International Journal of Computer Vision, 303-320 (1988)

A 'Complexity Level' Analysis of Immediate Vision

JOHN K. TSOTSOS
Department of Computer Science, University of Toronto, 5 King's College Road, Toronto, Ontario MS5 1A4, Canada

Abstract

This paper demonstrates how serious consideration of the deep complexity issues inherent in the design of a visual system can constrain the development of a theory of vision. We first show how the seemingly intractable problem of visual perception can be converted into a much simpler problem by the application of several physical and biological constraints. For this transformation, two guiding principles are used that are claimed to be critical in the development of any theory of perception. The first is that analysis at the 'complexity level' is necessary to ensure that the basic space and performance constraints observed in human vision are satisfied by a proposed system architecture. Second, the 'maximum power/minimum cost principle' ranks the many architectures that satisfy the complexity level and allows the choice of the best one. The best architecture chosen using this principle is completely compatible with the known architecture of the human visual system, and in addition, leads to several predictions. The analysis provides an argument for the computational necessity of attentive visual processes by exposing the computational limits of bottom-up early vision schemes. Further, this argues strongly for the validity of the computational approach to modeling the human visual system. Finally, a new explanation for the pop-out phenomenon so readily observed in visual search experiments, is proposed.

Introduction

The task of visual perception can be shown to be intractable for brute-force architectures in a straightforward manner, and in fact, subproblems, such as polyhedral scene labeling, have been shown to be inherently NP-complete [Kirousis and Papadimitriou 1985]. Yet, human vision is an effortless and exquisitely precise sense. How can this be? In the past, researchers have resorted to processing limits and attention in order to cope with this dilemma. Neisser, for example, first claimed that any model of vision that was based on spatial parallelism alone was doomed to failure, simply because the brain was not large enough [Neisser 1967]. This led him to his two-stage process of perception: a pre-attentive phase followed by an attentive phase. However, it is difficult to couch such a model in computational terms, there are so many missing details. Moreover, the reason for the need for attention is less than satisfactory. Stating that the brain is simply not large enough does not yield any useful constraints on the architecture of the visual system. Yet, Neisser's claim hints at the difficult issues of computational complexity that must be addressed. More recently, Feldman and Ballard concluded that time complexity considerations lead to massively parallel models being the only biologically plausible ones, since only they satisfy the 100-step rule [Feldman and Ballard 1982; Ballard 1986]. That is, since neurons compute at a rate of about 1000 Hz, and since simple perceptual phenomena do indeed occur in about 100 milliseconds, then biologically plausible algorithms can require no more than 100 steps. They did not, however, explain exactly how massive these networks must be (also, see Zucker [1985]). Feldman and Ballard also stress the importance of conservation of connections. Although the emphasis is correct, their application of this constraint leaves many questions unanswered and, in particular, they did not demonstrate that their set of conserving techniques is sufficient. Rumelhart and McClelland claim that the time and space requirements of a

theory of cognitive function are important determinants of the theory's biological plausibility [Rumelhart and McClelland 1986a]. However, they do not provide any details on how such constraints may be satisfied. In this paper, a simple demonstration leads to the conclusion that parallelism, on its own, of biologically plausible degree is insufficient to satisfy the time complexity constraints for vision, and it is reasonable to speculate that it is insufficient, on its own, for any cognitive task. In addition, this paper proposes that satisfaction of the space and time complexity constraints be one of the elements of the test that new theories of visual perception must pass.

Computational complexity issues are broad and pervasive in the development of a theory of perception. The key philosophy underlying the research to be described in this paper is that the complexity considerations of the nature 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 computational complexity issues, even as part of the computational level of his theory [Marr 1982]. According to Marr, the computational level of a theory addresses the questions: What is the goal of the computation? Why is it appropriate? and, What is the logic of the strategy by which it can be carried out? The representational and algorithmic level of the theory 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, the implementational level of the theory asks: How can the representation and algorithm be realized physically? Complexity issues span these three levels. Much past work in computer vision, motivated by Marr's philosophy, has tacitly assumed that the language of continuous mathematics is equivalent to the language of computation. Mathematical modeling is *not* equivalent to computational modeling. In proposing a mathematical solution for a problem, say that of solving optic flow equations, one has not also solved the problem computationally, even if simulated on a computer. There are still issues of representation, discretization, sampling, numerical stability, and computational complexity (at least) to contend with. The key first component of the computational level, in my mind, is the consideration of complexity issues, and Marr did not explicitly include this in his definition. Thus, I claim that there is another level of analysis required for any theory of perception—the *complexity level.*

This paper is concerned *only* with the complexity level. In particular, strategies for how a tractable solution can be achieved are discussed involving both time and space considerations. I will not attempt to ascribe functional significance to specific brain areas, I will not claim particular neural models, I will not propose representation schemes, and I will not choose a set of specific visual entities. There will be no algorithm proposed for a computer vision system. On the other hand, this exercise is claimed to be a critical one in the computational modeling of perception, and indeed, in the computational modeling of any aspect of intelligence. One of the key problems with AI solutions for tasks involving intelligence is that the solutions are so fragile with respect to 'scaling up' with problem size. That is, theoretical solutions are derived, usually without theoretical regard to the amount of computation required, and then if an implementation is produced, it is tried out on a few small examples only. The standard claim then 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 research. In particular, parallel solutions, such as those proposed by the connectionist community, although motivated by complexity considerations, seem to neglect detailed considerations of computational complexity altogether (see the collections of papers on the subject in Rumelhart and McClelland [1986b] and Feldman [1985]). For example, few if any deal with the time and space requirements of the relaxation procedures that they use, particularly in the context of time-varying input (but see Tsotsos [1987a] for empirical results on this). The issues raised by time and timing are in general not handled well [Tsotsos 1986]. If one is in the business of realizing systems, and proving that they behave in the required manner, the first requirement of a realizable system is that the task attempted and/or the proposed solution be computationally tractable.

The Model of Computation

A very simple abstract model will be proposed for the task of visual perception involving four main elements, keeping in mind that the interest is primarily in the complexity of the task:

1. 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.
2. At each array element, one or more tokens representing physical parameters of the scene may be computed. These tokens are of a given type, and for each type there are many possible token instances. Types are not necessarily independent. Tokens are distinct from measurements (usually taken to mean the output of some convolution operation), features (usually imply some level of interpretation), and primitives (nondecomposable elements), but are intended to be the elements that comprise the output of early vision. It is thus assumed that the output of early vision is retinotopic. A map is defined as a retinotopic representation of only one type of visual parameter. Maps are logical abstractions, and not necessarily physically separable entities. There are M maps in the system and the types will be left unspecified and abstract.
3. A knowledge base of visual prototypes, each one representing a particular visual object, event, scene, or episode. There are VP of these prototypes. Each prototype may be considered as an invariant description of a visual entity (invariant for size, location, rotation, and other parameters as appropriate).
4. A large pool of identical processors, each having the capability of choosing 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, and 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 match process is the basic operation of the model. 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. This is clearly not a simple task. The output of a processor is match success or failure, with an associated goodness of fit measure. It is assumed that the entire sequence of processing steps, regardless of what they may be, are collapsed into a single processing layer. Each processor completes this operation in PS seconds. The final output of the system is also available in PS seconds, and thus the actual time required for this process does not matter.

The specific representations do not matter for this discussion. The costs associated with this model of computation are in the number of retinotopic elements (P), tokens ($M \times P$), visual prototypes (VP), and processors (PP); and the connectivity costs both in number of connections and their total length.

A time complexity function will be formulated in such a way as to address the number of comparisons of location/token sets to prototypes that need to be performed within a single bottom-up processing pass, in the worst case, with no prior information about the scene. It is claimed that the output of a single bottom-up pass through the entire visual system corresponds to pre-attentive vision, and leads to the pop-out phenomenon observed in perceptual experiments for certain stimuli [Treisman 1985]. If a percept 'pops out,' the perception is immediate and effortless. Note that in these experiments, if a target is not provided, it is not the case that nothing is recognized—access to the subject's entire knowledge base is still required. Recognition in this case may not be complete nor unambiguous pre-attentively, however. Thus, the entire knowledge base of visual prototypes must be included in the analysis that is to be presented. Stimuli are not matched to targets only; targets may be the first candidates for matching; but if they all fail, then other elements of stored visual knowledge are considered. This definition of pre-attentive vision may be expanded by noting that further processing beyond the first bottom-up pass will not yield a different interpretation. A minimal set of

optimizations are then introduced to change the architecture of the system so that the timing constraint is satisfied. The implications of the resulting architecture and complexity function are then examined, and lead to many characteristics of primate visual systems as well as to several predictions.

The Nature of the Computational Task

Neisser, among others, claimed that a spatially parallel model of perception is inadequate quantitatively [Neisser 1967]. Neisser was motivated by the fundamental dilemma faced by all theories based on spatial parallel processing: If more than one item of the same kind is present in the visual field, how are they distinguished? In order to deal with the entire visual field at once as well as all the possible interpretations, one requires a much larger brain and too much experience. Neisser's claim is easy to demonstrate. Given VP visual prototypes, P elements of a retinotopic array, and M types representing visual parameters at each array element, then

$$VP \times 2^{P \times M} \tag{1}$$

operations are required in the worst case. The number of possible subsets of location/type pairs is the powerset of all locations times parameter types. (The null set is included here, but has little effect at this stage of the discussion. It will be deleted later when it will make a difference.) Another possible complexity function would include M as a multiplier of the powerset of locations, rather than in the exponent of the powerset. However, this implicitly makes the assumption that only one type of parameter is required to define a visual entity, and this is true only in very special circumstances. The expression in equation (1) allows an arbitrary subset of parameters to be required for any visual entity. The expression in equation (1) does not enumerate the number of images, rather it enumerates the number of data items that must be considered and comparisons that must be performed with those data items in the worst case. This is clearly combinatorially explosive. Interestingly, there has been a recent proof that the common 'blocks world' problem, addressed by so many researchers in the late 1960s and the early 1970s is inherently NP-Complete in the number of lines [Kirousis and Papadimitriou 1985]. The specific theorems that they proved are: (a) It is NP-Complete, given an image, to tell whether it has a legal labeling and, (b) It is NP-Complete, given an image, to decide whether it is realizable as the projection of a scene.

We can demonstrate the implications of this complexity measure by using a few relevant estimates for human vision for 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 brand names for common objects. Thus a conservative lower estimate for the number of prototypes is $VP =$ 100,000. A reasonable but arbitrary upper estimate would be $VP =$ 10,000,000, for demonstration purposes. 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; and a temporal derivative on at least some of these quantities. At the photoreceptors, all of these types are rolled up into a single continuous signal. An upper estimate on M of 12 will be used for demonstration purposes. P is the number of locations in the retinotopic representation. For illustrative purposes, three values will be used, an upper, a middle, and a lower value. 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 $\times$ 1K image) is the middle value, and the size of a 256 $\times$ 256 image is the lower value (65536 pixels). It will become apparent that the particular choices for these parameters have no effect on the general conclusions, and it must be emphasized that the numerical choices for these parameters are for demonstration purposes only.

A time-complexity function for a task ex-

presses an upper bound on its time requirements by giving for each possible input, the largest amount of time needed, in terms of the input length. A priori, there is no way to predict which portions of the visual field will represent an image-specific projection of a given prototype, and thus in the most simple brute-force algorithm, a single processor in the worst case must consider each receptive field against each stored prototype, and in the average case, half that number of comparisons. It should be obvious that a parallel scheme requires much serious consideration of the problems of communication, shared resources, synchronization, task scheduling, etc. It is assumed that they can be resolved—there would be no impact on the results claimed in this paper. However, the effective speed-up due to parallelism is clearly smaller than the number of available processors.

Given *PP* as the degree of effective speed-up due to parallelism, then the amount of time taken to perform the worst case number of operations as presented in equation (1) is given by:

$$PS = \frac{2^{P \times M} \times VP \times PS}{PP} \qquad (2)$$

where each processor requires *PS* seconds to complete one operation and the output of the system is also available in *PS* seconds. Table 1 gives values for *PP* for the estimates on *P*, *M*, and *VP* described above. The inescapable conclusion is that with this simplified architecture, the task is intractable: *parallelism alone is not the answer.* Although the problem cannot be solved without parallelism, it is interesting to note that the 'connectionist' community claims that massive parallelism is sufficient to satisfy the timing constraints. Moreover, it is easy to show that this problem is in the class NP. A nondeterministic scheme would simply guess which of the possible input data subsets would lead to a non-null search problem solution.

If a task is computationally intractable, then the only realizable solutions are approximating ones. Kirousis and Papadimitriou—even though they are not vision researchers—recognized the apparent contradiction contained in their theorems. Biological vision is an existence proof and thus, they claim, one of two possibilities arise: (a) vision is easier since other cues such as color can be used or (b) the probabilistic distribution of real scenes is biased for the development of ingenious fast algorithms. The first speculation of Kirousis and Papadimitriou is easily dismissed. If a richer world is considered, then not only are more cues available, but the space of possibilities is also dramatically increased, and thus the addition of other cues can only worsen the tractability of the task. This, of course, assumes that no information is available a priori. A drastic improvement may be possible for specific situations when the viewer has some knowledge of what to expect. The second claim is more believable, yet seems to be very difficult to prove. There are two more views possible on the nature of approximating solutions. One approach is to search for polynomial time algorithms for various specific visual computations (see, for example, Poggio [1982] and Mackworth and Freuder [1985]). Although progress has been made in this direction, we are far from the development of such an algorithm for the entire problem of vision. Finally, 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

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

PP	$VP = 10^5$		$VP = 10^7$	
P	$M = 1$	$M = 12$	$M = 1$	$M = 12$
130,000,000	$10^{39,133,905}$	$10^{469,608,805}$	$10^{39,133,907}$	$10^{469,608,807}$
1,000,000	$10^{301,035}$	$10^{3,612,365}$	$10^{301,037}$	$10^{3,612,367}$
65,536	$10^{19,733}$	$10^{236,744}$	$10^{19,735}$	$10^{236,746}$

such that performance degrades gracefully for the more complex situations. Then, one may ask the question 'How large a problem can the brain handle?' In part, this question motivated the approach in this research. Put differently, 'What are the limits of a bottom-up single pass early vision process?' Only by first answering this question can the computational need for attention be justified and a strategy for attentive processing be developed.

In formulating an answer for this question, one must employ some criterion for deciding between competing configurations. In computer science, 'computing power' is commonly used to rate various computer systems. This is defined as the number of operations performed per second. A similar decision principle can be stated for choosing 'best' configurations for vision systems:

The Maximum Power/Minimum Cost Principle

The power of a pre-attentive vision system is defined as the amount of data that may be processed per degree of parallelism (or per processor, for simplicity), within a single bottom-up pass. Power is increased by increasing VP, M, or P or by decreasing the required degree of parallelism PP. The cost of a system is a function of the number of units allocated for the maps, the number of processors, the required fan-in and fan-out, the number of total connections, and the total connection length. Preferred configurations are those that maximize power while minimizing cost. The goal is to maximize the richness of the visual world that is immediately accessible to the system, within the hardware constraints.

Demonstrating Complexity Sufficiency

A time complexity function has been formulated for a brute force architecture attacking the first, bottom-up pass of visual information processing. The goal now is to discover a sufficient set of global optimizations so that a biologically plausible architecture is obtained, that yields a sufficient, but not necessary, solution to the timing constraints. Then, space complexity considerations, using connectivity, will be presented that lead to specific predictions on the size of problem that human vision solves. The side effects of the complexity considerations will also be presented. The result will be an argument supporting computational modeling of human vision, and an argumen for a sufficient architecture for biologically motivated designs of vision systems.

Biologically plausible values for PP, P, and M are:

- For PP: The speed-up due to parallelism is clearly at least one, 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 presumably, a collection of neurons is required to accomplish this for each degree of speed-up. Since about 20% of the cerebral cortex is devoted to visual processing, the value of PP that is biologically plausible is significantly less than 10^9.
- For P: Stensaas and colleagues measured the size of striate cortex (V1) in humans, and found that its extent averaged approximately 2100 mm^2, and ranged from 1500—3700 mm^2. for each hemisphere [Stensaas et al. 1974]. Hubel and Wiesel claimed that the basic processing unit within the visual cortex is a hypercolumn, a localized collection of neurons, organized in columns, providing a complete set of processors for orientation, color, motion, etc. [Hubel and Wiesel 1977]. The receptive fields within each hypercolumn are all localized to the same region of visual space, thus the representation is retinotopic. Since a hypercolumn is approximately 1 mm^2 in extent, then V1 contains on average 2100 hypercolumns. Extrastriate areas are smaller, and have smaller hypercolumns (where hypercolumns hav ebeen found). It is assumed that the output of early vision corresponds to the output of the most abstract, retinotopic extrastriate areas. If, in an abstract sense, the elements of the retinotopic representations in this paper are equated with hypercolumns with respect to map resolution, then acceptable values for P are those less than 2100.

— For M: According to van Essen and Maunsell, the division between retinotopic and non-retinotopic areas, although fuzzy in general, may be placed after areas MT and V4 and before IT, area 7, and the frontal eye fields [van Essen and Maunsell, 1983]. Thus, at the assumed output of early vision, V4 seems to have three-separate representations of visual space, and MT has one [van Essen and Zeki, 1978]. Since the total number of visual areas is on the order of 20, and only some are retinotopic, and each may have several representations, acceptable values for the number of representations that comprise the output of early vision, that is M, are less than 20.

Additional Assumptions

Hexagonal images are assumed, packed with hexagonal pixels, of order N (i.e., N pixels per side). A hexagonal tiling of a hexagonal image was chosen for convenience, and for the resemblance to the retinal mosaic; however, much of the discussion is independent of the choice of image mosaic. Whenever the choice does have an impact on the results, it will be pointed out. The number of pixels in such a hexagonal image is $P_N = 3N^2 - 3N + 1$. In these hexagonal images pixels are uniformly distributed across the image. All receptive fields and prototypes are hardwired to the processors, and all data from a receptive field can be accessed simultaneously, as is also true for all data associated with a visual prototype. Connectivity will be consdiered in terms of other units, not synapses. All maps are assumed to be the same size.

A Sufficient Set of Optimizations

This section will develop a small, sufficient set of optimizations that will lead to a biologically plausible function relating P, PP, VP, and M. The development can be likened to a 'back of the envelope' computation, with estimates for variable values used only to guide the development. Once the function has been determined, conclusions will be drawn about the values of some of the variables.

Efficiency can be gained by attacking the prototype search through a process of successive refinement. This is quite a standard tactic in AI—'divide and conquer,' has been used in wide varieties of AI systems. As used here, the idea is similar to the perceptual '20 questions game' described in Richards [1982]. 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 of decomposition hierarchies found in the knowledge representation literature. Note that although a binary tree search is serial in nature, the key here is the number of operations, and the search will be parallelized later in the paper. Therefore,

$$PP = 2^{P \times M} \times \log_2 VP \tag{3}$$

This is a very minor improvement. On its own, the standard technique employed by all knowledge-based vision systems, namely prototype organization, is at best a small (but crucial) contributor in defeating the complexity problem of vision. Note that although a hashing scheme would be even faster, it would not be sufficient on its own to make a difference; and further, it is not clear what biologically plausible mechanism could implement hashing.

A critical observation on the physical world is that it is not the case that all 2^P possible combinations of locations are meaningful and thus reasonable to consider. Objects are not spread arbitrarily in 3-space, and events are not spread arbitrarily in the time dimension. Their physical characteristics are also similarly localized. Assuming a hexagonal image of order N, and that only hexagonal contiguous regions of whole array elements are considered as processor receptive fields, then some simple geometry yields N^3 receptive fields over the whole image or, in pixels, approximately,

$$\frac{P^{1.5}}{3\sqrt{3}} + \frac{P}{2} + \frac{5\sqrt{P/3}}{8} \tag{4}$$

Only this number of receptive fields need be considered. Figure 1 illustrates the receptive field structure. The degree of speed-up function for this third architecture is dramatically different:

$$PP = N^3 \times (2^M - 1) \times \log_2 VP \tag{5}$$

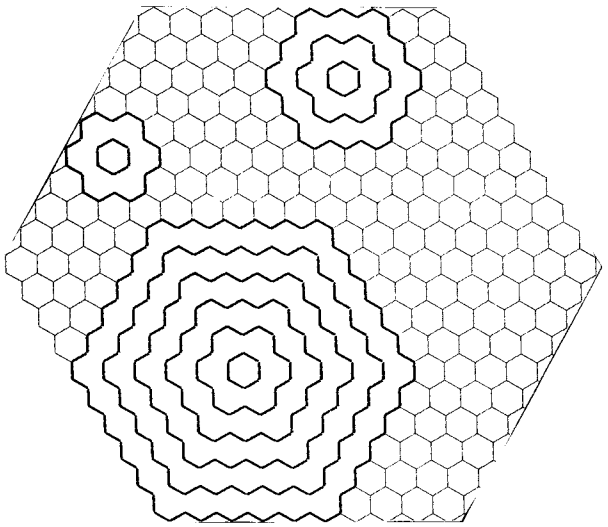

Fig. 1. The hexagonal retinotopic stimulus representation. The hexagon is of order N, that is, N elements per side. The diameter of the hexagon is $2N - 1$ elements. For three of the elements, the corresponding complete set of receptive fields that may centered on those elements are shown. Each element in the hexagon can be the center of a number of hexagonal receptive fields in this manner. In total, there are N^3 receptive fields.

The powerset of maps still remains in the expression because, a priori, it is not known which subset of maps is the correct one for a best image to prototype match; and, therefore, in the worst case all subsets must be examined. The null set has been removed however, since it may have a numerical effect. Table 2 gives the values for *PP* resulting from this expression. Although there has been significant change in the estimated degree of speed-up, the values are still not close to biologically plausible values. One important consequence of the localization of receptive fields is that no comparative relations between receptive fields may be computed. This is not worrisome since it has been observed in humans that the determination of spatial relations requires serial processing and is not a pre-attentive ability [Ullman 1983], and thus this optimization leads to an implication consistent with the observations. Another side-effect of this particular receptive field structure is that it does not permit as fine a selection of tokens across the receptive field as the first expression (equation (1)). In equation (1), some of the subsets could indeed

Table 2. PP for varying values of *P*, *M*, and *VP* for architecture 3.

PP	$VP = 10^5$		$VP = 10^7$	
P	$M = 1$	$M = 12$	$M = 1$	$M = 12$
130,000,000	$10^{12.68}$	$10^{16.29}$	$10^{12.82}$	$10^{16.43}$
1,000,000	$10^{9.5}$	$10^{13.12}$	$10^{9.65}$	$10^{13.26}$
65,536	$10^{7.73}$	$10^{11.35}$	$10^{7.88}$	$10^{11.49}$

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 are used for each location across the receptive field. This too is reasonable, since visual parameters display the same localization as the objects which exhibit them.

As a third potential optimization, we note that it is not the case that all visual stimuli involve all types of tokens. Let $\hat{M}$ represent the number of types of visual parameters that are 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 aassociated with each map that signals whether or not the map has been activated. The idea is borrowed from Treisman [1985]. A direct result is the logical segregation of types, an idea that arose in the 'intrinsic image' theory of [Barrow and Tenenbaum 1978] and also in the 'feature integration' theory of Treisman. Physical separation of types into physically distinct maps follows if connectivity lengths are considered. Cowey presents this reason for the evolution of physically separate visual maps: units that compute similar quantities need to communicate with one other for consistency purposes and thus need to be connected to one another [Cowey 1979]. The connectivity lengths would be prohibitive if the units were separated. The new expression for speed-up is

$$PP = N^3 \times (2^{\hat{M}} - 1) \times \log_2 VP \qquad (6)$$

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 *PP* are barely plausible biologically. Therefore, since pooled

response and map segregation does not lead to savings for all possible images, one may speculate that the role is that of speeding up the computation for the simpler inputs (the simplest and therefore fastest situation being for $\hat{M} = 1$), thus minimizing response time for simple inputs, and may be considered a mechanism for the system's graceful degradation with increasing complexity of the input. The issue of the number of maps and their functionality will be further elaborated in a later section.

Further efficiency could be gained by trading off precision, and this constitutes the only optimization that causes a degradation in the fidelity of the incoming signal. This can be achieved by reducing the resolution of the visual image and, simultaneously, abstracting the input in order to maintain its semantic content. The 'processing cone' representation of Uhr [1972] has the right flavor, but does not include the proper semantic abstraction. Abstraction implies that some data is lost, and thus, the 'filter' of attention theories has a strong computational counterpart.

Let $\hat{N}$ be the size of the new abstracted array. What is the largest array that leads to complete inspection within the time constraint? The expression for degree of speed-up is changed to

$$PP = \hat{N}^3 \times (2^{\hat{M}} - 1) \times \log_2 VP \qquad (7)$$

If VP is set to 10,000,000, $\hat{M}$ to 1, and PP to 1,000,000, then from this equation, $\hat{N}$ is 35, and $P_{\hat{N}}$ is 3571. For VP = 100,000, $\hat{N}$ is 39, and $P_{\hat{N}}$ is 4447. It is easy to see that variations in PP, $\hat{M}$, 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 less than 2100. This, then, is the satisfying architecture and is illustrated in Figure 2.

Exploring further, more insights can be obtained from equation (7). Figure 3 shows a family of curves of this relationship for $P_{\hat{N}}$ vs. $\log_{10} PP$ 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. Qualitatively, several conclusions can be drawn and verified analytically. If these are the basic performance relationships, then the designer of

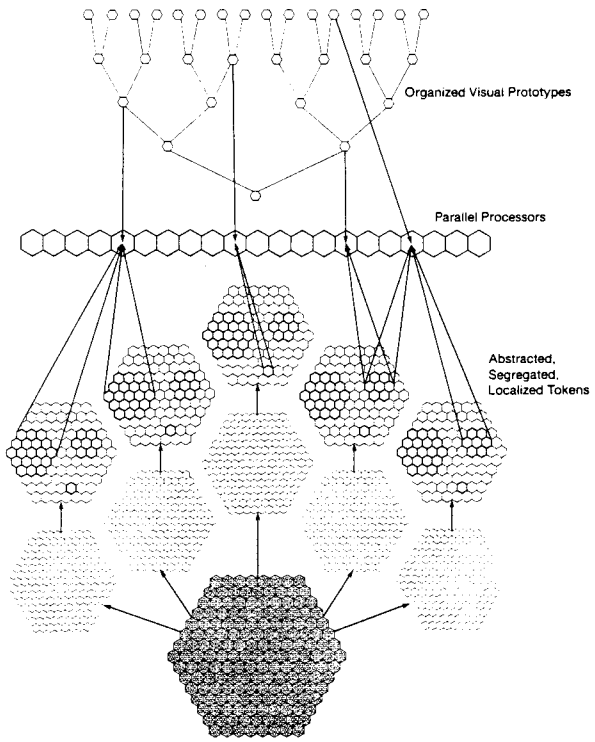

Fig. 2. The architecture that satisfies the timing constraint for a single bottom-up processing pass. The binary tree knowledge base of prototypes is at the top. The processor layer is in the middle, each processor accessing one visual prototype and one receptive field from one or more maps. The lower half of the diagram depicts the early abstraction hierarchy, with logically separated maps and decreasing resolution.

the visual system is faced with a few choices and trade-offs. First of all, 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 3. Changes in VP have a very small effect on PP, as can be easily seen from the partial derivative, $\partial PP/\partial VP = PP \times (\log_2 e)/(VP \times \log_2 VP)$. It is more expensive to use larger maps, since $\partial PP/\partial \hat{N} = PP \times 3/\hat{N}$. The largest expense is incurred for adding maps, because

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

If, for example, VP = 10,000,000, $PP = 10^{5.6}$, $\hat{M}$ = 1, and $\hat{N}$ = 26, then the derivative of PP for changes in $\hat{M}$ is 12 times steeper than for changes

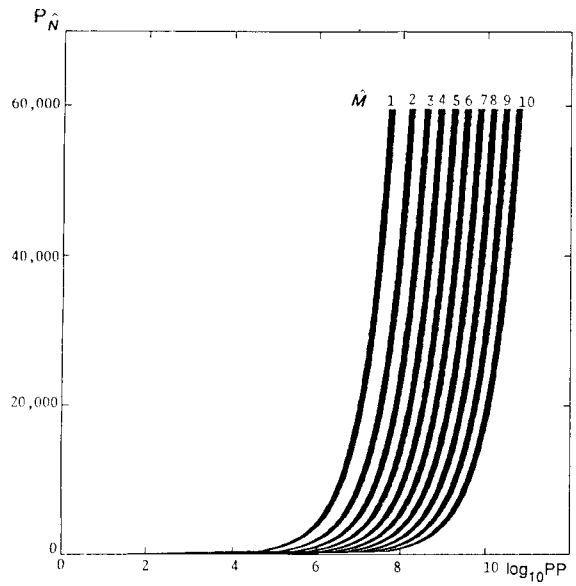

Fig. 3. The family of curves generated using equation (7) for varying values of *PP*, *VP*, $\hat{M}$, and $P_{\hat{N}}$. $P_{\hat{N}}$, ranging from 0 to 60000, is plotted against $\log_{10}PP$, ranging from 1 to 9. Each thick solid curve represents a value of $\hat{M}$, with $\hat{M} = 1$ the leftmost, and $\hat{M} = 10$ the rightmost. The thickness of each curve represents that fact that within it is the entire range of *VP*, from 100,000 to 10,000,000.

in $\hat{N}$ and 223,000,000 times steeper than for changes in *VP*. Thus, of the three variables, it is most critical that the value of $\hat{M}$ be set as low as possible.

Although equation (7) may lead to ballpark figures, it still remains to determine reasonable estimates for the configuration of the visual system. The maximum power principle can be used at this point to guide the search among all the reasonable values. A simple objective function may be formulated to embody some of the constraints of the principle. A relationship is defined that expresses the amount of data that must be processed in a single bottom-up pass, in the worst case. A different measure is required here, rather than using the one that has been the basis of all the analysis so far, since by definition, each processor is allowed time to perform only one of the basic matching operations defined earlier. Also, this measure must be totally divorced from the algorithm or organization used by the vision system in order to be an appropriate metric for comparison purposes. The input to the system is an array of $P_{\hat{N}}$ elements, $\hat{M}$ types of tokens at each elements. There are $P_{\hat{N}} \times \hat{M} \times VP$ data 'chunks' in total for the worst case. Or, in other words, this is the amount of raw data that must pass through the processors in the worst case. The greater the amount of data that each processor can process, the greater the system's power. The system power, then, may be expressed as

$$\text{power} = \frac{P_{\hat{N}} \times \hat{M} \times VP}{PP} \tag{8}$$

where *PP* is defined in equation (7). A system can increase its power by increasing input array size, the number of maps, and/or the size of the knowledge base and/or decreasing the speed-up required to satisfy the time constraints. This is precisely what the maximum power principle requires.

The best configuration satisfying the requirements of immediate perception is the one that maximizes the power principle. Unfortunately, there is no global maximum. System power increases with increasing values of *VP*, and thus it can be concluded that *VP* should be as large as possible. Power decreases with increasing values of $P_{\hat{N}}$, as it does with increasing values of $\hat{M}$, but much more slowly than it does for $\hat{M}$. It can be concluded that both of these parameters must be small. Further, it is critical that $\hat{M}$ be very small, since power decreases exponentially with increasing $\hat{M}$, much faster than for the other two parameters. Since $\hat{M}$ must be a positive integer, the best setting for this parameter in order to minimize the requirements for *PP* and to maximize power is $\hat{M} = 1$. It is more difficult to argue for specific settings of $P_{\hat{N}}$, since the constraints on this value so far are contradictory: (a) $P_{\hat{N}}$ must be large enough to ensure good image resolution; and, (b) $P_{\hat{N}}$ must be small to ensure high system power. The minimum cost principle for connectivity will decide this issue in a later section.

Implications and Predictions

Using the basic conclusions of the previous section, a number of characteristics may be derived for the resulting architecture. Some are confirmed by the known neuroanatomy of primate

vision; others will stand as predictions. The architecture, guided by the philosophy of the complexity level and the maximum power/minimum cost principle, exhibits columnar processor organization. Connectivity constraints predict the average sizes of the retinotopic maps, the number of maps that comprise the output of early vision, and the degree of required speed-up due to parallelism.

Decreasing System Speed with Increasing Data Complexity

From equation (7), and the conclusions of the previous section, a statement may be made about the speed of processing for increasingly complex input data. The simplest input data situation—that is, $\hat{M} = 1$, also corresponds to the maximum power situation, and fixes the degree of parallelism for the system, to:

$$PP = \hat{N}^3 \times \log_2 VP \tag{9}$$

and thus, the time for computation is $T_{\min}$, given by

$$T_{\min} = \frac{\hat{N}^3 \times \log_2 VP \times PS}{PP} = PS \tag{10}$$

As $\hat{M}$ increases, the time to compute increases exponentially with $\hat{M}$, up to a maximum given by $T_{\max}$, when all maps (M are active:

$$T_{\max} = \frac{\hat{N}^3 \times (2^M - 1) \times \log_2 VP \times PS}{PP}$$

$$= T_{\min} \times (2^M - 1) \tag{11}$$

The progressive increase in base computation time for increasingly complex preattentive tasks has been experimentally documented [Nakayama and Silverman 1986]. However, the exponential relation is not immediately apparent. It is rather difficult to examine this from reaction time comparisons, and in fact, there may be other optimizations that play a role in reducing this exponential increase. Obviously, the expression in equation (11) gives the computation time for the worst case with respect to searching the entire set of map subsets. The average case, which is what the visual search experiments of Nakayama and Silverman expose, requires half the computation time. Thus, the average computation time, if $\hat{M}$ maps are active is given by

$$T_{\text{ave}} = \frac{\hat{N}^3 \times (2^{\hat{M}} - 1) \times \log_2 VP \times PS}{2 \times PP}$$

$$= T_{\min} \times \frac{(2^{\hat{M}} - 1)}{2} \tag{12}$$

A further optimization may be that map subsets are internally ordered, and this would also lead to a speedier response. Finally, there is the issue of discriminability along the stimulus quality dimension. The closer two tokens are along a given type dimension, the more difficult it will be to separate them perceptually. For example, pop-out is faster for a display where the target is red and every other element of the stimulus is black than for a display where the target is orange and the distractors are red or yellow. It seems that arbitrary amounts of computation are required to distinguish tokens along the same dimension, depending on their perceptual similarity along that dimension. These are issues for further consideration and are not directly addressed by the theory in this paper.

Columnar Processor Organization

The question addressed by this section is: 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, $PP/P_{\hat{N}}$ processors. The number of processors, in the best configuration for immediate perception derived earlier, is given by setting $\hat{M}$ to 1 in equation (7). Using this, and the expression for $P_{\hat{N}}$ in terms of $\hat{N}$, the number of processors in an assembly is

$$\frac{\hat{N}^3}{3\hat{N}^2 - 3\hat{N} + 1} \times \log_2 VP \tag{13}$$

But, $\hat{N}^3/(3\hat{N}^2 - 3\hat{N} + 1)$ 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 set of processors a *receptive field assembly;* this will be the basic processing unit for the remaining discussion. Each of the receptive field assemblies must be connected to its relevant retinotopic elements; and

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 $Y - (K + 1)/2$. Thus, the diameter of the receptive field is K, an odd integer, and 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

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

It is assumed that processors are unit distance above the stimulus array, but this does not affect the result. loc is the location of the processor and could 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 connection length for those connections. The same is true of the two-dimensional case, since the situation is circularly symmetric. Thus 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, as is true in this situation, 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, then the remaining processors must be placed above or below this layer. Then, the same argument 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. The center pixel requires $\hat{N} \log_2 VP$ processors (or $\hat{N}$ receptive field assemblies), while the pixels on the rim require $\log_2 VP$ processors, (or 1 receptive field assembly). This structure is not unlike that of Hubel and Wiesel's hypercolumns in an abstract sense. If implemented on 'flat' hardware (like the cortex), this organization would exhibit inverse magnification, and this too is observed in the primate visual system [Daniel and Whitteridge 1961]. 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, in one time step, since 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 (matched pre-attentively). This is one way in which the serial nature of binary search is 'parallelized.'

Size and Number of Maps

Connectivity considerations, both in terms of number of connections and in terms of lengths, lead to many more predictions. 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 hardwired directly to the receptive field assemblies; there is no other way that a strictly bottom-up pass, without any a priori knowledge, can occur in parallel. Consider connectivity in the direction from the retinotopic representations to the processor layer. The first quantity to determine is the total number of wires that are required to connect each receptive field to its receptive field assembly. This will be computed by simply summing the number of pixels for each of the $\hat{N}^3$ receptive fields. Each point in the array is a member of a ring of points, each point of which is the center for the same number and sizes of receptive fields. The number of elements of each ring, at radius i is given by $P_i - P_{i-1}$. The receptive fields that are centered by each member of the ring are of sizes 1 through $\hat{N} - i + 1$. The sum of the elements in each of these size receptive fields for each element of each possible ring then gives the total number of wires required to hardwire all the receptive fields, independently of one

another. The following expression accomplishes this:

$$\sum_{i=1}^{\hat{N}} \left\{ (P_i - P_{i-1}) \sum_{j=1}^{\hat{N}-i+1} P_j \right\} = \frac{\hat{N}}{10}(3\hat{N}^2 + 1) \times (\hat{N}^2 + 1) \quad (15)$$

Call this the area (A) of each map. The total number of wires is $A \times M$, and the average fan-out from the retinotopic representations is $(A \times M)/(P_{\hat{N}} \times M)$, or $A/P_{\hat{N}}$. 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 $(A \times M)/\hat{N}^3$. Now if we use the biological constraint of fan-out and fan-in for neurons of approximately 1000, these two connectivity expressions yield predictions of

$\hat{N}$ = 21 or 22

$P_{\hat{N}}$ = 1261 or 1387

M = 7 or 6

average fan-out = 975 or 1118

average fan-in = 929 or 874

PP: for $VP = 10^7$: $PP = 10^{5.33}$ or $10^{5.4}$

for $VP = 10^5$: $PP = 10^{5.19}$ or $10^{5.25}$

Each of these values is completely consistent with the biologically plausible ranges described earlier, for areas MT and V4. Since each map was assumed to be the same size, $P_{\hat{N}} \times M$ is probably a better prediction for total number of data items in the output of early vision. It is more correct, however, to note that the values for M and $P_{\hat{N}}$ are lower bounds only and that the values for *PP* are thus upper bounds only. There may in fact be other connectivity optimizations that are present that would permit a larger number of elements to be connected. A simple calculation would also reveal that connectivity consisderations would predict that at least three layers are required in the input abstraction hierarchy to go from a retina of 130,000,000 receptors to a map of size 1300.

There are no connections from the processors to any of the larger maps in the input abstraction hierarchy. The number of such connections would be prohibitive. A back-of-the-envelope calculation can help here as well. Suppose that the processors are to be connected to M maps of high resolution, say 1K and 1K. 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 *PP* 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 connections of 10^{13}, this is clearly not how Nature implemented access to high resolution maps. If information is to be transmitted to the processors from the larger maps, then it must be done *through the input abstraction hierarchy,* 'attentively,' by tuning of the operators that compute the representation of the top-level maps. This conclusion supports the findings of Moran and Desimone [1985] by providing additional justification for a top-down tuning mechanism.

If the visual areas of the brain contain on the order of 10^9 neurons, then the prediction for *PP* implies that about 5000 neurons are required for each unit of incremental speed-up due to parallelism. What this analysis claims is that with connectivities of 1000 on average, it is not possible to hardwire more than 7 maps, or 7 types of early vision output parameters, to a set of processors, and even then, the maps can contain only about 1300 elements (or a 36 × 36 image). Further, the degree of speed-up due to parallelism that is required is on the order of $10^{5.3}$ for knowledge bases of prototypes that may possibly be large enough to reflect human performance. Thus, these figures may be considered as the limits on the capacity of early vision schemes.

The prediction of 6 or 7 maps in total now predicts the response time for the slowest preattentive worst-case performance of the system, using equation (11). However, intuitively the prediction of 6 or 7 maps seems small, remembering the earlier statement that perhaps many more parameters may be needed to completely characterize each point in visual space. Given that the uncertainty principle must play a role in the measurement of signal properties, some amount of inseparability seems necessary on those

grounds too. If each type is really along more than one dimension, then it is possible to have many more actual values, implying a coarse-coded representation. A coarse-coded representation at this level would certainly allow many more actual values to be extracted, thus leading to a visual system capable of much richer interpretations of the visual world. (See Hinton [1981] and Ballard, Hinton, and Sejnowski [1983] for discussions on coarse-coded representations for vision.)

A New Explanation for the Pop-out Phenomenon

An important conclusion of this research is that pre-attentive vision, as defined by Neisser [1967] and by Treisman [1985], can be shown to be just a special case of the entire process of visual perception, and not a process distinct from attentive vision. The reason is complex and involves careful consideration of the functionality of maps and the matching process required for finding targets in a display.

Treisman used two basic types of displays—conjunctive and disjunctive. In a conjunctive display, the target is defined by conjoining two stimulus qualities, such as shape with color. There is more than one element in the stimulus pattern that displays the same conjunction of types of stimulus qualities and each of these is called a distractor. There is only one target with the prespecified conjunction of tokens (for example, red with round). An increase in the number of distractors leads to the observed positive linear slope in response time. For such displays, where there is no target (called conjunction negatives), the slope is approximately twice as large, leading to the conclusion that a self-terminating search is taking place. On the other hand, a disjunction display requires a target defined by a single stimulus quality, such as shape. However, the definition of distractors is different and here means all other elements in the display. Only one element of the stimulus has this quality, and the response time curve is flat with respect to the number of elements in the display. The disjunction negative case does exhibit a positive linear slope with number of elements in display, perhaps pointing to the need for an exhaustive search to verify that no target is present.

The explanation Treisman and her colleagues propose is that for conjunction displays, separate feature maps must be brought into register using a spotlight of attention. This spotlight is necessarily serial so that a positive linear slope in response time versus distractors is found, due to a self-terminating search. Unfortunately, spatial registration cannot be the reason for the phenomenon as claimed by Treisman—if all elements of the maps are hardwired, registration is quite easily solved by the neural hardware. In the disjunction or pop-out case, Treisman claims that map activity is sufficient to signal that the target is present. This view assumes that there is a map for each feature type—unfortunately, a great many more feature types are used by Treisman than the number of maps in the brain—both those that are known as well as the number predicted by the theory of this paper. Moreover, why then is a linear search required to verify the disjunction negative case? Could not the same cue just as easily be used to signify the lack of a target? Response time slopes for both display types seem to be related to ease of stimulus quality discriminability. Why should this be the case for disjunction negatives? Finally, the response time for a conjunction display with only one element, namely the target, is larger than for a disjunction display of the target only. Treisman's explanation does not account for this. A different explanation for conjunction as well as for disjunction performance is required.

To this point, global decisions on image contents have not been discussed. Global decisions are dependent on the goals of the recognition task. As presented, the architecture in figure 2 includes $\hat{N}^3$ receptive fields, and there is sufficient processing time to compute the best association for each with a visual prototype from the knowledge base. Each of these associations is a candidate for satisfying the goals of the visual task, and may wholly or only partially identify the visual entity in the corresponding receptive field. Thus, there are $\hat{N}^3$ separate outputs, or candidates for matching with the goals of the recognition task. Assuming some kind of cooperative process within columns, at best there will be $P_{\hat{N}}$ candidates, rather than $\hat{N}^3$ candidates. Let the number of goals or a priori visual targets be G. In the worst case, $P_{\hat{N}} \times G$ matches of candidates to

targets are needed. Goals are dynamic, and are usually too few to try and organize, nor do they necessarily shape properties on which organization can be done. Thus, the optimizations used earlier cannot be applied here. Candidates, however, could be ordered by goodness-of-fit criteria, using strength of response. Thus, it seems that the best this architecture can accomplish is a serial, self-terminating search. This is what is observed, and in this view, the positive slope is due only to the number of candidates and goals.

What is the number of candidates that are extracted to be matched against the goals or targets, by the subjects in the visual search experiments? Moran and Desimone [1985] have discovered, in monkeys, that single neurons as early as V4 (as well as in IT, but not in V1) can be tuned so that separate stimuli within the same receptive field can be individually attended, via top-down control, depending on spatial location and/or stimulus quality. Effective and ineffective stimuli were determined for each neuron, and then each was presented in different portions of the receptive field simultaneously. Attention to the effective stimulus lead to a strong response, while attention to the ineffective stimulus lead to a significant attenuation of response, even though the effective stimulus was still in the field. The tuning was observed to be changeable from trial to trial. They claim that unwanted information is filtered from the receptive fields of neurons in extrastriate cortex as a result of selective attention. If the findings of Moran and Desimone are also true for human vision, then over a course of several trials, subjects can tune out nongoal responses at the map level. The tuning in the case of Treisman's experiments would not be on location, but only on stimulus quality. So for example, if the targets are either a brown T or a letter, for a given display, subjects could tune out nonbrown, non-T, and nonletter stimuli at very early processing levels. In a disjunction display, there would always be only one candidate remaining for matching against goals because of the definition of the display. In the conjunction case, the number of candidates would be exactly the number of distractors plus the target.

Earlier in the paper, it was concluded that the best solution for immediate perception was for $\hat{M} = 1$. Yet, it is was also shown that up to seven maps could be accomodated using connectivity constraints. There are two possibilities. Firstly, it could be the case that only one physical map is active and relevant. The evidence seems to go against this view however, due to the inseparable nature of neural processing. Thus, it would indeed be very special circumstances in which a single neuron would yield information about only one physical parameter (for example, color without shape, depth, or motion information). These are not the kinds of stimuli that are used to demonstrate pop-out. For the second explanation, recall that $2^{\hat{M}} - 1$ is the number of subsets of maps, not the actual number of maps. Since the receptive field assemblies can be directly connected to all maps, and pooled response could point to the active subset, then all tokens could conceivably be integrated in parallel, one after another. The resulting time to compute all these parameters would vary as in equation (12). Of course, stimulus quality discriminability also plays a role, and this affects both the determination of the candidates and the sequential selection of candidates. Earlier, it was demonstrated that there was a need for a coarse-coded representation at the map level. Thus, only under special stimulus conditions, would it be possible for consideration of a single subset of types to lead to single visual parameters that lead to unambiguous target-stimulus matches, and thus pop-out. If those special conditions do not exist, then more subsets must be considered. There is no spotlight required for this, just more work. The spotlight is required only for selecting from among competing candidates. This explanation predicts that the response time for pop-out displays will increase as the number of stimulus qualities increase. This increase in base computation load is most apparent in the results of Nakayama and Silverman [1986], who have tried a more complex set of disjunction experiments than Treisman, thus confirming the prediction (color with motion, for example). If on the other hand, it had been assumed that the maps were independent, then the extraction of all the data needed for the candidate-to-goal matching would occur in constant time regardless of the number of stimulus qualities. Using Treisman's experimental parameters, if we had assumed the independence of maps, the theory presented in

this paper would predict that the slope of the curve for response time vs. number of distractors in the display for conjunction negatives would be the same as that for feature negatives, and this slope would be twice that for conjunction positives. The disjunction positive curve would still be flat, yet the conjunction positive curve would intersect with the disjunction positive curve for size of display of one element. This is not what is observed. On the other hand, if the maps are not independent, the observed results are predicted: serial self-terminating search for conjunctions, flat response time for disjunction positives, a higher response time for conjunction displays with one element than for disjunction displays with one element, and a disjunction negative response slope that is smaller than for conjunction negatives.

The second issue is that of visual features or primitives. With the view presented above, Treisman's claimed features represent only what is 'tunable'—and is much larger than the set of 'default' features or primitives that the system computes. In the terminology of this paper, Treisman discovers tokens but not types. However, if Treisman's visual search paradigm rejects a feature, then it is indeed not even tunable. The conclusion is that visual search can be used only to reject candidate stimuli as features, and not to prove their existence. Using visual search in this mode only would be a very time-intensive task as well offering only indirect evidence for particular features.

Conclusions

The development of theories of visual perception lacks guiding principles, that is, a set of fundamental considerations that can both direct the creation of a theory and test its validity. Two such principles are proposed in this paper, the 'complexity level' of analysis and the maximum power/minimum cost principle. This research has demonstrated that significant conclusions about the architecture of biologically plausible visual systems can result by the faithful application of these principles.

The implications for computer vision are clear and quite important. The reason that many of the high-level vision proposals have not been entirely satisfactory (see Tsotsos [1987b] for a comprehensive overview), is that a strong argument for the computational need for high-level processing has never been presented. That need must be in terms of the basic computational inadequacies of spatially parallel, bottom-up visual architectures. The capabilities of such architectures have been derived in this paper for biologically motivated designs (and still apply in large part for nonbiologically motivated designs, but not entirely). The argument for high-level vision, and indeed, for computational modeling of human vision, is now on a solid foundation, and the results of this paper point to a very different style of high-level vision research than currently practised.

It has been shown that in addition to spatial parallelism, the other characteristics of a visual processing architecture that satisfies the timing constraints, are:

- hierarchical organization through abstraction of prototypical visual knowledge, in order to cut search time at least logarithmically
- localization of receptive fields, noting that the physical world is spatiotemporally localized and that objects and events, and their physical characteristics, are not arbitrarily spread over time and space
- 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
- hierarchical abstraction of the input token arrays, in such a way as to maintain semantic content yet reducing the number of retinotopic elements

These optimizations may be considered as sufficient, but not necessary, conditions to satisfy the time complexity constraint for the architecture of a visual system with performance comparable to human pre-attentive visual performance. Some of these particular optimizations do not represent new concepts (see Ballard [1986] for example), but their packaging using the thread of complexity and the resulting quantitative analysis is novel.

Applying connectivity constraints to this architecture, that is, determining how all the elements are to be connected, and the resulting cost of connection, many further characteristics of primate visual systems are implied and several others predicted:

- processor columnar organization
- tokens of visual parameters at high resolution cannot be directly accessed, rather must be obtained by tuning of computing units and through the input abstraction hierarchy
- token coarse coding
- physical separation of some maps
- predictions for the best architecture for immediate perception
- predictions for the overall configuration of the visual system in terms of the size and number of maps
- a top-down control mechanism, in the spirit of the experimental findings of Moran and Desimone.

Top-down control can be driven by both location and stimulus quality. This would intuitively require the matching of visual prototypes with abstracted retinotopic descriptions; the determination of differences and similarities between successful matches and perceptual goals; and, if goals are unsatisfied, determination of which computations to tune and how to tune them. In addition, the visual routines of Ullman [1983] necessarily play a role. The overall control may proceed in much the same fashion as in Tsotsos [1980, 1985].

Another conclusion of this work is that, contrary to current psychological theories, pre-attentive vision is shown to be simply a special case of the visual process, and not a component separable from attentive vision. Put simply, if a single bottom-up pass yields an unambiguous immediate match to the goals of the perceptual task, then one has the pop-out phenomenon observed in pre-attentive vision. In non-pop-out situations, it is not the case that a different sort of mechanism takes over. More time is required for integration of more types of parameters, and subsequent serial search is required to select candidates for matching with the targets. This leads to the observed serial nature of attentive vision tasks. Since in typical perceptual experiments, subjects are told what to expect in the stimuli, and in most experiments have a substantial amount of training on the stimulus set, the first bottom-up pass may be tuned to attenuate the responses to nontarget stimuli, as pointed to by the results of Moran and Desimone.

Acknowledgements

Many thanks are due to Allan Jepson and Steve Zucker for several productive discussions and for their useful suggestions. The author is a Fellow of the Canadian Institute for Advanced Research. This research was conducted with the generous assistance of the Natural Sciences and Engineering Research Council of Canada.

References

Ballard, D. [1986]. "Cortical connections and parallel processing: Structure and function," BEHAVIORAL AND BRAIN SCIENCES, vol. 9(1), pp. 67-90.

Ballard, D., Hinton, G., and Sejnowski, T. [1983]. "Parallel visual computation," NATURE, vol. 306(5938), pp. 21-26.

Barrow, H., and Tenenbaum, J.M. [1978]. "Recovering intrinsic scene characteristics from images." In COMPUTER VISION SYSTEMS, A. Hanson and E. Riseman (eds.). Academic Press: New York, pp. 3-26.

Corbeil, J.-C. [1986]. THE STODDART VISUAL DICTIONARY. Stoddart Publishing: Toronto.

Cowey, A., [1979]. "Cortical maps and visual perception," QUART. J. EXPER. PSYCHOLOGY, vol. 31, pp. 1-17.

Daniel, P., and Whitteridge, D. [1961]. "The representation of the visual field on the cerebral cortex in monkeys," JOURNAL OF PHYSIOLOGY, vol. 159, pp. 203-221.

Feldman, J. (special issue editor), [1985]. "Connectionist models and their applications," COGNITIVE SCIENCE, vol. 9(1), pp. 1-169.

Feldman, J., and Ballard, D. [1982]. "Connectionist models and their properties," COGNITIVE SCIENCE, vol. 6, pp. 205-254.

Hinton, G. [1981]. "Shape representation in parallel systems," PROC. SEVENTH INT. JOINT CONF. ARTIF. INTEL., Vancouver. pp. 1088-1096.

Hubel, D., and Wiesel, T. [1977]. "Functional architecture of macaque visual cortex," PROC. ROY. SOC. (LONDON), vol. B-198, pp. 1-59.

Kirousis, L., and Papadimitriou, C. [1985]. "The complexity of recognizing polyhedral scenes," 26th Annual Symposium on Foundations of Computer Science, Portland, Ore.

Mackworth, A., and Freuder, E. [1985]. "The complexity of some polynomial network consistency algorithms for constraint satisfaction problems," ARTIFICIAL INTELLIGENCE, vol. 25, pp. 65-74.

Marr, D. [1982]. VISION, W.H. Freeman: San Francisco.

Moran, J., and Desimone, R. [1985]. "Selective attention gates visual processing in the extrastriate cortex," SCIENCE, vol. 229, pp. 782–784.

Nakayama, K., and Silverman, G. [1986]. "Serial and parallel processing of visual feature conjunctions," NATURE, vol. 320(6059), pp. 264–265.

Neisser, U. [1967]. COGNITIVE PSYCHOLOGY. Appleton-Century-Crofts: New York.

Poggio, T. [1982]. "Visual algorithms," MIT AI Memo 683, Cambridge, Mass., May.

Richards, W. [1982]. "How to play twenty questions with nature and win," MIT AI Memo 660, Cambridge, Mass.

Rumelhart, D., and McClelland, J. [1986a]. "PDP models and general issues in cognitive science." In PARALLEL DISTRIBUTED PROCESSING, D. Rumelhart and J. McClelland (eds.). MIT Press: Cambridge, Mass., pp. 110–146.

Rumelhart, D., McClelland, J. (eds.) [1986b]. PARALLEL DISTRIBUTED PROCESSING. MIT Press: Cambridge, Mass.

Stensaas, S., Eddington, D., and Dobelle, W. [1974]. "The topography and variability of the primary visual cortex in man," JOURNAL OF NEUROSURGERY, vol. 40, pp. 747–755.

Treisman, A. [1985]. "Preattentive processing in vision," COMPUTER VISION, GRAPHICS AND IMAGE PROCESSING, vol. 31, pp. 156–177.

Tsotsos, J. [1980]. A FRAMEWORK FOR VISUAL MOTION UNDERSTANDING, PhD thesis; also, CSRI-TR-114, Dept. of Computer Science, Univ. of Toronto.

Tsotsos, J. [1985]. "Knowledge organization and its role in the interpretation of time-varying data: The ALVEN system," COMPUTATIONAL INTELLIGENCE, vol. 1(1), pp. 16–32.

Tsotsos, J. [1986]. "Connectionist computing and neural machinery: examining the test of 'timing,' BEHAVIORAL AND BRAIN SCIENCES, vol. 9(1), pp. 106–107 (commentary on [Ballard 1986]).

Tsotsos, J. [1987a]. "Representational axes and temporal cooperative processes." In VISION, BRAIN AND COOPERATIVE COMPUTATION, M. Arbib and A. Hansen (Eds.). MIT Press/Bradford Books: Cambridge, Mass., pp. 361–418.

Tsotsos, J. [1987b]. "Image understanding." In THE ENCYCLOPEDIA OF ARTIFICIAL INTELLIGENCE, S. Shapiro (Ed.). John Wiley & Sons: New York, pp. 389–409.

Ullman, S. [1983]. "Visual routines," MIT AI Memo 723, Cambridge, Mass.

Uhr, L. [1972]. "Layered 'recognition cone' networks that preprocess, classify and describe," IEEE TRANS. ON COMPUTERS, pp. 758–768.

van Essen, D., and Maunsell, J. [1983]. "Hierarchical organization and functional streams in the visual cortex," TRENDS IN NEUROSCIENCE, September, pp. 370–375.

van Essen, D., and Zeki, S. [1978]. "The topographic organization of rhesus monkey prestriate cortex," JOURNAL OF PHYSIOLOGY, vol. 277, pp. 193–226.

Zucker, S. [1985]. "Does connectionism suffice?" BEHAVIORAL AND BRAIN SCIENCES, vol. 8(2), pp. 301–302 (commentary on [Feldman 1985a]).