BEHAVIORAL AND BRAIN SCIENCES (1990) 13, 423–469
Printed in the United States of America

Analyzing vision at the complexity level

John K. Tsotsos
Department of Computer Science, University of Toronto and The Canadian Institute for Advanced Research, 10 King's College Rd., Toronto, Ontario, Canada M5S 1A4
Electronic mail: *tsotsos@ai.toronto.edu*

Abstract: The general problem of visual search can be shown to be computationally intractable in a formal, complexity-theoretic sense, yet visual search is extensively involved in everyday perception, and biological systems manage to perform it remarkably well. Complexity level analysis may resolve this contradiction. Visual search can be reshaped into tractability through approximations and by optimizing the resources devoted to visual processing. Architectural constraints can be derived using the minimum cost principle to rule out a large class of potential solutions. The evidence speaks strongly against bottom-up approaches to vision. In particular, the constraints suggest an attentional mechanism that exploits knowledge of the specific problem being solved. This analysis of visual search performance in terms of attentional influences on visual information processing and complexity satisfaction allows a large body of neurophysiological and psychological evidence to be tied together.

Keywords: attention; complexity; computation; cortex; matching; search; vision.

1. Introduction

The research described in this target article conforms to the computational paradigm for modeling biological vision. This paradigm has been interpreted in rather limited ways in the past decade, however. Computation seems to be associated with mathematical models that are (perhaps) simulated on a computer. Although this association is certainly within the computational paradigm, the set of available computational tools is much larger than the class of continuous mathematics. One such tool is complexity theory, which is concerned with the amount of computation required to solve a given problem and the number of elements (processors, connections, memory, and so forth) needed for its computation. The formal theory is concerned with the inherent difficulty of computation; complexity level analysis tries to match a proposed solution to a prespecified set of resources.

It is a natural consequence of the computational paradigm that all tools of computation should be considered and that their power should be brought to bear on the problem of vision. Complexity theory should reveal basic insights into the structure and performance of human vision; if successful, its effect on theories of visual perception would be great. It could delimit the space of permissible solutions in a formal and theoretical fashion. All theories and models would then have to fit within these theoretical limits or be clearly subject to the criticism that they were unrealizable.

1.1. Motivation and goals

The general task of visual search[1] will be shown to be inherently intractable in the formal sense. Given the ubiquity of visual search tasks in everyday perception, it may be true that visual perception in general is likewise intractable. Yet human vision is effortless and exquisitely precise. How can this be? Ancient philosophers were aware that humans could attend to the relevant and ignore the irrelevant. [See Näätänen: "The Role of Attention in Auditory Information Processing as Revealed by Event-related Potentials and Other Brain Measures of Cognitive Function" *BBS* 13(2) 1990.] More recently, psychologists have studied attention and proposed that some kind of processing limit in the brain must give rise to such a phenomenon. Neisser, for example, claimed that any model of vision that was based on spatial parallelism alone was doomed to failure because the brain was not large enough (Neisser 1967). This led him to his two-stage process of perception: a preattentive phase followed by an attentive phase. It is difficult to formulate such a model in computational terms, however; there are so many missing details. Moreover, the explanation of the need for attention is less than satisfactory. The idea that the brain is not large enough does not yield any useful constraints on the architecture of the visual system. Arguments to the effect that a given fixed resource is not large enough to accommodate a specified problem lead naturally to considerations of computational complexity. Neisser hinted at the need to address the difficult issues of computational complexity, even though in 1967 complexity theory was barely in its infancy.

Complexity considerations are commonplace in the computational vision literature. Many researchers (for example, Grimson 1986; Mackworth & Freuder 1985; Poggio 1982; and others) routinely provide an analysis of the complexity of their proposed algorithms – this is simply good computer science. It is important to demon-

0140-525X/90 $5.00+.00

strate that specific algorithms have tractable requirements in terms of computer size and execution time, but this is not the same as addressing the complexity issues for vision in general. In the realm of interdisciplinary theories, Feldman and Ballard (1982) concluded that complexity considerations about timing suggest that massively parallel models are the only biologically plausible ones, because only they satisfy the "100 step rule." That is, because most neurons compute at a maximum rate of about 1000 Hz, and because simple perceptual phenomena occur in about 100 milliseconds, biologically plausible algorithms can require no more than 100 steps. Feldman and Ballard did not explain exactly how "massive" these networks must be, however (also, see Zucker 1985). They also stressed the importance of conserving connections. Although their emphasis was correct, their application of this constraint leaves many questions unanswered and, in particular, Feldman and Ballard did not demonstrate that their set of conserving techniques was sufficient given the resources of the brain. Rumelhart and McClelland (1986a) claim that the time and space requirements of a theory of cognitive function are important determinants of the theory's biological plausibility. However, they do not provide a detailed analysis of how such constraints may be satisfied. A number of other papers (Pylyshyn 1984; Uhr 1980; and others) touch on the complexity issue and do not attack it head on; they do provide support for the utility of complexity considerations, however. Uhr, for example, gives the 100-step argument of Feldman and Ballard almost exactly, except that he uses 400 steps.

Serious consideration must be given to computational complexity in the computational modeling of perception and, indeed, in the computational modeling of any aspect of intelligence. One of the key problems with artificial intelligence is that the solutions proposed are fragile with respect to the question of "scaling up" with problem size: Theoretical solutions are usually derived without regard to the amount of computation required and then if an implementation is produced, it is tried out only on a few small examples. The standard claim is that if faster or parallel hardware were available, a real-time solution would be obtained. There is something very unsatisfying about this type of claim. In particular, parallel solutions, such as those proposed by the connectionist community, although motivated by complexity considerations, typically fail to demonstrate the computational sufficiency of their approaches (see the collections of papers on the subject in Rumelhart & McClelland 1986b and Feldman 1985). For example, few if any deal with the time and space requirements of the relaxation procedures they use, particularly in the context of time-varying input (but see Tsotsos 1987a for empirical results on this). Complexity analysis was not applied to determining the limits of the information processing capacity of the visual system in van Doorn, van de Grind, and Koenderink's (1984) comprehensive overview of techniques and approaches for studying the limits of perception. If one is committed to realizing systems and proving that they behave in the required manner, the first prerequisite would seem to be that the candidate system be computationally tractable.

The problem vision researchers from the many relevant disciplines face is that experiment results and explanatory theories from these disparate fields are not immediately compatible, and often appear contradictory. There has been very little work on "the big picture" which the individual results may fit.[2] Current theories are hence open to a kind of criticism that is, in one important sense, unfair at this stage of our knowledge. There is no test that can be applied to a theory to determine whether or not basic considerations are satisfied. Satisfying complexity constraints is one test that new theories of visual perception must pass.

The key principle underlying the research described here is that considerations about the computational complexity of the perceptual task are critical and lead directly to "hard" constraints on the architecture of visual systems, both biological and computational. It is surprising that Marr did not even mention the problem of computational complexity as part of the computational level of his theory (Marr 1982). According to Marr, the computational level of his theory addresses the questions: What is the goal of the computation? Why is it appropriate? What is the logic of the strategy by which it can be carried out? Marr called solutions at this level "in principle" solutions. At the representational and algorithmic level one asks: How can this computational theory be implemented? What is the representation for the input and output? What is the algorithm for the transformation? And, finally, at the implementational level one asks: How can the representation and algorithm be realized physically? [See Anderson: "Methodologies for Studying Human Knowledge" *BBS* 10(3) 1987.] Complexity spans these three levels. Yet considerations of efficiency or complexity are not just implementational details as Marr implies. If the task to be performed or the algorithm to be implemented is tractable, then perhaps efficiency is only an implementational detail. However, if the task is an intractable one, as vision in its most general form seems to be, complexity satisfaction is not simply a detail to contend with during implementation, just as discretization and sampling effects or numerical stability are not simply implementational details. Complexity satisfaction is a major constraint on the possible solutions of the problem. It can distinguish between solutions that are realizable and those that are not.

It is important to specify exactly what is meant by complexity level analysis: Given a task, a set of performance specifications, a fixed amount of input, and a fixed set of resources with which to accomplish the task, two related questions can be asked. The first is, "How much computation is required to accomplish the task?"; the second is, "Are the given resources sufficient to accomplish the task?" In general, the resources specified in a problem description do not necessarily match the required computation; there could be a mismatch between problem complexity (the answer to the first question) and the resources. This does not mean that no realization is possible – it means that further analysis is required to reshape the task or to optimize the resources so as to attain a satisfactory match. Note that reshaping the problem often means making approximations or being content with suboptimal solutions – aspects of the full generality of the problem must be sacrificed to obtain a realizable solution. This process of optimization toward matching the computational requirements of a problem with a given resource I call "analysis at the complexity level." The result of the analysis will show how much computation will actually be performed, what the nature of the actual problem solved is, and what the first-order performance characteristics of the realization are. This analysis will not provide answers to "how" questions – how the computation is actually carried out. Nevertheless, ascertaining how much computation can be performed will strongly constrain which computations are chosen to actually solve the problem.

As with many aspects of science, this analysis points to an iterative methodology. This target article will deal only with first-order complexity, analogous to building a house starting with the internal wood frame. Once the frame defines the skeleton of the house, one can begin to add detail. So too with this analysis; fine-scale considerations are not dealt with. When a problem is inherently intractable, one must reduce the intractability at the large scale before worrying about detailed considerations at finer scales. The process of design does not depend on only one type of building material but on many. We will consider only two types of material: complexity satisfaction and minimization of cost. The constraints we derive will be termed "sufficient" in one sense only: They are sufficient to satisfy the first-order complexity level analysis (discussed later). It must also be noted that the constraints are not formal necessary conditions.

An engineer is provided with a set of design specifications a new apparatus must meet. We are faced with an inverse problem and much more difficult one: discovering the specifications and design principles of an existing system whose performance and composition is still far from being understood. This lack of understanding would appear to make such an inverse analysis impossible. There are some elements of biological visual systems, however, that are better understood than others. For example, it seems well accepted that the average connectivities among neurons is about 1000 for both fan-out and fan-in, that there are 30 or so visual areas (van Essen & Anderson 1989), that there is a hierarchical organization connecting these areas (van Essen & Maunsell 1983), that responses of individual neurons may be affected by a spectrum of stimuli rather than a single one (Zeki 1978; Maunsell & Newsome 1987), and that the performance of a given neuron may be profoundly affected by attentional influences related to the task at hand (Moran & Desimone 1985; see also the papers by Allman et al. 1985; Crick & Asunama 1986; Desimone et al. 1985; van Essen et al. 1984 for more detail about the neuroanatomy and neurophysiology of the visual cortex in primates).

1.2. Visual search

Our analysis will concentrate on one important aspect of vision. Consider the experimental paradigm for *visual search:* Given a set of memory items or targets and a test display that contains several nontarget items and may or may not contain targets, measure the length of time a subject needs to detect a given number of the targets in the display. There are special cases of visual search. One can vary the number of elements in the test display up to the limit where there are only image textures rather than discrete items; one can vary the number of items in the memory set or the number of items that must be found in the test display. In each variant, however, matching one memory item to a test display seems to be a basic subproblem. This may be a basic subproblem of all visual tasks.

Visual search experiments measure the response times of subjects in recognition or detection tasks. These are assumed to reflect the amounts and kinds of visual information processing leading to the response. The connection of computational complexity analysis seems rather direct. One theory of visual search performance has arisen from the substantial body of work assembled by Treisman over the past decade. Treisman and colleagues have defined a framework they term "feature integration theory" (Treisman & Gelade 1980; Treisman & Schmidt 1982; Treisman 1985; Treisman & Souther 1985; Treisman & Gormican 1988; Treisman 1988; Treisman & Sato 1990). Two main categories of stimuli are used: disjunctive and conjunctive displays. In a disjunctive display, the target is identified by only one feature, such as color, whereas in a conjunctive display, the target is defined by more than one feature, such as color and orientation. A typical disjunctive display could be a field of blue vertical lines, with an embedded target consisting of a blue horizontal line. In these displays the target is found immediately and effortlessly; it "pops out." Response time is constant and independent of the number of display items. A conjunctive display would be a field of randomly selected colored letters where the target was, say, a red letter "A." Treisman claims that attention must be directed serially to each stimulus in the display whenever conjunctions of more than one feature are needed to correctly characterize or distinguish the possible objects presented. Response time is observed to be linear in the number of display items. If the target is not provided to subjects in advance and they are required to determine which stimulus item is the target, a third behavior results. The odd-man-out in the display is the target sought. This type of response seems to involve a longer and more difficult search because subjects search through possible combinations of features shared by subsets of display items.

If visual search is considered from a computational viewpoint, the following questions arise:

1. What is the inherent computational complexity of visual search?
2. How can algorithms for visual search be realized so that they fit into human brains?
3. What are the characteristics of those algorithms?
4. Could the differences in computational complexity among the three types of experiments described account for the resulting three distinct behaviors?
5. Can the infinite space of possible architectures for vision be bounded on the basis of complexity considerations and the size of the brain, and still solve the problem?
6. Could complexity satisfaction be one of the reasons visual processing structures evolved into their current form?

On the assumption that *BBS* readers and commentators will be familiar with the biological aspects of perception rather than the computational ones, the review of background material will focus on computational complexity theory.

1.3. Overview of complexity theory

Computational complexity is studied to determine the intrinsic difficulty of mathematically posed problems that arise in many disciplines.[3] Many of these problems involve combinatorial search, i.e., search through a finite but extremely large, structured set of possible solutions. Examples include the placement and interconnection of components on an integrated circuit chip, the scheduling of major league sports events, or bus routing. Any problem that involves combinatorial search may require huge search spaces to be examined; this is the well-known combinatorial explosion phenomenon. Complexity theory tries to discover the limitations and possibilities inherent in a problem rather than what usually occurs in practice. After all, the worst case does occur in practice as well. This approach to the problem of search diverges from that of the psychologist, physicist, or engineer. In the same way that the laws of thermodynamics provide theoretical limits on the utility and function of nuclear power plants, complexity theory provides theoretical limits on information processing systems. If biological vision can indeed be computationally modeled, then complexity theory is a natural tool for investigating the information processing characteristics of both computational and biological vision systems.

For a given computational problem C, how well, or at what cost, can it be solved?

1. Are there efficient algorithms for C?
2. Can lower bounds be found for the inherent complexity of C?
3. Are there exact solutions for C?
4. What algorithms yield approximate solutions for C?
5. What is the worst-case complexity of C?
6. What is the average complexity of C?

Before studying complexity one must define an appropriate complexity measure. Several measures are possible, but the common ones are related to the space requirements (numbers of memory or processor elements) and time requirements (how long it takes to execute) for solving a problem. Complexity measures in general deal with the cost of achieving solutions.

The study of complexity has led to more efficient algorithms than those previously known or suspected. Perhaps the most important use of complexity theory is illustrated by the following (adapted from Garey & Johnson 1979): Suppose you were assigned the task of designing and implementing a new piece of software. Your job is to construct a design that meets the specifications of this program. After months of work, you are unable to come up with any design that does substantially better than searching through all the possible options, but this would involve years of computation time, and is thus totally impractical. Under these circumstances it may be possible to prove that no efficient algorithm is possible – the problem is inherently intractable – and hence the specifications of the problem should be changed.

Complexity theory begins with a 1937 paper in which the British mathematician Alan Turing introduced his well-known Turing machine, providing a formalization of the notion of an algorithmically computable function. He postulated that any algorithm could be executed by a machine with an infinitely long paper tape, divided into squares, a printer that writes and erases marks on the tape, and a scanner that senses whether or not a given square is marked. This imaginary device can be programmed to find the solution to a problem by executing a finite number of scanning and printing operations. What is remarkable about the Turing machine is that in spite of its simplicity, it is not exceeded in problem-solving ability by any other known computing device. If the Turing machine is given enough time, it can in principle solve any problem that the most sophisticated computer can solve, regardless of serial/parallel distinctions or any other type of ingenious design. As a result, the fact that a problem can be solved by a Turing machine has been accepted as a necessary and sufficient condition for the solvability of the problem by algorithm. This thesis[4] states that any problem for which we can find an algorithm that can be programmed in any programming language running on any computer, even if unbounded time and space are required, can be solved by a Turing machine.

The Church/Turing thesis also led to impossibility proofs for computers. Turing proved that the problem of logical satisfiability – for a given arbitrary formula in predicate calculus, is there an assignment of truth values of its variables such that the formula is true? – cannot be decided by any algorithm in a finite number of steps. This provided the basis for other similar proofs of intractability. Once one could prove problems were inherently intractable, it was natural to ask about the difficulty of an arbitrary problem and to rank problems in terms of difficulty.

Certain intrinsic properties of the universe will always limit the size and speed of computers. Consider the following argument from Stockmeyer and Chandra (1979): The most powerful computer that could conceivably be built could not be larger than the known universe (less than 100 billion light-years in diameter), could not consist of hardware smaller than the proton (10^{-13} cm in diameter), and could not transmit information faster than the speed of light (3×10^8 per second). Given these limitations, such a computer could consist of at most 10^{126} pieces of hardware. It can be proved that, regardless of the ingenuity of its design and the sophistication of its program, this ideal computer would take at least 20 billion years to solve certain mathematical problems that are known to be solvable in principle. Because the universe is probably less than 20 billion years old, it seems safe to say that such problems defy computer analysis.

A more specific example is a well-studied problem in integer mathematics, the Knapsack Problem. In one form, the question is: Given a list of numbers and a "knapsack size," is there a subset of the listed numbers that adds up to the knapsack size? So, for the list of numbers: 4, 7, 13, 18, 25, 32, 42, 49, and a knapsack size of 89, the answer is yes because $4 + 18 + 25 + 42 = 89$. If the knapsack size were 90, the answer would be no. It has been shown that the only possible solution is to search through all possible subsets of numbers in the list and check whether or not they add up to the knapsack size. Given N numbers there are 2^N subsets, so in the worst case, that is, the case in which the subset that gives the right answer is the last one checked, 2^N operations are required. Even the average case would require $\frac{2^N}{2}$ operations, and is still exponential. Using a universe-sized computer as in the illustration earlier, 488 numbers in a knapsack would need more than 20 billion years of computing with 10^{126} computing elements operating in parallel, each requiring 1 millisecond to check one of the subsets. Four hundred and eighty-eight numbers in a knapsack lead to 6.3×10^{146} subsets. Even if the processor were speeded up by, say, 6 orders of magnitude, it would not help substantially. With processors that require 10^{-9} seconds to check each subset, there could only be 508 numbers in the list. Yet, the knapsack problem is clearly solvable in principle. This points to an important emendation of Marr's (1982) view of computational vision namely, that "in principle" solutions are not necessarily realizable and thus are not necessarily acceptable. A necessary condition on their validity is that they must also satisfy the complexity constraints of the problem and the resources allocated to its solution.

1.4. Some basic definitions

The following are some basic definitions in complexity theory (Garey & Johnson 1979). A *problem* is a general question to be answered, usually with several parameters whose values are left unspecified. A problem is described by giving a general description of all of its parameters and a statement of what properties the answer, or *solution* is required to satisfy. An *instance* of the problem is obtained by specifying particular values for the problem parameters. An *algorithm* is a general step-by-step procedure for finding solutions to problems. To *solve* a problem means that an algorithm can be applied to any problem instance and is guaranteed to produce a solution to that instance. The *time requirements* of an algorithm are conveniently expressed in terms of a single variable, N, reflecting the amount of input data needed to describe an instance. A *time complexity* function for an algorithm expresses its time requirements by giving, for each possible input length, an upper bound on the time needed to achieve a solution. If the number of operations required to solve a problem is an exponential function of N, the problem has *exponential time complexity*. If the number of required operations can be represented by a polynomial function in N, the problem has *polynomial time complexity*. Similarly, *space complexity* is defined as a function for an algorithm that expresses its space or memory requirements. *Algorithm complexity* is the cost of a particular algorithm. This should be contrasted with *problem complexity*, which is the minimal cost over all possible algorithms. These two forms of complexity are often confused. The dominant kind of analysis is *worst-case:* at least one instance out of all possible instances has this complexity. Although *average case* analysis may better represent the problems encountered in practice, it tells us little about performance for a particular problem instance. Moreover, the characterization of the average case has proven to be a difficult theoretical task. Worst-case analysis, on the other hand, places bounds on all instances.

The notion of a good algorithm and an intractable problem was developed in the mid-to-late 1960s. A *good* algorithm is one whose time requirements can be expressed as a polynomial function of input length. An *intractable* problem is one whose time requirements are exponential functions of problem length, or in other words, a problem that cannot be solved by any polynomial time algorithm for all instances. Note that the boundary between good and bad problems is not precise. A time complexity of N^{1000} is surely not very practical whereas one of $2^{0.001}$ is perfectly realizable. Yet empirical evidence seems to point to the fact that natural problems simply do not have such running times, and that the distinction is a useful one.

A critical idea in complexity theory is *complexity class* and related to it, *reducibility.* If a problem S is known to be efficiently transformed (or reduced) to a problem Q then the complexity of S cannot be much more than the complexity of Q. *Efficiently reduced* means that the algorithm that performs the transformation has polynomial complexity. The class **P** consists of all those problems that can be solved in polynomial time. If we accept the premise that a computational problem is not tractable unless there is a polynomial-time algorithm to solve it, then all tractable problems belong in **P**.

In addition to the class **P** of tractable problems, there is also a major class of presumably intractable problems. If a problem is in the class **NP**, then there exists a polynomial p(n) such that the problem can be solved by an algorithm having time complexity $O(2^{p(n)})$.[5] A problem is *NP-Complete* if it is in the class *NP*, and it polynomially reduces to an already proven NP-Complete problem. These problems form an equivalence class. Clearly, there must have been a "first" NP-Complete problem. The first such problem was that of "satisfiability" (Cook's 1971 Theorem).[6]

There are hundreds of NP-Complete problems – Knapsack is one of them. If any NP-Complete problem can be solved in polynomial time, then they can all be. Most computer scientists are pessimistic about the possibility that nonexponential algorithms for these problems will ever be found, so proving a problem to be NP-Complete is now regarded as strong evidence that the problem is intrinsically intractable. If an efficient algorithm can be found for any one (and hence all) NP-Complete problems, however, this would be a major intellectual breakthrough with immense practical implications.

What does a computer scientist do when confronted with an NP-Complete problem? A variety of approaches have been taken.

1. Develop an algorithm that is fast enough for small problems but would take too long with larger problems. This approach is often used when the anticipated problems are all small.

2. Develop a fast algorithm that solves a special case of the problem, but does not solve the general problem. This approach is often used when the special case is of practical importance.

3. Develop an algorithm that quickly solves a large proportion of the cases that come up in practice, but in the worst case may run for a long time. This approach is often used when the problems occurring in practice tend to have special features that can be exploited to speed up the computation.

4. For an optimization problem, develop an algorithm that always runs quickly but produces an answer that is not necessarily optimal. Sometimes a worst-case bound can be obtained on how much the answer produced may differ from the optimum, so that a reasonably close answer is assured. This is an area of active research, with suboptimal algorithms for a variety of important problems being developed and analyzed.

5. Use natural parameters to guide the search for approximate algorithms. There are a number of ways a problem can be exponential. Consider the natural parameters of a problem rather than a constructed problem length and first attempt to reduce the exponential effect of the largest valued parameters.

NP-completeness effectively eliminates the possibility of developing a completely satisfactory algorithm. Once a problem is seen to be NP-complete, it is appropriate to direct efforts toward a more achievable goal. In most cases, a direct understanding of the size of the problems of interest and the size of the processing machinery is of tremendous help in determining which are the appropriate approximations.

2. The computational nature of the visual search task

2.1. Complexity and visual search

It seems that the following question has never been asked: What are the computational requirements for experimental paradigms in biological studies of visual perception? In other words, how computationally difficult are the tasks presented to subjects? How many performance differences can be explained simply by considering the relative complexity of the tasks? How can a computational model be defined using experimental results from biology if one does not first understand fully the computational nature of the experiment itself? The experimental measurement of response time for visual search tasks (or other tasks in which the measurement of response time is a primary goal) is clearly connected to the speed of processing as well as the algorithm in the system, which in turn reflects the amount of processing machinery allocated to the task. All of this in turn has a very clear connection with computational complexity. Once one makes the connection between computational and biological studies of vision, these are very natural theoretical questions to ask.

The visual search task has not been defined in computational terms by the psychology community. According to the definition provided by Rabbitt (1978), which is consistent with other versions found in more recent papers (see the collection of papers on attention in Parasuraman & Davies 1984, for example), visual search is a categorization task in which a subject must distinguish between at least two classes of signal: goal signals which must be located and reported and background signals which must be ignored. This definition does not specify how signals are located, or represented, nor how goal and background are distinguished.

2.2. A computational definition of visual search

The general question of visual search is: Given a test image and a target image, is there an instance of the target in the test image? The general version of visual search seeks the subset of the test image that best matches the target; in its full generality it includes the possibility of noisy or partial matches. The problem is viewed as a pure information-processing task, with no assumptions about how the data may be presented or organized. The problem can also be of arbitrary size and may use arbitrary stimulus qualities. This captures some of the aspects of "all possible algorithms" that are required to determine problem complexity.

The question posed by visual search has two variants, one in which the target is explicitly provided in advance (say, as a picture), and another where the target is expressed only implicitly, perhaps by specifying relationships it must have with other stimulus items (say, as a command to find the odd-man-out). In the former case, the explicit knowledge of the target gives bounds in space and in stimulus quality to the search task, whereas in the latter case, no similar bounds are possible. Two key definitions are thus required to specify a computational formalism for visual search: *unbounded visual search*, in which either the target is explicitly unknown in advance or it is somehow not used in the execution of the search; and *bounded visual search*, in which the target is explicitly known in advance in some form that enables explicit bounds to be determined that can be used to limit the search process. These bounds may be in the form of the spatial extent of the target, feature dimensions that are involved, or specific feature values. They may be expressed either visually or verbally. The bounds affect the search process through an attentional mechanism; this is only one specific aspect of the broad notion of attention.

A test image containing an instance of the target is created by translating, rotating, and/or scaling the target and then placing it in the test image. The test image may also contain confounding information such as other items, noise, and occluding objects, or other processes may distort or corrupt the target. The variations in the problem are depicted in the following figures. Figure 1a gives the target and Figure 1b gives a test display that contains six variations of the target. Black forms the "figure," and white is the "ground." In Figure 1b, the items show:

A. the target with noise
B. a partial match made up of two separate shapes
C. a partial match
D. the target with additive occlusion
E. the target with subtractive occlusion
F. a perfect match

The solution to a visual search problem involves solving a subproblem, which we call visual matching. Visual matching and visual search are the same if no items require image rotation or scaling in order to match the target. Visual matching hence considers only the location of the target and its identity; it operates on normalized items in the images. This is a useful abstraction because all rotations and scalings that lead to unique images can be enumerated and searched linearly in the worst case. Although choosing which spatial transform to consider is a difficult problem in its own right, we assume that it is abstracted away. It does not have exponential time or space complexity and thus does not affect the main result. Also, much experimental work exists that does not require item rotations and scalings. An instance of the visual matching problem is specified as follows:

A test image **I**
A target image **T**, modified using a 2D spatial transformation
A difference function **diff(p)** for p ∈ **I**, **diff(p)** ∈ R_p^0 is the set of non-negative real numbers of fixed precision p)

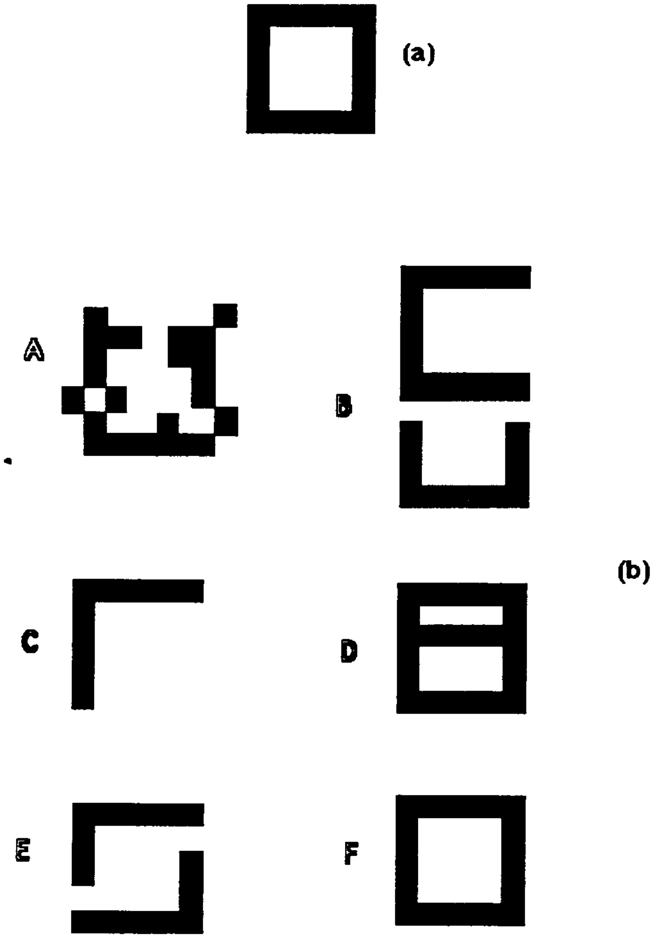


Figure 1. A sample target image (a) and a test image (b) that shows several variations of the target, illustrating the breadth of the matching problem encompassed by visual search. The only perfect match of target to test item is with item F; each of the other matches are partial. In the computational definition of visual search, the computation of the diff and corr functions would yield different values for each of the possible matches of target to test item.

A correlation function **corr(p)** for p ∈ **I**, **corr(p)** ∈ R^0_p
Two thresholds θ, φ, both positive integers

Here is how this particular collection of data can represent the visual match problem:

1. A test image **I** is the set of pixel/measurement quadruples (x,y,j,m_j). x,y specify a location in a Euclidean coordinate system, with a given origin. $\mathbf{M}_i$ is the set of measurement types in the image, such as color, motion, depth, etc., each type coded as a distinct positive integer. m_j is a measurement token of type j, represents scene parameters, and is a nonnegative real number of fixed precision, that is, with positive error due to possible truncation of at most p. (Only a finite number of bits may be stored.) $\mathbf{I}' \subseteq \mathbf{I}$ is a sub-image of **I**, i.e., a subset of quadruples. It is not necessary that all pixel locations contain measurements of all types. It is also unnecessary that the set of pixels be spatially contiguous. **I′** defines an arbitrary subset of pixels. For ease of notation, $i_{x,y,j}$ has value m_j. If j is not an element of $\mathbf{M}_i$ or if the x,y values are outside the image array then $i_{x,y,j} = 0$.

2. A target image **T** is a set of pixel/measurement quadruples defined in the same way as **I**. $\mathbf{M}_t$ is the set of measurement types in the target image. The types correspond between **I** and **T**, i.e., type 3 in one image is the same type in the other. The two sets of measurement types, however, are not necessarily the same. The coordinate system of the target image is the same as for the test image and the origin of the target image coincides with the origin of the test image. $t_{x,y,j}$ has value m_j. If j is not an element of $\mathbf{M}_t$ or if the x,y values are outside the image array $t_{x,y,j} = 0$.

3. The diff function will be the sum of the absolute values of the point-wise differences of the measurements of a subset of the test image with the corresponding subset of the goal image. It is expressed as follows for an arbitrary subset **I′** of the test image:

$$\sum_{p\in I'} \mathrm{diff}(p) = \sum_{p\in I'} \left[\sum_{j\in M_i} |t_{x,y,j} - i_{x,y,j}| \right] \le \theta$$

This sum of differences must be less than a given threshold θ for a match to be potentially acceptable. Note that other functions that minimize some other property may be as suitable. This threshold is a positive integer.

4. Because a null **I′** satisfies any threshold in the above constraint, as does a small subset of background pixels alone, we must enforce the constraint that as many figure matches must be included in **I′** as possible. 2D spatial transforms that do not align the target properly with the test items must also be eliminated because they would lead to many background-to-background matches. One way to do this is to maximize the point-wise product of the target and image. As it turns out, this is also the cross-correlation commonly used in computer vision to measure similarity between a given signal and a template. Therefore,

$$\sum_{p\in I'} \mathrm{corr}(p) = \sum_{p\in I'} \left[\sum_{j\in M_i} t_{x,y,j} \times i_{x,y,j} \right] \ge \phi$$

In the simple figures shown above, using the subsets of the test image that correspond directly to each item, the figure having a value of 1, and the background a value of 0, and the target item translated so that it fits perfectly over each test item, the diff and corr values are:

Test item	diff	corr
A	10	17
B	3	18
C	10	10
D	4	20
E	2	18
F	0	20

Because both constraints must be satisfied, depending on the choice of threshold, the best match is easily found. If the correlation constraint is set to 18 and the difference threshold to 3, items B, E, and F are the only possibilities. Tighter thresholds lead to different possibilities. Note that there is no claim here that the algorithm necessarily corresponds to human performance. The definition is given primarily for purposes of formal proof and is claimed to be a reasonable one. It is possible to provide other functions for difference and correlation and to reconstruct similar proofs using them.

2.3. The complexity of visual matching

The task posed by unbounded visual matching is:

Given a test image, a difference function, and a correlation function, is there a subset of pixels of the test image such that the value of the difference function for that subset is less than a given threshold and such that the value of the correlation exceeds some other threshold? In other words, is there a set $\mathbf{I}' \subseteq \mathbf{I}$ such that it simultaneously satisfies

$$\sum_{p \in I'} \mathrm{diff}(p) \leq \theta \text{ and } \sum_{p \in I'} \mathrm{corr}(p) \geq \phi\ ?$$

One point about this specification of visual matching must be emphasized: It forces a bottom-up approach to the solution of visual matching. The constraints given must be satisfied with subsets of the input image. The target image is neither given nor permitted to provide direction to any aspect of the computation.

This unbounded visual matching problem has exactly the same structure as a known NP-Complete problem, namely the Knapsack Problem. The relationship is clear if one examines their respective syntactic forms; the visual matching problem is formulated above, the Knapsack problem below. The relationship arises from the fact that both problems require solutions to simultaneously satisfy two constraints and a solution may be an arbitrary subset of the input data. The reader may recall a different version of the Knapsack Problem that was presented in an earlier section. A direct reduction (by local replacement) of Knapsack to visual matching is the appropriate proof procedure. The formal Knapsack Problem follows:

Knapsack
instance: Finite set U
for each u ∈U there is a function $s(u) \in Z^+$ (the set of positive integers)
and a function $v(u) \in Z^+$
positive integers B and C
question: Is there a subset $U' \subseteq U$ such that

$$\sum_{u \in U'} s(u) \leq \mathbf{B} \text{ and } \sum_{u \in U'} v(u) \geq \mathbf{C}\ ?$$

The following then is the main theorem:

Theorem 1: Unbounded visual matching is NP-Complete.

The proof is given by Tsotsos (1989). The task is inherently exponential in the number of pixels, or image size, $O(2^I)$. The proof can also be trivially extended to any number of sensory dimensions; it is thus claimed that any perceptual search task is NP-Complete in its unbounded form. Because this result is independent of the implementation, biological vision cannot be "general purpose" as seems to be widely believed. It only appears to be general purpose, perhaps because it is coupled to action and manipulation (see also Ballard 1989).

If we consider attentional optimizations, using the target item, it is easy to show that the problem has linear time complexity. The key is to base the computation of the difference and correlation functions on the target rather than the test image. If the task is restated as follows: Is there a subset $\mathbf{I}' \subseteq \mathbf{I}$ such that

$$\sum_{p \in T} \mathrm{diff}(p) \leq \theta \text{ and } \sum_{p \in T} \mathrm{corr}(p) \geq \phi\ ?$$

where

$$\sum_{p \in T} \mathrm{diff}(p) = \sum_{p \in T} \left[\sum_{j \in M_t} | t_{x,y,j} - i_{x,y,j} | \right]$$

and

$$\sum_{p \in T} \mathrm{corr}(p) = \sum_{p \in T} \left[\sum_{j \in M_t} t_{x,y,j} \times i_{x,y,j} \right]$$

a simple algorithm is apparent. The computation of the diff and corr functions is driven by the target image and measurements of its parameters rather than those of test image. First, center the target item over each pixel of the test image; compute the diff and corr measures between test and target image at that position; among all the positions possible, choose the solution that satisfies the constraints. The resulting worst-case number of multiplications and additions would be given by:

$$O\,[|\mathbf{I}| \times |\mathbf{T}| \times |\mathbf{M}_t|] \qquad (1)$$

In other words, the worst-case number of computations of the diff and corr functions is determined by the product of the size of the test image in pixels, the size of the target in pixels and the number of measurements at each pixel of the targets. Because at least one linear algorithm exists, this leads to the second theorem:

Theorem 2: Bounded visual search has linear time complexity.

The task has linear time complexity in the size of the image. This provides a strong hypothesis: Because actual psychological experiments on visual search with known targets report search performance as having linear time complexity and not exponential (Treisman 1985), the inherent computational nature of the problem strongly suggests that attentional influences play an important role.

2.4 The natural parameters of computation

As discussed earlier, there are only a limited number of options available when one is faced with a problem that is NP-Complete. One must first ensure that the actual values of the exponentials are large enough to merit special mechanisms, however. The unbounded visual search problem is exponential in the number of image pixels. This is certainly a large number; one accordingly moves on to the consideration of special strategies. One is to look for optimizations and approximations guided by the natural parameters of the task: The complexity of a problem may be more affected by some parameters than others. The parameters that naturally have large values should be considered for optimization before those with naturally small values. The natural parameters of the visual search problem correspond to the elements of the computational model used in the remainder of this paper. They are:

i. A stimulus array with P elements. This is a retinotopic representation, that is, one whose physically adjacent elements represent spatially adjacent regions in the visual scene.

ii. At each array element, one or more tokens representing physical parameters of the scene that may be computed. These tokens are of a given type, and for each type there are many potential token instances, yet only one instance can be associated with a type at any one time. Types are not necessarily independent. A map is defined as a retinotopic representation of only one type of visual parameter. All maps are assumed to be the same size and represent parameters that are computed with equal difficulty. Maps are logical abstractions, not necessarily physically separable entities. There are M maps in the system; the types will be left unspecified and abstract. The point is simply to count how many are possible at the output of early vision. The question of how many instances are possible for each type will not be addressed; it is not important for first-order complexity. It would be affected by the method of selecting tokens from among competitors. For example, a competitive scheme such as relaxation (Hummel & Zucker 1983) or winner-take-all (Feldman & Ballard 1982) could be assumed, and the complexity would be polynomial (on the assumption that the schemes converge). For this discussion it is assumed that although there may be many instances at the same location and time competing to represent a single type, some polynomial complexity mechanism chooses from among the competitors.

iii. A knowledge base of visual prototypes, each representing a particular visual object, event, scene, or episode. Let VP represent the number of prototypes. Each prototype may be considered an invariant description of a visual entity (invariant for size, location, rotation, and other parameters as appropriate).

iv. A large pool of identical processors, each able to choose a subset of the stimulus array locations, fetching a subset of the tokens representing physical characteristics at each location, accessing one visual prototype, and then matching the token set to the prototype. Collections of location/token elements are termed receptive fields; thus, a receptive field is defined as the area of the visual scene in which a change in the visual stimulus causes a change in the output of the processor to which it is connected. The matching process is the basic operation. Matching here means that the processor determines whether or not the collection of tokens over the selection of locations optimally represents an image-specific projection of the prototype. The output of a processor is matching success or failure with perhaps an indication of response strength. Each processor completes this operation in S seconds. The final output of the system is also available in S seconds; thus the actual time required for this process does not matter. The effective speed-up due to parallelism will be denoted by the variable Π. No difficulty is posed for determining first-order complexity by not specifying exactly what each processor does. As long as the complexity of each is polynomial rather than exponential, there is no change in complexity class. It is easy to see that the computation of the diff and corr functions presented earlier has polynomial time complexity.

2.5. Complexity level analysis and realizable visual search procedures

How can the fact that the visual search problem in its completely. general form is inherently exponential be reconciled with the fact that biological visual systems perform so well? Recall from the discussion in section 1.2 that the inherent computational complexity of a problem is independent of the implementation of its solution. Hence biology must face the same theoretical problem. But although it is very tempting to claim that the optimizations and approximations that will be presented here are indeed those that biology uses, a little caution is required. Strong arguments for this claim will be presented in the remainder of this paper. The model must be taken as only one of many possible ones, however. It is a challenge to determine which others satisfy the complexity level as described here and yet are more plausible biologically.

Neisser (1967), among others, claimed that a spatially parallel[7] model of perception is quantitatively inadequate: The exponential nature of unbounded visual search in its most general form was proved formally earlier, but it is also easy to demonstrate Neisser's claim quantitatively.

Because the unbounded case is being considered, we add into the determination of complexity the potential size of the target set, that is, the number of possible visual elements that may be present in an arbitrary image – the full visual prototype knowledge base. Given VP visual prototypes, P elements of a retinotopic array, and M types representing visual parameters at each array element,

$$\mathrm{VP} \times 2^{\mathrm{P} \times \mathrm{M}} \qquad (2)$$

matching operations are required in the worst case. If Π is the degree of effective speed-up due to parallelism, the amount of time taken to perform the worst-case number of operations as presented in equation (2) is given by:

$$\mathrm{S} = \frac{2^{\mathrm{P} \times \mathrm{M}} \times \mathrm{VP} \times \mathrm{S}}{\Pi} \qquad (3)$$

where each processor requires S seconds to complete one operation and the output of the system is also available in S seconds. From now on, the equation will be simplified by canceling the S terms:

$$\Pi = \mathrm{VP} \times 2^{\mathrm{P} \times \mathrm{M}} \qquad (4)$$

The number of possible subsets of location/type pairs is the power set of all locations times parameter types. The power set of a set includes all possible subsets of elements as well as the null set. The null set has no effect at this stage of the discussion and will be deleted later when it will make a difference. Each processor has a receptive field which is defined by a subset of pixel locations. Each prototype must be matched against each possible subimage. Another possible complexity function would include M as a multiplier of the power set of locations rather than in the exponent of the power set. However, this presupposes that only one type of parameter is needed to define a visual entity, and this is true only in very special circumstances. Equation (2) allows an arbitrary subset of parameters to be required for any visual entity. It does not provide an enumeration of the number of images; rather, it enumerates the number of data items that must be considered and the number of comparisons that must be performed with those data items in the worst case. This is clearly combinatorially explosive.

We can demonstrate the implications of this complexity measure by using a few relevant estimates for human vision of the amount of input data and the number of visual prototypes in memory. In the "Visual Dictionary" (Corbeil 1986), 25,000 items are included pictorially. The world categorized is one of black and white outline diagrams with little shading, no color, no motion, and no specializations or category names for common objects. Biederman (1988) claims that there are 30,000 readily identifiable objects in the world. These are individual objects; he does not include whole scenes or collections of objects. If these were included a very large number of possibilities would presumably result. Thus, a conservative lower estimate for the number of prototypes is VP = 100,000. A large but arbitrary upper estimate would be VP = 10,000,000.[8] M is surely 1 at the photoreceptors. An upper bound is rather difficult to estimate; one must answer the question: How many independent parameters are required to describe each point in visual space? Intuitively, there seem to be many: location in three dimensions; wavelength; energy; surface orientation; surface roughness; local gradients; and temporal derivatives on at least some of these quantities. At the photoreceptors, all of these variables are rolled up into a single continuous signal. Marr (1982) uses six different quantities in his primal sketch, from which all other required visual information can be derived[9]: relative depth; local changes in depth; discontinuities in depth; local surface orientation; local changes in surface orientation; and discontinuities in surface orientation. An upper estimate of M as 12 will be used. P is the number of locations in the retinotopic representation; an upper, middle, and lower value will be used. The number of receptors in the retina (130,000,000) is the upper value, the number of retinal ganglion cells (approximately 1,000,000 and roughly the same as the number of pixels in a 1K × 1K image) is the middle value, and the size of a 256 × 256 image is the lower values (65,536 pixels). It will become apparent that the choices for these parameters have no effect on our general conclusions; the numerical choices are for demonstration purposes only.

Table 1 gives values for Π for the estimates on P, M, and VP described above using equation (4). The inescapable conclusion is that with this simplified architecture, the task is intractable: Parallelism alone is not the answer, as Neisser correctly pointed out. But remember that complexity measures reflect worst-case situations. Suppose the brain is large enough to handle the sizes of problems that normally occur in the real world and is designed such that performance degrades gracefully for the more complex ones. Then one may ask "How large a problem can the brain handle?" In part, it is this question that has motivated this research. Biologically plausible values for Π, P, and M must also be determined if we are to use guidance from the natural parameters of the problem in the complexity level analysis. The speed-up due to parallelism is clearly significant, but it surely cannot be as large as the number of neurons in the brain, 10^{10}. Realizable parallel processing systems require considerations of local memory, synchronization, communication, and so on, and a collection of neurons is presumably required to accomplish this for each degree of speed-up. This collection is the unit of parallelism in which we are interested. Because about 20% of the cerebral cortex is devoted to visual processing, the value of Π that is biologically plausible is significantly less than 10^9.

Hubel and Wiesel (1977) discovered that primary visual cortex (also called area 17 or V1 in mammals) exhibited a distinct columnar architecture with some apparent functional significance: the hypercolumn. They proposed that the hypercolumn is the basic processing unit and that each contain a complete collection of neurons sensitive to and selective for all the basic visual properties (color, motion, orientation, binocular disparity, luminance). The receptive fields within a hypercolumn were all overlapping and specific for a given region of visual space. Crossing into a neighboring hypercolumn reveals the same collection of neural sensitivities, but for an adjacent region of visual space. Thus, the representation is retinotopic. A layer of such hypercolumns may be thought of as representing visual space with a resolution equivalent to that of an image in which each hypercolumn is represented by a pixel. We will think of a "unit of output" as being the set of outputs that leaves a hypercolumn. It is known that the area of each hemisphere of the primary visual cortex in humans is 1500–3700 mm^2, with the average approximately 2100 mm^2 (Stensaas et al. 1974), and that each hypercolumn is approximately 1 mm^2 in area. There are therefore 1500–3700 hypercolumns in primary visual cortex or 2100 on average. Therefore, the output of the most abstract, retinotopic extrastriate areas must have on the order of a small number of thousands of units.

There are several maps in the visual cortex. Each map represents at least a portion of visual space and each has its own distinct characteristics. There may be 30 visual areas or so in primates, but not all are organized retinotopically, and, even then, with varying degrees of retinotopy (Maunsell & Newsome 1987). Because many areas have more than one population of neurons, there are more *logical* maps than physical ones. The logical map is the unit of the parameter M that this discussion will address. The areas commonly accepted as being retinotopic include V1, V2, V3, MT, and V4, whereas the nonretinotopic ones include IT, posterior parietal cortex, and the frontal eye fields. According to van Essen and Maunsell (1983), the division between retinotopic

Table 1. *Values of Π for varying values of P, M, and VP for the basic architecture.*

Π	VP = 10^5		VP = 10^7	
P	M = 1	M = 12	M = 1	M = 12
130,000,000	$10^{39,133,905}$	—	$10^{39,133,907}$	—
1,000,000	$10^{301,035}$	$10^{3,612,365}$	$10^{301,037}$	$10^{3,612,367}$
65,536	$10^{19,733}$	$10^{236,744}$	$10^{19,735}$	$10^{236,746}$

and nonretinotopic areas, although fuzzy in general, may be placed after areas MT and V4 and before IT, area 7, and the frontal eye fields. Maps seem to be organized hierarchically, as a partial ordering, so that the greater the distance from the retina, the smaller the maps are, and the larger the receptive fields of their neurons. There is also more than one pathway from the retina to higher levels of processing (see, for example, Stone et al. 1979; Ungerleider & Mishkin 1982; van Essen & Anderson 1989). Because at any level of the map hierarchy there are no more than a handful or so of maps, the number of maps at the output of early vision is on the order of a handful.

The important point to keep in mind is that we are only concerned with how many of each of the units each parameter represents are possible. Now, beginning with equation (4), the problem will be reshaped and the allocated resources will be optimized so that a biologically plausible, computationally tractable realization can be achieved.

3. Demonstrating complexity sufficiency

3.1. Aspects of an idealized structure

To provide a structure that satisfies the basic complexity constraints of time and space we will begin with an idealized one that is consistent with Theorem 1 presented earlier. One of the important aspects of the definition of problem complexity (section 1.3) is that it is the minimum over all possible algorithms. By beginning this analysis with an ideal structure, one that contains only data and processors, with no commitment to any processing method or data organization, we start at the same point where the proof left off, namely, with an NP-complete problem. No target item is specified in an unbounded task; instead, we add in the entire knowledge base of visual prototypes in order to study its effect. Figure 2 illustrates this ideal structure. There are three major components: input data from an image; input data from world knowledge; and processors (each defined in section 2.4).

For quantitative purposes, hexagonal images of order N packed with hexagonal pixels (i.e., N pixels per side) and hexagonal tiling of a hexagonal image are assumed, much of the discussion is independent of the choice of image mosaic. Whenever the choice does have an effect on the results, it will be pointed out. Many researchers have advocated the use of hexagonal tilings for images (e.g., Watson & Ahumada 1987). The number of pixels in such a hexagonal image is $P_N = 3N^2 - 3N + 1$, each uniformly distributed across the image. All input data and prototypes are hard-wired to the processors. All maps are assumed to be the same size and can be computed with equal difficulty.

Note that these assumptions bear little resemblance to the actual implementation of biological visual systems. The retinal mosaic is not uniform, map sizes are not the same, retinotopy is variable, map contents may differ in computation time as well as in other parameters, and so forth, but such differences do not affect the first-order constraints derived in this paper.

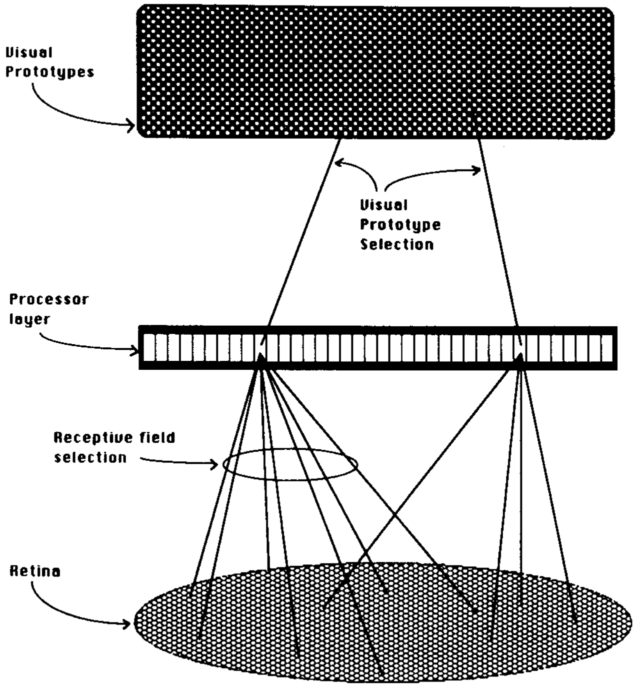


Figure 2. The idealized architecture of vision that corresponds to the NP-Complete problem of unbounded visual search. Each processor within the processor layer matches one subset of retinal locations and measurements with one visual prototype from memory.

3.2. A sufficient set of optimizations

Efficiency can be gained by attacking the search through all possible visual prototypes with a process of successive refinement. "Divide and conquer" is a standard tactic for designing complex systems.[10] Assume that we can build a binary tree whose leaves are the prototypes of the knowledge base and whose nodes are superclasses of prototypes. This is not unlike the specialization or decomposition hierarchies found in the knowledge representation literature. This transforms the linear search through the prototypes into a logarithmic search. Note that although a binary tree search is serial, the key here is the number of operations; the search will be "parallelized" later. We replace the linear term VP with the base 2 logarithm:

$$II = 2^{P \times M} \times \log_2 VP \qquad (5)$$

This is a very minor improvement. On its own this is at best a small contributor in defeating the complexity problem of vision.

Note that not all 2^P possible combinations of locations need to be considered in the model. Objects are not spread arbitrarily in 3-space, and events are not spread arbitrarily in the time dimension. Their physical characteristics are similarly localized. Simple optical arguments show that coherent objects in physical space remain coherent on an image. Localized operations are almost universal in computational vision since Rosenfeld (1962) used local texture measures for terrain identification.

The receptive field of a processor has so far been defined as an arbitrary collection of location/ measurement pairs. We now change this definition to a set of contiguous location/measurement pairs. Assuming a hexagonal image of order N, we only consider hexagonal contiguous regions of whole array elements as processor receptive fields.[11] Simple geometry yields N^3 receptive fields of the type described above over the whole image, or in pixels, approximately,

$$\frac{P^{1.5}}{3\sqrt{3}} + \frac{P}{2} + \frac{5\sqrt{P/3}}{8} \qquad (6)$$

This gives the total number of hexagonal, contiguous receptive fields of all sizes and centered at all locations in the image array, and these now take the place of the arbitrary power set of location/measurement tokens. These are accordingly the receptive fields that must be examined by the processors. Figure 3 illustrates the receptive field structure. Receptive fields clearly overlap and are of all possible sizes wholly within the retina. The degree of speed-up function for this third architecture is dramatically different:

$$\Pi = N^3 \times (2^M - 1) \times \log_2 VP \qquad (7)$$

The powerset of maps still remains in the expression because it is not known a priori which subset of maps is the correct one for the best image to prototype match; hence in the worst case all subsets must be examined. The null set has been removed, however, because it may have a numerical effect. Table 2 gives the values for Π resulting from this expression. Although there has been a significant change in the estimated degree of speed-up, the values are still not close to biologically plausible ones. A side effect of this particular receptive field structure is that it does not allow as fine a selection of tokens across the receptive field as equation (4). There, some of the subsets could indeed represent contiguous space, but the powerset of elements implied that over a contiguous space each element could be a different type of parameter. The new definition of receptive field requires that tokens for each selected type of parameter be used for each location across the receptive field. This is reasonable, because visual parameters display the same localization as the objects that exhibit them.

Table 2. *Values of Π for varying values of P, M, and VP for the basic structure.*

Π	VP = 10^5		VP = 10^7	
P	M = 1	M = 12	M = 1	M = 12
130,000,000	$10^{12.68}$	—	$10^{12.82}$	—
1,000,000	$10^{9.5}$	$10^{13.12}$	$10^{9.65}$	$10^{13.26}$
65,536	$10^{7.73}$	$10^{11.35}$	$10^{7.88}$	$10^{11.49}$

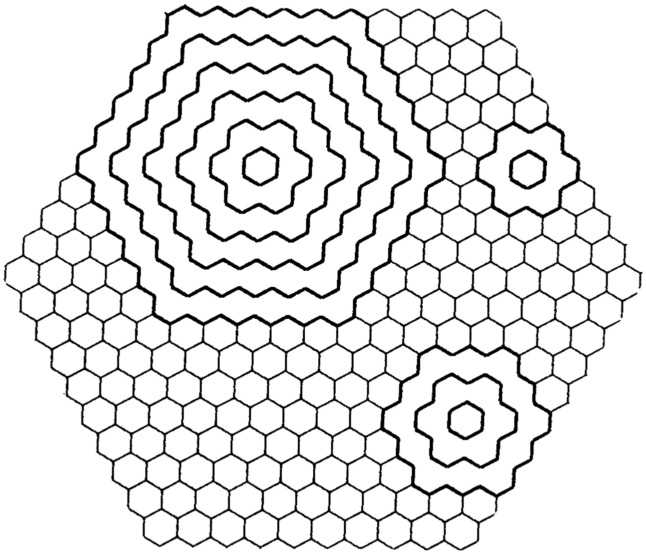


Figure 3. The hexagonal retinotopic stimulus representation. The hexagon is of order N, that is, N elements per side. The diameter of the hexagon is 2N-1 elements. For three of the elements, the corresponding complete set of receptive fields that can be centered on those elements is shown. In this way each element in the hexagon can be the center of a number of hexagonal receptive fields. There are N^3 receptive fields in all.

For the third optimization, we note that not all visual stimuli involve all types of tokens. Let $\hat{M}$ represent the number of types of visual parameters relevant for a given input. Thus, the number of possible subsets of types is $2^{\hat{M}} - 1$. This could be implemented via a computation of "pooled response," that is, an output associated with each map that signals whether or not the map has been activated. The idea is borrowed from Treisman (1985); in this use, it is assumed that it acts as a gating signal selecting which receptive fields are relevant for matching. A direct result is the logical segregation of types, an idea that arose in Barrow and Tenenbaum's (1978) theory as well as Treisman's "feature integration" theory. The new expression for speed-up is:

$$\Pi = N^3 \times (2^{\hat{M}} - 1) \times \log_2 VP \qquad (8)$$

The values for $\hat{M} = 1$, the simplest input, are found in Table 2 in the M = 1 column. Even for the smallest image, the values of Π are barely plausible biologically. Because pooled response and map segregation do not lead to savings for all possible images, their role may be to speed up the computation for the simpler inputs (the simplest and therefore fastest condition being for $\hat{M} = 1$). It is interesting to note that no search through subsets of maps is required at all if the target is permitted to influence the computation. The exponential term $2^{\hat{M}} - 1$ is replaced by 1. If the maps that would represent the target are known in advance, then only that complete set need be considered in the input. This point will be elaborated in a later section.

Separation of types into physically distinct maps follows if connectivity lengths are considered. According to Cowey (1979) physically separate visual maps evolved because units that compute similar quantities need to communicate with one other for consistency and thus need to be interconnected. The connectivity lengths would be prohibitive if the units were separated. Barlow (1986) gives another reason: The "new images" formed by reprojecting visual space are needed to allow similarities in distant parts of the image to be detected. Kaas (1989) proposes that the modular design is needed for ease in adding on more representations in an evolutionary pathway. We present a fourth possible reason, namely, that the logical organization of maps, if used appropriately, can lower the complexity of the task. It may be a mechanism for the system's graceful degradation with increasing complexity of the input. Algorithms whose performance degrades gracefully are preferable to fixed worst-case algorithms.

Further efficiency is gained by trading off precision. This can be achieved by reducing the resolution of the visual image and simultaneously abstracting the input to maintain its semantic content. Several proposals have appeared in the literature for such input abstraction. Among these are: the processing cone representation of Uhr (1972); the Gaussian pyramid of Burt & Adelson (1983); the hexagonal image pyramid of Watson & Ahumada (1987) which is perhaps closest to the basic hexagonal hierarchy representation of our work; and the hierarchical filter cascades of Fleet & Jepson (1989).

Let $\hat{N}$ be the size of the new abstracted array (illustrated in Figure 4) and change the expression for degree of speed-up to:

$$\Pi = \hat{N}^3 \times (2^{\hat{M}} - 1) \times \log_2 \mathrm{VP} \qquad (9)$$

We seek a value of P of about a few thousand. For $P_{\hat{N}}$ of say 3000, the value of $\hat{N}$ would be 32, and for VP = 100,000 and $\hat{M} = 1$, the required parallelism is $10^{5.74}$. This is wel' within the desired range. It is easy to see that variations in Π, $\hat{M}$, and VP lead to changes in $P_{\hat{N}}$, and that there are a great many possible configurations that lead to values of $P_{\hat{N}}$ that are in the low thousands.

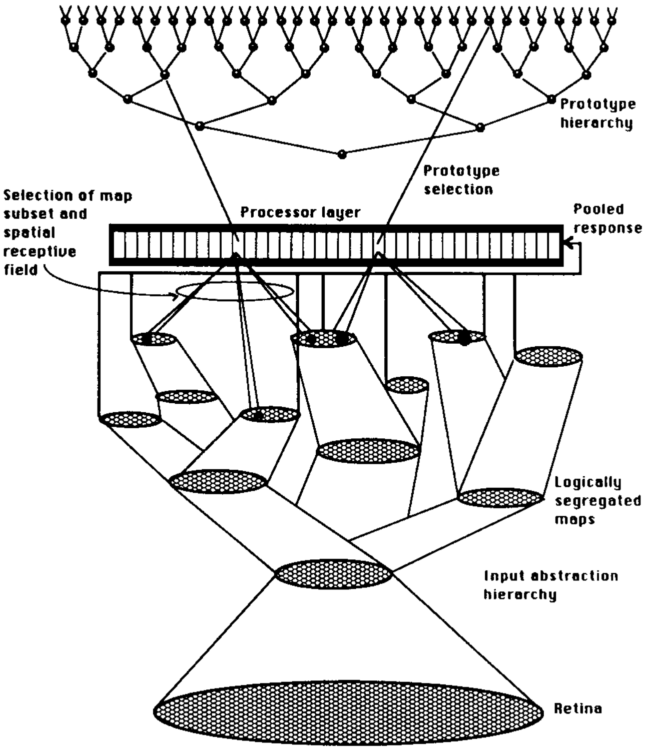

Prototype hierarchy
Prototype selection
Selection of map subset and spatial receptive field
Processor layer
Pooled response
Logically segregated maps
Input abstraction hierarchy
Retina

Figure 4. The architecture that satisfies complexity-level analysis. It includes an input abstraction hierarchy, logically segregated visual maps, a layer of parallel processors, a hierarchically organized set of visual prototypes, and a spatially contiguous definition of receptive field.

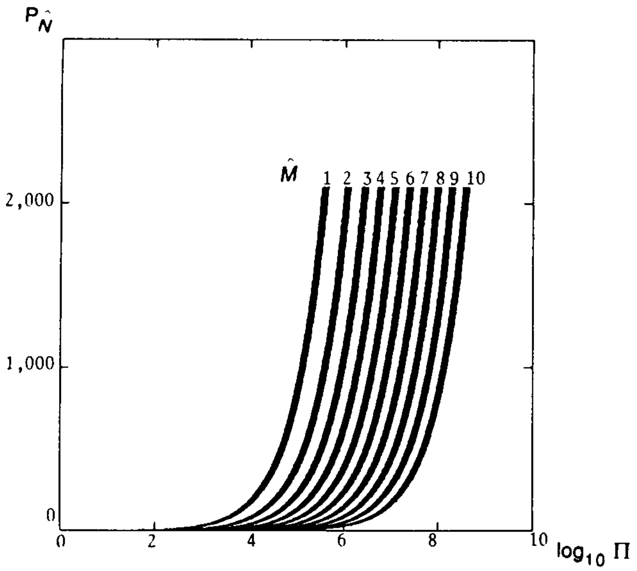


Figure 5. The family of curves generated using equation (9) for varying values of Π, VP, $\hat{M}$, and $P_{\hat{N}}$. $P_{\hat{N}}$, ranging from 0 to 60000, is plotted against $\log_{10}\Pi$, ranging from 1 to 9. Each thick solid curve represents a value of $\hat{M}$, with $\hat{M} = 1$ the leftmost, and $\hat{M} = 10$ the rightmost. The thickness of each curve represents the fact that within it is the entire range of VP, from 100,000 to 10,000,000.

3.3. The effects of parameter variations

More insights can be obtained from equation (9). Figure 5 shows a family of curves of this relationship for the values of $P_{\hat{N}}$ versus $\log_{10}\Pi$ for values of $\hat{M}$ ranging from 1 through 10, and for VP = 100,000 through 10,000,000. Thus, the thick solid curves, one for each value of $\hat{M}$, represent the family of curves for the same value of $\hat{M}$ for all values of VP between 100,000 and 10,000,000. Several qualitative conclusions can be drawn and verified analytically. If these are the basic performance relationships, the designer of the visual system is faced with a few choices and tradeoffs. First, there seems to be a 'hard complexity wall' on the number of processors. It is very cheap in terms of processors to incorporate a very large knowledge base of prototypes, as is clear from Figure 5. Changes in VP have a very small effect on Π, as can be seen easily from the partial derivative,

$$\frac{\partial \Pi}{\partial \mathrm{VP}} = \Pi \times \frac{\log_2 e}{\mathrm{VP} \times \log_2 \mathrm{VP}}.$$

It is more expensive to use higher resolution because

$$\frac{\partial \Pi}{\partial P_{\hat{N}}} = \Pi \times \frac{\frac{\sqrt{P}}{2\sqrt{3}} + \frac{1}{2} + \frac{5}{16\sqrt{3P}}}{\frac{P^{1.5}}{3\sqrt{3}} + \frac{P}{2} + \frac{5\sqrt{P/3}}{8}}.$$

The largest expense is incurred for adding maps, because

$$\frac{\partial \Pi}{\partial \hat{M}} = \Pi \times \frac{2^{\hat{M}} \times \log_e 2}{2^{\hat{M}} - 1}.$$

If, for example, VP = 10,000,000, $\Pi = 10^{5.6}$, $\hat{M} = 1$, and $\hat{N} = 26$, then the derivative of Π for changes in $\hat{M}$ is

60,114 times steeper than for changes in $\hat{N}$ and 223,000,000 times steeper than for changes in VP. Thus, it is most critical that the number of subsets to be searched be as small as possible (but this does not necessarily mean that M or $\hat{M}$ must be as small as possible).

4. Applying the minimum cost principle

4.1. The speed-up because of parallelism

Several arguments have been presented so far that affect the determination of the degree of parallelism. We have shown that bounded visual search has linear complexity, that is, if the target is used to direct the computation, the exponential nature of the unbounded approach can be avoided. Only one subset of maps need be considered. We have also shown that for the unbounded case, of all the variables the incremental cost incurred by additional maps on the required parallelism is greatest, because subset search is exponential. This implies that the number of subsets of maps considered at any one time must be small. This is conveniently solved by applying attentional information. If we apply the minimum cost principle to processors, the number of processors should be that dictated by the attentional version of the problem. The degree of parallelism in the system is therefore fixed to:

$$\Pi = \hat{N}^3 \times \log_2 \mathrm{VP} \qquad (10)$$

Only one subset of maps may be considered at any one time. Using the above expression, the minimum time for computation is $C_{\min}$, given by:

$$C_{\min} = \frac{\hat{N}^3 \times \log_2 \mathrm{VP} \times S}{\Pi} = S \qquad (11)$$

In the unbounded case, where the target is not known a priori, the exponential term is still needed and the time to compute increases exponentially with $\hat{M}$,

$$C = \frac{\hat{N}^3 \times (2^{\hat{M}} - 1) \times \log_2 \mathrm{VP} \times S}{\Pi} = C_{\min} \times (2^{\hat{M}} - 1) \qquad (12)$$

up to a maximum given by $T_{\max}$, when all maps (M) are active. This predicts that if the target is unknown, each of the possible subsets of the active maps is used in turn in matching.

4.2. Columnar processor organization

How are the processors connected to the retinotopic maps? At each array element of the most abstract maps we can define a *processor assembly:* A processor assembly contains, on average, $\Pi/P_{\hat{N}}$ processors. Using equation (10), and the expression for $P_{\hat{N}}$ in terms of $\hat{N}$, the number of processors in an assembly is:

$$\frac{\hat{N}^3}{3\hat{N}^2 - 3\hat{N} + 1} \times \log_2 \mathrm{VP} \qquad (13)$$

But $\frac{\hat{N}^3}{3\hat{N}^2 - 3\hat{N} + 1}$ is the average number of processor receptive fields at each location. Thus, there are $\log_2$VP processors for each receptive field at each location. Call this basic set of processors a *receptive field assembly:* Each must be connected to its relevant retinotopic elements. The design principle of minimum cost is involved here because stacking the assemblies over the centers of their receptive fields minimizes connection length. The proof is straightforward. Assume a one-dimensional receptive field whose center is at position Y and whose rims are at positions Y + (K − 1)/2 and Y − (K − 1)/2. The diameter of the receptive field is K, an odd integer; this is the number of units to which each processor must be connected. The total length of all connections for a single processor to this receptive field can be expressed by:

$$\sum_{x=Y-\frac{K+1}{2}}^{Y+\frac{K+1}{2}} \sqrt{1+(loc-x)^2} \qquad (14)$$

It is assumed that processors are a unit distance above the stimulus array, but this does not affect the result. The location of the processor is given by the variable *loc* and can take values between 1 and K. This function is minimized when *loc* = Y. Thus, in the one-dimensional case described above, placing the processor over the center of its receptive field minimizes total length for those connections. The same is true of the two-dimensional case, because the situation is circularly symmetric. It follows that for one layer of processors the configuration with minimal total connectivity is one where each processor is placed directly over the center of its receptive field. If there is more than one layer of processors, the same conclusion is reached. More than one processor cannot occupy the same physical space. If a layer is configured so that the processors are over the centers of their receptive fields, the remaining processors must be placed above or below this layer. The same argument then applies: The minimum connection length for this next layer of processors is achieved if the processors are centered over their receptive fields. This procedure is applied until all processors have been allocated.

There is a column of processor assemblies for each retinotopic element (or pixel) and within the column there is a receptive field assembly for each of the receptive fields centered on that pixel. This means that the full set of prototypes can be recognized at each position in the visual field.[12] Figure 6 illustrates the organization of processor assemblies. This structure is like an abstract version of Hubel & Wiesel's hypercolumns. In principle if the decision criteria for branching in the knowledge base search are known and one branch decision does not depend on the previous decision then the processors can categorize each receptive field – in parallel and in one time step – because there is one processor for each of the $\log_2$ VP branches for each receptive field. The result of each receptive field match would be available at the outputs of the corresponding receptive field assembly and the pattern of responses within a receptive field assembly points to the most appropriate prototype, because all are checked, at least in a coarse sense. This is one way the serial nature of binary search is "parallelized." The center pixel requires $\hat{N}\log_2$ VP processors (or $\hat{N}$ receptive field assemblies), whereas the pixels on the rim require $\log_2$ VP processors (or 1 receptive field assembly).

4.3. Processor layer inverse magnification

The fact that the cortex is flat but the columnar processor organization described using the principle of minimum

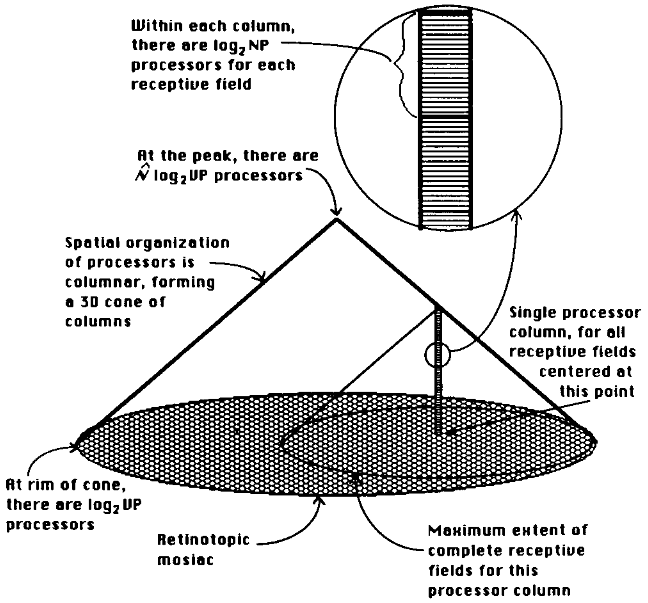

Within each column, there are $\log_2 NP$ processors for each receptive field
At the peak, there are $\hat{N} \log_2 VP$ processors
Spatial organization of processors is columnar, forming a 3D cone of columns
Single processor column, for all receptive fields centered at this point
At rim of cone, there are $\log_2 VP$ processors
Retinotopic mosiac
Maximum extent of complete receptive fields for this processor column

Figure 6. The columnar organization of processes that minimizes total length of receptive field to processor connections. The three-dimensional structure of the processors forms a cone. Within this cone are columns of processors and each column is centered on a specific pixel of the retinal image. Within each column is a complete set of processors for all of the receptive fields centered at that location in the image.

cost of connection length is conical is a good example of the conflict between design dimensions described in the introduction. Nature provided the flat package of the cortex; the processor structure must fit within it. The compromise is to sacrifice some (but not all) of the benefits of minimizing connection length. To implement such a structure on "flat" hardware and maintain at least some of the connectivity benefits of stacking the processors over their corresponding receptive field centers the set of processors must be flattened by pressing down on the cone's peak and redistributing the elements locally, as if the cone were made of putty. This leads to an inverse magnification, that is, more area of the processor layer per unit is devoted to central visual fields than to peripheral ones. A magnification function may be obtained by dividing the number of processors at a given radius R in the conical model by the average number of processors at each location over the whole cone. The value of R at the center is 1, while its value at the rim of the stimulus array is $\hat{N}$. Thus, the area magnification function is:

$$\frac{2(\hat{N} - \mathrm{R} + 1)\log_2 \mathrm{VP}}{(\hat{N} + 1)\log_2 \mathrm{VP}} = 2 - \frac{2\mathrm{R}}{\hat{N} + 1} \qquad (15)$$

This assumes uniform pixel distribution and uniform distribution of receptive fields, whereas the retina does not have uniform photoreceptor distribution, nor is the distribution of receptive fields uniform.

Daniel and Whitteridge (1961) measured cortical magnification factors for the monkey and discovered an inverse relationship between the location of a receptive field in the cortex and the corresponding location on the retina. This relationship was measured in terms of the amount of distance across the cortex that must be traversed in order to achieve a one degree traversal in visual space on the retina. A fit to the data was done by Schwartz (1977), who found that the relationship between cortical magnification and visual eccentricity was exponential. It is nevertheless interesting that receptive field localization and conservation of connection lengths alone lead to the negative slope across a flat processor layer. It should be emphasized that the layer of processors described here does not necessarily correspond to any area of the brain. This exercise was undertaken to show that complexity considerations will constrain an idealized structure so as to make resemblances with observed neuroanatomy difficult to ignore.

4.4. Size and number of maps

We are now ready to consider the question of the total size and number of maps. This will be answered using the structure derived so far, plus the observation that average connectivity in the cortex is about 1,000 for both fan-in and for fan-out. The analysis of this section will simply count how many receptive field assemblies in how many maps can be wired up given the observed average connectivity. Note that the numerical predictions of this section (and this section alone) depend on the hexagonal image assumption.

Each of the receptive fields must be hard-wired directly to the receptive field assemblies; there is no other way that parallel processing can occur. Consider connectivity in the direction from the retinotopic representations to the processor layer. The first quantity to determine is the total number of wires required to connect each receptive field to its receptive field assembly. This will be computed by simply summing for each of the $\hat{N}^3$ receptive fields the number of pixels it contains. Each point in the image is a member of a ring of points and each point of each ring is the center for the same number of receptive fields, all the same size. The number of elements of each ring at radius i is given by $\mathrm{P}_i - \mathrm{P}_{i-1}$. The receptive fields centered at each member of the ring are of sizes 1 through $\hat{N} - i + 1$. The sum of the elements in each receptive field in each element of each possible ring then gives the total number of wires required to hardwire all the receptive fields, independently of one another. This is given by:

$$\sum_{i=1}^{\hat{N}} \left\{ (P_i - P_{i-1}) \sum_{j=1}^{\hat{N}-i+1} P_j \right\} = \frac{\hat{N}}{10}(3\hat{N}^2 + 1)(\hat{N}^2 + 1) \qquad (17)$$

Call this the area (A) of each map. The total number of wires is A × M, and the average fan-out from the retinotopic representations is $(\mathrm{A} \times \mathrm{M})/(\mathrm{P}_{\hat{N}} \times \mathrm{M})$, or $\mathrm{A}/\mathrm{P}_{\hat{N}}$. At the receptive field assemblies, the average fan-in from the retinotopic representations is the total number of wires divided by the number of receptive field assemblies, or, $\mathrm{A} \times \mathrm{M}/\hat{N}^3$. Each neuron can receive input from about 1000 other neurons and can provide output for about 1000 other neurons, on average. The number of fan-out synapses ranges from a few to several thousand, whereas for fan-in, the range is from a few hundred to a few tens of thousands. Now if this biological constraint on fan-out and fan-in of approximately 1000 is used, the two expressions given above are equated to 1000, and then the equations are solved yielding:

The number of outputs per map is $P_{\hat{N}} = 1284$ ($\hat{N} = 21.185$)
The number of output maps that can be accommodated is $M = 7.4$
The required parallelism is: if VP = 10,000,000, $10^{5.34}$
if VP = 100,000, $10^{5.2}$
The range of connectivity: 7.4 to 9501.

Obviously, integer values of maps and connections are the only realizable values. Each of these values is completely consistent with the biologically plausible ranges described earlier. Note that the values for M and $P_{\hat{N}}$ are lower bounds only. Similarly, the value for Π is an upper bound. There may in fact be other optimizations at work that would permit a larger number of elements to be connected.

The number of outputs predicted poses a serious problem. It seems implausible that the visual system has such coarse spatial resolution. There are two choices: (a) each of the units of output as defined for this work really corresponds to a "bundle" of outputs that perhaps represent fine scale space; or (b) some other compensatory mechanism exists. If option (a) were the case then this would lead to a reexamination of the connectivity argument above. There would still be only a fixed number of connections – and the number of single outputs that could be wired up would be even smaller. This does not seem to be the right way to go. A proposal for a compensatory mechanism will be presented in Section 4.5.

The prediction of seven maps also seems small, given that it was earlier stated that many more parameters may be needed to completely characterize each point in visual space. There is much evidence indicating the inseparability of some early visual operations (but not all);[13] thus, although a single neuron provides a single value as a firing rate, that response may depend on more than one stimulus quality. If each of the M types of parameters is affected by more than one stimulus quality then it is possible to have many more actual values, implying a coarse-coded representation. A coarse-coded representation at this level would allow many more actual values to be extracted, thus leading to a visual system that is capable of much richer interpretations of the visual world and encodes visual information more efficiently (see Hinton 1981, and Ballard et al. 1983, for discussions of coarse-coded representations for vision). It should be clear that coarse-coding is not a necessary mechanism for all types of units; rather, it is an additional tool that can increase the total number of stimulus dimensions that can be coded by the ensemble of seven maps.

4.5. Task-directed influences

The effect of task-directed influences for expected map selection has already been described. This can reduce the remaining exponential component of the complexity function so that a linear function is achieved. In the previous section, lower bounds on the number of outputs per map were derived. The spatial resolution implied by those results was very coarse, seemingly too coarse to account for any aspect of perception. This section will present another connectivity argument that leads to yet another important use of task-directed information.

If the spatial resolution at the output of early vision is too coarse, the simple solution would be to allow access by the processors to each of the higher resolution layers of the input abstraction hierarchy. We could in fact, do this, easily in a computer with software; each element of the input hierarchy could be simply addressed and accessed. Computer memory is random access. There is no evidence for random access in the visual cortex, however. The other obvious solution is simply to connect each processor directly to each of the intermediate representations of the input hierarchy. In that way, each processor would have immediate access to information at all spatial resolutions. There can be no connections from the processors to any of the larger maps in the input abstraction hierarchy, however. The number of such connections would be prohibitive.

Suppose that the processors are to be connected to M maps or high resolution, say 1,000 by 1,000 pixels in extent. We can use the formulae developed earlier for receptive field fan-in to the processor layer, but this time, P = 1,000,000. The resulting additional number of connections per map would be approximately 10^{13}. The additional average fan-in at each receptive field assembly, if Π is on the order of 10^5, is on the order of 10^7. This calculation could be repeated for each of the layers of resolution as well. Given that the cortex contains 10^{10} neurons, with an estimated total number of connections of 10^{13}, this is clearly not how nature implemented access to high resolution maps. There are no known connections between areas IT and V1 and V2 (Maunsell & Newsome 1987), for example, and this analysis predicts that there should be none. If information is to be transmitted to the processors from the larger maps, it must be done *through the input abstraction hierarchy,* "attentively," by tuning the operators that compute the representation of the top-level maps. In this way, spatial resolution at the output of the hierarchy can be effectively increased because top-down tuning can select individual units at the input of the hierarchy, which have higher spatial resolution, for transmission through the hierarchy to the top. This conclusion supports the rapidly growing set of findings that describe attentional influences in extrastriate visual areas by providing additional justification for an attentional tuning mechanism (Fuster 1988; Haenny & Schiller 1988; Haenny et al. 1988; Maunsell et al. 1988; Moran & Desimone 1985; Motter 1988; Mountcastle et al. 1987; Spitzer et al. 1988).

Several important questions about attentional influence arise, however. Is the influence implemented as selective inhibition or as selective enhancement? How is the communication of the influence accomplished? Are all the variables that define a stimulus treated equally by the attentional mechanism? Is there a resolution limit either in space or along a stimulus quality dimension? These are not all new questions. For example, Downing and Pinker (1985) conclude that attention is sensitive to at least depth, visual angle, and retinal and cortical resolution. It is unfortunately impossible to provide answers to all of these questions at this time. Two suggestions will be made, however: An argument will be presented as to why selective inhibition may be more appropriate than selective enhancement; and a proposal for communication of the attentional influence will be sketched.

An attentional scheme has as its main goal the selection of certain aspects of the input stimulus while causing the effects of other aspects of the stimulus to be minimized. Let us assume that the output at any spatial location is determined by a winner-take-all competitive process that can be modeled by a weighted sum computation (e.g., Feldman & Ballard's 1982). Let us denote the weighted sum by Σ wu_i. The signal of interest will be identified as u_k; w is a weight factor; with no loss of generality we can assume that all weights are equal. Let us express the iterative computation of winner-take-all as follows. The (n + 1)th iteration for unit *i* is given by:

$$u^{n+1}_i = u^n_i - \sum_{j\neq i} wu_j \tag{18}$$

In a selective inhibition scheme, anything that is not part of the desired signal is attenuated by a factor of A. The resulting weighted sum is therefore

$$\sum_{i\neq k} w\frac{u_i}{A} + wu_k \tag{19}$$

In an enhancement scheme, the signal of interest is enhanced by A, so that the resultant weighted sum is

$$\sum_{i\neq k} wu_i + wAu_k \tag{20}$$

The stopping criteria for such an iterative scheme are important: One simple stopping test is to terminate when all but one of the signals falls below a threshold. If attentional inhibition is applied the first iteration would look like:

$$\text{for } i \neq k : u^1_i = \frac{u^0_i}{A} - \sum_{j\neq i, j\neq k} w\frac{u^0_j}{A} - wu^0_k \tag{21}$$

and

$$\text{for } k : u^1_k = u^0_k - \sum_{j\neq k,} w\frac{u^0_j}{A} \tag{22}$$

Let us compare the non-attentional with the attentional cases. Equation (22) causes the response at unit *k* to decay more slowly than its non-attentional equivalent (equation 18) because the contributions are attenuated. Equation (21) has a faster rate of decay than u_k in equation (22) because of the additional negative contribution from u_k. Also, each response begins the iteration attenuated by a factor of A. Thus, it should not take as many iterations to reach the threshold as in the nonattentional case, and because the rate of decay of u_k is smaller than in the non-attentional case, its final value will be larger. A short example may be useful:

Within the framework just described, set up 10 competing units. The signal of interest has value 1.0, the others 0.95. This is among the more difficult of discriminations because there is a great deal of similarity among unit values. Set the attenuation factor to 2, the weights to 0.01 and the stopping threshold to 0.1. The uninfluenced version requires 24 iterations to stop, and the *k*th unit has a final response of 0.1565. In the inhibited case, the process stops after only 13 iterations; the *k*th unit has value 0.6925. If an enhancement model is tried, with the same parameter values, 17 iterations are required, and the final value of the *k*th unit is 1.33. Inhibition is the fastest method. The final value is not important because the winner is still the winner regardless of magnitude; speed is the critical parameter for this optimization.

It is easy to see that the required number of iterations of winner-take-all increases with increasing similarity among competitors or with an increasing number of competitors; it depends on the distribution of competing signals. The greater the inhibition, the greater the speed-up of winner-take-all. The concept can be generalized, and it would still lead to similar characteristics. We extend the concept of attention as inhibition from spatial selection to the selection of maps and of feature ranges of interest. Based on the same argument as above, not only would inhibitory selection lead to faster responses but also to larger ones for the selected items. The parameter to which the argument is applied does not matter. Thus, inhibiting more maps, more spatial units, and more of the space covered by a given stimulus quality all have the effect of speeding up the winner-take-all process. Spitzer, Desimone, and Moran (1988) found with stimulus dimensions such as orientation, color, and size, that as discrimination difficulty increases, attention leads to larger responses than in unattended cases. It would be interesting to test whether these larger responses are also achieved in a shorter time.

How is this inhibition applied? Given an input abstraction hierarchy, which looks like a complex of interconnected truncated cones, (recall Figure 4), spatial attentional influence, is applied in spotlight fashion at the top; this appears in many other models (such as Treisman's 1985). This spotlight, however, must make its influence felt throughout the processing hierarchy to at least some appropriate depth. There is no other way the message of spatial selection could reach the items that are actually selected. The spotlight analogy is therefore insufficient, and we propose (keeping our metaphors optical) a "beam" that passes through the hierarchy. The input abstraction hierarchy has its root at the top, where the selection of space is made; that is also where the beam is rooted. The output of a given unit at the top of the hierarchy is directly affected by the subhierarchy for which it is the root. Thus, the beam must affect the branches and leaves of the selected subtree. The beam expands as it traverses the hierarchy, covering all portions of the processing mechanism that directly contribute to the output at its point of entry at the top. In other words, the effective receptive field of the output unit selected for attention is that entire subhierarchy, and the attentional beam must control that receptive field. The central portion of the beam allows the stimulus of interest to pass through unaffected, whereas the remainder of the beam selectively inhibits portions of the processing hierarchy that may give rise to interfering signals. The spatial selection aspect of the beam is illustrated in Figure 7. If the output at the top of the hierarchy is affected by multiple maps, the unit may be the root for a number of subhierarchies (or pathways; see Figure 8). Multiple beams then control these pathways. As mentioned earlier, attention in the form of inhibition was generalized beyond spatial attention to other dimensions of the stimulus such as the selection of features and feature ranges, and so forth. This generalization would require that the spatial beam have substructure that can be manipulated depending on attended information.

The cross-section of the beam will be modeled with a two-dimensional Gaussian profile whose central portion represents the pass zone and whose tails represent the inhibit zone. This will permit easy control of the location, size, and shape of the pass zone. The Gaussian weighting

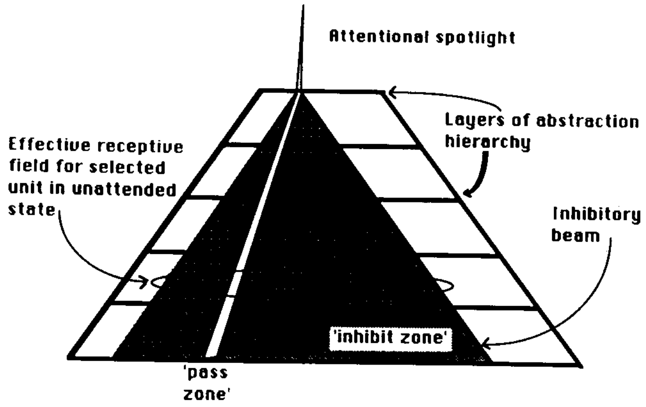

Attentional spotlight
Layers of abstraction hierarchy
Effective receptive field for selected unit in unattended state
Inhibitory beam
'inhibit zone'
'pass zone'

Figure 7. Inhibitory attentional beam operating on an idealized input abstraction hierarchy. In this purely spatial example, the input is abstracted with a single pyramidlike structure. Attention is applied to the hierarchy at the top as a spatial spotlight. As described in the text, the selected unit at the top is affected by a subhierarchy (shown in gray) and the attentional beam must control this subhierarchy. The beam affects the hierarchy by allowing a selected element at the base of the hierarchy to pass through while inhibiting all other elements that influence the receptive field of the selected item at the top of the hierarchy.

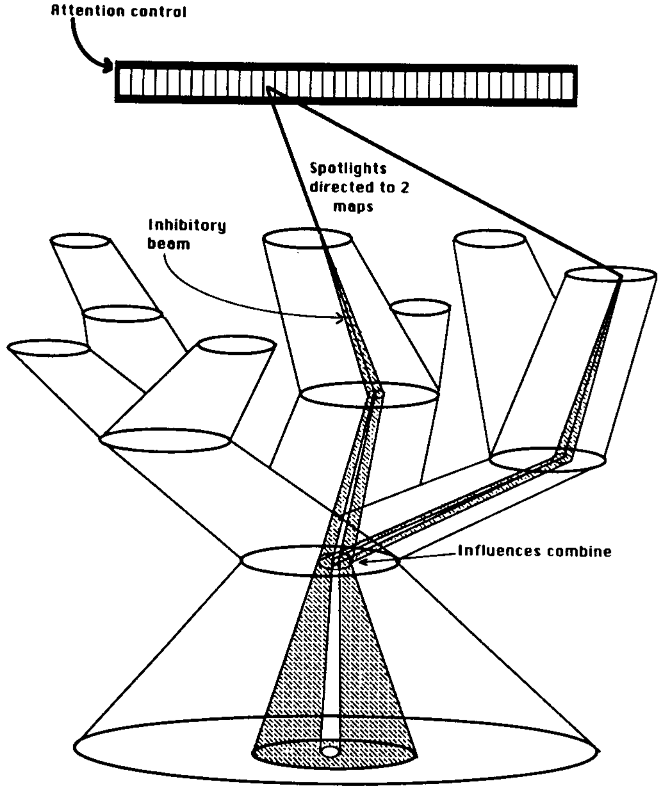


Figure 8. Inhibitory attentional beam applied to a biologically plausible input abstraction hierarchy. As in Figure 7, the attentional beam must affect all of the elements of the input abstraction hierarchy that affect the output of the selected unit at the top. In this case, attention is applied to two pathways in this hierarchy, say, one for shape and another for color. The beams eventually must merge and their influences must combine.

is applied multiplicatively; the pass zone has values near 1.0 whereas the inhibit zone has values that approach zero. This multiplication leads to inhibition of the nonattended signals.

There is experimental evidence supporting the proposal described in this section. First, the idea of selective inhibition as an expression of neural attention has been observed in the experiments of Moran and Desimone (1985) and Motter (1988), both recording from V4 in monkeys. Moran and Desimone discovered that single neurons in trained monkeys as early as V4 (but not in V1) can be turned so that separate stimuli within the same receptive field can be individually attended. They claim that unwanted information is filtered from the receptive fields of neurons in extrastriate cortex as a result of selective attention on either stimulus location and/or stimulus quality "almost as if the receptive field has contracted around the attended stimulus." This shrinking is the selection of the area of spatial interest in our proposal. The attenuation was quite pronounced in V4, somewhat smaller in IT, and not found in V1. In the words of Moran and Desimone, "the very structure of the receptive field, recently considered to be a fixed property of the neuron, can change from moment to moment in the behaving monkey depending on the immediate task and state of attention." Motter found that about 50% of neurons in V2 and V4 and 20% of neurons in V1 were affected by attention. He found both enhancement and inhibition in V1 and V2, and primarily inhibition in V4. In all three areas, the effect of attention increased with the number of stimuli in the display that had to be inhibited. Those neurons that were enhanced due to attention became more so, whereas those that were inhibited became more inhibited. This is nicely explained by the control over the distribution of winner-take-all units that the inhibitory beam exhibits. In the case of the observed enhancement, the winner-take-all example given earlier shows how the unit of interest has a larger response when compared to the unattended case. If inhibition is applied to a larger number of competitors, the inhibited ones represent a larger proportion of the competition and thus the effect is more pronounced. In V1 and V2 the effect of attended stimuli outside the receptive field was noticed, whereas no such effect was found in V4.

Fuster (1988) noticed the same effects for color-selective neurons in IT. In addition, the cells he studied seemed to store salient aspects of the target stimuli. Even more striking is the discovery that the target may be presented by touch as well as by vision, with similar attentional results (Haenny et al. 1988). The representation of the target may be independent of the actual stimulus. Wolfe, Cave, and Yu (1988) and Treisman and Gormican (1988) both describe marked performance decrements with increasing difficulty of discrimination whether from an increase in the number of stimulus features represented or a decrease in the distance between features along a given stimulus dimension. This can be explained by the speed-up in winner-take-all schemes because of increasing inhibition as described earlier. Note that the above results indicate both behavioral and neural effects of attention have been discovered.

Anderson and van Essen (1987) have recently proposed the idea of "shifter networks" to explain attentional effects in vision; there is some similarity with the inhibitory beam idea of this section. Although Anderson and van Essen maintain that the shifter circuit can be used to account for directed visual attention, scaling and blurring, motion compensation, and the registration problem in stereo vision, this comparison will focus only on attention. I suggest that my beam idea can accomplish the other tasks as well. The Anderson and van Essen proposal requires a two-phase process: First, a series of microshifts map the attentional focus onto the nearest cortical module; then a series of macroshifts switch dynamically between pairs of modules at the next stage, continuing in this fashion until an attentional center is reached. As stated earlier, the receptive field of the output unit selected for attention is the entire subhierarchy that affects that unit's response; any attentional mechanism must therefore necessarily control the subhierarchy. Shifter circuits seem to allow only for attentional focus shifts; there is no apparent method for control of the size and shape of the attentional focus. This is easy to accomplish in the beam proposal because the beam has internal structure that can be manipulated. Also, Anderson & van Essen do not describe how the observed inhibition and enhancement of attended neurons can be accomplished using the shifter circuit. How are the effects of nonattended regions of a receptive field eliminated? It would seem that not only is shifting required to move the attended region into some attentional unit, but it is also required to move the nonattended areas out of consideration. Anderson and van Essen do not describe how this could be accomplished. Both spatial shifts and selective inhibitions are needed, as in the beam proposal, to explain the experimental results described earlier.

Attentional influence on the computation of visual information has the following profound effects:

1. Inhibitory selection of aspects of the input that are not of interest lead to faster winner-take-all decisions among competing units and a final larger single-unit response than in nonattended cases.

2. Selection of maps of interest reduces the exponential nature of the time complexity function to a linear function. Note that if no task information is available then the complexity function is forced to its exponential version of equation (12) and the execution of visual tasks depends on the number of active maps.

3. Selection of the spatial region of interest has two advantages. First, it allows access to high-resolution input information as described above. Second, it selects which spatial regions can contain potential matches with a target. In the complexity function for bounded visual search given in equation (1), $|\mathbf{I}|$ represented the number of possible locations (pixels) of the test image that the target must be hypothesized to be centered on in the worst case. If a scheme that selects spatial regions of interest is available then there is an effect on both of these parameters. First, only the regions of interest need be considered as centering points for the target, rather than all pixels. The term $|\mathbf{I}|$ is replaced by the number of those regions $\mathbf{R}$. Second, if the regions are known, then a large class of translation transformations need not be tried; however, in each region, all the rotations and scalings must be tried. This will be represented by $|\mathbf{R}-\mathbf{S}|$. The resulting complexity function becomes:

$$O[\mathbf{R} \times |\mathbf{T}| \times |\mathbf{M_g}| \times |\mathbf{R} - \mathbf{S}|] \qquad (23)$$

If the task specifies that targets will appear at the same scale and without rotation in the test display, then the transformation term disappears altogether:

$$O[\mathbf{R} \times |\mathbf{T}| \times |\mathbf{M_t}|] \qquad (24)$$

Note that this function is linear in the number of regions of interest (test display items) and the slope is determined by the size of the target and the number of stimulus dimensions that must be considered. Of course, this does not explain how the spatial selection comes about; that is a topic of current research. A more detailed description of how the inhibitory attentional beam functions and its effects on the input abstraction hierarchy can be found in Tsotsos (in preparation). The implications of this for visual search performance will be discussed in the next section.

5. A new explanation for visual search performance

5.1. Preview

So far, this research has demonstrated the following points about the task of visual search:

1. Unbounded visual search is inherently NP-Complete.

2. Bounded visual search is inherently linear.

3. Even after approximations and optimizations, unbounded visual search still remains exponential in the number of active maps (although it is a rather small exponential).

4. Computing the number of output elements and connectivity constraints leads to the need for an attentional selection of spatial areas of interest.

5. The selection of maps of interest, spatial areas of interest, and ranges of interest for stimulus qualities leads to even greater computational savings.

6. Attentional selectivity is inhibitory and can be described as an attentional beam passing through the input abstraction hierarchy inhibiting the irrelevant and allowing the relevant to pass through.

7. Speed functions have been derived for cases where the target is unknown and known (Equations 12 and 23, respectively). If rotation and scale transforms are considered, an additional multiplicative term must be included in each of these equations (it is explicit in Equation 22 only).

The natural question to ask now is whether these conclusions can provide new insights about visual search performance in humans, or at the very least, whether any of their implications can be confirmed experimentally. Motivated by the lack of consensus on the nature of visual search performance, a new algorithm that ties together the above conclusions will be presented.

5.2. An algorithm for visual search

Recently, Treisman has "hedged her bets" (her words) by producing a different version of her theory alongside the one previously described (Treisman 1988; Treisman & Sato 1990). In this new version, top-down inhibition plays a role in providing a different explanation of conjunction search. She speculates that inhibitory attentional influences assist in eliminating distractors from consideration. In the disjunctive task case, only one item would remain, and in the conjunctive case only those items that share features with the target item remain. Several experiments are described that explore specific aspects of this new model. This explanation also appeared in Tsotsos (1987c; 1988) in an attempt to reconcile the experimental evidence of Treisman's older model with complexity arguments. There is a large body of experimental evidence that supports Treisman's old model, however, including Treisman's own data. Is there some way to tip the scale in favor of one of these? The algorithm given below appeared in a slightly different form in Tsotsos (1987c); the version here is an elaboration.

An algorithm for visual search performance in humans based on the seven points made in the preview to this section is presented in a step-by-step fashion (with a new section for each step). The bounded version of the problem is assumed, and changes required for the unbounded version are noted.

5.2.1. Specify targets and task. A set of targets or target images must be defined, a task must be described, and these must be stored in memory. There is growing evidence that there is a representation of the target in the cortex. Haenny et al. (1988) and Maunsell et al. (1988) have found individual neurons of V4 that seem to represent the orientation of a cue in an orientation selection task. Storing orientations is not the same as storing tasks, but this finding is very significant. The tasks that may be specified include detection (is one or more of the targets present in the display?), discrimination (are all the targets in the display the same?), counting (how many targets are there in the display?) [See also Davis & Perusse: "Numerical Competence in Animals: Definitional Issues, Current Evidence, and a New Research Agenda" *BBS* 11(4) 1988.]

5.2.2 Apply attentional map and feature range inhibition. The signals that provide the attentional effect of the task on the processing hierarchy must be generated using the inhibitory beam concept described earlier. This implies that nondistractor elements in a conjunctive display have no effect. For example, if the targets are either a brown 'T' or a letter, for a given display, subjects could attenuate responses to non-brown, non-T, and non-letter stimuli. In a disjunctive positive display, there would always be only one candidate remaining for matching against targets. In a disjunctive negative display, all responses are attenuated, and because there is no differentiation among candidates, all are considered candidates. This leads to the observed linear slope of response with the number of elements in the display. In the conjunctive positive case, say, with a target of a red letter "A," the subject could attenuate non-red and non-A responses, and thus the number of candidates would be exactly the number of distractors plus the target. The conjunctive negative case is similar to the disjunctive negative one; all elements are candidates and none of the candidates match any of the targets. In the unbounded case, there is no benefit from this step, because no a priori expectations are available.

Treisman, among others, has observed a "search asymmetry" in her experiments and claims that this may be used as a diagnostic for determining which features can be processed preattentively (Treisman & Souther 1985; Treisman & Gormican 1988). The basic reason for the asymmetry seems to be the difference between searching for the presence versus the absence of a feature. If a feature is present, then the search is preattentive and immediate, whereas if it is absent, the search is linear. The difference is attributed to the fact that single features can be detected by the mere presence of activity in the relevant feature map, whereas absence requires exhaustive search. Using this diagnostic, Treisman and colleagues have addressed the issue of visual primitives. They claim that color, brightness, terminators, blobs, closure, tilt, and curvature are good candidates, whereas intersection, juncture, number, and connectedness are improbable ones.

If attentional selectivity is applied and affects an early abstraction hierarchy in the manner described in the previous section, the features Treisman claims are primitive represent only what the system can be "tuned" to recognize. None of the units representing the target of the visual search task are left in their "default" state – hence they do not represent that default or primitive feature during the course of the experiment. The set of tunable features is much larger than the set of default features or primitives the system computes because each default feature can presumably be tuned into several slightly differing forms. For example, a unit whose default is to recognize a horizontal line segment of a particular length could be tuned to recognize line segments slightly off the horizontal or slightly longer than the default. In addition, the ability to select attentively the spatial subset of an individual unit means that it may be tuned to respond to a spatial subset of its default tuning. This may have no obvious correlation with that default. A face-recognition unit may also respond well to curved lines if tuned appropriately. In the terminology of this paper, Treisman discovers tokens but not types. However, if Treisman's visual search paradigm rejects a feature, then processing units can indeed not be tuned to recognize it. The conclusion is that visual search can be used only to reject candidate stimuli as features, not to discover their existence. Using visual search in this way only would be a very time-intensive task as well as offering only indirect evidence for particular features.

Therefore it seems search asymmetry is not the good diagnostic that Treisman claims it is. If a subject is told to search for the absence of a feature, the only possible tuning that could be applied is the logical "negation" of the feature. That is, all possible stimulus patterns except the feature would be attenuated. This would lead to no attenuation at all for any of the stimulus elements, because the spatial intersection of the feature with all other stimuli is so large. All elements of the stimulus remain as candidates and this leads to a serial search.

5.2.3. Compute map representations. The computation of map representations takes constant time and is accomplished by the input abstraction hierarchy.

5.2.4. Do steps 5 – 11 for each map subset considered. This step begins a loop that must be executed $2^M - 1$ times in the unbounded case. This is where the "fishing expedition" strategy of Treisman and Sato (1990) fits. If the target is known in advance and is present in the test image, this loop need be executed only once.

5.2.5. Compute receptive field-prototype associations. Associating a prototype with each receptive field takes constant time because there is one processor assembly of $\log_2 VP$ processors allocated for each receptive field. This step computes a necessarily coarse analysis of input receptive field contents and not a detailed one; the detailed resolution data are not available at first. Gleitman and Jonides (1976) and Jonides and Gleitman (1976) show that categorization requires less processing than identification, and thus an initial coarse analysis can be used to locate items for which detailed analysis should be performed. Reaction times are longer for the identification of within-category items than for between-category items. The receptive field-prototype associations, computed in parallel, may be one explanation for this observation.

5.2.6 Do steps 7 – 11 for each candidate. How is a candidate for matching to targets determined? One possibility is that all elements of a stimulus are candidates. In this case pop-out would never be observed, however. The definition of a candidate could be task-dependent. The attentional attenuation followed by coarse analysis described above plays an important role here. Stimulus items that have none of the characteristics of any of the targets are attenuated and those that remain have an associated category because of coarse analysis. The selection of a candidate can be based on a number of criteria: It could be simply a random selection from among the possibilities or the candidate with the largest response. In any case, once it is selected, the attentional spatial beam can be applied in order to access detailed information for a final comparison with the target. Hoffman et al. (1983) claim that automatic detection requires the allocation of spatial attention to the area of the target, in effect, selecting a candidate, even in the pop-out case. In the unbounded situation a similar subsequent attentional action takes place. A first processing pass coarsely locates items and analyses features; the attentional beam can then allow access to detailed information. Wolfe et al. (1989) propose a similar model they call guided search and provide supporting experimental data.

5.2.7. Transform candidate representation to target representation. There is no a priori reason to believe that the representation at the candidate level is the same as at the target level. Thus, there must be time allocated to the transformation process between representations. There is also a connectivity argument as to why an intermediate representation may be required. The number of connections between all of the targets and all of the candidates would be prohibitive, and certainly much greater than the observed average of 1,000. In particular, the worst-case number of connections from the processor layer to each target would be $\hat{N}^3 \times \log_2 \text{VP}$. An intermediate representation would solve the problem of connectivity, because not all receptive fields must be hard-wired to all targets. A switching network plus appropriate control would be required to select a candidate and to route candidate information into the intermediate buffer. Thus, a linear search of candidates is necessitated because each must be stored in this buffer before being matched to a target. Experimental evidence points to a similar conclusion. Duncan (1980) discovered that simultaneous targets interfere with one another, suggesting that there is a limit to the storage or processing of multiple targets. Broadbent and Broadbent (1987) claim that there is only one representational space for targets because they have found that the identification of one target impedes the next.

5.2.8. Do steps 9 – 12 for each target. Targets are dynamic and usually too few to organize; nor do they necessarily share properties on which organization can be done. Thus the optimizations used earlier cannot be applied here. It accordingly seems that the best this architecture can accomplish is a serial, self-terminating search. This is what is observed.

Serial search is required to identify and count more than one target in a pop-out display. This would be true because each of the qualifying targets would pass through attentional tuning and would be present as matching candidates. If, on the other hand a subject is required simply to determine whether a target is present, and there is more than one target, there should be no effect, and pop-out should proceed as if there were only one target. Sagi and Julesz (1986) demonstrated this indirectly by asking subjects to determine whether all targets in the stimulus were the same, given multiple possible targets. This revealed a serial search of targets. They called this a discrimination task, distinguishing it from a detection task, in which it was simply determined whether a target was present. They point out that their results disagree with Treisman's feature integration theory; the results do support the argument in this paper, however.

As the number of targets increases, so does the response time. The loop beginning with step 8 in the algorithm is performed for each target; and, in the worst case, that matching is performed for each candidate against each target. Without any a priori ordering of target preference, and with each candidate being equally strong and likely, the average time to pass through step 8 is proportional to half that number of matches in the positive case, and the full number of matches if there is no target in the stimulus. Thus, there are three contributors to the serial, self-terminating process observed in visual search: the loop beginning with step 4, which is executed more than once only in the unbounded case; the loop of the algorithm beginning with step 6, and the loop beginning with step 8.

5.2.9. Match candidate to target. The time requirements for matching in the bounded case are given in Equation 23 and point to a linear contribution due to the number of maps activated. This conclusion depends on the assumption made earlier that all maps are computed with equal ease. Treisman and Sato (1990) have quantified the contribution to the slope of conjunctive tasks made by different stimulus features: The effect is additive, with the smallest contributor being size, followed by color, motion, and then orientation. This implies that my assumption is not valid: The use of the variable M (and $\hat{M}$) should be replaced everywhere by a function ϕ of the maps. When this is done, the speed function derived earlier (Equation 23) for the single target bounded case becomes:

$$O[\mathbf{R} \times |\mathbf{T}| \times \phi(\hat{M})] \qquad (25)$$

remembering that **R** represents the number of candidates defined by step 6 of the algorithm. Because $|\mathbf{T}|$ is a constant for a given experiment, this leads to a response curve that is linear in the number of candidates and whose slope is determined by the size of the target image and the contribution of the active maps. If there is one candidate, regardless of the number of defining stimulus features, the curve is flat when plotted against distractors. This is precisely what is observed. One prediction arising from this is that the larger the target image, the steeper the slope. In addition, the response time for pop-out displays will increase as the number of stimulus qualities increases.

5.2.10. If match fails then try next candidate.

5.2.11. If match succeeds then continue to next target.

5.2.12. If targets exhausted exit. An exit with matching success means that the task is successfully completed, in cases of discrimination, detection, or counting. An exit with matching failure signals a failure to complete the task positively and that either the task is complete and a reply is negative or that further processing may be needed. This may include retuning the input abstraction hierarchy and the receptive field- prototype association process in order to try the task again. Such issues are beyond the scope of this paper.

It seems that given this algorithm, plus the results of the previous sections of the paper, the case for attention if targets are known a priori is very strong. Treisman's reluctance to accept one or the other version perhaps issues from her commitment to describing both unbounded and bounded search results with a common processing strategy. As is clearly shown above, the processing strategies vary dramatically.

6. Conclusions

Theories of visual perception lack basic principles to guide their development and to test their validity. Two such principles are proposed in this paper, the "complexity level" of analysis and the minimum cost principle. We have demonstrated that significant conclusions about the architecture and performance of biologically plausible visual systems can be derived from the faithful application of these principles.

The implications for computer vision are clear and quite important. The reason that many of the computer vision proposals that use attention have not been entirely satisfactory (see Tsotsos 1987b, for a comprehensive overview) is that a strong argument for the computational need for attentive processing has never been presented. That need must be based on the basic computational inadequacies of spatially parallel, nonattentive visual architectures. The capabilities of such architectures have been derived in this paper for biologically motivated designs (and still largely, but not entirely, apply to nonbiologically motivated designs). The argument for attentive vision, and indeed for the computational modeling of human vision, is now on a solid foundation.

It has been shown that unbounded visual search is inherently NP-Complete in the size of the image. This result is *independent of the implementation*, that is, whether one is considering the brain, a machine, or some yet to be discovered method of implementation, the inherent complexity of the general problem remains the same and an implementation must deal with it. Even after a reshaping of the problem and optimization of the resources, the problem remains exponential in the number of maps. Our analysis has, however, revealed a set of architectural constraints that permit the unbounded visual search problem to be solved within the resource limits of the brain and at observed performance rates. These optimizations are:

1. parallelism of sufficiently high degree;
2. hierarchical organization through the abstraction of prototypical visual knowledge in order to cut search time at least logarithmically;
3. localization of receptive fields, noting that the physical world is spatio-temporally localized and that objects and events, and their physical characteristics, are not arbitrarily spread over time and space;
4. the fact that maps are summarized via a pooled response, using the observation that not all visual stimuli require all possible parameter types for interpretation, and thus leading to separable, logical maps;
5. hierarchical abstraction of the input token arrays so as to maintain semantic content yet reduce the number of retinotopic elements.

Note that some of these optimizations do not represent new concepts and have been common in the literature for some time (see Ballard 1986, for an example of another recent model), but their interpretation in terms of complexity and the resulting quantitative analysis is novel.

It has also been shown that bounded visual search has linear time complexity; several optimizations of this complexity function has also been presented, along with experimental support for its final form.

Applying the minimum cost principle, many further characteristics of the visual system are predicted:

1. Processor columnar organization;
2. Inverse magnification within the processor layer with respect to the retinotopic array;
3. Tokens of visual parameters at high resolution cannot be directly accessed, but must be obtained by the tuning of computing units and through the input abstraction hierarchy;
4. Token coarse coding;
5. Predictions for the overall configuration of the visual system in terms of lower bounds on the size and number of maps and upper bounds on the required degree of parallelism;
6. An inhibitory attentional beam.

Finally, an algorithm at an abstract level of description was presented for the sequence of processing steps performed for visual search tasks that integrates many of the above conclusions.

Although this article dealt almost exclusively with visual search, its conclusions are not confined to visual search. More complex kinds of visual tasks are subject to the same kinds of complexity arguments and their analysis will yield very similar results. This is not surprising, because it is probably true that if any visual task is decomposed into subtasks, some form of visual search will be present in that decomposition. Hence, other more complex visual tasks will have complexity consistent with the complexity of their worst subtask.

ACKNOWLEDGMENTS

Many thanks are due to Allan Jepson, Steve Zucker, and Geoff Hinton for several useful discussions and suggestions. Anne Treisman and Patrick Cavanaugh helped me understand visual search. Greg Dudek and David Fleet read drafts and provided comments. The author is a Fellow of the Canadian Institute for Advanced Research. This research was conducted with the financial support of the Natural Sciences and Engineering Research Council of Canada, and the Information Technology Research Center, a Province of Ontario Center of Excellence.

NOTES

[1] The definition of the term "visual search" commonly used by psychologists is given in section 1.2; a computational definition is presented in section 2.2.

[2] See Dobson & Rose 1985 and Maxwell 1985 for excellent treatments of the methodological problems in both neuroscience and artificial intelligence.

[3] See Garey & Johnson 1979; Pippenger 1978; and Stockmeyer & Chandra 1979.

[4] The condition-for-solvability thesis was also put forward by another mathematician, Alonzo Church (1936) and is hence usually stated as the Church/Turing Thesis.

[5] The notation O( ) stands for "order."

[6] There are many more problem classes; they are not of immediate concern here.

[7] Neisser was motivated by the following question: If there is more than one item of the same kind in the visual field, how are they distinguished? One way is to duplicate processing resources everywhere across the image. If a model of perception were to deal with the entire visual field at once as well as with all the possible interpretations only by using parallel processors for each spatial possibility, it would require a much larger brain and too much experience.

[8] The number of discriminable objects is much larger. For example, a banana is still a banana even though it changes color as it ripens; during color changes, it can still be identified as a banana as opposed to an apple, but it can also be discriminated from other bananas at different stages of maturation.

[9] Surprisingly, Marr omits color and explicit temporal information.

[10] Perhaps the most eloquent argument for the use of hierarchies in defeating the complexity of large systems is Simon's (1982). The use of hierarchies has been pervasive in artificial intelligence for more than twenty years, and also in the neurosciences, at least since Hebb's model (1949).

[11] Hartline's (1940) concept of a receptive field has been extended to the retina by Kuffler (1953) and generalized for sensory processing by Mountcastle (1957), whose general view is adopted here.

[12] Not all scales are represented at each position, however; the instances recognized are smaller with higher positional eccentricity.

[13] Evidence for the inseparability of retinal measurement is summarized by Fleet et al. 1985. A summary of examples of inseparability in other areas is provided by Cowey 1979.

Open Peer Commentary

Commentaries submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.

The theory and practice of attention

Kyle R. Cave

Department of Psychology, University of California—San Diego, La Jolla, CA 92093

Electronic mail: *kcave@cogsci.uscsd.edu*

Tsotsos argues that complexity level analysis is a powerful but neglected theoretical tool, and he applies it to the problem of preattentive vision. He first concludes that bottom-up processing alone cannot cope with the complexity and then goes on to present a model of visual processing.

The inadequacy of purely bottom-up processing. Although Tsotsos is almost certainly right in concluding that some sort of selection is important in visual processing, I do have a question about how he reached this conclusion. In the architecture presented in Figure 2, the comparison between a single receptive field and a single visual prototype involves a number of operations, one for each possible combination of maps. The series of optimizations that Tsotsos applies leaves this situation essentially unchanged: Without knowledge about what type of target is being sought, a separate operation is necessary in each receptive field for every possible combination of maps. Is there not room here for an additional important optimization? For a given visual prototype, only a single subset of maps will be relevant. Why not include a list of the relevant maps with each prototype? For each combination of receptive field and prototype, the list would be consulted and a single comparison would be done, involving only the specified maps. This optimization could sharply reduce the number of comparisons when nothing is known in advance about the stimulus, and thus it brings into question Tsotsos's conclusion that selection is necessary.

Tsotsos goes on to elaborate his architecture in order to account for human performance in visual search tasks. Because he is claiming that complexity-level analysis has allowed him to formulate a new architecture that can account for many aspects of visual processing, it is appropriate to evaluate this architecture as we would any other, including considerations unrelated to complexity.

Similarity to other models. One of the most striking aspects of this architecture is its resemblance to existing models of visual search performance. Tsotsos mentions Neisser's (1967) and Treisman's (Treisman & Gelade 1980; Treisman 1986) distinction between early processes that operate over the entire visual field in parallel and later processes that are limited to a single region at any one time. He adopts the same principle for his system. Also, his parallel stage serves to identify potential targets whereas the serial stage confirms those that have been nominated, just as in guided search (Wolfe et al. 1989; Cave & Wolfe, in press), Duncan & Humphreys's (1989) similarity model and Hoffman's (1979) two-stage model. The resemblance goes further: The feature-range inhibition that eliminates elements with nontarget features is very similar to the top-down inhibition in guided search. (The main difference appears to be that Tsotsos's feature-range inhibition applies only to those elements that have *none* of the relevant features, whereas the top-down activation in guided search allows for a wide range of possible activations for each element, depending on how many of the relevant features it has.) As Tsotsos notes, Treisman (1988) has proposed adding such inhibition to feature integration theory as well. If Tsotsos's analysis stands up, its main contribution to visual attention may not be to produce a new explanation for visual search, as he might have hoped, but to provide new support for these ideas that have been proposed earlier.

Using bottom-up processing to direct attention. Even if the bottom-up, preattentive system is as limited by complexity as Tsotsos claims, it must be designed to produce the most useful information possible when nothing is known in advance about the stimulus. One possible strategy is to identify those locations of the visual field that exhibit large changes in feature values.

Adding mechanisms to detect these feature differences would allow attention to be quickly directed to those parts of the input that are most likely to be important. A demonstration of how these mechanisms can work can be found in the bottom-up component of guided search. The complexity of these mechanisms can be controlled by only allowing comparisons within maps, and also by limiting the distance between locations over which features are compared.

Where is the line between parallel and serial processing? If Tsotsos's architecture is to be a useful account of visual processing, it must give a clearer description of the limitations on parallel processing. For instance, in section 5.2.7. he introduces a separate representation that is used to match a single candidate element with the target. What visual operations are the processors in the original architecture capable of executing, and what operations require this new representation? Just what aspects of visual processing are handled by Tsotsos's array of parallel processors, and what processing tasks are passed off to the target-matching representation? This is just another way of asking the old question of which visual processing operations can be performed in parallel and which must be performed serially. Also, what computational advantages come from separating parallel operations into feature maps that operate independently, and does Tsotsos expect to find such independent maps in the visual system? (In section 3.2 he states that using physically distinct maps can lower complexity, whereas at the end of section 4.4 he suggests coarse coding across parameter types. Isn't there a conflict here?)

In summary, if this architecture is presented as a serious model of visual processing, then Tsotsos should specify its operation in more detail. In doing so, he will have to tackle more specific questions about which computational operations are necessary and how they are accomplished. Perhaps it is in constraining solutions to these specific questions that complexity analysis is likely to make its most important contributions to the study of visual attention.

Complexity at the neuronal level

Robert Desimone
Laboratory of Neuropsychology, National Institute of Mental Health, Bethesda, MD 20892
Electronic mail: *jcg@nihcudec and jcg@nihcu.bitnet*

Although I cannot comment on the formal aspects of Tsotsos's complexity analysis, I think few would disagree with his thesis that computational approaches to visual search immediately run up against a combinatorial problem. Tsotsos's work quantifies just how serious the combinatorial problem is, and he outlines a plan for how this problem can be solved. It is the latter aspect of his work that I will comment on, from a neurobiologist's point of view.

Tsotsos's general approach to the combinatorial problems of vision is to compress and abstract the data that arise from the retina. This is accomplished, in part, by first segregating visual features in separate feature maps that can subsequently be attended to individually. Next, the maps are arranged in a series of hierarchical levels, with an increase in receptive field size at each level. The latter idea – that the visual pathways are organized hierarchically, with increasing receptive field size as you move towards the "top" – is consistent with most neurobiological views on the organization of visual cortex. However, there is much less support for the notion that features are segregated in different maps, especially if the maps are supposed to correspond to different visual areas. Recent physiological and anatomical work suggests that, beyond the segregation that takes place between the "dorsal" spatial system and the "ventral" object recognition system (Ungerleider & Mishkin 1982), most of the segregation of features in the ventral system probably occur at the level of columns or modules within an area rather than across areas (for a review, see Desimone & Ungerleider 1989). Even given some degree of anatomical segregation within a cortical area, it is still not clear whether features such as color or shape are actually represented by different populations of cells within an area, or rather are represented in a "multiplex" fashion by cells sensitive to many different features. If the latter is true, it is even less clear how the visual system could select just one feature for processing at the expense of others, as Tsotsos's model requires. The representation of features is probably the murkiest area in the neurobiology of vision.

Large receptive fields are a second problem that any model of vision must handle, as Tsotsos points out. Although large receptive fields allow for generalization of coding across large retinal areas (and thus tremendously reduce the combinatorial problems of visual search), they incur a cost in resolution, particularly when there is more than one stimulus or feature located in a field. This general issue of how the visual system copes with multiple stimuli in receptive fields is sometimes referred to as the "binding problem" (see Wise & Desimone 1988, for a discussion of the relationship between the binding problem and attention). What is needed is an attentional mechanism that can select specific locations in a receptive field for processing.

In my own work (Moran & Desimone (1985), I have seen evidence for attentional selection of stimuli in receptive fields of neurons in areas V4 and the inferior temporal cortex (IT). If an animal attends to one stimulus and ignores another in the receptive field of a neuron in these areas, the neuron will respond primarily to the attended stimulus. That is, responses to unattended stimuli in the field are suppressed. The result is both a reduction of distracting information and an increase in spatial resolution. We have seen analogous effects outside the spatial domain. If an animal must discriminate a stimulus of one orientation or color from another very similar stimulus (presented at a different point in time rather than at a different point in space), the orientation or color tuning of cells in area V4 is sharpened, resulting in a behaviorally measurable increase in orientation or color "resolution" (Spitzer et al. 1988). As Tsotsos points out, his model of complexity reduction through attention is consistent with our results on both spatial and feature attention. In fact, I presume that his model is based in part on our results.

Although there are many areas of agreement between Tsotsos's model and our neurobiological data, I am somewhat bothered by his notion that a beam of attention is directed from the highest level of cortical abstraction (presumably IT) back to the lowest. At least in the case of spatially directed attention, recent work suggests that the neuronal mechanism is closely related to the oculomotor system. In our studies, we have found that local deactivation of small zones in the superior colliculus, a "classic" oculomotor structure, impairs an animal's ability to attend to a target in the presence of a distractor (Desimone et al. 1989). The effect is spatially specific to the receptive field location of the deactivated zone, that is, the animal is impaired only when the target appears inside the deactivated zone. I should add that we only see pronounced effects when there is a distractor in the visual field, and that the effects do not depend on eye movements. Others have also found evidence for a role of the superior colliculus in spatial attention in the absence of eye movements (Albano et al. 1982; Posner et al. 1985; Kertzman & Robinson 1988). Although it may seem surprising at first that the oculomotor system would be involved in covert shifts of attention, both systems require the targeting of stimuli and so might usefully share some common "hardware." In fact, the effects of a shift of gaze and a shift of covert attention are nearly the same on the visual system; both cause visual processing to be dominated by new input. A computational model that incorporated some of these neurobiological findings would be very useful.

Computation, complexity, and systems in nature

Bradley W. Dickinson

Department of Electrical Engineering, Princeton University, Princeton, NJ 08544-5263

Electronic mail: *bradley@princeton.edu*

The view that dynamic phenomena arising in nature result from computational processes of various kinds has led researchers to formulate and study a wide variety of fascinating problems. For systems whose behavior is described by physical theories, as confirmed by reproducible experiments, the corresponding physical laws of computation may be studied. In this way, ultimate limits on computational devices may be examined, particularly in the sense of various performance bounds. Crucial differences in perspective – for example, the important distinction between the nondeterministic models for ideal computing machines and the probabilistic models used in physical theories for microscopic systems (quantum mechanics) and complex macroscopic systems (statistical mechanics) – are the source of important practical and philosophical problems.

Turning this approach around, interesting questions about the behavioral constraints on physical systems that perform particular computations may be investigated. For example, it might be argued that the Church-Turing thesis should be included in any list of fundamental physical principles, along with such things as the first and second laws of thermodynamics, because it can be viewed as providing a basic limitation on the behavior of (some suitable class of) physical systems. Such considerations can lead to rather subtle conclusions about the form of mathematical models suitable for describing the dynamic behavior of physical systems (Pour-El & Richards 1981; 1982).

Add complexity considerations to those already mentioned and a much richer, and probably more relevant, class of problems arises. The space and time complexity properties of computational problems provide insights about the practical capability of providing solutions in the sense that important *resource* requirements (usually taking the worst case over all problems of a fixed size) are characterized. For certain intractable problems (i.e., for NP-Complete problems according to the widely held hypothesis that P ≠ NP), the time requirements for solving problems grow faster than any polynomial function of the problem size, whereas exponentially growing computation time suffices. According to a strengthened version of the Church-Turing thesis, with an appropriate notion of the resources required by a physical computing system, NP-Complete problems remain intractable for such systems (Vergis et al. 1986). For physical systems whose behavior is governed by mathematical models that can be efficiently simulated by a digital computer, the strengthened thesis can be proved. For other systems, such as those modeled using probability or chaotic behavior, verification of the strengthened thesis remains as an open problem.

To determine the general laws of biological computation may be viewed as the ultimate challenge of biocybernetics. For most physiological systems of interest to neuroscientists – including the human visual system, and in particular the visual search system dealt with in Tsotsos's target article – observed behavior, often described in broad categorical terms, is described in terms of empirical, external (i.e., input/output) models; structural information about the nature of the underlying internal models (perhaps corresponding to a partial description of the neural "wiring diagram" and a description of behavior at the single neuron level) should be expected to provide only a crude guide to computational limitations. More significantly, the validity of some general system design principle (Tsotsos's minimum cost principle falls in this category) is usually not clear, but rather remains an assumption invoked for analytical convenience. Finally, the requirement to pose biological computations in a form compatible with the precise mathematical formulation used in computational complexity analysis can easily lead to results that depend in essential ways on immaterial, unrecognized combinatorial intricacies.

From my perspective, the calculations and reasoning that lead Tsotsos to his conclusions about visual search are not particularly persuasive, although I am not equipped to evaluate the conclusions in relation to the body of physiological and psychological evidence. I suspect that the complex processes of adaptation appearing at even the lowest level of visual information processing lead to a much broader range of possibilities than is suggested by Tsotsos's emphasis on an attentive versus nonattentive dichotomy of search processes. Knowledge representation, viewed from a symbolic "AI" perspective or from a connectionist "PDP" perspective, must certainly be important in guiding search processes. I am not convinced that computational complexity considerations are crucial to an understanding of human visual search capabilities, or that they provide a good motivation for the kind of conceptual algorithm for the visual search process that is developed in Tsotsos's article.

In conclusion, I wish to emphasize that my comments are intended to counter the general impression I got from Tsotsos's article that computational complexity considerations are *the* key to the development of an understanding of visual search. Tying together a broad range of physiological and behavior evidence with models of various kinds is a desirable goal of all efforts to "reverse engineer" this process. The conceptual models developed in Tsotsos's article can be expected to benefit those who might attack the problem from other perspectives – computational learning theory is one that comes to mind – as well as researchers in computer vision.

Task-dependent constraints on perceptual architectures

Roy Eagleson

Centre for Cognitive Science, University of Western Ontario, London, Ontario, Canada, N6A 5C2

Electronic mail: *elroy@uwo.ca, elroy@uwovax.bitnet*

There is little to criticise in Tsotsos's target article taken as an argument for certain 'top-down' influences in perception, and in visual search tasks in particular. The historical support for attentional mechanisms is well known in the psychology and psychophysics of perception. In computational vision, Tsotsos's complexity-level analysis suggests new arguments against certain strictly bottom-up models.

Search tasks provide an exemplary arena for investigating the complexity of vision, partly because of the amount of empirical research that has been done on this perceptual task. Applying complexity-level analysis to this task also raises the question of the tractability of whole classes of perceptual problems, because any general model of perception has a search-space component that immediately raises the problem of intractability. Thus, Tsotsos has highlighted an issue that is often swept under the rug. This commentary will focus on how he has generalized this treatment of visual search to encompass perception in general.

In vision, the dimensionality of the input "feature space" is high. To make the problem of visual perception tractable, this space must be systematically pruned in a way that effectively limits access primarily to relevant information. However, when dealing with the complete manifold of sensory inputs, it is in general not possible to state a priori that any particular part of the available information will not be relevant to the problem. This, in microcosm, is the ubiquitous "frame problem" that always haunts general AI (Pylyshyn 1987). Clearly, the search space must be restricted. Tsotsos describes an architecture that restricts the search space in a way that is effective in certain tasks. However, as a general solution it is open to criticism.

Citing a "minimum cost principle," Tsotsos proposes that computational efficiency in vision is increased by the use of a parallel matching process, and by a restriction on the resolution and spatial extent of feature maps. This is accomplished through the use of a hierarchical organization of both the input maps and the matching prototypes. Attention, on the other hand is viewed as a selective tuning to relevant feature maps "through the input abstraction hierarchy" in order to increase their resolution. This additional constraint takes the form of an inhibitory beam that restricts access to information that lies in a local region of the feature space. These constraints are shown to be sufficient to reduce the complexity of the visual search tasks selected by Tsotsos. A question remains, however; do they apply to more general visual search tasks, or to problems like computing the spatial relationships of features?

The low-level transduction of features can be formalized using the compact set-theoretic notation adopted by Bennett et al. (1989). On the abstract space X (spanning the arbitrary perceptual dimensions relevant to the task), μ_x is a measure class for some observation that is applied to the input signal by the mapping $f \mapsto \int f d\mu$, defined over the receptive field of this abstract "feature detector." The specification of these measures is the primary goal of vision research on the transducer level, namely, how to specify the primitive features. In addition to asking what measures exist, a bottom-up tradition would specify the transformations or operations that would map these measures to more abstract representations. Using this terminology, top-down influences would amount to a control structure, or operations, on these mappings.

There is an additional use for attention that Tsotsos does not consider, namely, its role as a spatial index. An indexing mechanism is required to encode the conjunction of different feature types in visual search tasks, or to encode the spatial relations among a number of features. This requirement arises in tasks as basic as contour parsing. For example, in the perception of topological spaces (i.e., contours, surfaces, and so forth), a spatial index provides the role of a parametric variable; it is required in order to bind together entities that correspond to a continuous connection over the response manifold of feature detectors at each point. (For example, a contour can be described by a systematic connection over the response of appropriate discontinuity detectors.)

This type of formulation characterizes a class of perceptual problems that are not solved by cascaded transformations on the input space. They require an algorithm for describing how to encode the scene (e.g., what primitive features are connected in order to form a pattern). They are required for perception in general, as well as for visual search. They are needed, in particular, for executing what Ullman (1984) calls visual routines, for encoding the spatial relations among places in an image containing arbitrary features as well as for tracking moving patterns (Pylyshyn 1989; Pylyshyn & Storm 1988). Extending this argument in more general pattern recognition, Pylyshyn and Biederman (1988) informally refer to the techniques used by expert classifiers as "knowing where to look" in the input feature space. This can be specified by the current data available in the scene, and through acquired procedural knowledge. This type of task-dependent search constraint is fundamentally different from reducing the search space to local regions by an inhibitory mechanism.

In the case of visual search, Tsotsos's constraints are appropriate for feature matching, when the problem *can* be solved by a hierarchy of transformations. In problems where this is not the case, there must be a provision for task-dependent constraints in the perceptual architecture. Thus, this commentary pertains to the general applicability of Tsotsos's "minimal cost principle," which is not the major thrust of his paper. Tsotsos's general argument is not weakened by questions about the form of the imposed constraints. Theoretical and empirical research should be combined in his framework to evaluate perceptual models intractability.

ACKNOWLEDGMENTS

I would like to thank Zenon Pylyshyn for suggestions on this commentary and related research which was supported by the Canadian Institute for Advanced Research, the Ontario Information Technology Research Centre, and by NSERC operating grant number A2600.

What are the insights gained from the complexity analysis?

Jan-Olof Eklundh

Department of Numerical Analysis and Computing Science, Royal Institute of Technology, S-100 44 Stockholm, Sweden

Electronic Mail: *joe@bion.kth.se*

There is no doubt in my mind that the task of visual processing in general should and could be studied from the viewpoint of complexity. Nevertheless, such an analysis has not really been performed in the literature on computational vision, except for earlier attempts by Tsotsos himself. There seems to have been an implicit assumption that a high degree of spatial parallelism would overcome all the complexity issues, even though it was pointed out early (e.g., by Neisser 1967 and later by Ullman 1984) that such a view could not explain the performance of all visual tasks. The analysis in the target article is accordingly welcome. What can be discussed is the insight we gain from it. Is the analysis providing us with important constraints and an architectural design, as the author claims? Let us consider this question of what the analysis really buys us by going through the target article step by step.

The very first substantial reduction of complexity obtained, in section 3.2, stems from the constraint that the receptive fields are contiguous. This observation certainly can not be contested, but it is not startling. Tsotsos starts out with a formulation of the search problem that is based on arbitrary collections of location/measurement pairs, a formulation that fails to capture *any* structural properties of a world of coherent physical matter. The insight gained here is perhaps that there *exist* architectures that will create complexity problems, but this is not much to spend words on.

The next factor to be considered in formula (4) reflects the exponential dependence on the number of visual maps. We are now coming to a problem for which the complexity argument bears substantially more weight. First, there is no obvious structure in the problem or in the environment that allows us to avoid the exponential structure trivially. Second, the discussion leading tc the suggestion about pooled responses and abstraction hierarchies is indeed plausible. It is perhaps not entirely motivated to say that optimizations due to complexity predict these features, but the complexity analysis certainly supplies us with good arguments.

Similar remarks can be made about the conclusions based on the minimum cost principle in section 4. The predictions made are supported by the complexity argument, even though they hardly *follow* in a strict logical sense, because other architectures are in principle possible (as is also pointed out early in the target article). For example, consider the suggestion about an inhibitory attentional beam. From a computational viewpoint, this idea is very interesting and satisfies the stated criteria. However, in its present form it hardly captures the dynamics of a computational visual system responding to continuously varying input, which is the real-life situation. So the proposed model is not as much a prediction as a model that satisfies the conditions given.

To conclude, I think that it is both useful and necessary to apply complexity arguments in modeling a visual system in computational terms. Much insight can thereby be gained. However, it is not equally clear in which directions the implications go. Can we make precise predictions about the possible architectures or can we just check which models (obtained by other means) satisfy the complexity bounds? Be that as it may, the approach should be pursued.

Is unbounded visual search intractable?

Andrew Heathcote and D. J. K. Mewhort

Department of Psychology, Queen's University at Kingston, Kingston, Ontario, Canada K7L 3N6.

Electronic mail: *heathcot@qucdn.bitnet and mewhortd@qucdn.bitnet*

Tsotsos defines two types of visual search: *bounded,* in which a representation of the target is used in the search, and *unbounded,* in which such a representation is not used, as in an odd-man-out or unique-item problem. Tsotsos equates *unbounded* visual search with unbounded matching when using an invariant prototype. Because unbounded matching is NP-Complete, he claims that unbounded search is also NP-Complete. His claim permits him to apply an analysis of matching to search.

We cannot prove that, in general, unbounded matching is not equivalent to unbounded visual search. However, we can show that Tsotsos's only example of unbounded search can be solved by a tractable algorithm. Specifically, for each pixel in the image, calculate the absolute difference between that pixel's measure and each of the other pixels' measure. Sum the resulting absolute differences. Repeat the operation for each relevant dimension, and sum the sums across dimensions. Pixels from the unique item will have the largest total (see Cave & Wolfe, in press, for a similar algorithm).

The complexity of our algorithm is O(M × N × (N-1)), where M is the number of types of measure and N is the number of pixels. Hence, Tsotsos's only example of unbounded search can be computed in polynomial time: It is not NP-Complete like unbounded matching. Tsotsos has not given us an example of intractable search; hence, the optimizations he developed may be useful to promote efficiency, but are not necessary to ensure tractability.

What does complexity analysis tells us about bounded visual search? Tsotsos suggests that bounded search can be solved by matching target templates to the image, where each template is centered over each pixel. He claims that "the inherent computational nature of the problem strongly suggests that attentional influences play an important part." Treisman's (1988) model could execute this algorithm with a serial attentional operator and Tsotsos appears to equate his algorithm with such a serial instantiation. However, the complexity analysis does not imply serial processing. A plausible alternative is spatially parallel detection by limited capacity filters. Tsotsos acknowledges a role for adaptive filters in his proposed image-abstraction hierarchy; he thereby provides a mechanism by which a parallel detection algorithm could be instantiated. If we stipulate that the filters are tuned by practice, we can avoid the combinatorial explosion implicit in proposing parallel filters at each point in the image for all target types. Thus, the "inherent computational nature of the problem" suggests little about the role of attention in bounded search.

The algorithm for bounded visual search has "linear time complexity in the size of the image." Tsotsos cites the experimental results reviewed by Treisman (1985) in support. Treisman's results, however, were linear in the number of objects (or groups of objects, see Treisman 1982) rather than in the number of pixels. Tsotsos uses number of pixels and size of image interchangeably (see sect. 2.3, para. 4). Hence, he cannot use Treisman's data for support. In addition, search times do not appear to be strictly linear under fine-grained analysis (Pashler 1987).

Complexity analysis reveals that most of the computational cost of visual search results from the large number of possible pixel sets. Tsotsos reduces the number of pixel sets to be searched in two ways. First, only sets containing contiguous pixels are considered. Second, an image-abstraction hierarchy produces a compact representation of the input image "by reducing the resolution of the visual image and simultaneously abstracting the input to maintain its semantic content." The compact image contains fewer pixels than the original; hence, searching it is cheaper.

Preprocessing of the image to produce a compact representation will result in the loss of some information and hence should increase the likelihood of errors in the search process. To increase accuracy, Tsotsos suggests that potential matches found by searching the compact image are checked by analyzing the corresponding area in the original image. Access to the original image is gained by changing the processing carried out by the image-abstraction hierarchy to produce an inhibitory winner-takes-all "spotlight." Note that this scheme is similar to Wolfe et al.'s (1989) guided search model: Both use a coarse parallel analysis of the image to direct a fine-grained serial analysis.

A final word of caution. Tsotsos claims that considerations of "computational complexity . . . lead directly to 'hard' constraints on the architecture of visual systems." We agree that tractability is a real constraint. When comparing tractable algorithms, however, Tsotsos clearly prefers simplicity and assumes that simplicity has the logical status of tractability. For example, he chooses an inhibitory rather than excitatory winner-takes-all scheme for attentive selection (both being tractable) because of a savings in computation. When comparing tractable algorithms, however, the less complex algorithm is not necessarily the one that will be used by biological vision. "Because a simple task could, theoretically, be handled by a simple mechanism does not mean in fact that the brain handles it that way . . . in the complex brain of a higher animal other mechanisms may insist on getting into the act" (Hebb 1958, p. 453). What is economical for one problem may lead to costs for another.

Analyzing vision at the complexity level: Misplaced complexity?

Lester E. Krueger and Chiou-Yueh Tsav

Department of Psychology, Ohio State University, Columbus, OH 43210-1222.

Electronic mail: *ts0340@ohstmvsa.ircc.ohio-state.edu*

According to Tsotsos, the computational constraints on visual search "argue for an attentional mechanism that exploits knowledge of the specific problem being solved" (Tsotsos's abstract). In this commentary, we argue that Tsotsos's analysis is questionable in three respects. First, it is conceivable that noncomputational processes ("smart" perceptual mechanisms: Runeson 1977) enable the perceiver to escape some of the constraints Tsotsos cites. Second, the decision processes involved in interpreting the matching or mismatching features obtained by the perceptual processes are more complicated than Tsotsos allows. Thus, Tsotsos has misplaced the complexity; the truly complex processes in visual search may be at the decisional level (Eriksen et al. 1982), rather than at the perceptual level. Third, attention may rarely produce a top-down tuning of feature extraction or comparison, as depicted by Tsotsos, and when it does the effect may involve perceptual enhancement more often than inhibition (Proctor 1981). Each of these three points will be addressed in turn.

1. Noncomputational perceptual processes. Tsotsos states that the general problem of visual search is computationally intractable, and that visual search is only made tractable and biologically plausible through approximations and by optimizing the resources devoted to visual processing. Tsotsos implicitly rules out the possibility that noncomputational processes may provide viable alternatives to computational ones. Runeson (1977) has described tools, such as the polar planimeter (used to measure the area of irregular shapes), that perform no explicit computation, yet extract complex variables from the environment. Similar smart mechanisms within the perceiver may "directly register complex variables" (Runeson, p. 172). Thus, in visual search, brain mechanisms may "resonate to" the matching items in the display (Gibson 1966; 1979). Although smart perceptual mechanisms are more likely to operate in an analog than a digital manner, they are not required to do so (Runeson 1977); their key property is that they "capitalize on the peculiarities of the situation and the task, i.e., use shortcuts, etc." (Runeson, p. 174); for example, "there are many systems in which global minima can be found using only local interactions" (Marr 1982, pp. 186–87). Highly efficient, hardwired, specialized modules (Fodor 1985; Marr 1982) or smart mechanisms (Runeson) may thus obviate the need for the softwired, general-purpose attentional mechanisms proposed by Tsotsos.

An important constraint cited by Tsotsos is that "each neuron can receive input from about 1,000 other neurons and can provide output for about 1,000 other neurons, on average" (sect. 4.4, para. 2). However, if brain fields or other large entities are active and effective, then such a constraint need not apply. Evoked neuromagnetic fields, for example, have been mapped in the visual, auditory, and somatosensory cortices (Kaufman et al. 1984; Okada et al. 1984); it is conceivable that these patches of neuromagnetically active brain tissue provide smart mechanisms for perceiving, attending, and related activities.

Tsotsos writes that "the speed-up due to parallelism is clearly significant, but it surely cannot be as large as the number of neurons in the brain, 10^{10}" (sect. 2.5, para 6). However, an alternative view would take as the basic unit not the neuron, but the far more numerous molecule, atom, or subatomic particle. Smart mechanisms may make use of the mechanical, chemical, neuromagnetic, and/or electrical properties of portions of the brain in undreamed-of ways, so as to perform, in effect, certain computations, and to do so virtually instantaneously (Runeson 1977). The cochlea provides a good example of such a possibility. As depicted in the place theory of pitch perception (Bekesy 1956), mechanical properties (stiffness, width at various intervals) determine where on the cochlea a particular auditory frequency will produce its maximum displacement.

Tsotsos depicts perceptual processes that might well be simplified and streamlined by smart mechanisms. He postulates (section 2.2) that both the sum of differences, diff(p), between the test item and target item, and the cross-correlation of matching features, corr(p), are considered by the perceiver. These two measures vary in an apparently independent manner (see table at end of section 2.2) for the Test Items A to F shown in Figure 1 of the target article, as would be expected if the two measures were sensitive to truly independent factors. However, for a fixed size of stimulus, the number of matching features (sameness count) is necessarily inversely related to the number of mismatching features (difference count). The reason that Tsotsos's two measures are not more closely related is that in computing the cross correlation, he assigns a value of 1 to darker (figural) elements, and a value of 0 to lighter (background) elements. Thus, only the cross product for matching darker (figural) elements can lead to an increment in the cross-correlation measure; the cross product is zero for positions in which the lighter (background) elements match. However, in visual search, figural or target items may be either lighter or darker than the background elements; either type of figure can be detected about equally well. Thus, Tsotsos's cross-correlation measure is rather dubious. A highly efficient (i.e., smart) mechanism would probably rely on only a single measure – either the sameness count (Eriksen et al. 1982) or the difference count (Krueger 1978) – in deciding whether a match was present.

2. Decision processes. Would that visual search were (as Tsotsos depicts it) so simple that it ended as soon as a sufficiently high sameness count and/or a sufficiently low difference count was obtained. In actuality, the perceiver faces a far more difficult situation. In odd-man-out comparisons, for example, the true difference count between two adjacent mismatching items may be underestimated early in processing, owing to delayed features (Eriksen et al. 1982; Krueger & Chignell 1985), and then overestimated somewhat later in processing, owing to the effect of internal noise (Krueger 1978). Tsotsos's model "includes the possibility of noisy or partial matches" (sect. 2.2, para. 1), but it is primarily deterministic in tone and orientation. Tsotsos takes no account of stochastic processes that produce variability in the arrival rate or time of features early in processing and perturbations in feature extraction and comparison processes somewhat later. He concedes (section 5.2.12) that failure to detect a match in the display may lead to rechecking (see Krueger 1978), but he states that "such issues are beyond the scope of this paper." Such issues may be quite important for his model, however, because they may determine the extent to which visual search is computationally tractable. In particular, variability may hobble the attentional mechanisms that Tsotsos uses to make visual search biologically plausible.

3. Cognitive impenetrability of perception. Whether an inhibitory attentional beam operates in a top-down manner so as to select stimulus qualities (patterns or feature maps; spatial areas) of interest, as Tsotsos claims, is doubtful. Top-down, cognitive processes may affect where a perceiver is attending in space, but they typically have little effect on what features are extracted or compared (Fodor 1985; Krueger 1989; Marr 1982). A seeming exception to the latter rule is Proctor's (1981) priming principle, which posits sensory tuning at feature extracting or comparison owing to the prior presentation of a (priming) character (for supporting evidence see, e.g., Chignell & Krueger 1984; Proctor & Rao 1983). However, this tuning may be due to a lower-level sensory or perceptual persistence rather than to attention per se. Furthermore, the tuning involves not an inhibition of nonfavored patterns, as Tsotsos favors in his model, but a facilitation or enhancement in the encoding of the expected (primed) pattern (see, e.g., Chignell & Krueger 1984; Proctor & Rao 1983).

Complexity is complicated

Paul R. Kube

Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093.

Electronic mail: *kube%cs@ucsd.edu*

In sections 1 and 2 of his paper, Tsotsos defines "unbounded visual search" (UVS) and presents arguments that it is intractable; in sections 3 through 5, he uses this result as part of an analysis of trade-offs among resources in a biologically plausible architecture for vision. I think that the arguments in the first sections are somewhat unsatisfactory for reasons I will discuss below. However, I think much of the complexity-satisfaction analysis in the later sections is useful and important, and so I will also suggest a way it can proceed independently of the prolegomena.

I will argue that

1. Under a reasonable assumption about realizable photoreceptors, UVS is not NP-Complete for a visual system;
2. Even if UVS were NP-Complete, it can't be claimed to be exponential;
3. In any case, UVS is not a satisfactory model of a hard visual search problem.

I will consider each of these points in turn and close with some constructive comments about how Tsotsos's project can proceed in spite of them.

1. Tsotsos is considering the question: "How computationally difficult are the tasks presented to subjects?" (sect. 2.1, para. 1). Because his goal is to deduce constraints on the architecture of the visual system from the requirement that it solve hard problems as best it can, it is reasonable to consider its behavior in the face of a problem that will challenge its computational resources. It should not seem germane to the analysis of a visual system's architecture that a task might be hard merely because it requires making brightness discriminations that are beyond the dynamic range of its photosensors, for example, whereas it is germane that the task require discriminating spatially complex brightness patterns. To this end, Tsotsos defines the "unbounded visual search" problem (UVS) and states (sect. 2.3) that it is NP-Complete, by reduction from the Knapsack problem.

However, it is important to note that Knapsack is not NP-Complete "in the strong sense" (Garey & Johnson 1979, p. 95); it has a known polynomial-time algorithm if the magnitude of the integers in the set isn't allowed to grow as an exponential function of the size of the set (Bellman 1954). Similarly, UVS is NP-Complete only if the fixed-precision numbers m_j representing image parameters are unbounded by any polynomial of the size of the test image **I.** Now retinal photoreceptors have a limited dynamic range, and the best video cameras deliver only 10 or 12 bits per pixel independent of total sensor area; the requirement that the dynamic range of the individual photosensors grow exponentially (or even polynomially!) with retinal area seems truly unrealizable. But if it can't be met, then the system behind the sensors can't even be *given* NP-Complete UVS problems to solve, so there is no reasons for its architecture to be optimized to attempt to solve them. It would seem preferable, in this context, to consider a visual problem that is NP-Complete in the size of the retina alone; but see below.

2. Tsotsos appears to need an exponential visual search problem to start his complexity-satisfaction analysis in section 2.5 (see Equation (2)); and he claims (sect. 2.3, para. 4; sect. 2.5, para. 1; sect. 2.5, para. 2; and elsewhere) to have shown that UVS is inherently exponential. However, even if UVS were NP-Complete in the relevant sense, to show that it has an exponential lower time bound on a deterministic machine would be to answer the most important unsolved question in complexity theory – viz. whether P = NP – and Tsotsos hasn't done that. Of course, Tsotsos is aware of this (e.g., sect. 1.4, para. 5); but remarks such as "it has been shown that the only possible solution [to the NP-Complete Knapsack problem] is to search through all possible subsets of numbers in the list" (sect. 1.3, para. 7), though simply false, suggests that he does want a provably exponential problem for his analysis. If so, given the present state of the art in the theory of complexity, he needs one that is *harder* than NP-Complete. One could describe a more or less plausible, provably exponential search problem on a binary retina (eliminating the photoreceptor dynamic-range problem); but below I will argue instead that such a problem is not needed for the most interesting parts of his complexity-satisfaction analysis.

3. Besides not being plausibly NP-Complete and not presently provably exponential, UVS appears to be an unnatural visual search problem. In section 2.2, paragraph 2, Tsotsos says that unbounded visual search is search "in which either the target is explicitly unknown in advance or it is somehow not used in the execution of the search," and this seems reasonable. However, his formal description of UVS has neither of these properties, because it requires for its solution that the target **T** be both known and used – in particular, the spatial structure of **T** must be known in order to compute the functions **diff** and **corr.** Given this, the requirement that the spatial structure of the target not be used in the obvious way to prune the search seems an unsatisfyingly artificial restriction on the solution to the problem. And it's hard to see what the result has to do with the intuitive idea of "odd-man-out" search, where the problem appears to lie in computing what the target should be, not finding which subset of the image best fits a single, already given target.

Two classes of more appropriate hard visual search problems suggest themselves: Those in which the system is really required to search every subset of the image, thus guaranteeing exponential (not just NP-Complete) time complexity; and those in which the problem turns on finding the best match between the image and members of a large set of stored visual prototypes. Intuitively, these are mutually exclusive problem classes, because the structure of each visual prototype should permit computing a match while restricting image accesses to a polynomial in image size (that UVS tries to have it both ways is what makes it unsatisfactory). But if the use of prototypes for search precludes the need to search all image subsets, then Equation (2), which relates them,is inappropriate as a starting point.

I suggest that the second class of problem – unrestricted, prototype-directed visual search – is the right choice, and that Equation (7), not Equation (2), should be taken as the starting point of complexity-satisfaction analysis. If prototypes represent shape only, then before being matched with an image each one must be transformed by scaling, translation, and perhaps rotation, and it must have specified for each which subset of the set of possible image features defines the contrasts that determine its structure. (An edge can, in general, be determined by contrasts in any combination of brightness, motion, stereo disparity, color, texture, and so forth (see, e.g., Cavanagh 1987). Ignoring rotation, this immediately gives as a complexity function for unrestricted search $O(N^3 \times 2^m \times VP)$, which is identical to Equation (7) on a hexagonal retina of order N. The analysis of complexity satisfaction can be resumed from this point. Note that in this framework, all the differences in complexity between "odd-man-out" and target-directed search are due to the difference between potentially having to search the entire transformed visual prototype knowledge base or not; and this will typically be a substantial difference.

The result is a problem that isn't exponential, or even NP-Complete, in image size. But that's all right. It is still a hard enough problem to put interesting constraints on the architecture of a system that wants to solve it, as the rest of Tsotsos's analysis shows.

Probability theory as an alternative to complexity

David G. Lowe

Computer Science Department, University of British Columbia, Vancouver, B.C., Canada V6T 1W5.

Electronic mail: *lowe@vision.cs.ubc.ca or lowe%ubc@relay.cs.net*

To create a scientific theory of vision – as opposed to a theory of a particular biological or computer visual system – one must understand the computational constraints faced by any system performing particular visual tasks. Tsotsos's target article is one of the few to take such a general approach, and it has deservedly attracted considerably attention and interest.

Unfortunately, complexity theory can give us only rather weak constraints on the design of a practical system. It is only used to rule out the exact solution of the most difficult possible (worst-case) instance of a problem. There is no requirement that biological systems have perfect performance in all possible cases, so they are not generally constrained by complexity theory. As Tsotsos points out, complexity theory is still useful for ruling out certain problem *definitions* as intractable, and this is the useful role that it plays in his paper. However, he overstates his case by saying that "after all, the worst case does occur in practice as well." In fact, the worst case addressed by complexity theory typically has a vanishingly small chance of occurring during a system's lifetime and is therefore unlikely to influence practical design.

In the second half of the article, Tsotsos moves away from complexity theory to the stronger constraints embodied in algorithmic analysis. He uses this more specific analysis to make many valid points about the computational importance of attention. The one missing ingredient here is a justification for the importance of visual search. It is simple to set up a computationally challenging laboratory task, but its constraints on the design of visual systems depend on whether the task is necessary for achieving useful goals in the interpretation of natural images. Some forms of visual search are clearly of great value for higher-level tasks (e.g., recognition), but it may not be necessary to perform them in the full generality specified by Tsotsos.

There is an alternative theoretical tool for studying the general computational constraints on vision that I have chosen to use in my own work. This is to analyze visual performance in terms of probability theory, which allows for explicit trade-offs between visual goals and computational constraints. The probabilistic approach is based on the assumption that perfect visual performance is impossible (due to inherent ambiguities in the data) and that therefore the goal of a system is to maximize the probability of making correct interpretations (Lowe 1990). If successes and failures are weighted according to their survival value to the animal, then it can be argued that this is the design criterion imposed on biological visual systems by evolution. To the extent that all visual systems face the same ambiguities in their input, they will all need to incorporate the same probabilistic inferences to optimize performance.

It might seem that the probabilistic approach requires too much knowledge about every possible input to a visual system. However, there are many cases in which the properties of vision itself (e.g., the projection function from 3-D to 2-D) constrain the probabilistic distribution of inputs. For example, grouping elements in a local neighborhood rather than an entire image can be justified because the probability that two features arise from the same object varies as the inverse square of their separation (see Lowe 1985; 1987, for a full analysis). This provides a justification for the local "spotlight of attention" that shows it can enhance performance as well as merely limit computation. Similar methods can show the value of grouping image features according to connectivity, collinearity, parallelism, and symmetries. Full system optimization requires trading off the probabilistic value of each grouping with the computational cost of its derivation. It is notable that connectionist learning procedures can also be seen as optimizing a very similar probabilistic function over a space of inputs (Hinton 1989) and that they thereby function as an empirical approach to deriving the same set of probabilistic inferences.

Tsotsos has taken an important step in the difficult task of deriving general constraints on the computational structure of any visual system. The logical extension of this work is to consider not only worst-case performance, but also the average case that is addressed by probabilistic analysis.

Support for an intermediate pictorial representation

Michael Mohnhaupt and Bernd Neumann
Universität Hamburg, Fachbereich Informatik, D-2000 Hamburg 50, West Germany.
Electronic mail: *mohnhaupt@rz.informatik.uni-hamburg.dbp.de and neumann@rz.informatik.uni-hamburg.dbp.de*

First, we will make a general comment about the importance of complexity-level analysis for vision; second, we will draw some important conclusions from the visual search scheme proposed in Tsotsos's target article.

Complexity-level analysis is a logically necessary consequence of adopting the information-processing paradigm for vision and cognition. One cannot think in information-processing terms without recognizing the relevance of complexity barriers. Complexity-level analysis can contribute to our understanding: (1) because it is a powerful instrument for choosing among alternative architectures as impressively shown by the target article, and (2) because its implications offer a strong basis for falsification. The potentially dramatic implications have become evident recently when Chen (1982; 1989) hypothesized that topological features can be extracted by the visual system faster than simple geometric features. If this were true, the information-processing explanation would be doomed as the computation of topological properties is known to be more complex and the computation of local geometric features (see Minsky & Papert 1969). Fortunately, in Rubin & Kanwisher (1985) and Liu et al. (1989) counter explanations for Chen's experiments are provided.

In the following we focus on one important point of the proposed algorithm for visual search in Tsotsos's section 5.2. A crucial step is the transformation of candidate representations to target representations (section 5.2.7). This is particularly difficult if the algorithm is applied to a more general situation in which the target is not given explicitly in the stimulus, but through other information sources. For example, it might be given through spatiotemporal context, general expectations about the scene, or language cues. This includes situations where only prototype information is available and the detailed shape of the target is unspecified.

The target article points out the need for intermediate representations to match items against targets as a consequence of complexity-level analysis. This ties in with interesting results concerning the type of information that might be contained in such an intermediate representation. From the work of Rosch et al. (1976) and Rosch (1978), it is well known that information about basic-level categories can facilitate perception, but priming with a superordinate category does not lead to a significant speed up. Basic-level categories are the highest level of abstraction for which there is a clearly definable visual shape. Rosch and coauthors conclude that top-down control is performed by forming mental images, which cannot be generated from superordinate categories. Hence, the intermediate representation for matching targets against items, which is a necessary consequence of the complexity level analysis, should not be considered above the mental-image abstraction level.

We view this as additional support for a spatial or spatiotemporal pictorial buffer and associated local and parallel processes whose main task is to combine bottom-up and top-down information. There is growing evidence for a shared imagelike representation from psychological (see, e.g., Finke 1985) and psychophysical experiments (see, e.g., Farah 1985). In addition, there is recent work in artificial intelligence investigating the computational properties of pictorial representations and local and parallel processes working on such representations (see Larkin & Simon 1987; Steels 1988; Gardin & Meltzer 1989; Mohnhaupt & Neumann 1990). In this research a pictorial buffer is used for top-down motion recognition, for learning from observation, path-planning and several aspects of spatiotemporal reasoning. Ullman (1989) presents an approach to object recognition by aligning pictorial descriptions. It is commonly argued in all the approaches that a pictorial representation and its local processes are computationally advantageous for tasks related to concrete visual objects. (Tsotsos would probably call this a "second order complexity consideration.") The work provides an operational definition of "pictorial thinking": (1) fill the buffer with the necessary information, typically with object views, (2) apply local spreading-activation processes to perform certain operations, and (3) read off the answer.

From this perspective we view Tsotsos's work as strong support for investigating mental images and associated functions in biological systems, as well as pictorial representations and their computational properties in artificial systems. In artificial intelligence these questions have been underrated, despite some promising early work (see, e.g., Sloman 1975; Kosslyn 1980; Funt 1980).

Is it really that complex? After all, there are no green elephants

Ralph M. Siegel

Thomas J. Watson Research Laboratory, International Business Machines, Yorktown Heights, NY 10598.
Electronic mail: *axon@ibm.com*

The brain solves very complex problems. There is no question that even the simplest cognitive problems handled by the brain can have combinatorial possible solutions. If the derivations of the target article are correct, the exact solution of such visual problems is NP-Complete. Yet, in spite of the difficulties encountered by the brain, it is able to find a workable solution to cognitive problems quickly, either by elegant Marr-like solutions (1982) or by Ramachandran-like (1985) shortcuts.

Evolution has provided us with functioning neural systems. Many researchers, as is well documented in the target article, have analyzed the real biological constraints embedded in the neural wetware. The target article admirably tries to mix biological knowledge with theoretical complexity analysis, bringing into focus cortical mappings and attentional mechanisms.

But Tsotsos perhaps ignores the most essential facets of real brain organization. Real neural networks consist of millions of neurons. Each neuron is highly nonlinear in its temporal and spatial properties (e.g., Llinas & Yarom 1986). The power of an individual neuron in integrating incoming action-potential traffic is slowly becoming clear (Miller et al. 1985; Llinas & Yarom 1986; Gamble & Koch 1847; Shepard & Brayton 1987). And each neuron is heavily interconnected to many others. The beauty of the brain is that because of all this implementation complexity, it can solve the really *tough*, perhaps even NP-Complete problems of the real world. This view of vision contrasts with Tsotsos's "analysis at the complexity level" (sect. 1.1, para. 6).

The target article suggests, following on Neisser's (1967) results, that sheer neuronal parallelism is not enough to solve visual matching problems based on some back-of-the-envelope computations (Equations 2 through 4, and Table 1). The number of "matching operations" (sect 2.5, para. 3) to perform visual search is extremely large. The number, however, is based on (1) the introspective argument that there is something like a "visual dictionary" in the brain's memory (parameter VP), (2) the idea that something equivalent to pixels exists in the brain (parameter P), and (3) the idea that there is an actual number of independent parameters of the visual image (parameter M). What happens to the meaning of these numbers if the brain does not use something like "visual prototype" or "dictionary" and is not making comparisons between internal templates and the external environment? What if the visual system computes in analog without individual elements like pixels? Finally, what if certain combinations of visual parameters just do not occur in the real world? *There are no green elephants.* How does one alter the equations to implement these bottom-up constraints? The visual system has evolved; it develops and functions in a domain of parameters given by the real environment. The numbers derived for the requisite "matching operations," and corresponding amounts of parallelism (π) do not make sense in light of the real stuff of the brain.

How much processing can be performed by the dynamic properties of the real neurons? A number of workers (Zucker 1983; Skarda & Freeman 1987; Hopfield & Tank 1986; Sporns et al. 1989) talk of solving problems of pattern recognition by relaxing the dynamics of neural systems to either a steady state or strange attractors. Experimental evidence for such dynamics is now being sought (Skarda & Freeman 1987; Eckhorn et al. 1988; Gray & Singer 1989; Siegel 1990). Where is the interface between these bottom-up approaches to the visual system and the top-down approach of complexity theory?

Clearly a proper exposition of the problems of vision is important and, in this regard, the target article is quite valuable. However, it remains to be seen whether it is possible to derive solutions to real-world problems from such principles. The construction of models for vision and, by extension, brain function, is more likely to draw on the vast range of experimental findings concerning how the brain actually works. More may be understood from a study of the modes of behavior of real and richly complex biological systems as they solve the difficult problems of visual perception, cognition, and motor activity.

Algorithmic complexity analysis does not apply to behaving organisms

Gary W. Strong

College of Information Studies, Drexel University, Philadelphia, PA 19104.
Electronic mail: *strong@duvm.bitnet or strong@duvm.ocs.drexel.edu*

Tsotsos describes his methodology for analyzing vision by making an analogy with building a house (sect. 1.1): First you must start with "the internal wood frame"; then you "can begin to add detail." It is unfortunate if the house plan is predetermined by a particular architectural approach, however, because then the frame and the details are constrained by the approach rather than by the needs of the occupants. In an analogous fashion, I believe that Tsotsos's analysis of vision is constrained more by computational theory than by human vision; as a result, it does not meet the needs of those wishing to understand vision. Although there are a few other problems with the target article, such as the unclear relationship between the two Knapsack problems (sects. 1.3 & 2.3) and Tsotsos's unwarranted use of the neuroscience term "columnar organization" for a purpose other than it is generally used, my comments will be restricted to a criticism of the particular architecture by which Tsotsos constructs his house.

In brief, I am not convinced that computational complexity theory, as currently conceived, is adequate for modeling the information processes of biological organisms. There are problems with analyzing a behavior into its time and space requirements as Tsotsos does. The most inappropriate assumption is that all of the complexity is within the organism. Consider the case where a behaving organism uses the complexity of the (external) stimulus array to defray the cost of internal processing complexity (the "don't carry anything you can readily find later" principle). The impression I have of my own vision is that off-fovea information is extremely degraded, but that it is there if I wish to process it and that, until I process it, it exists in my head in only a very simple form, rather than displaying all its complexity. Tsotsos's goal of making "no assumptions about how the data may be presented or organized" (sect. 2.2) is therefore inappropriate. It is equivalent to putting the entire problem in the head, which would seem to be absurd for a perception problem. Tsotsos asks "How can a computational model be defined using experimental results from biology if one does not first understand fully the computational nature of the experiment itself?" (section 2.1) I second this question.

Besides the fact that the problem itself has been inappropriately framed, I believe that there are additional problems with basing a complexity analysis of perception on a propositional account of images. ["A test image I is the set of pixel/measurement quadruples (x,y,j,m_j)" (section 2.2).] Not only does this analysis assume that the entire image is in the head at the same time, it also assumes that measurements used in perception are all local. They are not. Such an account leads to difference and correlation functions that do not relate well to human behavior.

For example, Figure 1 would produce a large difference and a zero correlation using Tsotsos's formulae. A human, however, would consider it a very good match to the target image of Figure 1(a) of the target article. It is unfortunate that Tsotsos's whole argument requires one to accept the computational characterization; otherwise, the reader is still free to infer that human visual search has linear time complexity in the unbounded case. Tsotsos is aware of this criticism, because he states (sect. 2.2) that "there is no claim here that the algorithm necessarily corresponds to human performance." This is an unfair claim, because most of the target article seems to indicate otherwise: Terms such as perception and attention are used with reference to human performance – unless the reader is to infer that Tsotsos means only machine perception and attention (which would be misuse of the terms).

If we suspend disbelief on the issue of whether or not such a complexity analysis can be applied to vision, we can see how Tsotsos might claim it has value anyway. In section 2.5, it is claimed that the number of matching operations (which is the basis for complexity evaluations) is related to the "power set of all locations times parameter types." This results in a very large value for the number of operations. Instead, consider a "brain" that factors measurements into types, as does the Tsotsos's model, but where the types are not independent. For example, color discrimination and location discrimination may be compensatory, in that fine-grained analysis of color may come at the expense of a loss of discrimination in location. Intuitively, this seems true in humans. (Try to match colors with the tiny color chips supplied by some paint stores.) In order to reduce the complexity of the independence assumption, Tsotsos defines a receptive field in section 3.2 that "requires that tokens for each selected type of parameter be used for each location across the receptive field." Thus, the complexity analysis has led Tsotsos into a consideration of what seems to be a human limitation of visual perception. (See Strong & Whitehead 1989, for a neural network model that uses global feature detectors to construct representations.)

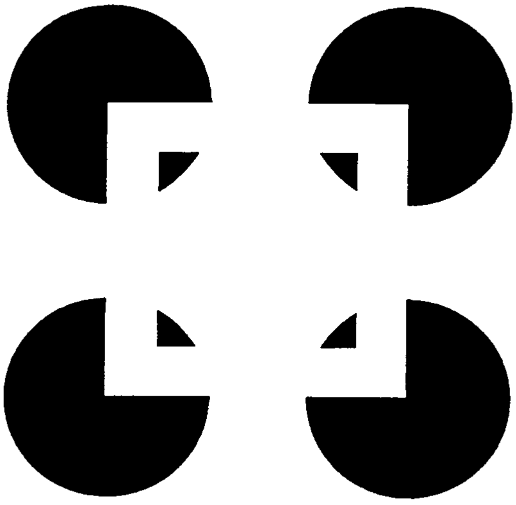


Figure 1. (Strong). An illusion of a rectangular frame over four black circles.

Such an approach is fraught with difficulty, however, because of the seemingly ad hoc consideration of ways to limit complexity. For example, I suggest that hierarchy is not the best way to defeat complexity, in opposition to Simon (1962), who is cited by Tsotsos in section 3.2, note 10. A better way might be Holland's (1975) "intrinsic parallelism," which is based on overlapping subsets, not hierarchies (see also Goldberg 1989). With Holland's approach a hierarchical binary tree of explicit prototypes is not necessary because optimization can be achieved without explicit search by using Holland's genetic algorithm expressed in terms of cell assemblies. (See Davis 1989 for a mapping of Holland's formalism onto neural networks.) An additional advantage of such an approach is that the connectivity explosion referred to by Tsotsos (Section 4.4) can be addressed. Neurons don't have to be connected one-on-one within cell assembly populations; there only need be enough connections for feedback to maintain activity within cell assembly over a period of time. His concern for numbers of connections causes Tsotsos to come to a conclusion that is at variance with the neuroscience literature: that "there can be no connections from the [receptive field] processors to any of the larger maps in the input abstraction hierarchy" because "the number of such connections would be prohibitive" (section 4.5). Contrary to this view, there appear to be feedback connections throughout the cortex, including back down the "abstraction hierarchy" (Goldman-Rakic 1988). There must therefore be some yet unexplained mechanism that combines constraints in the reverse direction. Tsotsos needs this himself when he considers the recombination of multiple pathways by the intersection of attentional "beams" from above in his bounded-vision case.

Search and the detection and integration of features

Anne Treisman

Department of Psychology, University of California, Berkeley, CA 94720
Electronic mail: *treisman@violet.berkeley.edu*

Tsotsos raises some important issues concerning computational complexity in visual coding, from which he derives a model of visual attention and search. He relates his conclusions to behavioral data, some of which I reported, and compares and contrasts his theoretical suggestions with mine, among others. My commentary has a narrow focus relative to this ambitious scheme. I will restrict my remarks to clarifying two issues related to my account of visual attention and feature integration. The first concerns the general architecture we proposed for the visual system and for the role of attention; the second concerns the use of search to throw light on visual features.

Tsotsos suggests that I have put forward two different models for feature integration and am reluctant to choose between them. I believe there is a confusion here, both about the alternative hypotheses and about the selection I made. I wrote (not too seriously) "I have hedged my bets on where to put the master-map of locations by publishing two versions of the figure! In one of them, the location map receives the output of the feature modules (Treisman 1986) and in the other it is placed at an earlier stage of analysis" (Treisman 1988, p. 203–4). This ordering of levels has nothing to do with the idea that attention may be modulated by inhibition from feature maps. It concerns whether visual stimuli are initially represented as conjunctions in a location-addressable but not content-addressable form, and subsequently analyzed into separate feature maps, or whether the initial representations are already separated along dimensional lines. In the more recent account I selected the version with the location map first, although I emphasized that there are reciprocal connections between the feature maps and the location map. This makes the initial order in most respects inconsequential.

Independent of the choice of sequential order, I recently suggested a separate modification to the theory, introducing the possibility of inhibitory control when the feature maps coding the target are sufficiently distinct from those coding the distractors (Treisman 1988; Treisman & Sato 1990; see also Wolfe et al. 1989 for a similar model). For example, if the target is a red vertical bar and the distractors are green vertical and red horizontal bars, inhibition from the feature map representing green and from the feature map representing horizontal could reduce or eliminate activity both in locations containing green and in locations containing horizontal elements. The only location remaining unaffected would be that containing the red vertical target. Tsotsos's claim that inhibition leaves all the elements in conjunction search as equal candidates is therefore incorrect. Feature inhibition could eliminate all the distractors in the conjunctive as well as in the disjunctive search condition, thus explaining the recently reported cases of parallel search for conjunction targets. We contrast this feature-based inhibition with a spatial "window of attention" that directly selects locations independent of what they contain. There is no need to "tip the scale in favor of one" of these modes of selection (Tsotsos, sect. 5.2, para. 1); they are intended to be complementary and to act together whenever both are applicable to the task at hand.

I find it difficult to determine whether our modified account is equivalent to that proposed by Tsotsos. His attentional beam inhibits all but one or more selected features in all but one location. The problem is to understand how this can be achieved through units at the highest level, which are those with the largest receptive fields and which therefore no longer retain information about specific locations. This seems a strange place to put a "spatial spotlight." To select on the basis of both locations, the attentional "beam" must be focused at the level(s) at which the neural codes retain the relevant specificity.

The top-down beam in Tsotsos's model combines spatial selection with the selection of properties within a location. There is behavioral evidence that may conflict with this view: When attention is directed to an object in a particular location, it is in fact difficult to attend selectively to one of its features and ignore others (see, for example, the interference in various forms of the Stroop test; Kahneman & Treisman [1984] also review other relevant evidence). This is one reason why in my model I separated the spatial spotlight from other attentional control mechanisms and made feature inhibition operate via the location map.

One other important difference between our models for conjunction search concerns Tsotsos's initial assumption that identification depends on a dedicated processor for each possible prototype at each possible location, and that all input data and prototypes are hard-wired to the processors. He later modifies this to suggest a switching network that routes candidate inputs serially to the processors that compare them to target prototypes. This does not solve the problem of the arbitrarily large number of particular targets for which search is possible, however. We can search not just for a monkey but for a yellow monkey with purple spots, one eye closed, and so forth. We need the possibility of creating temporary ad hoc representations (such as the "object files" described by Kahneman & Treisman 1984) which are not hard-wired elements in a prelearned visual dictionary. Allowing candidates to queue for processors, as Tsotsos proposes, reduces the number of location-specific processors needed, but seems not to solve the problem of search for arbitrary conjunctions.

The second topic for clarification concerns our use of the search paradigm to help define visual primitives coding stimuli at early preattentive levels (Treisman 1985; Treisman & Gormican 1988). Our subjects were never instructed "to search for the absence of a feature"; they were always shown the two stimuli that would function as the target and as the distractors in each task, so they were never required to search for a "logical negation." We inferred from their performance which of the pair (if either) had a feature for which the visual system had (or could generate) appropriately tuned detectors. Note that separable features, in our terms, are not discrete stimuli: Separability is a relation between two dimensions of variation or between values on a dimension. Like Tsotsos, we emphasized that "no search task allows direct inference to the complete code for a particular stimulus in any absolute sense" (Treisman & Gormican 1988, p. 40).

However, search performance can both "reject candidate stimuli" (as Tsotsos says) and also "discover their existence" when search is parallel, although it does not exactly specify their identity. For example, if a line tilted 18 degrees from vertical "pops out" of a display of otherwise identical vertical lines, we infer that some feature that correlates with the orientation difference must be detected in parallel. When we find that a vertical line is more difficult to detect against 18-degree tilted distractors, we infer that it does not possess a unique distinguishing feature. Search asymmetries of this kind may suggest which end of an asymmetric dimension functions as the standard or reference value and which is signalled as the presence of an additional feature (e.g., tilt, curvature, convergence).

Further experiments can help to specify more exactly the nature of the features that mediate parallel detection. For example, by using a tilted frame, we found that the relevant description is probably "frame-aligned versus misaligned" rather than (or as well as) "vertical versus tilted." Pairs of oriented dots appear to share the same orientation codes as connected lines when they share the same direction of contrast (Treisman 1985), but probably not when one dot is darker and one lighter than the background (O'Connell & Treisman, in preparation).

Tsotsos suggests as an alternative explanation of search asymmetries that the features that are found more easily are those that can be "tuned" away from their default state. Why would this predict faster detection for deviating values? Could the standard stimuli not be found simply by leaving the detectors in their untuned state, so that only the target is effective? There is certainly physiological evidence for some flexibility of response from individual units, but it is not obvious how this explains the search asymmetries.

Some important constraints on complexity

Leonard Uhr
University of Wisconsin-Madison, 1210 West Dayton Street, Madison, WI 53706.

This kind of confrontation of computer vision with complexity theory is potentially of great importance, for it can suggest how a scalable system that successfully recognizes a few relatively simple objects in a small image might be enlarged using bigger computers to recognize more and more. The following are a few elaborations and extensions I think are worth making.

A. It is important to consider and minimize both time and space complexity. Indeed, time is crucial – not only because brains are incredibly fast but also because objects move fast – which is what has driven brains to be so fast. It has been known experimentally for many years that we humans need roughly 300 to 800 msec to push a button that indicates we have searched for, recognized, and successfully found some complex object. Because neurons need roughly 1.5 msec to fire, and the time needed to do such things as transduce (break down rhodopsin), carry signals from eye to cortex and from cortex to muscle, and activate muscle is at least 100 msec, this suggests a serial depth of processing on the order of 100 to 500. Amazing recent experiments with monkeys have found individual neurons that respond selectively to (among other things) a particular person's face in from 70 to 200 msec. Thus, serial depth appears to be reduced even further, to only 25 to 75 or so. That is, images of objects that are resolved on reinas with 10^6 to 10^8 rod and cone sensor nodes are recognized by brains with 10^{10} to 10^{11} neurons in fewer that 10^2 steps (time).

This is mind-boggling speed – orders of magnitude faster than any of today's computer vision systems even when they have been parallelized (ignoring the fact that their performance today is far poorer than the brain's). But it jibes very suggestively with two other important facts: (1) The major pathways from retina through areas of the cortex involved with vision have a serial depth of roughly 15 to 40. (2) The kinds of logarithmically converging structure that Tsotsos's target article espouses would have similar serial depths (e.g., $\log_4 10^6$ is 10). Only massively parallel systems could so much in so little time.

B. Complexity is actually $O(V^I)$, where V is the number of possible values that might be input at each pixel spot in the image. That's certainly a quibble, because 2^I is impossible enough, and high-quality images can be gotten with V only 2^{24}, using only 24 bits (e.g., 8 for each of 3 primary colors). But it helps hammer home the point that we must focus on small, finite, realizable real-world sizes; we cannot hope to handle the "general" recognition problem moving into ever-larger sizes.

C. Multilayer logarithmic convergence is probably the key to reducing the combinatorial explosion. Rather than have the impossibly large but necessary worst-case number of V^I nodes in a single inner layer, successive combinings, abstractings, and reducings of information can lower this to feasible size, although we probably won't know for sure that this is true until we have actually achieved it by building successful perceivers. (The fact that brains are successful proves that feasible implementations are possible.)

D. Complexity results are even more appropriately applied to neural (connectionist) networks (NN). Because NN are Turing-equivalent universal computers if they have at least one layer of interior (hidden) units between input and output units, many people appear to feel that they only need to be made bigger and they will become as general and as powerful as desired. Most of today's NN input images to an 8-by-8 array (obviously such images are oversimple), and use only 1 to 4 interior layers and a total of a few hundred nodes. Each node links into all the nodes in the next layer, and learning consists of changing weights associated with links in the direction of reducing the backpropagated error signal (the difference between the output vector and the correct output). If there exists a successful state, this kind of learning will (under certain circumstances) achieve it. But the complexity for NN is just the same as for any other kind of artificial intelligence system, except that one cannot take advantage of the kinds of heuristically promising structures that Tsotsos develops in his paper.

E. Worst-case complexity is much too stringent a measure. They expected case – what the real world actually demands that the recognizer do – poses the real problem. Unfortunately, nobody has clearly specified the expected case for perceptual recognition, and it seems unlikely that anybody ever will. Possibly the best that can be said is to point to the number of different objects that humans can recognize (probably in the hundreds of thousands; possibly in the millions or more) and the number of different variant instances of each (a rough indication is the minimal resolution – probably on the order of 100-by-100 – needed to recognize a complex object). This still given an overbound, but one that is far less exaggerated than V^I.

F. Although it is clearly true that brains are not 100% general purpose when confronted by worst-case complexities, they are surely far more general purpose than any other known system. Consider the following recognition problem: Generate random images and assign these at random as the instances of different object-classes. Most of these instances will be highly disconnected, with slight and subtle differences that are crucial to recognition. The different instances of the same object will be completely unrelated. Humans would do very poorly – we ignore slight differences, strongly favor features that reveal well-connected shapes, and use similarity measures to generalize. We are almost certainly pretty poor at recognizing this kind of "general" set of objects, although we are fabulously good at recognizing the very general sets of object-classes that guided natural evolution.

On brains and models

William R. Uttal

Department of Psychology, Arizona State University, Tempe, AZ 85287.
Electronic mail: *aowru@asuacad.bitnet*

Congratulations are in order for Tsotsos's astute analysis of the complexity issue and for his reminder about some of the practical difficulties arising from computational economics. Unfortunately, I don't think he has gone far enough in dealing with the problem of the *meaning* of our formal models of visual and other cognitive processes. Tsotsos shows that, for practical reasons, a bottom-up approach is unlikely to provide satisfactory explanations of such phenomena. There are many related reasons why it is not just practical matters of complexity and computational costs that prevent neuroreductionist models from meeting the tests of necessity and sufficiency required for their rigorous validation or rejection.

The first of these supplementary reasons comes from automata theory. In a little remembered, but very important paper, Moore (1956) showed that the internal mechanisms solely of a closed system could never be uniquely analyzed on the basis of the relationship between its inputs and outputs. There would always be more possible mechanisms than possible discriminating experiments. There should be nothing surprising in this proof. Moore's theorem is consistent with many other scientific principles, and complies with a long tradition in psychology. Indeed, Tsotsos explicitly acknowledges this point by emphasizing only the *sufficiency* of his particular multilayer, "attentional-beam" model rather than its uniqueness or necessity. Hence his model, admittedly a first approximation, is, but one of many (possibly infinitely many) that might meet the conditions of his analysis.

The distinction I make here is between models that provide *analogies* with the system being modelled and those that provide *homologies*. These terms are borrowed from biology. Analogous systems behave the same but use completely different mechanisms. A bird's and an airplane's wings are classic examples of such a "process analogy." The second order differential equation's ability to model a wide variety of oscillatory systems, all of which produce identical behavior by means of what may be completely different mechanisms, is another example of analogy. By contrast, a model that uniquely described the specific internal mechanisms of a process would be, in this context, a homologous *explanation* rather than an analogous *description*. I believe that Tsotsos's model and all other artificial intelligence programs can only be descriptive analogies, not the reductive explanations they are often portrayed as.

Modern developments in chaos theory further suggest that even given a miraculously efficient electrophysiological research tool and sufficiently powerful computers to overcome the complexity or combinatorial problem, we would not be able to derive cognitive processes from neural processes (or, for that matter, neural processes from cognitive processes). The implications of this profound development in mathematics have not I think, percolated deeply enough into the thoughts of those of us who are interested in the possibility of neuroreductionist explanations of cognitive processes. According to chaos theory, small uncertainties in even a deterministic universe can quickly pyramid to produce behavior that is both unpredictable and unanalyzable. The reason it is not possible to analyze or reduce molar behavior to its initial state or underlying mechanism is that information about the history of the system is no longer available in a chaotic system – the interactions among its microscopic components are apparently random even though the overall behavior of the system is not random. Our inability to predict the behavior of a chaotic system given the initial state of the components emerges from the fact that the pyramiding of small uncertainties quickly produces apparently random interactions. Thus, according to chaos theory, one cannot reproduce the specific sequence of events that led to a molar state by studying its microscopic elements or vice versa.

The laws of thermodynamic irreversibility also suggest that complex systems cannot be run backward in the way many think would be required to develop a necessary and sufficient neural model of a cognitive process. Cognitive phenomenology presents a case (as does Tsotsos) for an inferential (or top-down) kind of processing in the brain that is not well modelled by any of the mechanisms so far proposed. The very nature of mathematics is itself an argument for explanatory irreducibility. Mathematics, it is often overlooked, is neutral about how it is instantiated; hence it is more a descriptive than an explanatory or reductive tool. These supplementary arguments are all subjects to interpretation and open to challenge; they do not, in my opinion, have the rigor of the combinatorial (complexity), automaton, or chaos theoretical arguments.

What this all seems to suggest is that no mathematical, neural, or computational model can ever be validated or even tested as a truly explanatory, reductionist theory of a cognitive process (in the sense of specifying its exact instantiation). Rather, such models are, at best, process descriptions of the time course of the systems they represent. Such descriptions can still be very useful. A more realistic view of what we are doing when we model is that we are describing processes using the terminology of one or another analogous system. We are not reductively explaining unique internal mechanisms. The most important general conclusion is that *in principle* constraints are probably operating on theory production that may forever prevent us from crossing the barrier between the neural microcosm and the cognitive macrocosm in either direction. These *in principle* constraints may be much more serious than the *in practice* one that Tsotsos has so capably considered.

Complexity, guided search, and the data

Jeremy M. Wolfe

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139.
Electronic mail: *wdfe@psyche.mit.edu*

The most useful aspect of Tsotsos's target article is that it gives rigor to the claim that parallel processing by itself will not be able to solve the visual search problem. This demonstration of the necessity of attention is welcome. I have more difficulty with the specifics of Tsotsos's account of visual search. In its broad form, it resembles our guided search model (Wolfe et al. 1989; Wolfe & Cave, in press; Cave & Wolfe, in press). However, in a number of places it seems to be at variance with the data or with reasonable inferences made from those data. I will briefly discuss three such areas.

Bottom-up processing. Tsotsos wishes to make a clear distinction between the searches for a known target and searches for an unknown target. In our model, we make a similar distinction between "top-down" and bottom-up" processing. Tsotsos does not have much hope for bottom-up searches. However, searches for an *odd-man-out* on a homogeneous field of distractors are easy. In section 4.1, Tsotsos argues that search time increases exponentially with the number of relevant maps. Let me suggest an experiment that I do not think has been done. Imagine a distractor set of small, red, Xs moving to the left in the frontal plane. A target would be any unique item created by altering one or more of the features defining the distractors. On each trial, the unique item could be different in size, color, shape, motion, and so forth. Is there any doubt that this task would be done efficiently with little practice and with no information about the identity of the target? Moreover, one could imagine the distractors varying randomly from trial to trial without significant degradation of performance. As long as all of the distractors are the same on a given trial and as long as the difference between distractors and the odd man out is sufficiently large, search will be easy, even if the observer must monitor a multitude of maps.

The nature of parallel guidance. Tsotsos proposes what he calls "feature range inhibition" as a way to make search more efficient. Though the description is a little unclear, this seems to be similar to our "top-down" parallel guidance. In top-down guidance, information about basic features present in the target is used to limit the deployment of attention to locations likely to contain the target item. Tsotsos does not appear to use this type of guidance in an optimal way. Consider his example, in section 5.2.2, of a search for a red A. In his version, feature range inhibition excludes items that are "non red **and** non A." This leaves "exactly the number of distractors plus the target." This is not much of an improvement. We can do much better if we exclude all non red **or** non A items, leaving only the target item. In our model, we use excitation rather than inhibition to perform the equivalent operation of taking the intersection of all red items **and** all A items. This, too, will restrict search to the red A. Once guidance is proposed, it seems a p:ty not to use it to best advantage.

Noise. Actually, our version of guidance works *too* well. If one could restrict to the intersection of the set of red items and the set of As, conjunction searches would be as efficient as feature searches. We and others (McLeod et al. 1988; Nakayama & Silverman 1986; Treisman & Sato 1990) have shown that conjunction searches need not be serial, self-terminating nor are they strictly parallel. We assume that the parallel guidance of attention is limited by noise and thus is not perfect. That is, the effort to produce the intersection of the set of red items and the set of As significantly reduces the set of candidate targets but some distractors are still checked and discarded by attention. The recognition that perceptual systems are noisy is important in models of this sort but this does not appear to be part of Tsotsos's account. His model appears to be largely deterministic. As noted, this makes it difficult to account for the data on conjunction searches. It also makes it difficult to account for the data from blank trials (trials containing only distractor items). Tsotsos's model would appear to require serial exhaustive search in the case of blank trials. This is a reasonable model for true serial, self-terminating searches but it does not explain when a trial should be terminated in parallel (feature) search or in what we call guided search through a subset of items. In these cases, slopes on blank trials and the target trial/blank trial slope ratios vary significantly across subjects. We see termination of search in these cases as a signal detection problem. Subjects search through items in order of decreasing probability that a given item is the desired target. In the face of some variable amount of internal noise, each subject sets a termination threshold based on a desire to respond quickly and yet miss few target items.

None of these problems could be said to represent a "fatal flaw" in Tsotsos's work. Rather, each represents an area for possible revision of his current model. We find it encouraging that overall, Tsotsos's model, driven by complexity theory, already bears considerable resemblance to our data-driven guided search model.

Adaptation and attention

Steven W. Zucker
Research Center for Intelligent Machines, McGill University, Montreal, Quebec, Canada.
Electronic mail: *zucker@larry.mcrcim.mcgill.edu*

Tsotsos argues that "an attentional scheme has as its main goal the selection of certain aspects of the input stimulus while causing the effects of other aspects of the stimulus to be minimized." Moran and Desimone (1985) empirically discovered that "the very structure of the receptive field, recently considered to be a fixed property of the neuron, can change from moment to moment in the behaving monkey depending on the immediate task and state of attention." The process of focusing attention is thus connected to dynamic variation in receptive field properties, a seemingly novel connection. But is this truly a novel phenomenon, and, if so, how might the mechanism be understood? We submit that analogous phenomena exist in a more primitive form as adaptation, and that the roots of attention can be illuminated by exploring the analogy with adaptation.

Adaptation exists in two forms: (1) intensity adaptation, by which the central excitatory region of a circular-surround (retinal ganglion) receptive field expands or contracts at the expense of the inhibitory surround as a function of photon intensity (Barlow et al. 1957); and (2) the effective operating range of cells in the visual cortex varies as a function of contrast (Sclar et al. 1989). That is, both (1) receptive field structure and (2) activity levels can vary as a function of stimulus properties. In this case the stimulus properties are physically based, and functionally they extend the sensitivity of the visual system to a broader range of operating environments.

The analogy between adaptation and attention arises as follows. Visual cortical neurons respond to contrast-encoded stimulus features, and exhibit a sigmoidal operating characteristic (plot of firing rate versus contrast). If all contrasts were in the saturated range then no structure would be visible. Adaptation is a primitive mechanism for preventing this, that is, for adjusting the operating range so that orientation (say) structure is detectable (pops out?) from the background. Analogously, attention is a mechanism for adjusting the feature context so that more complex structures are detectable from the background feature clutter. The increased responses to attended stimuli (Desimone et al., in press) would imply that the visual system has "adapted to" the unattended stimuli.

Although the analogy between adaptation and attention provides perspective, much remains before biologically plausible mechanisms can be specified. Adaptation is largely a *bottom-up* process, whereas attention may well be a myriad of processes many of which are *top-down*. Nevertheless, Tsotsos argues for a model of attention as an inhibitory process within a pyramidal beam, extending from large abstract receptive fields to tiny, low-level ones. A key reason for this is the positional accuracy to which attentional effects can be measured. It is almost as if the attended stimuli are described within a finer coordinate system than the unattended ones. This accuracy may well reside in the visual descriptions being selected, however, and not in the details of the attentional beam. There is no evidence that abstract descriptions or their features are continuously distributed over the retinotopic array, with the attentional "beam" highlighting a well-delimited retinotopic subfield. In contrast, one might speculate that visual inferences are carried out by multistage processes, with feedforward and feedback loops between them. The initial stages could be coarse, local ones, and the latter stages precise, global ones. Attention could act as a gate between the early and later stages, effectively adapting away the unattended stimuli. There would then be no need to postulate a "beam" running through a pyramid of receptive fields.

Author's Response

A little complexity analysis goes a long way

John K. Tsotsos
Department of Computer Science, University of Toronto, Toronto, Ontario Canada M5S 1A4.
Electronic mail: *tsotsos@ai.toronto.edu*

1. Introduction

Commentators misunderstood several points in the target article. I will deal with these before addressing the many important and substantive issues raised in the commentaries.

I did not claim that computational complexity is "the key" to vision, as **Dickinson** states, only that it is an important and heretofore neglected dimension of study. This is explicit throughout the introductory section of the paper. Complexity analysis can reveal insights that no other method of analysis can, but it cannot even begin to address certain other equally important issues. Dickinson's commentary was dedicated to countering a view that was not expressed in the target article.

Siegel mistakenly believes that complexity theory is a top-down approach to vision. Marr (1982) describes the use of Laplacian operators; would Siegel consider Laplace's equation a bottom-up approach to vision? I hope not. Both complexity theory and Laplacian functions are tools. Complexity theory led me to develop a theory that has a significant top-down component; complexity provides its mathematical foundation. **Strong** likewise makes this unusual connection between tool and model, claiming that complexity theory as currently conceived is not adequate for modeling biological information processes even though in his own work he develops computational models of biological information processing. Complexity theory is one of the theoretical underpinnings of computation. Complexity theory does not model; it is a tool that provides a source of constraints for a model.

Heathcote & Mewhort provide an algorithm for unbounded visual search that they claim solves my only example in polynomial time. I did not give such an example in the paper, however. The example in section 2.2 was provided as part of the general discussion of visual match to illustrate the definition; it preceded the formal definition of unbounded visual search. That definition does not include the target pattern and although Heathcote & Mewhort are not specific, I assume that their algorithm requires the target. **Kube** also makes this mistake (his third point). The target is not specified in the unbounded problem; the values of the functions are given only as mappings. Heathcote & Mewhort distinguish between matching and search incorrectly; as described in section 2.2, the former is a subproblem of the latter, and thus its difficulty must be included within the difficulty of the latter. If unbounded visual matching is NP-Complete, as Heathcote & Mewhort seem to agree, then unbounded visual search is necessarily NP-Complete, too.

A major component of the representations I use is the hierarchy – the simple variety that everyone understands. I did not claim that the simple hierarchy is the "best" mechanism for beating complexity (**Strong**). The argument was for sufficiency only, as clearly stated in section 3.1. Perhaps intrinsic parallelism (Holland 1975) is indeed more efficient. Although the simple hierarchy is logarithmically time-bounded, Holland's scheme is exponential in the worst case. The worst case would be when the search creates the subset that contains the answer only after all other possible subsets have been examined. As **Uhr** points out, the logarithmic convergence in my model may be the best attack on the complexity problem; it is also biologically plausible.

Krueger & Tsay claim that I have misplaced the complexity and that the truly complex processes may be at the decisional level, yet they do not indicate whether or not they believe that the lower-level processes I have considered are tractable. I therefore assume that although they agree that perceptual processes are complicated, they believe that decisional ones are even more so. This may be so, but it is not obvious that perceptual processes, such as those in visual search, which require specific choices between response actions, require no decisional process.

Krueger & Tsay also note that I have not considered how "smart nonoccupational perceptual mechanisms" could eliminate complexity altogether and thus obviate my analysis. I cannot seriously entertain this suggestion. I have great difficulty in determining what "noncomputational" means in this context. The polar planimeter is not "noncomputational," and Runeson (1977) does not label it so in his paper. Moreover, he notes that he is only proposing an analogy, lest anyone mistakenly infer any closer ties between the planimeter and the brain. The polar planimeter is indeed computational because there is a precise mechanistic algorithm for using it with well-defined input and output. It would be easy to simulate its operation with a computer program. Digital computers are only one manifestation of a device that computes; one must not equate computation with computers alone.

Krueger & Tsay argue that a really smart process would use only one measure for matching, either sameness or difference, not both, as I propose. They seem to have misunderstood the correlation measure, which is not just a measure of sameness but ensures the maximality of the match. A single correctly matching pixel would pass a sameness or difference test, but not the two tests together as I have set them up. In addition, lightness and darkness have nothing to do with the validity of the computation. The formalism is capable of dealing with any type of physical measurement of a visual stimulus.

Krueger & Tsay go on to point out that vision may operate in "all kinds of undreamed of ways." **Siegel** also makes the "what if things aren't like this?" argument. This kind of criticism is easily leveled at any theory when the critics have no empirical counter evidence and offer no viable alternative theory. Is it really that complex? Siegel asks. After all, there aren't any green elephants. Have you never watched cartoons nor enjoyed abstract art? Siegel describes his view as contradicting mine, yet I agree completely that "the beauty of the brain is that . . . it can solve really tough problems." I am simply proposing a way of determining how tough the problems are and how they may be solved. More to the point, it is the very toughness of the problem that may force the brain to use the kinds of solutions I propose.

Both **Cave** and **Strong** claim that I have defined feature maps that operate independently. In section 2.4, I state that "types are not necessarily independent." A map represents one type of visual parameter; maps are physically independent, but the types of parameters they represent are not necessarily so. Many physical visual maps have now been documented, and within each a variety of visual parameters seem to be represented, not all independent of each other (Maunsell & Newsome 1987).

Cave concludes his commentary by claiming that my model is not a serious one unless more detail of operation is provided. I had stated explicitly, however, that I would not address the operational level in this paper. All I intended was to provide a source of constraints and hypotheses.

Uhr points out what he calls a minor quibble – that the complexity is really $O(V^I)$. This is the correct order for the number of distinct images, not the number of data groupings. My analysis does not include the number of possible values of each type of visual information. This is of course an important issue, but the $O(2^{|I|})$ stands, for the purpose of my analysis.

Strong claims that my account requires one to assume that the entire image is in the head during processing and that this is a bad assumption. "Don't carry anything you can readily find later," he says. First of all, in the typical visual search experiment, there is no time to wait until later – the trial is over in a few hundred milliseconds. Second, how do you know what to discard and what to carry if you have not analyzed, at least to some degree, in the first place? Strong goes on to argue that this bad assumption leads to performance that does not agree with human data. He provides a figure as an example and claims that any human would see a perfectly good match to the target I define in Figure 1 of the target article. This is highly unusual because I certainly cannot see my target, an open black rectangle, in his figure!

Heathcote & Mewhort write that I cannot use Treisman's data for comparison with my results because I use pixels and Treisman uses display items. In the early parts of the paper, I may have been unclear with this comparison; however, the relationship becomes explicit and clear in the description of the variables for Equation 23.

2. On computational modeling and visual science

Several commentators point out the impossibility of explaining biological phenomena with computational models. This argument has been made since the early days of artificial intelligence. Many have claimed that there is something special about implementations that are brainlike. This objection comes from at least two sources: those who follow Searle (1990) and those who work on neural networks. In the former case, the argument is rather nonspecific; in the latter case, it seems misguided. As **Uhr** points out, massive parallelism leads to greater speed and the ability to conceive radically different architectures than if one considers only von Neumann architectures. Most neural network research, however, is implemented on serial machines! Does this cause a problem? No, neural networks are Turing-equivalent, again as Uhr points out, and they are subject to the same results about computational complexity and computational theory as any other implementation. (See section 1.3 of the target article about the Church-Turing thesis.) It is important to note that relaxation processes are specific solutions to search problems in large parameter spaces and nothing more. Neural networks use variations of general search procedures called optimization techniques. If relaxation (or other optimization processes) is indeed the process by which real neurons perform some of their computation as **Siegel** suggests, it is subject to precisely the same considerations of computational complexity as any other search scheme.

Uttal points out that my particular theory cannot be an explanation of biological behavior and that it would at best be an analogy. Is there any other type of explanation? In physics, cosmology, or chemistry explanations and theories are put forward and the only requirement for their validity is that they account for the experimental observations. Would a cosmologist be required to create a universe in order for his theories to be taken seriously? Or a biologist, life? A theory that accounts for more observations than another is a better theory. Theories whose predictions are falsified are modified or rejected. In addition, computation itself plays a large role in modern theory construction even in the above disciplines. Computer simulation in particular has been a very powerful tool in the physical sciences. Yet, no cosmologist would claim that he is creating a universe and no one would criticize him for not doing so.

Is simulation of information processing particularly menacing for some reason? Or is it that in AI we have concentrated too much on toy examples and have not developed falsifiable theories and a solid experimental tradition, as in other disciplines? It is hard to say. The work presented in the target article, however, is intended to be one dimension of a framework for developing a general theory of biological and artificial perception. I have considered the dimension of computational complexity only, but other dimensions must also play a role, as many commentators have correctly noted. I have developed constraints that apply to all theories of perception and have tried to show one possible path of development that would satisfy those constraints.

Eklundh wonders whether the sort of analysis I propose can yield precise predictions or only provides constraints on the model space. He is right to ask. Complexity level analysis only yields constraints, as I point out in the target article. My analysis is followed by one possible model conforming to those constraints. Competing models are encouraged; such models do lead to precise predictions.

The target article was guided by observations in psychology, neurophysiology, and neuroanatomy. The lifetime of my results will be determined by experimental confirmation or refutation from those disciplines and by their usefulness in designing machine-based perceptual systems. Many predictions were made in the target article, most of which no commentator criticized. Most of the predictions conform very closely to known findings in biology. I am pleased to see that such investigators as **Desimone, Cave, Wolfe,** and **Treisman** find such a close resemblance between my suggestions and their own. That was the whole point! A large interdisciplinary set of observations was tied together using the thread of complexity analysis.

3. Visual search within vision

It is suggested by **Lowe** that I have not shown the importance of visual search for vision in general. Indeed, I only state that visual search may be a very basic problem that is found in most other types of visual information processing. Elaboration is in order. Basic bounded visual search task seems to be precisely what any model-based computer vision system has as its goal: Given a target or set of targets (models), is there an instance of a target in the test display? Lowe's own work certainly falls into this category (Lowe 1987). Even basic visual operations, such as edge-finding, are also in this category: Given an edge-detection model (e.g. Ballard & Brown 1982), is there an instance of this edge in the test image? It is difficult to imagine any vision system that does not involve similar operations. My remark about the ubiquity of search in vision therefore seems to have merit. The point has not been rigorously proved, of course, but it is clear that these types of operations appear from the earliest levels of vision systems to the highest.

4. Complexity is even *more* complicated

Strong wonders about the relationship between the two Knapsack problems I present, one as an example in section 1.3 and one with a formal definition in section 2.3. The complexity literature indicates that the same problem can be formulated in various ways. Different instances will share certain basic features. So it is with the Knapsack problem. Many different statements of it are given in Garey & Johnson (1976). The example in section 1.3 was found in Rosenkrantz & Stearns (1983) as an easily understood example for a noncomputational readership. The intractability claim that **Kube** disputes came from that article. As defined, the statement is true; more on this later.

Kube proposes that the theorems I present concerning the intractability of unbounded visual search do not hold; he provides conditions under which Theorem 1 does not hold, noting that the Knapsack problem is not NP-Complete in the "strong sense." He is right; however, he goes on to say that unbounded visual search is consequently not NP-Complete either. This is simply wrong. The problem is still NP-Complete and has exponential time complexity as defined, that is, with no a priori assumptions or bounds. My proof for unbounded visual search has been duplicated twice so far, each proof with slightly differing problem formulations (by Bart Selman, 1989, in our own department and by Ron Rensink at the University of British Columbia, personal communication, 1989).

Let us examine this a bit further. First, some definitions must be presented. Define two functions over the nonzero integers, *Length* and *Max*. The former is a function that maps any instance **I** of a problem to an integer corresponding to the number of symbols used to describe the instance under some reasonable encoding scheme for all instances. The latter maps an instance to an integer corresponding to the magnitude of the largest number in the instance. An NP-Completeness result does not necessarily rule out the possibility of solving a problem with a "pseudopolynomial" time algorithm. This is true only for "number problems," such as Knapsack. A problem is a number problem if there exists no polynomial **p** such that **Max[I]** $\leq$ **p(Length[I])** for all **I.** According to **Kube,** I assume that the magnitude of image values must increase exponentially with retinal size. I make no such assumption. Moreover, by definition, the relationship cannot be polynomial. Kube's comment does not fit the definitions. An algorithm that solves a problem is a pseudopolynomial-time algorithm if its time complexity function is bounded above by a polynomial function of the two variables **Length[I]** and **Max[I].** Kube points out that Knapsack has a known polynomial-time algorithm if an assumption can be made about the magnitude of the numbers; but this is not the same as the problem being inherently polynomial. If it were, it would have proved that all NP-Complete problems have polynomial solutions, disproving the conjecture on which the entire theory of NP-Completeness depends. It turns out that this is a common mistake, but to show why one must determine the complexity function for the proposed solution and the length of an instance of unbounded visual search.

The polynomial-time solution to which **Kube** refers is presented by Dantzig (1957) based on a method first proposed by Bellman (1954). Lawler (1976) provides a different algorithm for Knapsack also based on Bellman's equations. Bellman motivates his solution by pointing out that practical experience with the problems is put to use. I wished to conduct an analysis that did not depend on such experience. After all, there was no experience to draw on before our visual systems had evolved. Dantzig carefully notes that although algorithms for approximate solutions also exist using techniques of linear programming, the solution by Bellman is intended for the derivation of exact solutions. As such, it is recommended when there are only a few items in the knapsack and only one kind of limitation. Moreover, Bellman says that because of the nonlinear functional relationships inherent in his equations, only special cases of them can be solved and, even then, solutions will not necessarily be unique. The algorithm, which relies on dynamic programming, seems to require $O(\theta\cdot|\mathbf{I}|)$ operations where θ corresponds to one of the thresholds of the unbounded visual search problem defined in section 2.3, and $|\mathbf{I}|$ is the number of elements in the test image set. To encode an instance of unbounded visual search $O(|\mathbf{I}|\cdot\log_2\mathbf{Max[I]})$ bits are needed. The number of operations, $O(\theta+|\mathbf{I}|)$ is not bounded by any polynomial function of $|\mathbf{I}|\cdot\log_2\mathbf{Max[I]}$ and thus the general problem does not have a polynomial-time algorithm.[1] It is still NP-Complete. The NP-Completeness depends on large inputs.

What sizes of numbers are present in unbounded visual search? This problem has three kinds of numbers: the values of the test image, and the values of the ***diff*** and ***corr*** functions. The human eye can discriminate over a luminance span of about 10 billion to 1 (Dowling 1987). Thus, image values should have this as a range; similarly, the **diff** function has this range while the **corr** function has a range of 1 to 10^2 billion because it is a product of two image values. Thus **Max[I]** is at least 10^{20}. The retina has about 130 million photoreceptors. To binary encode one instance of unbounded visual search for humans would require $O([\log_2 10^{20}]\cdot 1.3\cdot 10^8\cdot 3)$ bits or more than 20 million bits! This is certainly too large to be biologically plausible. According to the definition given earlier, an algorithm is pseudopolynomial if it has a time-complexity function bounded from above by a polynomial function of **Length[I]** and **Max[I].** Using the estimates for Max and Length derived here, such a time-complexity function is of little help. This in fact exhibits a property of number problems that are NP-Complete yet have a pseudopolynomial-time solution: They display exponential behavior with large input numbers.

There is an additional problem with the pseudopolynomial time algorithm for Knapsack. That solution, together with all solutions based on Bellman's initial formulation[2] use the following clever observation: If we wish to solve a problem of size N, first determine the solution to same problem but of size N-1; the cost of determining the solution to the original problem then becomes easy because the decisions that must be made are only for the additional element. This line of reasoning can be extended from problems of size N all the way down to size 1. With this technique the number of operations becomes very small. Such solutions are known as recursive; each decision depends on decisions made for the problem of the next size down. This recursiveness poses a serious problem for biological plausibility. Bellman's functions are nonlinear; the algorithm that uses them involves two nested if-then-else conditions to decide which functions are used for each step based on the magnitude of the values determined in the previous step. Even though the solution may require polynomial rather than exponential time, it does not appear to be parallelizable because of the strong dependence of each step in the solution on the previous step.[3] In a retina size problem this solution may necessarily require 130 million sequential steps.

Lowe, Krueger & Tsay, Uttal, and **Wolfe** all describe the importance of noise and probabilities in vision. I agree that research must pursue these considerations. Probabilistic complexity is not quite so well understood, however. Lowe and Uhr question the use of worst-case complexity. I first point out that worst case does indeed occur in practice. In any problem of fixed length, not necessarily large, it is quite possible that a search method will find a solution only after all other possible solutions have been tried. That is just as much a worst-case scenario as is the largest possible problem. Perceptual algorithms must be *time-bounded* to be useful to a perceiving system. Worst-case complexity can provide this bound. Worst-case analysis can tell us about all instances of the problem; average-case analysis can only tell us about the average case; it is unclear what the average case could be for vision. Average-case and probabilistic analyses should also be attempted once the techniques are developed and we get a good enough idea of what the average visual input could be.

5. Complexity equations and the data

Several comments were made about the algorithm and explanation for visual search. Four experimental scenarios are addressed by the algorithm in section 5:

Type I: The target is the only item in the display to exhibit a specific feature; the target is known in advance.

Type II: The target is the only item in the display to exhibit a specific feature; the target is not known in advance.

Type III: The target is the only item in the display to exhibit a specific feature combination (two or more features); the target is known in advance.

Type IV: The target is the only item in the display to exhibit a specific feature combination (two or more features); the target is not known in advance.

Type I is the usual version of disjunctive search found in the literature; similarly, Type III is the usual version of conjunctive search. The target article is a bit vague about odd-man-out searches (**Wolfe** is justified in his criticism). In my defense, I have not seen too many experiments with Type II or IV conditions, Treisman and Sato (1990) being the only example. To help clarify the conclusions of the algorithm for visual search, I will give the time-complexity function for each of these conditions and comment on the relationship to the experimental data, where possible.

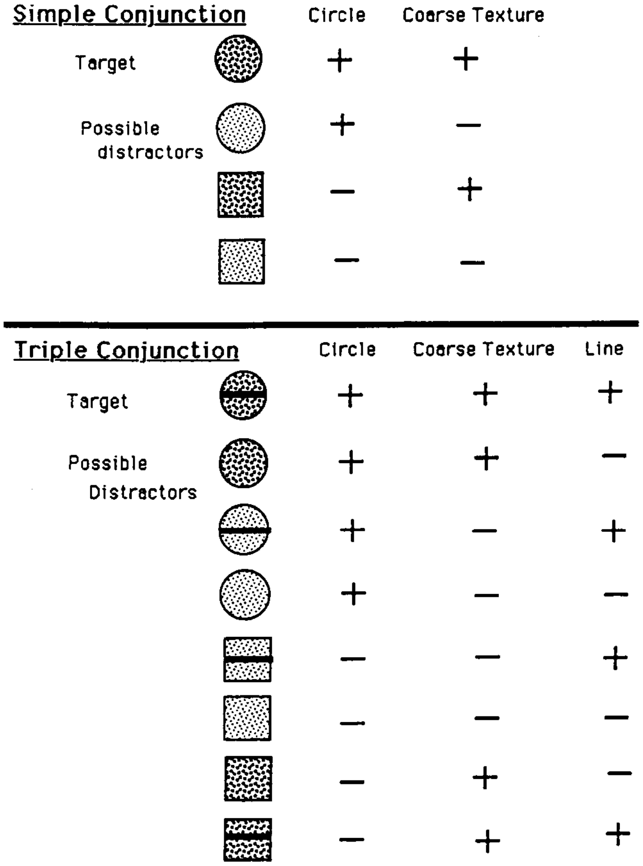
Type I: Response Time varies as $|T| \cdot \Phi(\hat{M})/2$

Type II: Response Time varies as $|R_t| \cdot |T| \cdot (2^{\Phi(\hat{M})} - 1)$

Type III: Response Time varies as $|R_a| \cdot |T| \cdot \Phi(\hat{M})/2$

Type IV: Response Time varies as $|R_t| \cdot |T| \cdot (2^{\Phi(\hat{M})} - 1)$

where $|R_t|$ stands for the total number of items in the display and $|R_a|$ represents the number of candidates left for matching after inhibitory tuning is applied. The other variables are as defined in the target article. In each case, the target may be present or absent in the test displays. Two targets rather than one would lead to a doubling of time to compute the visual response. Quinlan & Humphreys (1987) report similar effects. The story is not quite so neat, however. In section 5.2.6 I point out that the selection of candidates for matching may depend on their relative response strength. In other words, the ordering of candidates may be in descending order of response. Section 4.5 points out that inhibitory tuning based on the characteristics of the target leads to computational savings as well as larger responses and that the inhibition should be applied using a Gaussian weighting function over the feature dimension of interest, applying this weighting function multiplicatively. This mechanism manipulates the relative ranking of candidates for a search task. Consider the example in Figure 1. In the top half of the figure, the possible elements of a simple conjunction task are shown. For the given target, inhibition would rank the possible distractors depending on which features they possess (and to what degree). I have specified only a "+/−" scheme here; this is not to say that the ranks are equally spaced. The response depends on the relative strength of the item and the amount of inhibition applied. This, in turn, depends on the distance of the distractor feature from the comparable target feature along the same dimension. The weaker the distractor relative to the target, the smaller its final response; the farther away a feature from a target's feature along the same dimension, the weaker its final response. The fact that features may not be independently computed (coarse-coding, or neurons that are selective for both color and orientation, for example) complicates the determination of "same" dimension.



Simple Conjunction	Circle	Coarse Texture
Target	+	+
Possible distractors	+	−
	−	+
	−	−

Triple Conjunction	Circle	Coarse Texture	Line
Target	+	+	+
Possible Distractors	+	+	−
	+	−	+
	+	−	−
	−	−	+
	−	−	−
	−	+	−
	−	+	+

Figure 1. A comparison of possible relative effects of inhibitory tunings with known targets for simple and triple conjunction experiments. The "−" implies inhibition, the "+" denotes no change, both with respect to the relevant feature dimension. The magnitude of inhibition is not considered here; it would have an important effect in the actual ranking of candidates.

In a typical conjunction display, some combination of target and distractors is presented. Each display poses a potentially different distribution of relative rankings of candidate elements. It cannot be assumed that each display is of precisely the same difficulty. This is even more evident in a triple conjunction where the possible distributions of candidate rankings are even more varied, as shown in the triple conjunction example of Figure 1. If search does proceed by selecting candidates in order according to response strength, then it is easy to see how triple conjunctions may be faster than simple conjunctions. All that is required is to ensure that the ranking always leaves the target on top and that the distractors, even if ranked second, be distant seconds.

The proposal described in the preceding paragraph would lead to the observations of Egeth et al. (1984), who found that subjects can eliminate a feature dimension from consideration if instructed to do so. Treisman & Sato (1990) found that triple conjunctions can be fast if the target differs from distractors in two dimensions (representing two sources of inhibition for distractors) but that this is harder than a simple conjunction if the difference is only on one dimension. It also predicts the observations of Wolfe et al. (1989). Wolfe et al. always use size in their triple conjunctions and the target is always larger by at least double. It is easy to see how inhibition selective for scale can strongly favor the large element over the small. If Wolfe et al. repeated their experiments with small targets, my proposal predicts slower searches. It is odd that size plays such a large role in their experiments because Cave & Wolfe (in press) say that stereo and size are very effective for top-down guidance. Treisman & Sato (1990) report that conjunctions involving large size are faster than with small size. Burbeck & Yap (1990) recently reported that scale seems selectable based on context, with the largest response dominating. Further support for the proposal comes from Quinlan & Humphreys (1987), who observed that target-distractor discriminability influences the rate of conjunctive search. Another way of influencing the selection of candidates is to precue for location. Treisman (1985) reports a large advantage to precueing for location in conjunctive search whereas it is irrelevant in disjunctive searches. This, too, is consistent with the proposal.

Wolfe describes an odd-man-out problem that is surely "easy," i.e., parallel, with practice. If a unique item is created by one or more differences over the defining distractors, but all distractors are the same, then I must agree that the search appears easy, especially with practice.

It seems that my predictions for Type I and III agree well with observations. I have only one set of experimental data with which to compare with equations for Type II and IV; data supplied graciously by Anne Treisman (Treisman & Sato 1990). In that experiment, targets were unknown to subjects and displays were created with 4, 9 or 16 items. Targets consisted of (a) large items; (b) small items; (c) large-colored items; (d) small-colored items; (e) large-oriented items; and (f) small-oriented items. Using the standard method, the response times for each of these six conditions are plotted against display size and lead to linear relationships of varying slopes. These data are really three-dimensional, however, with the third dimension the feature dimension. My predictions for Type II and IV call for an exponential relationship in this dimension and a linear one in the display size dimension as observed. But how should we plot this feature dimension? It will not do to simply enumerate the features; it cannot be assumed that they are computed with equal ease. I fit exponential curves of the type predicted leaving the y-intercept and constraining the $\Phi(\hat{M})$ to have the same value across all display sizes for the same feature combination. The result is shown in Figure 2. The fit is very good.[4] The values of the exponents for each condition are: (a) 4.54; (b) 4.66; (c) 4.73; (d) 4.82; (e) 5.77; and (f) 6.12. The y-intercepts are: for 4 items, 564; for 9 items,

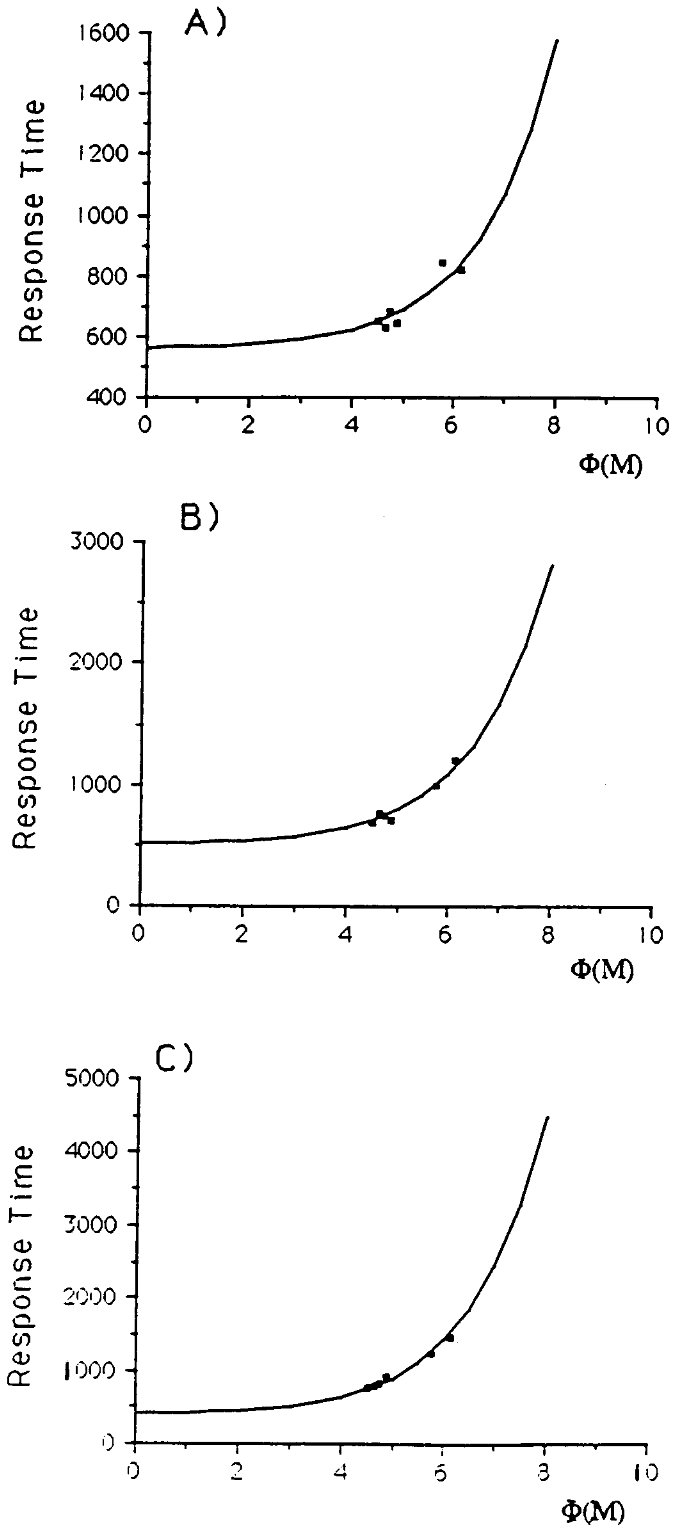


Figure 2. Plots of response time versus feature dimension $\phi(M)$ for unknown target experiments of Treisman & Sato (1990). (A) 4 items in display; smooth curve is $RT = 564 + 4 \cdot (2^{\phi(M)} - 1)$. (B) 9 items in display; smooth curve is $RT = 516 + 9 \cdot (2^{\phi(M)} - 1)$. (C) 16 items in display; smooth curve is $RT = 411 + 16 \cdot (2^{\phi(M)} - 1)$. The values of $\phi(M)$ found for each of the display types are: large size: 4.54; small size: 4.66; large-color: 4.73; small-color: 4.82; large-orientation: 5.77; small-orientation: 6.12. These are common across display size.

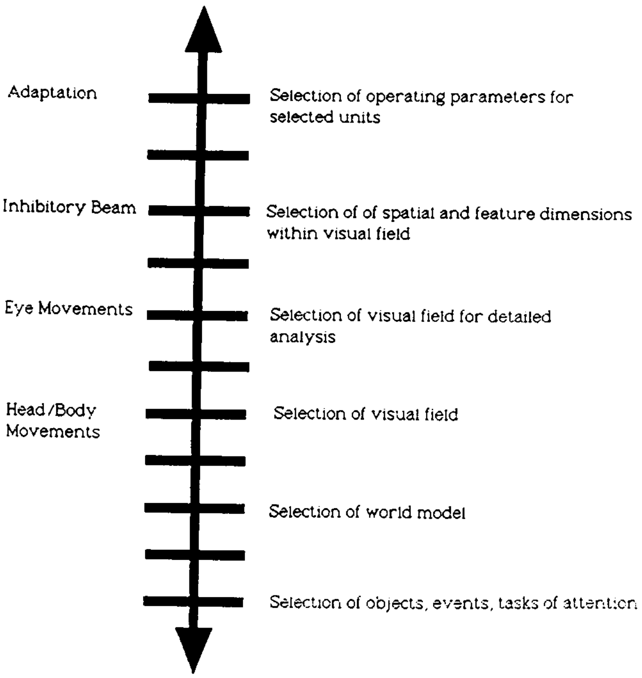

Adaptation
Selection of operating parameters for selected units
Inhibitory Beam
Selection of of spatial and feature dimensions within visual field
Eye Movements
Selection of visual field for detailed analysis
Head/Body Movements
Selection of visual field
Selection of world model
Selection of objects, events, tasks of attention

Figure 3. The spectrum of mechanisms for visual attention.

516; and for 16 items, 411. One set of curves alone does not confirm predictions with certainty; I hope further experimental data will test this prediction.

6. The inhibitory beam within attention

Zucker, Dickinson and **Desimone** all argue that other manifestations of attention should be studied, namely, adaptation and the oculomotor system. I could hardly disagree. In addition, they argue against the concept of the inhibitory attentional beam that I describe, as does **Strong. Eagelson,** too, points out that I have not considered other aspects of attention, such as spatial indexing. Before discussing this further, it is useful to try to place different attentional mechanisms into context. In Figure 3, I have tried to show (incompletely) the spectrum of attentional mechanisms. The spectrum is organized by the size of the "space" selected by attention, where space does not refer only to the three-dimensional world. The largest selection is that of task, then of the world model within which the task is to be solved, then the 3D-visual space that is relevant, then subsets of visual space, then subsets of computing units to apply, and, finally, the operating parameters for each unit. Adaptation is the lowest form of attentional manifestation in this categorization, my beam idea is next, and actions of the oculomotor system are above that. There is no claim in the target article that the beam is sufficient to account for all manifestations of attention. In visual search, the task, world model, and visual space are all preselected as part of the experimental conditions. Adaptation may be relevant for the experiment as a whole rather than on a case-by-case basis. Once the subject has adapted to the experimental conditions, adaptation may play a lesser role. Selection of subunits is relevant during the course of the experiment, however, and for this reason the beam falls out naturally as an important attentional mechanism. Of course, the others have influence, but they seem to be already fixed by the time the experiment is well underway.

Strong presents another criticism of the beam idea, namely, that I have contradicted the observed neuroanatomy with respect to back projections among visual areas in the cortex. Not at all; I recognize fully the extent of such connections. Maunsell & Newsome (1987) present a wiring diagram for the several areas in visual cortex. By definition, areas are grouped into layers depending on the specific connections they have with other areas; both forward and backward connections are included. They show seven layers that organize 16 areas into a hierarchy.[5] No layer includes more than six maps. No connection spans more than four layers with the majority spanning only one layer. Specifically, there are no connections from the highest layer to the lowest. This hierarchy is consistent with the bounds on the number of maps I derive for the most abstract retinotopic layer and, further, the logarithmic hierarchical convergence I propose would lead to similar numbers of layers, as **Uhr** points out. Van Essen and Anderson (1989) cite evidence of 15 to 30 resolvable steps for the size of the effective window of visual attention irrespective of retinal position and spatial scale of interest. My prediction for the size of maps, the width of the hexagon being about 21 and thus the existence of 21 different sizes of receptive fields, fits well with the number of steps of attention if attention must pass through these receptive fields.

Strong misses the point of the argument in section 4.5; I could probably have provided more discussion on this point. My argument is the following. I claim that high spatial resolution cannot be achieved by providing direct connections from the processor layer to the early layers of the input abstraction hierarchy. The sheer number of such connections makes this solution impossible. The only solution that is consistent with the neuroanatomy is indirect connection, i.e., access through the hierarchy as I proposed. This, of course, necessitates backward projections.

The connectivity pattern observed in primates is completely consistent with this reasoning. The total average connectivity with the beam in place increases by a factor of only 2. It would have also been more correct for me to return to the calculations of section 4.4 on N and M and to revise those figures based on an average connectivity of 500 rather than 1,000. P becomes 817 and M becomes 6. Maunsell & van Essen (1987) point out that area MT*[6] has a spatial extent of about 33 to 80 mm^2 in the macaque. They found a linear relationship between body weight and area to be the strongest single factor influencing the size of MT*.[7] If we extrapolate these results to humans of about 65 kg, human MT* would have an area of about 910 mm^2. If hypercolumns are about 1 mm^2, then this puts the number of spatial units of resolution in the range predicted by my calculations (John Maunsell, personal communication, 1989).

It is probably useful to summarize the role I put forward for the attentional beam concept. The beam arose out of the need to provide access to high-resolution information in a biologically plausible manner. High resolution is needed not only in space, but also in the features associated with each spatial location. So although the initial motivation for the beam arose out of the contradiction created by the determination of lower bounds on map size, the need is identical for the feature dimension. In the algorithm given in section 5.0, attention is applied in two key places, step 2 (section 5.2.2) and step 6 (section 5.2.6). In the first instance, it is used to "tune" the entire input hierarchy to expect features that are specified by the task and does not involve selection of spatial elements. One of the major effects of this tuning is to manipulate the distribution of competitors in the winner-take-all processes that are responsible for decision making. This manipulation changes the response characteristics of those processes, leading to enhanced response values achieved in shorter time. Haenny, Maunsell and Schiller (1988), as well as Desimone (1990), have observed a change in time course of response for attended units. Enhanced response is therefore a side-effect of an inhibitory mechanism. It is not necessarily the case that enhanced response implies an enhancement mechanism, as **Krueger & Tsay** suggest. Krueger & Tsay go on to point out that there is little evidence for any attentional effects on what features are extracted or compared. Their view is out of date and incompatible with the observations of Moran & Desimone, Haenny, Maunsell & Schiller and many others as cited in the target article.

The second instance of attention is the application of the beam as it was originally motivated, namely, to select spatial candidates for matching. The mechanism for this is currently being investigated. **Treisman** correctly criticizes the lack of detail for the implementation of the beam. I had intended it as an analogy only for the purpose of the target article, however, and a future paper will provide detail (Tsotsos, in preparation). Treisman points out that it is difficult to attend selectively to one feature at a particular location and to ignore others. How can one then account for the results of Moran & Desimone (1985) and other similar findings? In those experiments, exactly this type of selection occurs within individual receptive fields. I would suggest that some other mechanism must play a confounding role for the interferences Treisman cites. For example, coarse-coding of features would cause this effect to fall out naturally. Suppose a single unit codes both color and shape to some degree in a coarse-coded fashion. If only spatial selectivity is applied to that unit, it may indeed be difficult to select color over shape. This would be confirmed by my explanation too, as long as the only dimension of attention was location. Moran & Desimone and others showed that features selected can also be within units. My beam idea thus includes both of these aspects and the specification of the task determines which is used, if not both.

I still maintain that detectors in the early abstraction hierarchy are almost never in their "untuned" state; vision is almost always purposeful and if it is, the visual system will attempt to tune its resources in the direction most suited for that purpose. Optimizing the tuning of detectors for expected features will lead to faster responses for at least two reasons: competing items in the display are attenuated, and, winner-take-all results are speeded up as shown in section 4.5. **Treisman** points out that my alternative explanation for search asymmetries cannot be correct because I misinterpreted the "instructions" given to subjects. I did not misunderstand, but certainly misstated my understanding in the article. I realize that subjects only see the stimuli and nothing else and are not told to search for a logical negation of features. Attentional tuning for search asymmetry cases would inhibit any detectors that would respond to the features that were not part of the target. In an attempt to put the requisite decisional process into computation terms, I used the term "logical negation" and thus created some confusion about my meaning. Treisman asks why "could the standard stimuli not be found simply by leaving the detectors in their untuned state, so that only the target is effective?" Does that mean during the course of an experiment? How is it proposed to turn off attentional or top-down influences in a conscious human subject? I agree, however, with Treisman's clarification of the difference between "discovering existence" and "specifying identity."

7. Representation

Desimone and **Zucker** question the need for certain representations of features within the framework that I have presented. They (along with several other commentators but on different issues) must be reminded of the caveat I made early in the target articles. The derived constraints and framework resulted from complexity considerations **alone.** This is the view of vision that complexity alone can yield – a considerable one, all things considered, but certainly not complete. Moreover, in my definition of features, I deliberately left them unspecified because the goal was simply to count how many were possible. This led to a lower bound for the number of physical feature maps. I fully recognize that feature representation is a "murky" area; I do not think I have contributed much to it other than to place constraints on numbers of features.

Zucker points out that the beam idea requires continuous representation of features across space. In the idealized framework I present this is true.[8] I might point out that Zucker's own work, on curvature for example, also has this requirement and does not reflect the spatially fragmented nature of representations in the cortex. Do those breaks and gaps in representation have functional value, or are they artifacts of evolution or some other mechanism? We do not know at this point. I know of no model that has intentionally included the seemingly random gaps and anomalies of representation one finds in biology. How could it? We do not yet understand what a complete representation could be doing let alone one that seemingly cannot cover visual space adequately.

Treisman points out the need for object files, temporary ad hoc representations that are not hard-wired in a prelearned visual dictionary; she claims that my visual search algorithm has a problem because it does not include this. In step 1 of the algorithm (section 5.2.1), I describe the need to store a representation of the target. In step 7 (section 5.2.7), I argue for the need for a buffer representation, precisely because the wiring requirements would be too great. These temporary representations serve the same purposes (roughly) as Treisman's object files. **Mohnhaupt & Neumann** also point out the need for such an intermediate representation in vision, citing much relevant research.

8. Concluding remarks

The research described in the target article was first published as a technical report at the University of Toronto dated September 1987 (Tsotsos 1987a), shortly before that paper was submitted to *Behavioral and Brain Sciences.* The arguments of section 3 first appeared at the International Conference on Computer Vision, London, June 1987 (Tsotsos 1987b). **Cave**'s comment about the newness of the results is not correct; at the time of submission, none of the researchers currently espousing visual search explanations involving inhibitory guidance (**Wolfe, Cave, Treisman**) were doing so. The idea was indeed new back then, as was the explanation for visual search.

Throughout my development, I attempted to include only minimal assumptions and very simple optimizations within the framework. **Heathcote & Mewhort** believe that I confuse simplicity with tractability. If tractability can be achieved simply, then the result is all the more powerful. If you need to hang a picture frame, do you use a jackhammer for that nail or a simple tack hammer? In addition, I certainly do not propose that further optimizations are not possible. Some commentators objected to my pointing out that the best use must be made of the tools provided or that I could have chosen more powerful or extensive optimizations (**Strong, Cave,**[9] **Wolfe**[10]). One should remember that not all dimensions of a problem can be optimized simultaneously. How to choose which dimensions should be optimized and by how much is a judgment call – my intuition versus yours. I opted for a principle of least commitment. Who is right? Time will tell, of course. Science has always favored simple explanations for complex phenomena and it is our challenge to find them. If a solution is indeed too simple, then it should be easy to demonstrate this because it will not account for the experimental observations as well as another more sophisticated theory. This is how science progresses. I cannot claim at this point that I have found the complete and correct explanation; I can only hope that I have provided some useful constraints that delimit the future search for the solution and some hypotheses for one possible model. I was actually quite surprised to see how much can be explained with simple mechanisms and the single dimension of study on which I embarked.

Finally, I wish to emphasize strongly that complexity theory is as appropriate for the analysis of visual search specifically and of perception in general as any other analytic tool currently used by biological experimentalists. Experimental scientists attempt to explain their data and not just to describe it; it is not surprise that their explanations are typically well thought out and logically motivated, involving procedural steps or events. In this way, a proposed course of events is hypothesized to be responsible for the data observed. There is no appeal to nondeterminism or to oracles that guess the right answer or to undefined, unjustified, or "undreamed-of" mechanisms that solve difficult components. Can you imagine theories that do have these characteristics passing a peer-review procedure? They wouldn't pass such a procedure (at least not in our current view of science!). In proposing an explanation, experimental scientists attempt to provide an *algorithm* (using the definition of algorithm provided in my section 1.1) whose behavior leads to the observed data. Because biological scientists provide algorithmic explanations, computational plausibility is not only an appropriate but a necessary consideration. One dimension of plausibility is satisfaction of the constraints imposed by the computational complexity of the problem, the resources available for the solution of the problem, and the specific algorithm proposed.

ACKNOWLEDGMENTS

I thank Bart Selman, Greg Dudek, and Joachim von zur Gathen for useful discussions.

NOTES

[1] This line of reasoning is borrowed from Garey & Johnson (1976, pp. 90–91), who demonstrated that even though the Partition problem has a pseudopolynomial-time algorithm, it is still NP-Complete. The proof for the NP-Completeness of Knapsack involves a reduction from Partition.

[2] It seems that the great majority of pseudopolynomial-time algorithms for NP-Complete number problems are derived using the methods outlined by Bellman (1954) and Dantzig (1957).

[3] For example, no algorithms are known for linear programming that are parallelizable (Dobkin et al. 1979). Linear programming is used for approximate solutions to Knapsack.

[4] Error data, etc. were unavailable for proper statistical analysis of the fit.

[5] Van Essen and Anderson (1990) note that 24 visual areas are currently known.

[6] The heavily myelinated zone of the superior temporal sulcus area that is direction-selective receiving input from striate cortex.

[7] MT* (sq.mm.) = 14·body weight (kg.)

[8] Bob Desimone points out that recent experimentation has found a rather continuous representation of feature values along a given dimension, say color, at a given spatial location in V4 (personal communication).

[9] **Cave**'s suggestion for encoding the relevant maps with each prototype leaves open the problem of how to recognize colored objects in a black and white image, normally stationary objects that are moving, and other such exceptions from default settings.

[10] **Wolfe**'s "optimal" use of top-down guidance leads him and his colleagues to appeal to undefined noise effects to "fix" their model because it "works too well."

References

Letters "a" and "r" appearing before authors' initials refer to target article and response respectively.

Albano, J. E., Mishkin, M., Westbrook, L. E. & Wurtz, R. H. (1982) Visuomotor deficits following ablation of monkey superior colliculus. *Journal of Neurophysiology* 48:338–51. [RD]

Allman, J., Miezin, F. & McGuinnis, E. (1985) Stimulus specific responses from beyond the classical receptive field: Neurophysiological mechanisms for local-global comparisons in visual neurons. *Annual Review of Neuroscience* 8:407–30. [aJKT]

Anderson, C. & van Essen, D. (1987) Shifter circuits: A computational strategy for dynamic aspects of visual processing. *Proceedings of the National Academy of Science* USA 84:6297–6301. [aJKT]

Ballard, D. (1986) Cortical connections and parallel processing: Structure and function. *Behavioral and Brain Sciences* 9(1):67–90. [aJKT]

(1989) Animate vision. *Proceedings of the International Joint Conferences on Artificial Intelligence*, Detroit. [aJKT]

Ballard, D. & Brown, C. (1982) *Computer vision*. Prentice-Hall. [rJKT]

Ballard, D., Hinton, G. & Sejnowski, T. (1983) Parallel visual computation. *Nature* 306(5938):21–26. [aJKT]

Barlow, H. (1986) Why have multiple cortical areas? *Vision Research* 26(1):81–90. [aJKT]

Barlow, H., Fitzhugh, R. & Kuffler, S. (1957) Change of organization in the receptive fields of the cat's retina during dark adaptation. *Journal of Physiology* 137:338–54. [SWZ]

Barrow, H. & Tenenbaum, J. M. (1978) Recovering intrinsic scene characteristics from images. In: *Computer vision systems*, ed. A. Hanson & E. Riseman. Academic Press. [aJKT]

Bellman, R. (1954) Some applications of the theory of dynamic programming: A review. *Operations Research* 2:275–88. [PRK, rJKT]

Bennett, B., Hoffman, D. & Prakash, C. (1989) *Observer mechanics*. Academic Press. [RE]

Biederman, I. (1988) Aspects and extensions of a theory of human image understanding. In: *Computational processes in human vision*, ed. Z. Pylyshyn. Ablex Publishing Corp. [aJKT]

Broadbent, D. & Broadbent, M. (1978) From detection to identification: Response to multiple targets in rapid serial visual presentation. *Perception and Psychophysics* 42(2):105–13. [aJKT]

Bureck, C. & Yap, Y. (1990) Spatial-filter selection in large-scale spatial-interval discrimination. *Vision Research* 30(2):263–72. [rJKT]

Burt, P. & Adelson, E. (1983) The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications* 31:4:532–40. [aJKT]

Cavanagh, P. (1987) Reconstructing the third dimension: Interactions between color, texture, motion, binocular disparity, and shape. *Computer Vision, Graphics, and Image Processing* 37:171–95. [PRK]

Cave, K. R. & Wolfe, J. M. (in press) Modeling the role of parallel processing in visual search. *Cognitive Psychology*, [KRC, AH, JMW, rJKT]

Chen, L. (1982) Topological structure in visual perception. *Science* 218:699. [MM]

(1989) Topological perception: A challenge to computational approaches to vision. In: *Perspective in connectionism*, ed. R. Pfeiffer. Elsevier Science Publisher. [MM]

Chignell, M. H. & Krueger, L. E. (1984) Further evidence for priming in perceptual matching: Temporal, not spatial, separation enhances the fast-*same* effect. *Perception & Psychophysics* 36:257–65 [LEK]

Church, A. (1936) An unsolvable problem of elementary number theory. *American Journal of Mathematics* 58:345–63. [aJKT]

Cook, S. (1971) The complexity of theorem-proving procedures. *Proceedings of the 3d Annual ACM Symposium on the Theory of Computing*. New York. [aJKT]

Corbeil, J.-C. (1986) *The Stoddart visual dictionary*. Stoddart Publishing Co. [aJKT]

Cowey, A. (1979) Cortical maps and visual perception. *Quarterly Journal of Experimental Psychology* 31:1–17. [aJKT]

Crick, F. & Asunama, C. (1986) Certain aspects of the anatomy and physiology of the cerebral cortex. In: *Parallel distributed processing*, ed. D. Rumelhart & J. McClelland. MIT Press. [aJKT]

Daniel, P. & Whitteridge, D. (1961) The representation of the visual field on the cerebral cortex in monkeys. *Journal of Physiology* 159:203–21. [aJKT]

Dantzig, G. B. (1957) Discrete-variable extremum problems. *Operations Research* 5:266–77. [rJKT]

Davis, L. (1989) Mapping classifier systems into neural networks. In: *Advances in Neural Information Processing Systems 1*, ed. D. S. Touretzky. Morgan Kaufmann. [GWS]

Desimone, R. (1990) Untitled abstract presented at Visual Search: Segmentation, Attention and Identification, January 19–21, 1990, Irvine, CA. [rJKT]

Desimone, R., Chein, S., Moran, J. & Ungerleider, L. (1985) Contour, color, and shape analysis beyond the striate cortex. *Vision Research* 25(3):441–52. [aJKT]

Desimone, R., Moran, J. & Spitzer, H. (in press) Neural mechanisms of attention in extrastriate cortex of monkeys. In: *Competition and cooperation in neural nets 2*, ed. M. A. Arbib. [SWZ]

Desimone, R. & Ungerleider, L. G. (1989) Neural mechanisms of visual processing monkeys. In: *Handbook of Neuropsychology, vol. II.*, ed. E. Boller & J. Grafman. Elsevier Press. [RD]

Desimone, R., Wessinger, M., Thomas, L. & Schneider, W. (1989) Effects of deactivation of lateral pulvinar or superior colliculus on the ability to selectively attend to a visual stimulus. *Society for Neuroscience Abstracts* 15:162. [RD]

Dobkin, D., Lipton, R. & Reiss, S. (1979) Linear programming is log space hard for P. *Information Processing Letters* 8:96–97. [rJKT]

Dobson, V. & Rose, D. (1985) Models and metaphysics: The nature of explanation revisited. In: *Models of the visual cortex*, ed. D. Rose & V. Dobson. John Wiley & Sons. [aJKT]

Dowling, J. (1987) *The retina*. Belknap Press. [rJKT]

Downing, C. & Pinker, S. (1985) The spatial structure of visual attention. In: *Attention and Performance XI*, ed. M. Posner & O. Marin. Lawrence Erlbaum. [aJKT]

Duncan, J. (1980) The locus of interference in the perception of simultaneous stimuli. *Psychological Review* 87(3):272–300. [aJKT]

Duncan, J. & Humphreys, G. W. (1989) Visual search and stimulus similarity. *Psychological Review* 96:433–58. [KRC]

Eckhorn, R., Bauer, R., Jordon, W., Brosch, M., Kruse, W., Munk, M. & Reitboeck, H. J. (1988) Coherent oscillations: A mechanism of feature detection in the visual cortex? Multiple electrode and correlation analysis in the cat. *Biological Cybernetics* 60:121–30. [RMS]

Egeth, H., Virzi, R. & Garbart, H. (1984) Searching for conjunctively defined targets. *Journal of Experimental Psychology: Human Perception and Performance* 10(1):32–39. [rJKT]

Eriksen, C. W., O'Hara, W. P. & Eriksen, B. A. (1982) Response competition effects in *same-different* judgments. *Perception & Psychophysics* 32:261–70. [LEK]

Farah, M. J. (1985) Psychophysical evidence for a shared representational medium for mental images and percepts. *Journal on Experimental Psychology: General* 114:91–103. [MM]

Feldman, J. (1985) Connectionist models and their applications. *Cognitive Science* 9(1):1–169. [aJKT]

Feldman, J. & Ballard, D. (1982) Connectionist models and their properties. *Cognitive Science* 6:205–54. [aJKT]

Finke, R. A. (1985) Theories relating mental imagery to perception. *Psychological Bulletin* 98:236–59. [MM]

Fleet, D., Hallett, P. & Jepson, A. (1985) Spatio-temporal inseparability in early visual processing. *Biological Cybernetics* 52:153–64. [aJKT]

Fleet, D. & Jepson, A. (1989) Hierarchical construction of orientation and velocity selective filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. II, 3:315–25. [aJKT]

Fodor, J. A. (1985) Précis of *The modularity of mind*. *Behavioral and Brain Sciences* 8:1–42. [LEK]

Funt, B. V. (1980) Problem solving with diagrammatic representations. *Artificial Intelligence* 13:201–30. [MM]

Fuster, J. (1988) Attentional modulation of inferotemporal neuron responses to visual features. *Proceedings of the Society of Neuroscience*, Toronto. [aJKT]

Gamble, E. & Koch, C. (1987) The dynamics of free calcium in dendritic spines in response to repetitive synaptic input. *Science* 236:1311–15. [RMS]

Gardin, F. & Meltzer, B. (1989) The analogical representation of naive physics. *Artificial Intelligence* 38:139–59. [MM]

Garey, M. & Johnson, D. (1979) *Computers and intractability: A guide to the theory of NP-completeness*. W. H. Freeman & Co. [arJKT, PRK]

Gibson, J. J. (1966) *The senses considered as perceptual systems*. Houghton Mifflin. [LEK]

(1979) *The ecological approach to visual perception*. Houghton Mifflin. [LEK]

Gleitman, H. & Jonides, J. (1976) The cost of categorization in visual search: Incomplete processing of targets and field items. *Perception & Psychophysics* 20:(4):281–88. [aJKT]

Goldberg, D. E. (1989) *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley. [GWS]

Goldman-Rakic, P. S. (1988) Changing concepts of cortical connectivity: Parallel distributed cortical networks. In: *Neurobiology of neocortex*, ed. P. Rakic & W. Singer. John Wiley & Sons. [GWS]

Gray, C. M. & Singer, W. (1989) Stimulus specific neuronal oscillations in orientation columns. *Proceedings of the National Academy of Science* 86:1698–1702. [RMS]

Grimson, W. E. L. (1986) The combinatorics of local constraints in model-based recognition and localization from sparse data. *Journal of the Association for Computing Machinery* 33(4):658–86. [aJKT]

Haenny, P., Maunsell, J. & Schiller, P. (1988) State dependent activity in monkey visual cortex II. Retinal and extraretinal factors in V4. *Experimental Brain Research* 69:245–59. [arJKT]

Haenny, P. & Schiller, P. (1988) State dependent activity in money visual cortex I. Single cell activity in V1 and V4 on visual tasks. *Experimental Brain Research* 69:225–44. [aJKT]

Hartline, H. (1940) The receptive fields of optic nerve fibers. *American Journal of Physiology* 130:690–99. [aJKT]

Hebb, D. (1949) *The organization of behavior*. John Wiley & Sons. [aJKT]

Hebb, D. O. (1958) Alice in Wonderland or psychology among the biological sciences. In: *Biological and biochemical bases of behavior*, ed. H. F. Harlow & C. N. Woolsey. University of Wisconsin Press. [AH]

Hinton, G. (1981) Shape representation in parallel systems. *Proceedings of the International Joint Conference on Artificial Intelligence*, Vancouver. [aJKT]

Hinton, G. E. (1989) Connectionist learning procedures. *Artificial Intelligence* 40:185–234. [DGL]

Hoffman, J., Nelson, B. & Houck, M. (1983) The role of attentional resources in automatic detection. *Cognitive Psychology* 51:379–410. [aJKT]

Hoffman, J. E. (1979) A two-stage model of visual search. *Perception & Psychophysics* 25:319–27. [KRC]

Holland, J. H. (1975) *Adaptation in natural and artificial systems.* University of Michigan Press. [GWS, rJKT]

Hopfield, J. J. & Tank, D. W. (1986) Computing with neural circuits: A model. *Science* 233:625–33. [RMS]

Hubel, D. & Wiesel, T. (1977) Functional architecture of macaque visual cortex. *Proceedings of the Royal Society of London* B 198:1–59. [aJKT]

Hummel, R. & Zucker, S. (1983) On the foundations of relaxation labeling processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 5:267–87. [aJKT]

Jonides, J. & Gleitman, H. (1976) The benefit of categorization in visual search: Target location without identification. *Perception & Psychophysics* 20(4):289–98. [aJKT]

Kaas, J. H. (1989) Why does the brain have so many visual areas? *Journal of Cognitive Neuroscience* 1(2):121–35. [aJKT]

Kahneman, D. & Treisman, A. (1984) Changing views of attention and automaticity. In: *Varieties of attention*, ed. R. Parasuraman & J. Beatty. Academic Press. [AT]

Kaufman, L., Okada, Y., Tripp, J. & Weinberg, H. (1984) Evoked neuromagnetic fields. In: *Brain and information: Event-related potentials*, ed. R. Karrer, J. Cohen, & P. Tueting. *Annals of the New York Academy of Sciences* 425:722–42. [LEK]

Kertzman, C. & Robinson, D. L. (1988) Contributions of the superior colliculus of the monkey to visual spatial attention. *Society for Neuroscience Abstracts* 14:831. [RD]

Kosslyn, S. M. (1980) *Image and mind.* Harvard University Press. [MM]

Krueger, L. E. (1978) A theory of perceptual matching. *Psychological Review* 85:278–304. [LEK]

(1989) Cognitive impenetrability of perception. *Behavioral and Brain Sciences* 12(4):769–70. [LEK]

Krueger, L. E. & Chignell, M. H. (1985) *Same-different* judgments under high speed stress: Missing-feature principle predominates in early processing. *Perception & Psychophysics* 38:183–93. [LEK]

Kuffler, S. (1953) Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology* 16:37–68. [aJKT]

Larkin, J. H. & Simon, H. A. (1987) Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science* 11:65–99. [MM]

Lawler, E. (1976) *Combinatorial optimization: Networks and matroids.* Holt, Rinehart & Winston. [rJKT]

Liu, L., Zhao, N. & Bian, Z. (1989) Can early stage vision detect topology? *Proceedings of the International Joint Conference on Artificial Intelligence* 11:1591–95. [MM]

Llinas, R. & Yarom, Y. (1986) Oscillatory properties of guinea-pig inferior olivary neurones and their pharmacological modulation: An *in vitro* study. *Journal of Physiology* 376:163–82. [RMS]

Lowe, D. G. (1985) *Perceptual organization and visual recognition.* Kluwer Academic Publishers. [DGL]

(1987) Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence* 31:355–95. [DGL, rJKT]

(1990) Visual recognition as probabilistic inference from spatial relations. In: *AI and the eye*, ed. A. Blake & T. Troscianko. Wiley. [DGL]

Mackworth, A. & Freuder, E. (1985) The complexity of some polynomial network consistency algorithms for constraint satisfaction problems. *Artificial Intelligence* 25:65–74. [aJKT]

Marr, D. (1982) *Vision: A computational investigation into the human representation and processing of visual information.* W. H. Freeman. [arJKT, LEK, RMS]

Maunsell, J. (1989) Personal communication, June. [rJKT]

Maunsell, J. & Newsome, W. (1987) Visual processing in monkey extrastriate cortex. *Annual Review of Neuroscience* 10:363–401. [arJKT]

Maunsell, J., Sclar, G. & Nealey, T. (1988) Task-specific signals in area V4 of monkey visual cortex. *Proceedings of the Society of Neuroscience*, Toronto. [aJKT]

Maunsell, J. & Van Essen, D. (1987) Topographic organization of the middle temporal visual area in the macaque monkey: Representational biases and the relationship to callosal connections and myeloarchitectonic boundaries. *Journal of Comparative Neurology* 266:535–55. [rJKT]

Maxwell, N. (1985) Methodological problems of neuroscience. In: *Models of the visual cortex*, ed. D. Rose & V. Dobson. John Wiley & Sons. [aJKT]

McLeod, P., Driver, J. & Crisp, J. (1988) Visual search for conjunctions of movements and form is parallel. *Nature* 332:154–55. [JMW]

Miller, J. P., Rall, W. & Rinzel, J. (1985) Synaptic amplification by active membrane in dendritic spines. *Brain Research* 325:325–30. [RMS]

Minsky, M. & Papert, S. (1969) *Perceptrons.* MIT Press. [MM]

Mohnhaupt, M. & Neumann, B. (in press) Understanding object motion: Recognition, learning and spatio-temporal reasoning. *Journal of Robotics and Autonomous Systems.* North Holland. [MM]

Moore, E. F. (1956) Gedanken-experiments on sequential machines. In: *Automata studies*, ed. C. E. Shannon & J. McCarthy. Princeton University Press. [WRU]

Moran, J. & Desimone, R. (1985) Selective attention gates visual processing in the extrastriate cortex. *Science* 229:782–84. [arJKT, RD, SWZ]

Motter, B. (1988) Responses of visual cortical neurons during a focal attentive task. *Proceedings of the Society of Neuroscience*, Toronto. [aJKT]

Mountcastle, V. (1957) Modality and topographic properties of single neurons of cat's somatic sensory cortex. *Journal of Neurophysiology* 20:408–34. [aJKT]

Mountcastle, V., Motter, B., Steinmetz, M. & Sestokas, A. (1987) Common and differential effects of attentive fixation on the excitability of parietal and prestriate (V4) cortical visual neurons in the Macaque monkey. *Journal of Neuroscience* 7(7):2239–55. [aJKT]

Nakayama, K. & Silverman, G. H. (1986) Serial and parallel processing of visual feature conjunctions. *Nature* 320:264–65. [JMW]

Neisser, U. (1967) *Cognitive psychology.* Appleton-Century-Crofts. [aJKT, KRC, J-OE, RMS]

Okada, Y. C., Tanenbaum, R., Williamsonn, S. J. & Kaufman, L. (1984) Somatotopic organization of the human somatosensory cortex revealed by neuromagnetic measurements. *Experimental Brain Research* 56:197–205. [LEK]

Parasuraman, R. & Davies, D., eds. (1984) *Varieties of attention.* Academic Press. [aJKT]

Pashler, H. (1987) Detecting conjunctions of color and form: Reassessing the serial search hypothesis. *Perception and Psychophysics* 41:191–201. [AH]

Pippenger, N. (1978) Complexity theory. *Scientific American* 238(6):114–24. [aJKT]

Poggio, T. (1982) Visual algorithms. *AI Memo 683.* MIT Press. [aJKT]

Posner, J. I., Choate, L. S., Rafal, R. K. & Vaughn, J. (1985) Inhibition of return: Neural mechanisms and function. *Cognitive Neuropsychology* 2:211–28. [RD]

Pour-El, M. B. & Richards, I. (1981) The wave equation with computable initial data such that its unique solution is not computable. *Advances in Mathematics* 39:215–39. [BWD]

(1982) Noncomputability in models of physical phenomena. *International Journal of Theoretical Physics* 21:553–55. [BWD]

Proctor, R. W. (1981) A unified theory for matching-task phenomena. *Psychological Review* 88:291–326. [LEK]

Proctor, R. W. & Rao, K. V. (1983) Evidence that the *same-different* disparity in letter matching is not attributable to response bias. *Perception & Psychophysics* 34:72–76. [LEK]

Pylyshyn, Z. (1984) *Computation and cognition.* MIT Press/Bradford Books. [aJKT]

(1987) *The robot's dilemma: The frame problem in AI.* Ablex. [RE]

(1989) The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition* 32:65–97. [RE]

Pylyshyn, Z. & Biederman, I. (1988) *Computational processes in human vision: An interdisciplinary perspective.* Ablex. [RE]

Pylyshyn, Z. & Storm, R. (1988) Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision* 3:179–97. [RE]

Quinlan, P. & Humphreys, G. (1987) Visual search for targets defined by combinations of color, shape and size: An examination of the task constraints on feature and conjunction searches. *Perception and Psychophysics* 41(5):455–72. [rJKT]

Rabbitt, P. (1978) Sorting, categorization and visual search. In: *Handbook of perception: Perceptual processing*, vol. IX, ed. E. Carterette & M. Friedman. Academic Press. [aJKT]

Ramachandran, V. S. (1985) Guest editorial: The neurobiology of perception. *Perception* 14:1–14. [RMS]

Rensinck, R. (1989) Personal communication, University of British Columbia, September. [rJKT]

Rosch, E. (1978) Principles of categorization. In: *Cognition and categorization*, ed. E. Rosch & B. B. Lloyd. Erlbaum. [MM]

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M. & Boyes-Bream, P. (1976) Basic objects in natural categories. *Cognitive Psychology* 8:382–439. [MM]

Rosenfeld, A. (1962) Automatic recognition of basic terrain types from aerial photographs. *Photogrammetric Engineering* 28:115–32. [aJKT]

Rosenkrantz, D. & Stearns, R. (1983) NP-complete problems. In: *The encyclopedia of computer science and engineering*, 2nd ed., ed. A. Ralston & E. Reilly. Van Nostrand Reinhold Co. [rJKT]

Rubin, J. & Kanwisher, N. (1985) Topological perception: Holes in an experiment. *Perception and Psychophysics* 37. [MM]

Rumelhart, D. & McClelland, J. (1986a) PDP models and general issues in cognitive science. In: *Parallel distributed processing*, ed. D. Rumelhart & J. McClelland. MIT Press. [aJKT]

(1986b) *Parallel distributed processing*. MIT Press. [aJKT]

Runeson, S. (1977) On the possibility of "smart" perceptual mechanisms. *Scandinavian Journal of Psychology* 18:172–79. [LEK, rJKT]

Sagi, D. & Julesz, B. (1986) "Where" and "What" in vision. *Science* 228:1217–19. [aJKT]

Schwartz, E. (1977) Spatial mapping in the primate sensory projection: Analytic structure and relevance to perception. *Biological Cybernetics* 25:181–94. [aJKT]

Sclar, G., Lennie; P. & DePriest, D. (1989) Contrast adaptation in striate cortex of macaque. *Vision Research* 29:747–55. [SWZ]

Searle, J. (1990) Is the brain's mind a computer program? *Scientific American* 262(1):26–31. [rJKT]

Selman, B. (1989) Personal communication, Department of Computer Science, University of Toronto, January. [rJKT]

Shepard. G. M. & Brayton, R. K. (1987) Logic operations are properties of computer simulated interactions between excitable dendritic spines. *Neuroscience* 23:151–66. [RMS]

Siegel, R. M. (in press) Non-linear dynamical system theory and primary visual cortical processing. *Physica D*. [RMS]

Simon, H. (1962) The architecture of complexity. *Proceedings of the American Philosophical Society* 106:467–82. [aJKT, GWS]

Skarda, C. A. & Freeman, W. J. (1987) How brains make chaos in order to make sense of the world. *Behavioral and Brain Sciences* 10:161–96. [RMS]

Sloman, A. (1975). Afterthoughts on analogical representations. In: *Proceedings on Theoretical Issues in Natural Language Processing* 164–68. [MM]

Spitzer, H., Desimone, R. & Moran, T. (1988) Both behavioral and neuronal performance are improved by increased attention. *Proceedings of the Society for Neuroscience*, Toronto. [aJKT]

(1988) Increased attention enhances both behavioral and neuronal performance. *Science* 240:338–40. [RD]

Sporns, O., Gally, J. A., Reeke, G. N. & Edelman, G. M. (1989) Reentrant signaling among simulated neuronal groups lead to coherency in their oscillatory activity. *Proceedings of the National Academy of Science* 86:7265–69. [RMS]

Steels, L. (1988) Steps towards common sense. *Proceedings ECAI* 88:49–54. [MM]

Stensaas, S., Eddington, D. & Dobelle, W. (1974) The topography and variability of the primary visual cortex in man. *Journal of Neurosurgery* 40:747–55. [aJKT]

Stockmeyer, L. & Chandra, A. (1979) Intrinsically difficult problems. *Scientific American*, May. [aJKT]

Stone, J., Dreher, B. & Leventhal, A. (1979) Hierarchical and parallel mechanisms in the organization of the visual cortex. *Brain Research Reviews* 1:345–94. [aJKT]

Strong, G. W. & Whitehead, B. A. (1989) A solution to the tag assignment problem for neural networks. *Behavioral and Brain Sciences* 12:381–433. [GWS]

Treisman, A. (1982) Perceptual grouping and attention in visual search for features and for objects. *Journal of Experimental Psychology: Human Perception and Performance* 8:194–214. [AH]

(1985) Preattentive processing in vision. *Computer Vision, Graphics and Image Processing* 31:156–77. [arJKT, AH, AT]

1986) Features and objects in visual processing. *Scientific American* 255:144B-125. [KRC, AT]

(1988) Features and objects: The fourteenth Bartlett memorial lecture. *The Quarterly Journal of Experimental Psychology* 40A(2):201–37. [aJKT, KRC, AH, AT]

Treisman, A. & Gelade, G. (1980) A feature-integration theory of attention. *Cognitive Science* 12:99–136. [aJKT, KRC]

Treisman, A. & Gormican, S. (1988) Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review* 95(1);15–48. [aJKT, AT]

Treisman, A. & Sato, S. (1990) Conjunction search revisited. *Journal of Experimental Psychology: Human Perception and Performance*, 16. [arJKT, AT, JMW]

Treisman, A. & Schmidt, H. (1982) Illusory conjunctions in the perception of objects. *Cognitive Psychology* 14:107–41. [aJKT]

Treisman, A. & Souther, J. (1985) Search asymmetry: A diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General* 114:285–310. [aJKT]

Tsotsos, J. (1987a) Representational axes and temporal cooperative processes. In: *Vision, brain and cooperative computation*, ed. M. Arbib & A. Hansen. MIT Press/Bradford Books. [aJKT]

(1987b) Image understanding. In: *The encyclopedia of artificial intelligence*, ed. S. Shapiro. John Wiley & Sons. [aJKT]

(1987c) *Analyzing vision at the complexity level: Constraints on an architecture, an explanation for visual search performance, and computational justification for attentive processes*, RBCV-TR-87-20, Department of Computer Science, University of Toronto, September. [arJKT]

(1988) A "complexity level" analysis of immediate vision. *International Journal of Computer Vision* 1(4):303–20. [arJKT]

(1989) The complexity of perceptual search tasks. *Proceedings of the International Joint Conference on Artificial Intelligence*, Detroit. [aJKT]

(in preparation) Attentional influences in vision: Applying an inhibitory attentional beam to intermediate level vision. [aJKT]

Turing, A. (1937) On computable numbers with an application to the Entscheidungs problem. *Proceedings of the London Mathematical Society* 2(43):230–65. [aJKT]

Uhr, L. (1972) Layered "recognition cone" networks that preprocess, classify and describe. *IEEE Transactions on Computers* C-21:758–68. [aJKT]

(1980) Psychological motivation and underlying concepts. In: *Structured computer vision*, ed. S. Tanimoto & A. Klinger. Academic Press. [aJKT]

Ullman, S. (1984) Visual routines. *Cognition* 18:97–159. [RE, J-OE]

(1989) Aligning pictorial descriptions: An approach to object recognition. *Cognition* 32:193–254. [MM]

Ungerleider, L. & Mishkin, M. (1982) Two cortical visual systems. In: *Analysis of visual behavior*, ed. D. Ingle, M. Goodale & R. Mansfield. MIT Press. [aJKT, RD]

Van Doorn, A., van de Grind, W. & Koenderink, J., ed. (1984) *Limits in perception*. VNU Science Press. [aJKT]

Van Essen, D. & Anderson, C. (in press) Information processing strategies and pathways in the primate retina and visual cortex. In: *Introduction to Neural and Electronic Networks*, ed. S. Zornetzer, J. Davis & C. Lau. Academic Press. [arJKT]

Van Essen, D. & Maunsell, J. (1983) Hierarchical organization and functional streams in the visual cortex. *Trends in Neuroscience* 6:370–75. [aJKT]

Vergis, A., Steiglitz, K. & Dickinson, B. (1986) The complexity of analog computation. *Mathematics and Computers in Simulation* 28:91–113. [BWD]

Von Bekesy, G. (1956) Current status of theories of hearing. *Science* 123:779–83. [LEK]

Watson, A. & Ahumada, A. (1987) An orthogonal oriented quandrature hexagonal image pyramid. *NASA Technical Memorandum 100054*. [aJKT]

Wise, S. & Desimone, R. (1988) Behavioral neurophysiology: Insights into seeing and grasping. *Science* 242:736–41. [RD]

Wolfe, J., Cave, K. & Franzel, S. (1989) Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance* 15:419–33. [arJKT, KRC, AH, AT, JMW]

Wolfe, J., Cave, K. & Yu, K. (1988) Direction attention to complex objects. *Proceedings of the Society for Neuroscience*, Toronto. [aJKT]

Zeki, S. (1978) Uniformity and diversity of structure and function in rhesus monkey prestriate visual cortex. *Journal of Physiology* 277:273–90. [aJKT]

Zucker, S. (1985) Does connectionism suffice? *Behavioral and Brain Sciences* 8(2):301–2. [aJKT]

Zucker, S. W. (1983) Cooperative grouping and early orientation selection. In: *Physical and biological processing of images*, ed. A. Sleigh. Springer. [RMS]