A 'Complexity Level' Analysis of Vision

John K. Tsotsos

Dept. of Computer Science, and
The Canadian Institute for Advanced Research,
University of Toronto,
Toronto, Ontario, Canada M5S 1A4

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. This transformation converts the problem into a deterministic, biologically plausible one, with the specific transformations having direct biological counterparts, with an approximating solution. 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.

Overview

The task of visual perception can be shown to be intractable in a straightforward manner, and in fact, subproblems, such as polyhedral scene labeling have been shown to be NP-complete [Kirousis & 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 preattentive 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 & Ballard 1982]. 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. Rumelhart and McClelland too claim that the time and space requirements of a theory of cognitive function are important determinants of the theory's biological plausibility [Rumelhart & 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.

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]. 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. There are still issues of discretization, sampling, and computational complexity (at least) to contend with. Complexity issues span Marr's three levels of analysis. 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 a '0th' 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. If one is in the business of real-

CH2465-3/87/0000/0346$01.00 © 1987 IEEE

izing 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. This paper is an abbreviated version of [Tsotsos (in press A)].

The Model of Computation

A very simple 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:

- 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.
- 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. Tokens are distinct from measurements (usually taken to mean the output of some convolution operation), features (usually imply some level of interpretation), and primitives (non-decomposable 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. The types will be left unspecified and abstract.
- 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).
- 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 measurements 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, where PS is taken to be 100ms.

The specific representations do not matter for this discussion. A very simple way of thinking about this model is that it performs 'template matching' - this captures most of the processing style, but because of the invariant nature of the prototypes, is not a sufficient analogy. The costs associated with this model of computation are in the number of retinotopic elements, tokens ($M \times P$), visual prototypes, and processors that may fit within the system, such that the configuration satisfies the time and space complexity constraints.

A time complexity function will be formulated in such a way as to address the number of comparisons of tokens to prototypes that need 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. Thus, the entire knowledge base of visual prototypes must be included in the analysis that is to be presented. This definition 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 of 100ms will be 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 tokens representing visual parameters at each array element, then:

$$VP \times 2^{P \times M} \qquad (1)$$

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. 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 60's and the early 70's is NP-Complete in the number of lines [Kirousis & Papadimitriou 1985]. The specific theorems that they proved are: 1) It is NP-Complete, given an image, to tell whether it has a legal labeling; and, 2) It is NP-Complete, given an image, to decide whether it is realizable as the projection of a scene. It is thus no surprise that Waltz [Waltz 1975] and others could not completely solve this problem.

We can be more concrete about this complexity measure by using a few relevant values for human vision. What are appropriate estimates 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 bound for the number of prototypes is VP=100,000. An upper bound would surely be less than the number of neurons; a reasonable (but arbitrary) upper bound is VP=10,000,000. 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 bound 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 bound, a middle value and a lower bound. The number of receptors in the retina (130,000,000) is the upper bound, 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 bound (65536 pixels). It will become apparent that the particular choices for these parameters have no effect on the general conclusions.

A time complexity function for a task expresses 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 algorithm, a single processor in the worst case must consider each visual 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:

$$100\text{ms} = \frac{2^{P \times M} \times VP \times PS}{PP} \qquad (2)$$

Table 1 below gives values for PP for the bounds on P, M, and VP described above.

PP	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	$10^{19,733}$	$10^{236,744}$	$10^{19,735}$	$10^{236,746}$

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 Feldman and Ballard claim that massive parallelism is sufficient to satisfy the timing constraints.

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: 1) vision is easier since other cues such as color can be used; or 2) 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 & 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.

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, based on system throughput, can be stated for choosing 'best' configurations for vision systems:

The Maximum Power / Minimum Cost Principle
The power of a vision system is defined as the amount of data that can flow through the system per degree of parallelism (or per processor for simplicity), within a single bottom-up pass. The amount of data is characterized by the worst case number of data elements required for the worst case number of image-prototype matches that must be performed. Power is maximized by maximizing VP, M, and P and simultaneously, minimizing 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, and to maximize the variety of interpretations that can be associated with images, 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 argument 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 cortex, 10^{10}. Realizable parallel processing systems require considerations of local memory, synchronization, communication, etc., and presumably, a collection of neurons is required to accomplish this for each degree of speed-up. Since about 20% of the 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 sq. mm., and ranged from 1500 - 3700 sq. mm. 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, colour, motion, etc. [Hubel & 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 sq. mm. in extent, then V1 contains on average 2100 hypercolumns. Extrastriate areas are smaller, and have smaller hypercolumns (where hypercolumns have been 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 areas of influence, 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 & 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 & 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 processing is collapsed into one layer of identical processors, and thus the exact time required for the first bottom-up pass is not important. The figure of 100ms is a standard one for this; thus PS and 100ms cancel each other out in the complexity expression of equation (2). Each processor has an opportunity to complete one match only. 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 considered in terms of other units, not synapses. Where more than one map is used, all 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 biologically plausible quantitative ranges for P, PP, and M. The development can be likened to 'a back of the envelope' computation; thus, most of the exact quantitative conclusions in this section have little meaning.

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 or 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:

$$PP = 2^{P \times M} \times \log_2 VP \qquad (3)$$

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,

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

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

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

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

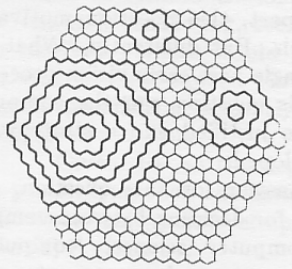

Figure 1.

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 below gives the values for PP resulting from this expression.

PP	$VP=10^5$		$VP=10^7$	
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	$10^{7.73}$	$10^{11.35}$	$10^{7.88}$	$10^{11.49}$

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 represent contiguous space, but the powerset of elements implied that over the 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 associated 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 & 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 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 $\hat{M}=1$ are found in Table 2. Even for the smallest image, the values of PP are barely biologically plausible. Therefore, since 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, thus minimizing response time for simple inputs.

N (or P_N) is still too large. Efficiency could be gained by trading off precision. 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 [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. The expression for degree of speed-up is changed to:

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

What is the largest array that leads to complete inspection within the time constraint? 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 a VP=100,000, $\hat{N}$ is 39, and $P_{\hat{N}}$ is 4447. It is easy to see that variations in PP and in VP lead to a great many possible configurations that with 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, that are also verified analytically. If these are the basic performance relationships, then the designer of the visual system is faced with a few choices and tradeoffs. First of all, there seems to be a 'hard complexity wall' on the number of processors. It is also very cheap in terms of processors to incorporate a very large knowledge base of prototypes. Changes in VP have a very small effect on PP, as can be easily seen from the partial derivative, $\frac{\partial PP}{\partial VP} = PP \times \frac{\log_2 e}{VP \times \log_2 VP}$. It is also relatively inexpensive to use larger maps, since as map size increases, again the effect on PP is inversely decreasing, i.e., $\frac{\partial PP}{\partial \hat{N}} = PP \times \frac{3}{\hat{N}}$. However, there is a relatively large and constant expense incurred for adding maps, because $\frac{\partial PP}{\partial \hat{M}} = PP \times \log_e 2$. From this analysis of the partial derivatives, it can be concluded that VP is the least expensive dimension in terms of increasing PP, while $\hat{M}$ is the most expensive.

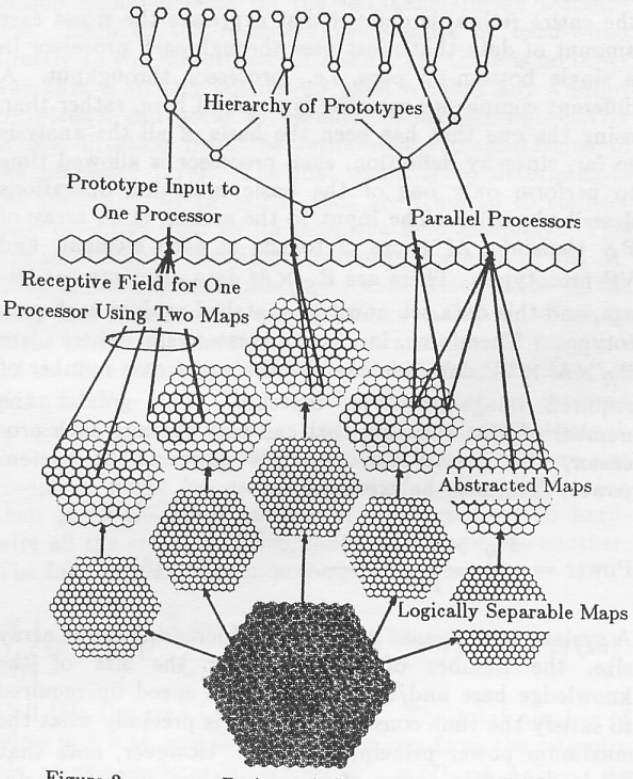

Figure 2.

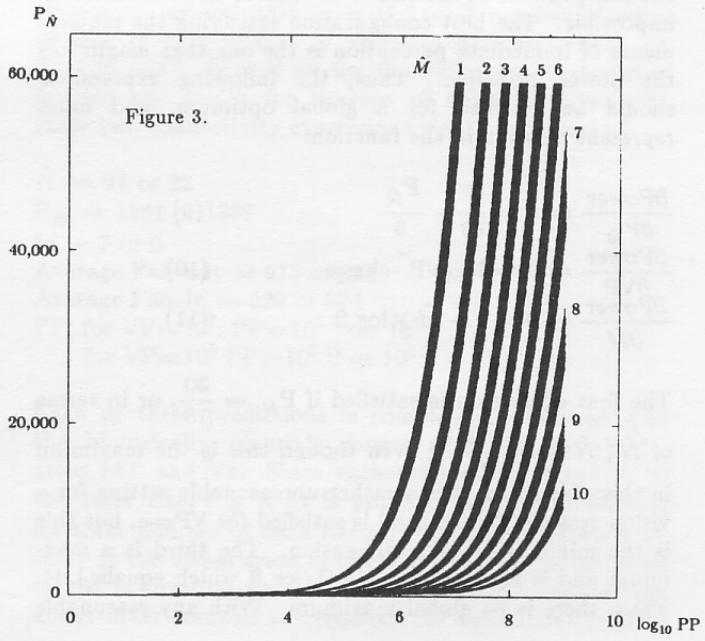

Figure 3.

Applying the Maximum Power Principle

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. An image-based complexity measure (that is, only using one receptive field, the size of the entire retina) is defined that expresses the worst case amount of data that must pass through each processor in a single bottom-up pass, i.e., processor throughput. A different complexity 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 described earlier. The input to the system is an array of $P_{\hat{N}}$ elements, $\hat{M}$ types of tokens at each element, and VP prototypes. There are $P_{\hat{N}} \times \hat{M}$ data elements per image, and this data set must be matched against each prototype. Therefore in the worst case there are $P_{\hat{N}} \times \hat{M} \times VP$ data 'chunks' for the worst case number of required image-prototype matches. The greater the number of data elements that can pass through each processor, the greater the system's power. The system power, then, may be expressed as:

$$\text{Power} = \frac{P_{\hat{N}} \times \hat{M} \times VP}{PP} \qquad (8)$$

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. However, note that PP is defined in terms of $P_{\hat{N}}$, and thus, unfortunately, the independent optimization of both of these variables is impossible. The best configuration satisfying the requirements of immediate perception is the one that maximizes the power principle. Thus, the following expressions should be satisfied for a global optimum, and must represent maxima in the function:

$$\frac{\partial \text{Power}}{\partial P_{\hat{N}}} = 0 = \frac{5}{16} - \frac{P_{\hat{N}}}{6} \qquad (9)$$

$$\frac{\partial \text{Power}}{\partial VP} = 0 = \log_2 VP - \log_2 e \qquad (10)$$

$$\frac{\partial \text{Power}}{\partial \hat{M}} = 0 = 1 - \hat{M} \times \log_e 2 \qquad (11)$$

The first expression is satisfied if $P_{\hat{N}} = \frac{30}{16}$, or in terms of $\hat{N}$, $\hat{N} = \frac{\sqrt{3}+1}{\sqrt{3}}$; even though this is the maximum in this dimension, it is a rather unreasonable setting for a vision system. The second is satisfied for VP=e, but this is the minimum in this dimension. The third is a maximum and is satisfied for $\hat{M} = 1/\log_e 2$ which equals 1.44. Thus, there is no global maximum. With any reasonable value of $P_{\hat{N}}$ and VP, system power would increase with increases in VP, and would decrease with increases in $P_{\hat{N}}$. However, regardless of the specific settings of VP and $P_{\hat{N}}$, system power will always peak at $\hat{M} = 1.44$. Since $P_{\hat{N}}$ and VP are the inexpensive dimensions of the system, it is satisfying to know that there is a peak in the expensive dimension, $\hat{M}$. VP should be as large as possible, and since types are indivisible in this theory, $\hat{M} = 1$, in order to maximize power. Good resolution is desirable, and a tradeoff between making P large and PP small is required, and preferred configurations are those with higher power.

Pre–Attentive Vision is a Special Case It has been shown that the most powerful configuration for immediate vision is one that analyzes input from only one map, or in other words, only from one of the many output representations of early vision. This is most powerful in terms of maximizing the richness of the visual world immediately accessable to the system and simultaneously minimizing the required degree of required parallelism and processing time. Treisman's (and others') visual search experiments discover stimulus characteristics that 'pop-out', i.e., are immediately perceived, and thus cause the system to exhibit this special case performance. Any other stimulus types (conjunctions, for example) require larger values of $\hat{M}$ and thus longer processing times. The explanation is detailed in [Tsotsos (in press) A].

Applying the'Minimum Cost Principle

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, and others will stand as predictions. The architecture, guided by the philosophy of the complexity level and the maximum power / minimum cost principle, implies that so-called preattentive vision is simply a special case of the entire process and not a separable compoenet. Further, the processor layer exhibits a columnar 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.

Columnar Processor Organization How are the processors connected to the retinotopic maps? At each input array location, 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:

$$\frac{\hat{N}^3}{3\hat{N}^2 - 3\hat{N} + 1} \times \log_2 VP \qquad (12)$$

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 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 their 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, 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 processor to this receptive field can be expressed by:

$$\sum_{x=Y-\frac{K+1}{2}}^{Y+\frac{K+1}{2}} \sqrt{1 + (loc - x)^2} \qquad (13)$$

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. This structure is not unlike that of Hubel and Wiesel's hypercolumns in an abstract sense. 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 (perhaps only approximately) 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. This is one way in which the serial nature of binary search is 'parallelized'. 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).

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 hard-wired 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 for each of the $\hat{N}^3$ receptive fields, its number of pixels. 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 hard-wire all the receptive fields, independently of one another. The following expression accomplishes this:

$$\sum_{i=1}^{\hat{N}} \left\{ (P_i - P_{i-1}) \sum_{j=1}^{\hat{N}-i+1} P_j \right\} = \frac{\hat{N}}{10}(3\hat{N}^2 + 1)(\hat{N}^2 + 1) \quad (16)$$

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 $(A \times M)/(P_{\hat{N}} \times M)$, 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 predictions 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. 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 incre-

mental speed-up due to parallelism. These figures may be considered as the limits on the capacity of early vision schemes that are spatially parallel, and entirely bottom-up, with no a priori information.

The prediction of 6 or 7 maps, or representations of the output of early vision, intuitively seems small, given that it was earlier claimed 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 distributed representation. A distributed 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 et al. 1983] for discussions on distributed representations for vision). A simple calculation would also reveal that connectivity considerations would predict that more than two 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 by 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 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 [Moran & Desimone 1985]. Moran and Desimone 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.

Summary

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 spatio-temporally 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 logically separable maps; and,
- 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.

Applying the maximum power principle leads to the conclusion that immediate, or 'pre-attentive' vision, is a result of only one map being required for an unambiguous match to a target, Thus, immediate perception is a special case of the whole process, contrary to current theories that separate pre-attentive from attentive vision.

Applying the minimum cost principle, particularly for connectivity, to this architecture, many further characteristics of visual systems are implied and several 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 inseparability (or coarse coding);
- physical separation of some maps;
- predictions for the overall configuration of the visual system in terms of the size and number of maps.

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 that can 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 (in press) B] 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 greatly apply for non-biologically 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.

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 financial assistance of the Natural Sciences and Engineering Research Council of Canada.

References

Ballard, D., Hinton, G., Sejnowski, T., "Parallel Visual Computation", *Nature 306-5938,* p21 - 26, 1983.

Barrow, H., Tenenbaum, J.M., "Recovering Intrinsic Scene Characteristics from Images", in *Computer Vision Systems,* ed. by A. Hanson and E. Riseman, Academic Press, 1978.

Corbiel, J.-C., *The Stoddart Visual Dictionary,* Stoddart Publishing Co., 1986.

Cowey, A., "Cortical Maps and Visual Perception", *Quarterly Journal of Experimental Psychology 31,* p1 -17, 1979.

Feldman, J., Ballard, D., "Connectionist Models and their Properties" *Cognitive Science 6,* p205 - 254, 1982.

Hinton, G., "Shape Representation in Parallel Systems", Proc. IJCAI, Vancouver, 1981.

Hubel, D., Wiesel, T., "Functional Architecture of Macaque Visual Cortex", *Proc. Royal Society of London B 198,* p1 - 59, 1977.

Kirousis, L., Papadimitriou, C., "The Complexity of Recognizing Polyhedral Scenes", 26th Annual Symposium on Foundations of Computer Science, Portland, Ore., Oct. 1985.

Mackworth, A., Freuder, E., "The Complexity of Some Polynomial Network Consistency Algorithms for Constraint Satisfaction Problems", *Artificial Intelligence 25,* p65 - 74, 1985.

Marr, D., *Vision,* W.H. Freeman, 1982.

Moran, J., Desimone, R., "Selective Attention Gates Visual Processing in the Extrastriate Cortex", *Science 229,* p782 - 784, 1985.

Neisser, U., *Cognitive Psychology,* New York, Appleton-Century-Crofts, 1967.

Poggio, T., "Visual Algorithms", MIT AI Memo 683, May 1982.

Richards, W., "How to Play Twenty Questions with Nature and Win", MIT AI Memo 660, 1982.

Rumelhart, D., McClelland, J., "PDP Models and General Issues in Cognitive Science", in *Parallel Distributed Processing,* ed. by D. Rumelhart and J. McClelland, MIT Press, 1986b.

Stensass, S., Eddington, D., Dobelle, W., "The Topography and Variability of the Primary Visual Cortex in Man", *Journal of Neurosurgery 40,* p747 - 755, 1974.

Treisman, A., "Preattentive Processing in Vision", *Computer Vision, Graphics and Image Processing,* 1985.

Tsotsos, J., "How Does Human Vision Beat the Computational Complexity of Human Visual Perception", (in preparation - to appear in *From Features to Objects,* ed. by Z. Pylyshyn, Ablex Press) A.

Tsotsos, J., "Image Understanding", in *The Encyclopedia of Artificial Intelligence",* ed. by S. Shapiro, John Wiley and Sons, (in press) B.

Uhr, L., "Layered 'Recognition Cone' Networks that Preprocess, Classify and Describe", *IEEE Transactions on Computers,* p758-768, 1972.

Ullman, S., "Visual Routines", MIT AI Memo, No. 723, June 1983.

van Essen, D., Maunsell, J., "Hierarchical Organization and Functional Streams in the Visual Cortex", *Trends in Neuroscience,* p370 - 375, 1983.

van Essen, D., Zeki, S., "The Topographic Organization of Rhesus Monkey Prestriate Cortex", *Journal of Physiology 277,* p193 - 226, 1978.

Waltz, D., "Understanding Line Drawings of Scenes with Shadows", in *The Psychology of Computer Vision,* ed. by P. Winston, Mcgraw-Hill, 1975.