An Attentional Prototype for Early Vision

Sean M. Culhane and John K. Tsotsos

Department of Computer Science, University of Toronto,
Toronto, Ontario, Canada M5S 1A4

Abstract.

Researchers have long argued that an attentional mechanism is required to perform many vision tasks. This paper introduces an attentional prototype for early visual processing. Our model is composed of a processing hierarchy and an attention beam that traverses the hierarchy, passing through the regions of greatest interest and inhibiting the regions that are not relevant. The type of input to the prototype is not limited to visual stimuli. Simulations using high-resolution digitized images were conducted, with image intensity and edge information as inputs to the model. The results confirm that this prototype is both robust and fast, and promises to be essential to any real-time vision system.

1 Introduction

Systems for computer vision are confronted with prodigious amounts of visual information. They must locate and analyze only the information essential to the current task and ignore the vast flow of irrelevant detail if any hope of real-time performance is to be realized. Attention mechanisms support efficient, responsive analysis; they focus the system's sensing and computing resources on selected areas of a scene and may rapidly redirect these resources as the scene task requirements evolve. Vision systems that have no task guidance, and must provide a description of everything in the scene at a high level of detail as opposed to searching and describing only a sub-image for a pre-specified item, have been shown to be computationally intractable [16]. Thus, task guidance, or attention, plays a critical role in a system that is hoped to function in real time. In short, attention simplifies computation and reduces the amount of processing.

Computer vision models which incorporate parallel processing are prevalent in the literature. This strategy appears appropriate for the vast amounts of input data that must be processed at the low-level [4, 19]. However, complete parallelism is not possible because it requires too many processors and connections [11, 17]. Instead, a balance must be found between processor-intensive parallel techniques and time-intensive serial techniques. One way to implement this compromise is to process all data in parallel at the early stages of vision, and then to select only part of the available data for further processing at later stages. Herein lies the role of attention: to tune the early visual input by selecting a small portion of the visual stimuli to process.

This paper presents a prototype of an attentional mechanism for early visual processing. The attention mechanism consists of a processing hierarchy and an attention beam that guides selection. Most attention schemes previously proposed are fragile with respect to the question of "scaling up" with the problem size. However, the model presented here has been derived with a full regard of the amount of computation required. In addition, this model provides all of the details necessary to construct a full implementation that is fast and robust. Very few implemented models of attention exist. Of those, ours is one of the first that performs well with general high-resolution images. Our implemented attention beam may be used as an essential component in the building of a complete real-time computer vision system.

Certain aspects of our model are not addressed in this investigation, such as the implementation of task guidance in the attention scheme. Instead, emphasis is placed on the bottom-up dimensions of the model that localize regions of interest in the input and order these regions based on their importance.

The simulations presented in this paper reveal the potential of this attention scheme. The speed and accuracy of our prototype are demonstrated by using actual *256* × *256* digitized images. The mechanism's input is not constrained to any particular form, and can be any response from the visual stimuli. For the results presented, image intensity and edge information are the only input used. For completeness, relationships to existing computational models of visual attention are described.

2 Theoretical Framework

The structure of the attention model presented in this paper is determined in part by several constraints derived from a computational complexity analysis of visual search [17]. This complexity analysis quantitatively confirms that selective attention is a major contributer in reducing the amount of computation in any vision system. Furthermore, the proposed scheme is loosely modelled after the increasing neurophysiology literature on single-cell recordings from the visual cortex of awake and active primates. Moreover, the general architecture of this prototype is consistent with their neuroanatomy [17, 18].

At the most basic level, our prototype is comprised of a hierarchical representation of the input stimuli and an attention mechanism that guides selection of portions of the hierarchy from the highest, most abstract level through to the lowest level. Spatial attentional influence is applied in a "spotlight" fashion at the top. The notion of a spotlight appears in many other models such as that of Treisman [15]. However, if the spotlight shines on a unit at the top of the hierarchy, there seems to be no mechanism for the rest of the selection to actually proceed through to the desired items.

One way to solve this problem in a computer vision system is to simply address the unit of interest. Such a solution works in the computer domain because computer memory is random access. Unfortunately, there is no evidence for random access in the visual cortex. Another possible solution is to simply connect all the units of interest directly. This solution also fails to explain how the human visual cortex may function because the number of such connections is prohibitive. For instance, to connect all possible receptive fields to the units in a single 1000 × 1000 representation, 10^{13} connections are needed to do so in a brute force manner [1]. Given that the cortex contains 10^{10} neurons, with an estimated total number of connections of 10^{13}, this is clearly not how nature implements access to high resolution representations.

The spotlight analogy is therefore insufficient, and instead we propose the idea of a "beam" – something that illuminates and passes through the entire hierarchy. A beam is required that "points" to a set of units at the top. That particular beam shines throughout the processing hierarchy with an *inhibit zone* and a *pass zone*, such that the units in the pass zone are the ones that are selected (see Fig. 1). The beam expands as it traverses the hierarchy, covering all portions of the processing mechanism that directly contribute

---

[1] see Tsotsos 1990 [17] for this derivation

to the output at its point of entry at the top. At each level of the processing hierarchy, a winner-take-all process (WTA) is used to reduce the competing set and to determine the pass and inhibit zones [18].

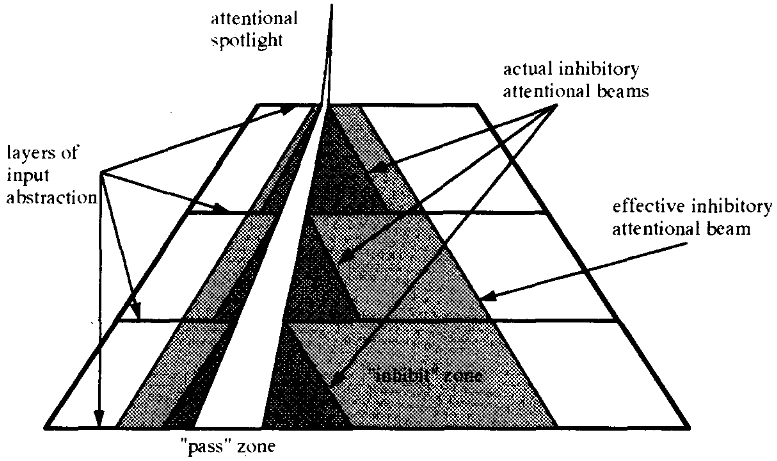
attentional spotlight
actual inhibitory attentional beams
layers of input abstraction
effective inhibitory attentional beam
"inhibit" zone
"pass" zone

Fig. 1. The inhibitory attentional beam concept. Several levels of the processing hierarchy are shown. The *pass zone* of the beam encompasses all winning inputs at each level of the hierarchy. The darkest beams represent the actual ***inhibit zones*** rooted at each level of the hierarchy. The light-grey beam represents the effective inhibit zone rooted at the most abstract level.

3 The Attention Prototype

The proposed attention prototype consists of a set of hierarchical computations. The mechanism does not rely on particular types of visual stimulus; the input only considers the magnitude of the responses. Connectivity may vary between levels. Each unit computes a weighted sum of the responses from its input at the level below. The weighted response used in this paper is a simple average; but in general the distribution of weights need not be uniform and may even be different at each level. Processing proceeds as dictated by Algorithm 1. An inhibit zone and a pass zone are delineated for a beam that "shines" through all levels of the hierarchy. The pass zone permeates the winners at each level and the inhibit zone encompasses those elements at each level that competed in the WTA process. This algorithm is similar to the basic idea proposed by Koch and Ullman [5]. One important difference is that our scheme does not rely on a saliency map. Another distinction is that we use a modified WTA update rule that allows for multiple winners and does not attenuate the winning inputs [2]. Also, the final stage of the algorithm is not simply the routing of information as Koch and Ullman claim, but rather a recomputation using only the stimuli that were found as "winners" at the input level of the hierarchy.

For illustrative purposes, the attention scheme is shown with a one-dimensional representation and illustrated in Fig. 2; the extension to two dimensions is straightforward. If a simple stimulus pattern is applied to the input layer, the remaining nodes of the

---

[2] The WTA updating function and a proof of convergence are described in Tsotsos 1991 [18]

```
1.  Receive stimulus at the input layer.
2.  Do 3 through 8 forever.
3.          Compute the remaining elements of the hierarchy based on the
            weighted sum of their inputs.
4.          Do 5 through 6 for each level of the hierarchy, starting at the top.
5.                  Run WTA process at the current level.
6.                  Pass winner's beam to the next level.
7.          Recompute based on winning input.
8.          Inhibit winning input.
```

Algorithm 1

hierarchy will compute their responses based on a weighted summation of their inputs, resulting in the configuration of Fig. 2(a). The first pass of the WTA scheme is shown in Fig. 2(b). This is accomplished by applying steps 5 and 6 of Algorithm 1 for each level of the hierarchy. Once an area of the input is attended to and all the desired information is extracted, then the winning inputs are inhibited. The attention process continues "looking" for the next area. The result is a very fast, automatic, independent, robust system. Moreover, it is a continuous and reactive mechanism. In a time-varying image, it can track an object that is moving if it is the item of highest response. In order to construct such a tracking system, the input would be based on motion.

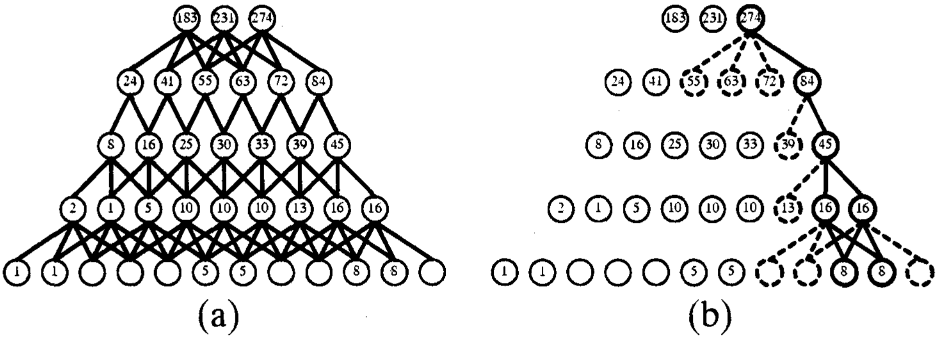

Fig. 2. A one-dimensional processing hierarchy. **(a)** The initial configuration. (b) The most "important" item is selected – the beam's pass (solid lines) and inhibit zone (dashed lines) are shown.

A number of the prototype's characteristics may be varied, including the number of levels in the processing hierarchy and the resolution of each level. The elements that compete in the WTA process are termed *"receptive fields"* (RF) after the physiological counterpart. In our implementation, a minimum RF *(minRF)* and a maximum RF *(maxRF)* are specified in terms of basic image units such as pixels. All rectangular RFs from *minRF* × *minRF* to *maxRF* × *maxRF* are computed and compete at each position in the input. RF shapes other than rectangular are possible. In general, a set of RFs are chosen that are appropriate for the current input computation.

There is an issue to consider when RFs of different sizes compete. If a small RF has a response of Ψ and a larger competing RF has a response $(\Psi - \varepsilon)$, then for a sufficiently small ε, the larger RF should "win" over the smaller one. For example, consider a RF

R_1 of size *2* × *2* that has a weighted average of 212, and a competing RF R_2 of size *20* × *20* that has a weighted average of 210. Since R_2 is 100 times the size of R_1 and over 99% the intensity, it seems reasonable to favour R_2 over R_1. Formally, this is exactly one of the constraints proposed in Tsotsos 1989 [16] for visual search: given more than one match with approximately the same error, choose the largest one. In the implementation of the attention model, this favouring of larger RFs of comparable value is accomplished by multiplying the weighted averages of all RFs by a normalizing factor that is a function of the size of the RF.

Marr suggests the following selection criterion for RF sizes: choose a given size if receptive fields of slightly smaller size give an appreciably smaller response and receptive fields that are larger do not give appreciably larger responses [7]. Marr notes, however, that more than one receptive field size may satisfy this requirement. For instance, consider a normalizing function that is linear. Also consider a 256 × 256 image-sized RF whose weighted average is 128. In such an instance, the largest possible RF should be weighted considerably less than two times the smallest possible RF. For this to hold, a linear function would have a slope less than 0.000015. Therefore, for two small competing RFs with similar sizes, the weighting is insignificantly small. Clearly a linear normalization function is not acceptable.

In the experiments presented in this paper, a normalization function whose rate of change is greatest for small RFs, without weighting very large RFs excessively is established. Since ε depends on RF size, it is smaller for small RF sizes. Thus, small RF sizes must be weighted more than larger RFs. This means that an acceptable function has a steep slope for small RF sizes and shallow slopes for the larger RF sizes. A good fit to this point distribution is the function $1/(1+e^{-x})$. In the experiments conducted, a similar compensating function of a more general form is used: [3]

$$\mathcal{F}(x) = \frac{\alpha + 1}{\alpha + \beta^{-\sqrt{x}}} ,$$

where, x represents the number of basic elements in the receptive field. Varying α affects the absolute value of the function's asymptote; varying β affects the steepness of the first part of the function. It was found empirically that values of $\alpha = 10$ and $\beta = 1.03$ generally give good results in most instances (see Figure 3).

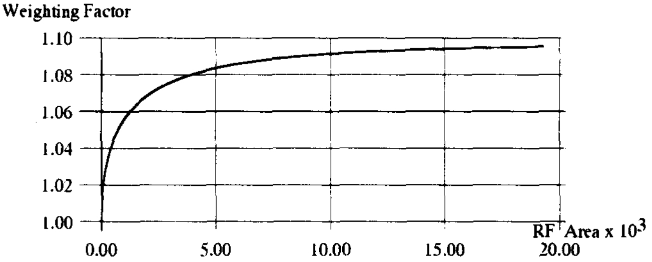

Fig. 3. F(x) for $\alpha = 10$, $\beta = 1.03$.

---

[3] The number 1 in the numerator is a result of normalizing $\mathcal{F}(x)$ for $x = 0$. The $\sqrt{x}$ is used to account for the *area* of the RF.

4 Experimental Results

We have implemented this attention prototype in software on a Silicon Graphics 4D/340 VGX. Simulations have been conducted using a wide variety of digitized 256 x 256 8-bit grey-scale images. In this paper, only brightness and edge information computed from the images are used as input to the prototype. Further research is required to determine on what other computations this attention beam should be applied.

This prototype lends itself to an implementation that is very fast, especially on hardware that supports parallel processes, such as the SGI 4D/340 VGX. In particular, the calculation of each element in a given level of the hierarchy is independent of all other elements at the same level. Therefore, the calculation of the hierarchy may be completed in parallel for each level. Furthermore, the WTA calculations at each time iteration are independent and may be done in parallel. In addition, the WTA process converges very quickly, typically taking less than ten iterations to determine the winner.

A simulation of the implementation for brightness is shown in Fig. 4. The lowest level of the processing hierarchy is the digitized image, and each successive level is a simple average of the previous level. This averaging computation has the effect of making each level appear as a smaller "blurred" version of the previous level. The WTA process is performed at the top of the hierarchy, and the pass zone is dictated by the RF that is "brightest". At each successively lower level, the WTA only operates on the RFs that fall within the beam from the previous level. Once the attention beam has located the winning RF and the surrounding inhibit zone in the input level, and all the information that is required is gathered from that focus of attention, the area is inhibited. In the simulations presented here, the region inhibited at the input layer is defined by the inhibit zone of the attention beam, contrary to the one-dimensional example in Sect. 3 where only elements in the pass zone are inhibited. In practice, once a region of the input is processed, or "foveated", it need not be considered again. The prototype then looks for the next "bright" area, starting by recalculating the processing hierarchy with the newly-inhibited image as its input. In this particular instance, the time taken to attend to each area in the input is approximately 0.35 seconds.

Following the movement of the pass zone on the input layer for successive fixations produces scan paths like the one shown in Fig. 5. The scan paths are interesting from a computational perspective because they prioritize the order in which parts of the image are assessed. The attention shifts discussed throughout this paper have been *covert* forms of attention in which different regions of the visual input have been attended. It is experimentally well established that these covert attention shifts occur in the humans [12]. In a similar way, the human visual system has special fast mechanisms called saccades for moving the fovea to different spatial targets (*overt* attention). The first systematic study of saccadic eye movements in the context of behaviour was done by Yarbus [20]. A future area of research is to discover a possible correlation between the scan paths of our attention beam and the scan paths of Yarbus.

A simulation using edge information was also conducted. At the bottom of the hierarchy is the output of a simple difference operator and again, each successive level is a simple average of the previous level. The WTA process successively extracts and then inhibits the most conspicuous items. Corresponding scan paths are displayed in Fig. 6. The results of this simulation using edges are interesting in several respects. The focus of attention falls on the longest lines first [4]. In effect, the strongest, or most salient, features

---

[4] In this instance, *maxRF* was set to 100 pixels so that only a portion of the longest line was attended to at first

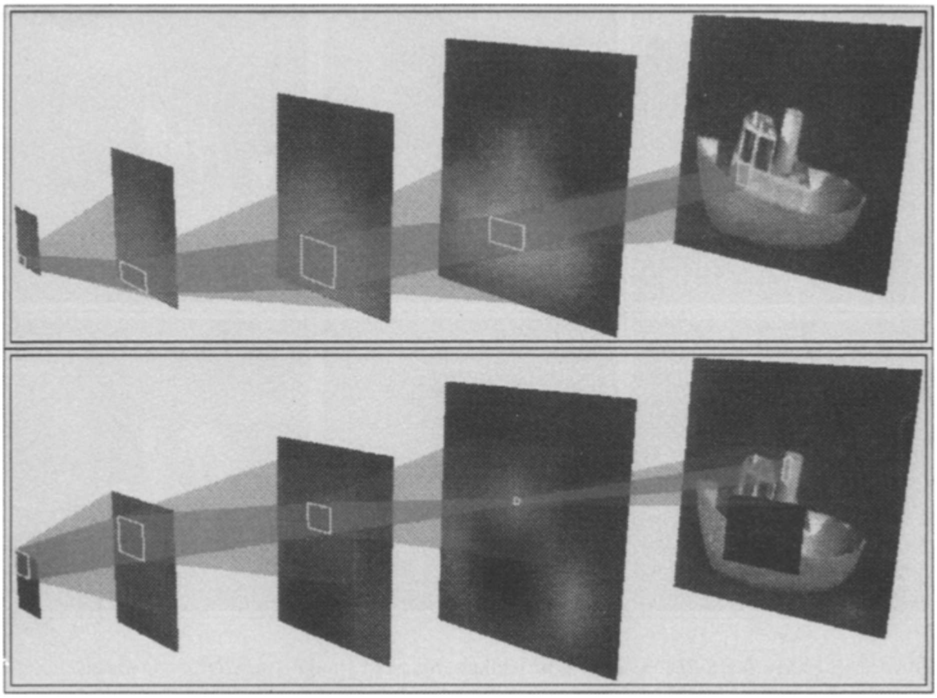

Fig. 4. Processing hierarchy and attention beam at two time intervals. The input layer is a 256 × 256 8-bit image. The beam is rooted at the highest level and "shines" through the hierarchy to the input layer. The darker portion of the attention beam is the pass zone. Once a region of the input is attended to, it is inhibited and the next "bright" area is found. The black areas in the input layer indicate the regions that have been inhibited.

are attended to in order of the length of the line, much like Sha'ashua and Ullman's work on saliency of curvature [14].

5 Discussion

The implementation of our attention prototype has a number of important properties that make it preferable to other schemes. For example, Chapman [3] has recently implemented a system based on the idea of a pyramid model of attention introduced by Koch and Ullman [5]. Chapman's model places a log-depth tree above a saliency map. Similar to our model, at each level nodes receive activation from nodes below. It differs, however, in that Chapman's model only passes the maximum of these values to the next level. There are several difficulties with this approach, the most serious being that the focus of attention is not continuously variable. The restriction this places on Chapman's model is that it cannot handle real pixel-based images but must assume a prior mechanism for segmenting the objects and normalizing their sizes. Our scheme permits receptive fields of all sizes at each level, with overlap. In addition, the time required for Chapman's model is logarithmic in the maximum number of elements, making it impractical for high-resolution images. Further, the time required to process any item in a sensory field

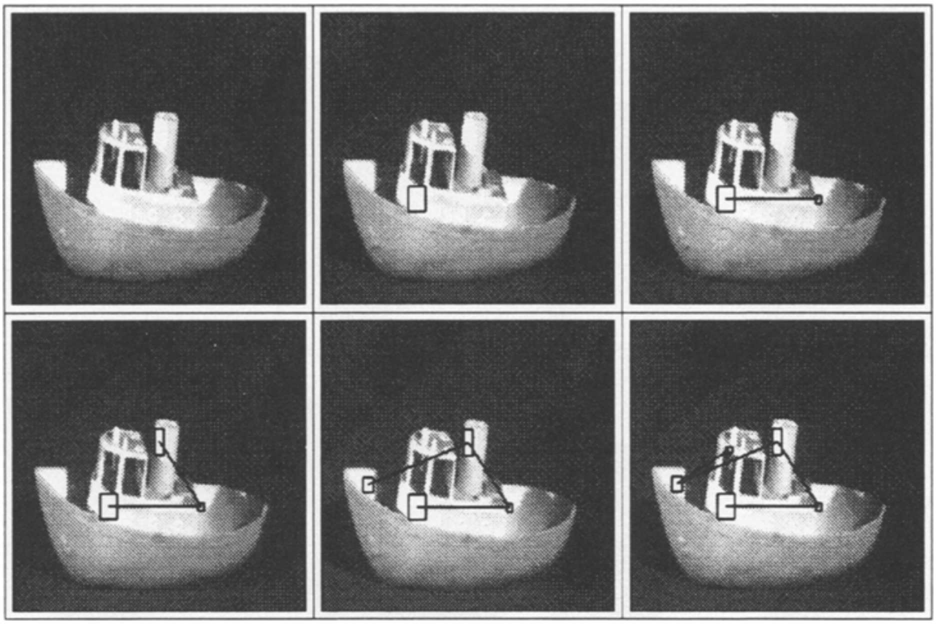

Fig. 5. Scan paths for a 256 × 256 8-bit image digitized image *(minRF = 5, maxRF = 40)*. The paths displays a priority order in which regions of the image are assessed.

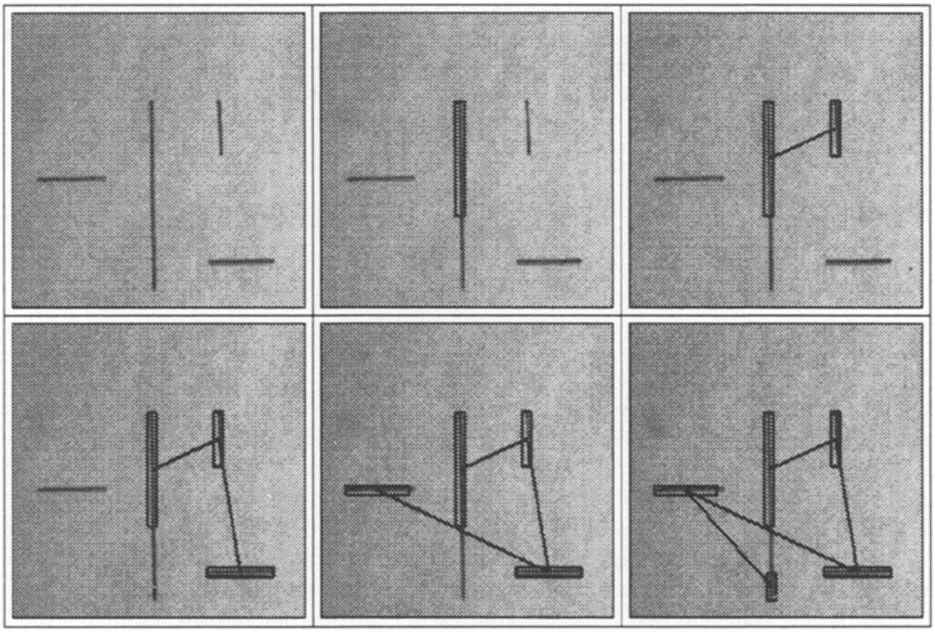

Fig. 6. Scan paths for a 256 × 256 8-bit image digitized image consisting of horizontal and vertical lines *(minRF = 10, maxRF = 100)*. The path displays a priority order in which regions of the image are assessed. The focus of attention falls on the longest lines first (only a portion of the longest line is attended to first in this example because *maxRF = 100*).

is dependent on its location, which is contrary to recent psychological evidence [6]. In our model, constant time is required irrespective of the locations of the sensory items.

Anderson and van Essen have proposed the idea of "shifter networks" to explain attentional effects in vision [1]. There is some similarity between their model and the inhibitory beam idea presented here. The Anderson and van Essen proposal requires a two-phase process. First, a series of microshifts map the attention focus onto the nearest cortical module, then a series of macroshifts switch dynamically between pairs of modules at the next stage, continuing in this fashion until an attentional centre is reached. A major drawback to this scheme is that there is no apparent method for control of the size and shape of the attention focus. This is easily accomplished in our beam proposal because the beam has internal structure that may be manipulated. Also, Anderson and van Essen do not describe how the effects of nonattended regions of a receptive field are eliminated. Finally, the shifting operation is quite complex and time consuming; whether this sort of strategy can account for the extremely fast response times of human attention is unclear.

Califano, Kjeldsen and Bolle propose a multiresolution system in which the input is processed simultaneously at a coarse resolution throughout the image and at a finer resolution within a small "window" [2]. An attention control mechanism directs the high-resolution spot. In many respects, our scheme may be considered a more general expansion of the Califano model. Our model, however, allows for many resolutions whereas Califano's is restricted to two. Moreover, our model allows for a variable size and shape of the focus of attention, whereas both are fixed in Califano's model. The size and shape of their coarse resolution representation are also fixed. These restrictions do not allow a "shrink wrapping" around an object, as it is attended to, from coarser to finer resolutions; nevertheless, our model performs this, as also observed in monkey visual cortex by Moran and Desimone [8].

Several attentional schemes have been proposed by the connectionist community. Mozer describes a model of attention based on iterative relaxation [9]. Attentional selection is performed by a network of simple computing units that constructs a variable-diameter "spotlight" on the retinotopic representation. This spotlight allows sensory information within it to be preferentially processed. Sandon describes a model which also uses an iterative rule but performs the computation at several spatial scales simultaneously [13]. There are several shortcomings of iterative models such as these. One problem is that the settling time is quite sensitive to the size and nature of the image. The time required may be quite long if there are similar regions of activity that are widely separated. For example, Mozer reports that his scheme took up to 100 iterations to settle on a 36 × 6 image [10]. These schemes are clearly not suited to real-world high-resolution images.

Summary

We have argued that an attention mechanism is a necessary component of a computer vision system if it is to perform tasks in a complex, real world. A new model for visual attention was introduced whose key component is an attentional beam that prunes the processing hierarchy, drastically reducing the number of computations required. The parallel nature of the hierarchy structure further increases the efficiency of this model. This efficiency was shown empirically with simulations on high-resolution images. The results confirm that our model is one that is highly suited for real-world vision problems.

Acknowledgements

Niels daVitoria Lobo provided helpful suggestions. This research was funded by the Information Technology Research Centre, one of the Province of Ontario Centres of Excellence, the Institute for Robotics and Intelligent Systems, a Network of Centres of Excellence of the Government of Canada, and the Natural Sciences and Engineering Research Council of Canada.

References

1. C.H. Anderson and D.C. Van Essen. Shifter circuits: A computational strategy for dynamic aspects of visual processing. In *Proceedings of the National Academy of Science, USA*, volume 84, pages 6297–6301, 1987.
2. R. Califano, A. Kjeldsen and R.M. Bolle. Data and model driven foveation. Technical Report RC 15096 (#67343), IBM Research Division - T.J. Watson Lab, 1989.
3. D. Chapman. *Vision, Instruction and Action.* PhD thesis, MIT AI Lab, Cambridge, MA, 1990. TR1204.
4. J.A. Feldman. Four frames suffice: A provisional model of vision and space. *The Behavioral and Brain Sciences*, 8:265–313, 1985.
5. C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.
6. B. Kröse and B. Julesz. The control and speed of shifts of attention. *Vision Research*, 29(11):1607–1619, 1989.
7. D. Marr. Early processing of visual information. *Phil. Trans. R. Soc. Lond.*, B 275:483–524, 1976.
8. J. Moran and R. Desimone. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229:782–784, 1985.
9. M.C. Mozer. A connectionist model of selective visual attention in visual perception. In *Proceedings: 9th Conference of the Cognitive Science Society*, pages 195–201, 1988.
10. M.C. Mozer. *The Perception of Multiple Objects: A Connectionist Approach.* MIT Press, Cambridge, MA, 1991.
11. U. Neisser. *Cognitive Psychology.* Appleton-Century-Crofts, New York, NY, 1967.
12. Y. Posner, M.I. Cohen and R.D. Rafal. Neural system control of spatial ordering. *Phil. Trans. R. Soc. Lond.*, B 298:187–198, 1982.
13. P.A. Sandon. Simulating visual attention. *Journal of Cognitive Neuroscience*, 2(3):213–231, 1990.
14. A. Sha'ashua and S. Ullman. Structure saliency: The detection of globally salient structures using a locally connected network. In *Proceedings of the Second ICCV*, pages 321–325, Tampa, FL, 1988.
15. A. Treisman. Preattentive processing in vision. *Computer Vision,Graphics, and Image Processing*, 31:156–177, 1988.
16. J.K. Tsotsos. The complexity of perceptual search tasks. In *Proceedings, IJCAI*, pages 1571–1577, Detroit, 1989.
17. J.K. Tsotsos. Analyzing vision at the complexity level. *The Behavioral and Brain Sciences*, 13:423–469, 1990.
18. J.K. Tsotsos. Localizing stimuli in a sensory field using an inhibitory attentional beam. Technical Report RBCV-TR-91-37, University of Toronto, 1991.
19. L.M. Uhr. Psychological motivation and underlying concepts. In S.L. Tanimoto and A. Klinger, editors, *Structured Computer Vision.* Academic Press, New York, NY, 1980.
20. A.L. Yarbus. *Eye Movements and Vision.* Plenum Press, 1967.