing such high-amplitude salience modulation in the even frames of the Blaser et al. (1999) attention experiment enabled very sensitive measurement of attention-induced salience modulations that otherwise might have remained below threshold (Lu and Sperling, 1999). In the present case, the product amplification principle was applied to measuring attention-induced salience modulations in the red–green gratings of the odd frames. The same amplification principle in apparent motion offers the possibility of efficient and sensitive measurements of salience in other contexts, such as visual search and short-term memory.

See Color Plate

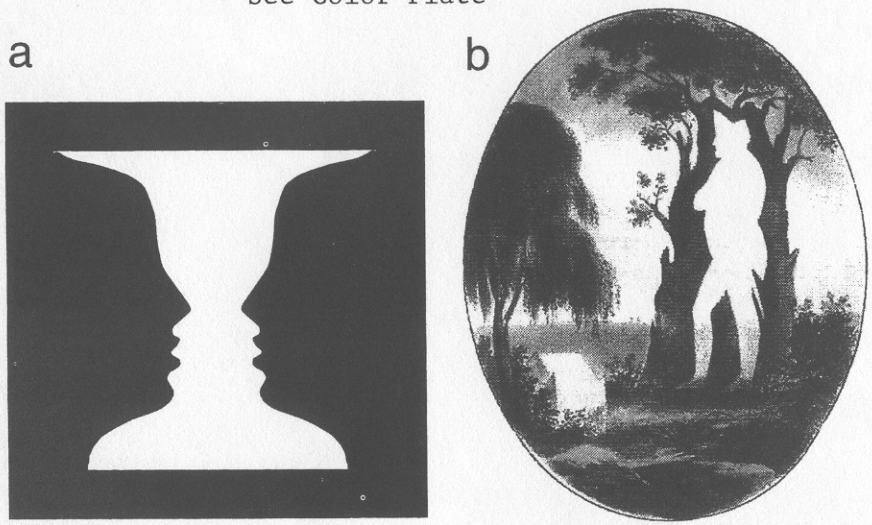


Figure 10.14
Figure–ground ambiguities. (*a*) Ambiguous profiles-vase, after Rubin (1915). (*b*) Forest scene with Napoleon. Normally, trees are seen as figure and the intervening space as ground. However, the intervening space can also be seen as figure when it is attended or has a meaningful shape. (Y&B Associates, after Currier and Ives, ca. 1835.) (See plate 9 for color version.)

Figure–Ground and Pattern Recognition Much has been written about figure–ground segregation, much of it inspired by Rubin's (1915) famous illustration of a perceptually bistable vase—pair of face profiles (see figure 10.14, plate 9). Following his example, most of the literature has focused on the experiential nature of distinction between figure and ground. Figure is seen as more important than ground, it seems nearer in depth, boundaries seem to belong to the figure, and so on.

Lu and Sperling (1995a) suggested that the salience map is the mechanism that determines what parts of the visual field are sent to shape-recognition processes. For example, when mapmakers produce maps of the continents, they use little graphic devices that cause the continents to be seen as figure and the oceans as ground. The continents have lots of details; the oceans are plain. The continents have varieties of colors and features; oceans are homogeneous in color. Consequently, in the United States, most persons feel they know the shapes of North and South America, but very few know or would recognize the shapes of the oceans. Maps designed for sailing and oceanography practice just the opposite principle, keeping land areas very plain and putting the detail and livelier colors in the ocean.

Mapmakers take advantage of bottom-up salience processes. However, top-down processes also have a strong influence on salience. A good example is a forest scene. Normally, the trees are perceived as figure, and the space between the trees as ground. That is, the shape system computes the shape of the trees and not of the spaces between the trees. However, when running away from something, it becomes essential to compute the shape of the space between the trees: Will we fit? Will what is chasing us fit? The salience map interpretation of this process is that there is top-down enhancement in the middle of the space between the trees. This enhancement needs only to be sufficiently precise to cause the salience map to mark the space between the trees as figure and, consequently, for the shape system to compute its shape. A nice example of computing the shape between the trees is illustrated in figure 10.14B.

The example of "Napoleon in the trees" (figure 10.14B) embodies a well-documented principle of figure–ground segmentation: familiar shapes are more likely to be perceived as figure. This in turn suggests a top-down influence of figure–ground segmentation, which is more complex than anything considered in the present salience model but is the kind of vertical interaction in the processing hierarchy encompassed by the model of Tsotsos and colleagues (chapter 14 in this volume).

Guided Search Perhaps most work on attention theory has been undertaken in the context of visual search tasks. In these tasks, an observer views an array of items comprising one or more targets and several distractors (nontargets). The search process is assumed to control the access of the to-be-searched items to pattern recognition processes either serially or in parallel, or in some more complex combination of both. Theories of search involve the strategic allocation of processing resources (Cave and Wolfe, 1990; Koopman, 1957; Sperling and Dosher, 1986). The sequence of items searched is determined by priorities that are assigned to spatial locations, to features, and to other stimulus properties that discriminate between stimuli and targets. Automatic, bottom-up factors are very important in locating targets that differ greatly from their surround (Cave and Wolfe, 1990); top-down factors may be equally important, such as the known probability of finding targets in particular locations. Because it combines both bottom-up and top-down influences, the salience map would provide an ideal mechanism to implement this kind of guided search.

Access to Memory Of all the processes discussed, access to memory is the most restrictive. The partial report paradigm (Sperling, 1960) offers a simple example. Observers were briefly exposed to 3×3 or 3×4 arrays of letters and asked to report just one,

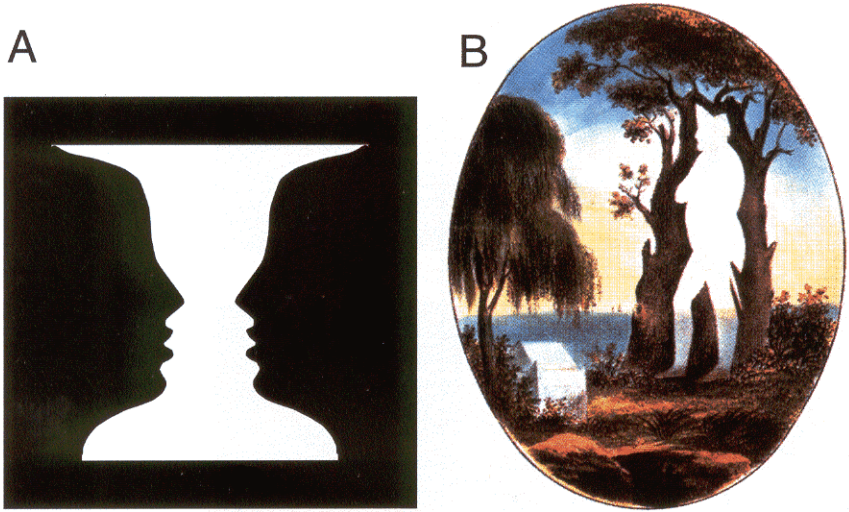


Plate 9 Figure-ground ambiguities. (*a*) Ambiguous profiles-vase, after Rubin (1915). (*b*) Forest scene with Napoleon. Normally, trees are seen as figure and the intervening space as ground. However, the intervening space can also be seen as figure when it is attended or has a meaningful shape. (Y&B Associates, after Currier & Ives, ca. 1835.) See chapter 10.

(randomly) selected row. The cue to report a row was coded as a high-, medium-, or low-pitched tone, so that it could be interpreted very quickly. The cued row was reported quite accurately, even when the instruction occurred several hundred milliseconds after the exposure was terminated.

Performance is nicely accounted for by a model that assumes there is an initial, default state of attending to the middle row, and that there is a quick transition to the row indicated by the tone when it occurs (Gegenfurtner and Sperling, 1993). The salience map provides an obvious mechanism for this spatial shift of attention, which controls access of visual input to visual short-term memory. It is quite analogous to the temporal attention window that was the subject of the first part of this chapter, and to the attentional amplification of color that was the subject of the second part. Just like a cue to attend to a particular color, which must be interpreted at a higher, cognitive level but takes effect at a much lower, perceptual level, a tonal cue to attend to a particular region in space also ultimately takes effect at a lower level to control inputs to the salience map.

Constraints on Top-Down Control of Salience With the eyes fixated, observers nevertheless can attend selectively to areas of visual space according to attention instructions. This is well known and has been amply confirmed. What are the constraints on the shape of the area to which observers can attend? The attention-modulation functions obtained from experiments on attention to motion provide one means of answering this question. Suppose that attention-modulation functions are determined primarily by limitations of the salience map, rather than by the specific stimuli used in our experiments. In this case, the spatial frequency filter functions of the model of Blaser and colleagues (1999) would describe the attentional constraints. That is, any request to distribute attention according to a particular spatial function, could be executed only to the level of accuracy permitted by the attentional filters. Whether the attention system could actually achieve this resolution limit is an empirical question.

A related question concerns the concurrent action of attention to color and attention to space. The attention system would be much simpler if there were two separate, independent attention processes—one allocating attention to a particular color in all of visual space, and the other allocating attention to a particular part of visual space. But this would imply that attention to location and to color are separable. One could attend to red in a certain location, but one could not attend to red in one location and to green in another. Results of preliminary experiments by Tse, Lu, and Sperling (2000) suggest that this is indeed true, implying that attention to color and to location are indeed separable. Obviously, such constraints and their dynamics need to be embodied in more detailed attention models than the ones proposed here.

10.4 Summary and Conclusions

Two models of attention have been proposed, each accounting for a significant set of experimental data and for important incidental observations. The first model shows how attention windows are constructed in successive attention episodes and how such attention windows control access to short-term memory. Once they are in memory, items acquired within a single attention episode lose their time stamp, and their order is coded simply in terms of their memory strength. It takes about 100 to 200 ms to self-generate an attention window in response to an attention cue, and the window width is several hundred milliseconds.

The second model describes the salience map, one of the most important mechanisms by which attention exerts its effects. This model is derived from experiments using a sensitive assay method involving third-order motion. Attention was found to amplify the salience of attended colors by, typically, about 30%. The model draws an important distinction between attentionally amplifying the salience of an attended color while leaving the appearance of the color itself unchanged. The salience map was proposed as the probable mechanism for a variety of tasks, including access to short-term memory, guided search, and pattern perception mechanisms. Both models made accurate and efficient predictions of significant data sets.

Acknowledgment

This research was supported by AFOSR, Life Science Directorate, Visual Information Processing Program.

Notes

1. In fact, pairs in which only one member of the pair is reported were included in the analysis because it was assumed that the other member would eventually have been reported if the response had not been artificially truncated after four reported numerals. Except for having more data to analyze by including partnerless items in an implicit pairing, there was no difference in any comparison or conclusion that depended on including or not including single-item pairs.

2. There also are measurement-theoretic inequalities involving *iBj*s that prove the data can be described by a strength model (see Reeves and Sperling, 1986: p. 189ff.).

3. In a later, more detailed model, the time to detect and interpret the attention-shift cue is represented not merely by the mean detection time but also by a distribution with a mean and variance. The same variance accounts for the variance of motor reaction times and for internal correlations in attention-shift data (Sperling and Shih, 1998).

4. This kind of attention hardware has some interesting difficulties in making accurate judgments of intermodal temporal order, like those baseball umpires attempt to make when judging the order of occurrence of a runner's foot touching a specific point and the sound of a baseball striking the fielder's glove.

5. How the mean luminance l_0 is computed and what constitutes the spatiotemporal neighborhood in which it is computed are complex questions that are of considerable interest in deriving an accurate theory of visual processing, but they are secondary to issues of attention. To simplify the estimation of mean luminance in the experiments described here, the stimuli were constructed so that the expected luminance was locally and globally the same everywhere in every frame.

6. Complications: (1) First-order vision is organized into channels, computations that are carried out within a particular spatial frequency band, typically one to two octaves wide. First-order motion is computed in all channels (spatial frequency bands), and each has a flow field. How these channel outputs are ultimately combined has not yet been resolved. (2) The Reichardt model—as well as all other equivalent or nearly equivalent models of first-order motion—computes only direction, not velocity directly. There are two proposed classes of velocity theories (temporal frequency counting and detector combination), but the brain's algorithm for computing velocity has not been determined.

7. Cavanagh (1992) presented observers with two superimposed gratings, a first-order (luminance) grating and an isoluminant color grating, moving in opposite directions. Observers perceived motion in the first-order direction. However, selective attention to an area of the colored grating could produce apparent motion consistent with the color-grating direction. The apparent motion of the color grating was assumed to be produced by the movement of attention in the process of tracking the moving area. The displays produced by Lu and Sperling (1995a) were much too quick to permit attentional tracking. Third-order motion is a more primitive process than attentional tracking, perhaps even a necessary precursor to the attentional tracking of moving objects.

8. A third observer, who was able to complete only half the experiment, had a red amplification factor α_r of 1.17, which indicates that selective attention to red more than doubled her red salience.

References

Adelson, E. H., and Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *J. Optical Soc. America* A2: 284–299.

Adelson, E. H., and Bergen, J. R. (1986). The extraction of spatio-temporal energy in human and machine vision. In *Motion: Representation and analysis* (pp. 151–155). Washington, DC: IEEE Computer Society Press.

Ahmad, S., and Omohundro, S. (1991). *Efficient visual search: A connectionist solution.* International Computer Science Institute Technical Report tr-91-040. Berkeley: University of California.

Anstis, S., and Cavanagh, P. (1983). A minimum motion technique for judging equiluminance. In J. D. Mollon and E. T. Sharpe (eds.), *Colour vision* (pp. 155–166). New York: Academic Press.

Blaser, E., Sperling, G., and Lu, Z.-L. (1999). Measuring the amplification of attention. *Proc. Nat. Acad. Sci. USA* 96: 8289–8294.

Burt, P. (1988). Attention mechanisms for vision in a dynamic world. In *Proceedings ninth international conference on pattern recognition, Beijing, China* (pp. 977–987).

Cavanagh, P. (1992). Attention-based motion perception. *Science* 257: 1563–1565.

Cave, K. R., and Wolfe, J. M. (1990). Modeling the role of parallel processing in visual search. *Cog. Psychol.* 22: 225–271.

Cheal, M. L., and Lyon, D. (1989). Attention effects on form discrimination at different eccentricities. *Q. J. Exp. Psychol.* 41A: 719–746.

Chubb, C., and Sperling, G. (1988). Drift-balanced random stimuli: A general basis for studying non–Fourier motion perception. *J. Optical Soc. America* A5: 1986–2006.

Chubb, C., and Sperling, G. (1989a). Second-order motion perception: Space–time separable mechanisms. In *Proceedings: Workshop on visual motion. (March 20–22, 1989, Irvine, California.)* (pp. 126–138). Washington, DC: IEEE Computer Society Press.

Chubb, C., and Sperling, G. (1989b). Two motion perception mechanisms revealed by distance driven reversal of apparent motion. *Proc. Nat. Acad. Sci. USA* 86: 2985–2989.

Chubb, C., and Sperling, G. (1991). Texture quilts: Basic tools for studying motion-from-texture. *J. Math. Psychol.* 35: 411–442.

Dosher, B. A., Landy, M. S., and Sperling, G. (1989). Kinetic depth effect and optic flow: I. 3D shape from Fourier motion. *Vis. Res.* 29: 1789–1813.

Gegenfurtner, K., and Sperling, G. (1993). Information transfer in iconic memory experiments. *J. Exp. Psychol. Hum. Percep. Perf.* 19: 845–866.

Koch, C., and Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Hum. Neurobiol.* 4: 219–227.

Koopman, B. O. (1957). The theory of search. III. The optimum distribution of searching effort. *Oper. Res.* 5: 613–626.

Krauskopf, J., and Li, X. (1999). Effect of contrast on detection of motion of chromatic and luminance targets: Retina-relative and object-relative movement. *Vis. Res.* 39: 3346–3350.

Kuffler, S. W. (1953). Discharge pattern and functional organization of mammalian retina. *J. Neurophysiol.* 16: 37–68.

Lu, Z.-L., Lesmes, L. A., and Sperling, G. (1999). The mechanism of isoluminant chromatic motion perception. *Proc. Nat. Acad. Sci. USA* 96: 8289–8294.

Lu, Z.-L., and Sperling, G. (1995a). Attention-generated apparent motion. *Nature* 377: 237–239.

Lu, Z.-L., and Sperling, G. (1995b). The functional architecture of human visual motion perception. *Vis. Res.* 35: 2697–2722.

Lu, Z.-L., and Sperling, G. (1999). The amplification principle in motion perception. *Invest. Ophthalmol. Vis. Sci.* 40: S199.

Mozer, M. (1991). *The perception of multiple objects: A connectionist approach.* Cambridge, MA: MIT Press.

Prinzmetal, W., Amiri, H., Allen, K., and Edwards, T. (1998). Phenomenology of attention: I. Color, location, orientation, and spatial frequency. *J. Exp. Psychol. Hum. Percep. Perf.* 24: 261–282.

Reeves, A. (1977). The detection and recall of rapidly displayed letters and digits. Unpublished doctoral dissertation, City University of New York.

Reichardt, W. (1961). Autocorrelation, a principle for the evaluation of sensory information by the central nervous system. In W. A. Rosenblith (ed.), *Sensory communication.* New York: Wiley).

Rubin, E. (1915). *Edgar Rubin synsoplevede figurer: Studien i psykologisk analyse.* Copenhagen: Gyldendalske Boghandel. German trans.: *Visuell wahrgenommene Figuren: Studien in psychologischer Analyse.* Copenhagen: Gyldendalske Boghandel, 1921.

Shih, S., and Sperling, G. (2000). Measuring and modeling the trajectory of visual spatial attention. *Psychol. Rev.,* in press.

Solomon, J. A., and Sperling, G. (1994). Full-wave and half-wave rectification in 2nd-order motion perception. *Vis. Res.* 34: 2239–2257.

Sperling, G. (1960). The information available in brief visual presentations. *Psychol. Monog.* 74, no. 11 (whole no. 498). Pp. 1–29.

Sperling, G., and Dosher, B. (1986). Strategy and optimization in human information processing. In K. Boff, L. Kaufman, and J. Thomas (eds.), *Handbook of perception and performance,* vol. 1, Chapter 2 (pp. 1–65). New York: Wiley.

Sperling, G., and Lu, Z.-L. (1998). A systems analysis of visual motion perception. In T. Watanabe (ed.), *High-level motion processing* (pp. 153–183). Cambridge, MA: MIT Press.

Sperling, G., and Weichselgartner, E. (1995). Episodic theory of the dynamics of spatial attention. *Psychol. Rev.* 102: 503–532.

Sperling, G., and Reeves, A. (1980). Measuring the reaction time of a shift of visual attention. In R. Nickerson (ed.), *Attention and Performance VIII* (pp. 347–360). Hillsdale, NJ: Erlbaum.

Tse, C.-H., Lu, Z.-L., and Sperling, G. (2000). Attending to red and green concurrently in different areas reduces attentional capacity. *Invest. Ophthalmol. Vis. Sci.* 41: S42.

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N., and Nuflo, F. (1995). Modeling visual attention via selective tuning. *Art. Intell.* 78: 507–545.

van Santen, J. P. H., and Sperling, G. (1984). Temporal covariance model of human motion perception. *J. Optical Soc. America* A1: 451–473.

van Santen, J. P. H., and Sperling, G. (1985). Elaborated Reichardt detectors. *J. Optical Soc. America* A2: 300–321.

Watson, A. B., and Ahumada, A. J. (1985). Model of human visual-motion sensing. *J. Optic. Soc. Am. A* 1: 322–342.

Werkhoven, P., Sperling, G., and Chubb, C. (1993). The dimensionality of texture-defined motion: A single channel theory. *Vis. Res.* 33: 463–485.

Werkhoven, P., Sperling, G., and Chubb, C. (1994). Perception of apparent motion between dissimilar gratings: Spatiotemporal properties. *Vis. Res.* 34: 2741–2759.