Human Neurobiol (1985) 4:219–227

Human Neurobiology


# Shifts in selective visual attention: towards the underlying neural circuitry

C. Koch[1] and S. Ullman[1,2]

[1] Artificial Intelligence Laboratory and Center for Biological Information Processing, MIT, E 25–201, Cambridge, MA, 02139, USA and [2] Department of Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel, and Department of Psychology, MIT, Cambridge, MA, USA

**Summary.** Psychophysical and physiological evidence indicates that the visual system of primates and humans has evolved a specialized processing focus moving across the visual scene. This study addresses the question of how simple networks of neuron-like elements can account for a variety of phenomena associated with this shift of selective visual attention. Specifically, we propose the following: (1) A number of elementary features, such as color, orientation, direction of movement, disparity etc. are represented in parallel in different topographical maps, called the early representation. (2) There exists a selective mapping from the early topographic representation into a more central non-topographic representation, such that at any instant the central representation contains the properties of only a single location in the visual scene, the *selected* location. We suggest that this mapping is the principal expression of early selective visual attention. One function of selective attention is to fuse information from different maps into one coherent whole. (3) Certain selection rules determine which locations will be mapped into the central representation. The major rule, using the conspicuity of locations in the early representation, is implemented using a so-called Winner-Take-All network. Inhibiting the selected location in this network causes an automatic shift towards the next most conspicious location. Additional rules are *proximity* and *similarity preferences.* We discuss how these rules can be implemented in neuron-like networks and suggest a possible role for the extensive back-projection from the visual cortex to the LGN.

**Key words:** Attention – Selective attention – Conjunction – Shifts in attention – Winner-take-all

A number of psychophysical studies concerning the detection, localization and recognition of objects in the visual field have suggested a two-stage theory of human visual perception. The first stage is the "preattentive" mode, in which simple features are processed rapidly and in parallel over the entire visual field. In the second, "attentive" mode, a specialized processing focus, usually called the focus of attention, is directed to particular locations in the visual field. The analysis of complex forms and the recognition of objects are associated with this second stage (Neisser 1967; Bergen and Julesz 1983; Treisman 1983; Ullman 1984; Julesz 1984). The computational justification for such a hypothesis comes from the realization that while it is possible to imagine specific algorithms performing tasks such as shape analysis and recognition at specific locations, it is difficult to imagine these algorithms operating in parallel over the whole visual scene, since such an approach will quickly lead to a combinatorial explosion in terms of required computational resources (Poggio 1984; Ullman 1984). This is essentially the major critique of Minsky and Papert to a universal application of perceptrons in visual perception (Minsky and Papert 1969). Taken together, these empirical and theoretical studies suggest that beyond a certain preprocessing stage, the analysis of visual information proceeds in a sequence of operations, each one applied to a selected location (or locations).

Experimental evidence for "selective attention" derives mainly from two different sources, psychophysics and physiology. The psychophysical evidence for a specialized processing focus, related but not identical to the fovea, which can be shifted around a visual scene, can be divided into two classes of experiments. First, experiments by Treisman and her collaborators (Treisman and Gelade 1980; Treisman 1982, 1983) showed that visual search for targets defined by a single feature (such as looking for a green line among many red ones) occurs in parallel across a spatial display (the so-called *pop-out* effect), whereas search for a conjunctive target defined in terms of several features (e.g. searching for a line which is both red and vertical, among lines which can be either red or blue, and horizontal or vertical) requires a serial, self-terminating scan through distracting items present in the display. Thus, while the search latency versus number-of-distractors curve is essentially flat for disjunctive features, it increases linearly for the conjunctive search. Julesz has also shown in his studies of texture discriminations that only a limited set of features, termed *textons,* can be detected in parallel (Bergen and Julesz 1983; Julesz 1984). Among the features that can be detected in parallel are color, orientation of line segments and certain shape parameters such as curvature (Treisman 1983; Julesz and Bergen 1983) and possibly stereo (Nielsen and Poggio, unpublished experiments). Second, in a series of different experiments, subjects were asked to detect a given target and were given a cue about the expected position of the target. Thus subjects "attended" the expected location, without fixating or foveating it (since they were required at all times to fixate a test spot). The performance was generally superior to the situation where no such pre-cuing occurred, suggesting the notion of an advance shift of the processing

focus to a particular spatial location (Eriksen and Hoffman 1972; Posner 1980; Bashinski and Bacharach 1980; Remington and Pierce 1984).

Phenomena related to the selective processing of visual information were also found in physiological studies. In a series of recording from awake, behaving monkeys, Goldberg and Wurtz (1972) found that the response of cells in the superficial layers of the superior colliculus to visual stimuli was enhanced if the monkey intended to use its receptive field as a target for a fast eye movement (saccade). These cells did not respond to eye movement per se, since they did not discharge to saccades in the dark. Nor is the discharge of these neurons synchronous with the onset of saccadic eye movements. The effect is highly spatial selective, since saccades to other areas of the visual field did not induce such an enhancement. In the posterior parietal cortex (Brodman's area 7), Bushnell et al. (1981) observed that visual responses are enhanced whenever the animal uses the receptive field of the neuron under study for some behavior, such as reaching out to touch the stimulus, initiating an eye movement towards the stimulus etc. Again, the effect is spatially selective (see also Mountcastle et al. 1981). Recently Haenny et al. (1984) have demonstrated selective gating in area V4 of the monkey. In their experiment the monkey was required to release a dial if it detected an agreement between tactile and visual stimuli (if the orientation of line grooves on the dial paralleled the orientation of a visually presented grating). While some cells responded to a specific visual cue, independent of the tactile one, e.g. they always responded to a horizontal grating, some discharged only if there was no discrepancy in the orientation of the two patterns. In summary, single cells in certain parts of the visual system respond differently to the same physical stimulus, enhancing their response as a function of the visual task being performed (see also Moran and Desimone 1985).

These results and the notion that visual analysis can be directed to selected locations raise a number of interesting questions. What are the operations that the visual system can apply to a selected locations? How does the selection proceed? That is, what determines the next location to be processed, and how does the processing shift from the current to the next selected location? In this paper we will explore some of these issues by first defining the problem and then exploring possible mechanisms and their implementation in simple neuron-like networks.

## The problem

We will now proceed to set the general framework for our subsequent discussion, emphasizing our assumptions and the exact nature of the problem being considered. As a starting point, we will suggest a framework for discussing selective attention in terms of cellular physiology. The experimental evidence indicates that selective visual attention plays already an important role at early processing stages, and therefore an attempt to relate selective attention to the physiological level seems justified.

We assume that selective visual attention operates on what we call the *early representation,* a set of topographical, cortical maps encoding the visual environment (Zeki 1978; Barlow 1981). The early representation includes a variety of different maps for different *elementary features* such as orientation of edges, color, disparity and direction of movement. For each location in these maps there are a number of dimensions, such as different colors or orientations. Neighborhood relations are preserved in these maps, i.e. nearby locations in the visual scene project to nearby locations in the map. Local, inhibitory connections, mediating lateral inhibition, occur either at an earlier stage or within the feature maps. Thus, locations that differ significantly from their surrounding locations are singled out at this level. The state of each of these maps therefore signals how conspicuous a given location in the visual scene is: a red blob surrounded by similar red blobs will certainly be less conspicuous than a red blob surrounded by green blobs. It should be emphasized that the different maps do not necessarily have to be in physically different locations, but may be intermixed. Moreover, these maps may possibly exist at different scales, i.e. at different spatial resolutions, in accordance with the evidence for multiple spatial channels (Campbell and Robson 1968; Wilson and Bergen 1979).

When an observer *selectively attends* a particular location, the properties associated with the selected location will be mapped into a higher, more abstract, non-topographic representation.[1] Knowing the properties of the selected location in the visual scene is equivalent to knowing the properties in the non-topographical representation (Fig. 1). This framework is compatible in general terms with the notion of a hierarchy of cortical areas devoted to the processing of different features.

Given this framwork, we can now ask some specific questions regarding the operation of selective visual attention. At least two problems must be solved here. (1) The *Winner-Take-All* problem: making sure that only one location in each map is active out of the many that are initially active. (2) The *spatial register* problem: that is aligning the different feature maps with respect to each other. Combining the information of the different feature maps or retrieving information relevant to a single location, presupposes a fast and reliable pathway to address the same location in different feature maps. Where is the anatomical correlate of this path-

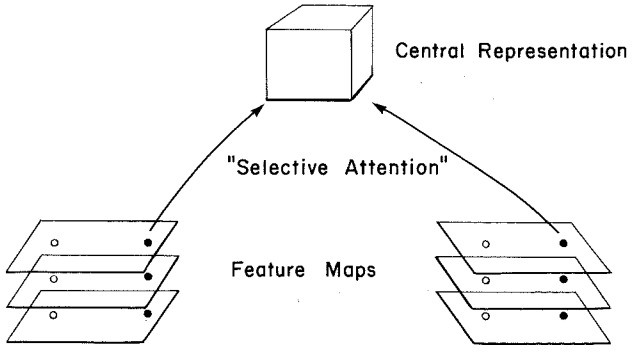


**Fig. 1.** A very schematic drawing defining what we mean by selective visual attention. The different features of objects across the visual field, such as color (to the left) and orientation (to the right), are represented in topographical maps having possibly different dimensions (e.g. for the different orientations). Selective attention is a mapping of the properties of a given, the "selected" location, into a higher, non-topographic representation

[1] By higher, we denote some stage of cortical processing further removed from the periphery

way? (3) The *shift* problem: How does the processing focus shift to another location? In the following we suggest some answers to these questions.

## Two mechanisms subserving selective visual attention

In order to understand how the selective mapping of the properties of the attended location may occur, we will introduce two intuitively quite plausible mechanisms; one to yield a simple measure of the conspicuity of a location in the visual scene and the other to select the single most active unit among a large number of active units. Formulating the operation of selective attention in terms of these mechanism, rather than in the language of higher cognitive concepts, has the advantage that specific predictions concerning the anatomy and electrophysiology of the specialized cortical regions involved in attention can be derived. The main point we wish to make is not that the particular mechanisms we propose are necessarily implemented in the brain, but that the shift of selective visual attention and related visual operations can be explained using simple mechanisms compatible with cortical physiology and anatomy.

### *The saliency map*

Given the different elementary feature maps, it seems plausible to assume that the conspicuity of a location in the visual scene determines the level of activity of the corresponding units in the different elementary feature maps. The higher their activity, for instance their firing frequency, the higher the saliency of the corresponding location in the visual field. Thus, the different feature maps code for the conspicuity *within* a particular feature dimension. In order to assess the global, overall conspicuity of a location, we will assume the existence of another topographical map, termed the *saliency map,* which combines the information of the individual maps into one global measure of conspicuity. The points corresponding to one location in the elementary feature maps project onto a unit in the saliency map. The exact nature of the projection is not relevant here, as long as increased conspicuity in the feature maps corresponds to an increased conspicuity in the saliency map. This map gives then a "biased" view of the visual environment, emphasizing interesting or conspicuous locations in the visual field. Since the saliency map is still a part of the early visual system, it most likely encodes the conspicuity of objects in terms of simple properties such as color, direction of motion, depth and orientation. Saliency at a given location is determined primarily by how different this location is from its surround in color, orientation, motion, depth, etc. It is possible, however, that the relative weight of the different properties contributing to this representation can be modulated by the activity of some higher cortical centers, as for instance during prolonged practice with a particular set of targets and distractors (Schneider and Shiffrin 1977).

### *Selective mapping*

Next, we have to make sure that only the properties corresponding to the most conspicuous location are mapped from the early representation into the more central one. We therefore postulate a "switch" that routes the properties of a single location, the *selected* or *attended* location, into the central representation. Note that the computations required to abstract certain properties from the visual input are performed within the early representation, i.e. prior to the selection process, and not subsequent to it. This distinction is important, for instance in the computation of color. As has been demonstrated psychophysically (e.g. Land 1983; Land et al. 1983), the computation underlying color perception is a global process, that requires the entire visual field (or a large portion of it). It is therefore reasonable to assume that the computation of color and other properties proceeds within the early representation, prior to the selection of a location for further processing. This view of attention is in accordance with the fact that it is not possible for visual attention to be allocated simultaneously to two different positions in space (Posner et al. 1980).

*The basic mechanism.* The operations underlying the selective routing of information from the early representation to the central one can be performed by two complementary cellular networks (Fig. 5). One such network, called the *Winner-Take-All* network (WTA network; see Feldman 1982, who introduced this term; Feldman and Ballard 1982) localizes the most active unit in the saliency map while the second network relays the properties of the selected location to the central representation. At any given time only one location is selected from the early representation and copied into the central representation. The WTA network, equivalent to a maximum finding operator, operates on the output $x_i$ of the units in the saliency map. In a neuronal network $x_i$ can be interpreted as the electrical activity (intracellular voltage or spiking rate) of the unit at location $i$. The WTA mechanism maps this set of input units onto an equal number of output units, described by $y_i$, using the transformation rule:

$$y_i = 0 \quad \text{if } x_i < \max_j x_i$$
$$y_i = f(x_i) \text{ if } x_i = \max_j x_j \qquad (1)$$

where $f$ is any increasing function of $x_i$ (including a constant). All output units are set to zero except the one corresponding to the most active input unit (Fig. 2).

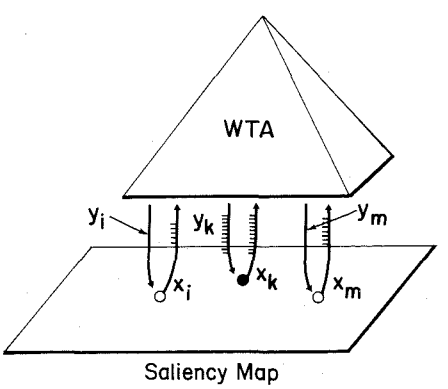


**Fig. 2.** A schematic drawing illustrating the Winner-Take-All (WTA) network computing the maximum $x_k$ of a set of $n$ input units in the salient map. It localizes the most conspicuous point by a number of parallel operations and activates the corresponding output line (in this case unit $x_k$) after at most $2log_m n$ time steps (if $m$ units can be compared simultaneously)

*The Winner-Take-All network.* Building a WTA network may appear as a straightforward task, but complications arise when the intrinsic properties of biological hardware are taken into account. Depending on the underlying hardware, two extremes for computing the maximum of a given set can be envisioned. On a serial machine, the simplest algorithm is a sequential search for the largest number through the entire input set. The drawback to this method is that for $n$ inputs, $n$ basic time steps are required (by basic time step we always refer to the time required to execute an elementary operation such as comparing two numbers). A highly parallel machine with $n$ processors, each one having direct access to the other $n-1$ processors, can compute the maximum in one time step by comparing simultaneously the value of each processor with the values of all the other processors.[2] A simple implementation one may suggest is a *mutual inhibitory network* of the type studied by Hadeler (1974), where every unit inhibits every other unit. In these networks, neurons are assumed to be linear summation devices, followed by a threshold operation (see for instance McCulloch and Pitts 1943). Such networks will be unable, however, to implement the WTA computation for arbitrary inputs $x_i$, since there is no guaranty that the network will converge (for more details, see Koch and Ullman 1984). Moreover, the requirement that each unit in these networks is connected to every other unit seems prohibitive in terms of the total number of connections required ($n^2 - n$ if the connections are uni-directional). We propose a more feasible implementation of a WTA network, based on two biologically motivated assumptions.

1. Except for some long-range excitatory connections, most connections, whether excitatory or inhibitory, are local.

2. Each elementary processing unit only performs some simple well-specified operation, such as addition or multiplication. In particular, the basic processing units are unable to use any symbolic information, such as addresses.

In the first network, every unit $i$ has associated with it a variable $y_i$. Every unit receives a constant and non-negative input $x_i$. The state equation for $y_i$ is given by

$$\frac{dy_i}{dt} = y_i (x_i - \sum_j x_j \cdot y_j), \tag{2}$$

where the sum is taken over all $j$, from 1 to $n$. The equation itself is due to K. P. Hadeler. With the initial condition $\Sigma_j y_j(0) = 1$, the solution is given by

$$y_i(t) = \frac{y_i(0)e^{x_i t}}{\Sigma_j y_j(0)e^{x_j t}} \quad . \tag{3}$$

By inspecting this equation we can see immediately that if $x_i$ is the maximum among all $x_j$, the corresponding $y_i$ will tend asymptotically to 1, while all other $y_j$'s decay to 0. Formally, the $y_i$'s correspond to a discrete probability distribution, since $\Sigma_j y_j(t) = 1$ for all times $t$. $\Sigma_j x_j y_j$ then corresponds to the average activity of the network. Notice, that the speed of convergence of $y_i$ depends on the strength of the input $x_i$. For large inputs, the time-constant $1/x_i$ is small and convergence is rapid while for a $x_i$ marginally larger than $x_j$, convergence will be slow. This scheme just requires one very fast processing unit, computing the global activity $\Sigma_j x_j y_j$ of the network and projecting to every other unit. If the input $x_i$ varies on a slower time-scale than the typical convergence time of the network, then the solution in the time-varying case is

$$y_i(t) \simeq \frac{y_i(0)e^{x_i(\epsilon t)t}}{\Sigma_j y_j(0)e^{x_j(\epsilon t)t}} \tag{4}$$

where $\epsilon << 1$ governs the time scale of the input. However, once the system "converges" to a particular solution $x_i$ at time $t$ and the input $x$ changes, the new solution, lets say $x_j$, will take considerable time to converge, since initially $x_i \to 1$ and $x_j \to 0$. The second drawback of this form of a WTA is that the update rule (2) is rather difficult to implement in "real" neurons. Finally, this network does not ex-

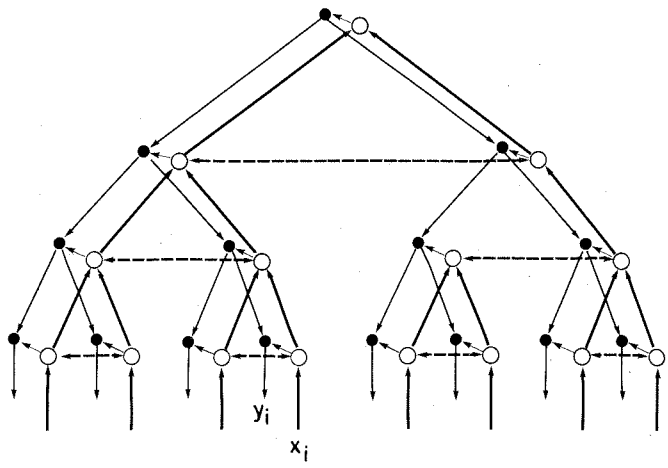


**Fig. 3.** A second implementation of the Winner-Take-All network with $n = 8$ input units. The local comparison takes place between $m = 2$ units. The more active unit inhibits the less active one and excites the unit on the next level. The auxiliary units, drawn in black, are only activated if they receive conjointly excitation from their associated main unit and from the auxiliary unit at the higher level. The auxiliary unit $y_i$, corresponding to the most active unit $x_i$ in the saliency map, will be activated after at most $2log_m n = 8$ time steps. In order to insure stability against noise and to enforce neighborhood relations between all neighboring points (for instance between the two middle units, belonging to two different subtrees) additional connections (and units) can be added between (and within) levels. We have just shown the most sparse implementation of a WTA network

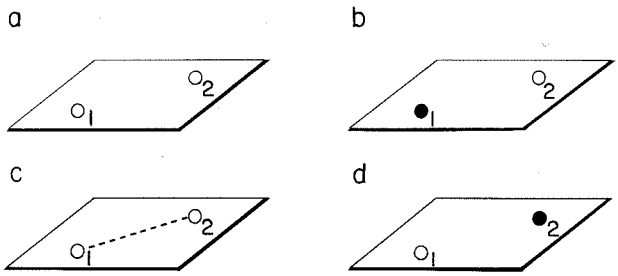


**Fig. 4.** Shifting visual attention within the salient map. Once the most conspicuous location (point 1) has been detected and examined, its corresponding output $x_1$ decays and the WTA mechanism shifts to the next most salient location, 2. The time needed to find the next location increases with increasing distance between locations 1 and 2

[2] This is essentially the mechanism Feldman and Ballard (1982) proposed for their implementation of a WTA network

hibit the distance effect discussed under "Shifting the processing focus". For these reasons, this particular implementation of a WTA network, although possible, does not appear probable.

The second implementation of a WTA network shows a hierarchical structure.[3] The network operates in a highly parallel fashion by computing the maximum of a small number $m$ of units across the whole input set. Next, comparisons are made among these local maxima to compute again the most active unit. These comparisons are repeated $k = log_m n$ times until the global maximum has been determined. Figure 3 shows one particular implementation of the WTA network with $m = 2$. The more active unit inhibits the less active unit and transmits its activity onto the next higher level. Here, among $n/2$ units, the process is repeated.[4] Under the assumption that the connections between the levels transmit faithfully the activity of the units, the top-most unit in the pyramid will hold the activity $x_i$ of the global maximum after $k$ time steps. However, it is the location of the maximum and not its absolute value which is of relevance for the selection process. The location of the corresponding unit in the saliency map can be obtained by the use of the second pyramid, having a reversed flow of information with respect to the first pyramid. It "marks" the path of the most active unit through the first pyramid, activating finally the output $y_i$ of the WTA. This is done with the help of an *auxiliary* unit associated with every unit in the first pyramid (called the *main* unit). An auxiliary unit is only activated if it receives conjoint excitation from its main unit and from the auxiliary unit at the next higher level. Since at every level the most active (main) unit in a local comparison suppresses the activity of the other $m - 1$ (main) units, the associated auxiliary units as well as all auxiliary units in the subtree below them can never be activated. After another $k$ time steps, the output $y_i$, corresponding to the most active unit in the saliency map, will be activated, while the rest of the output units remain silent. Except for the pathological case when two or more inputs are exactly equal, the WTA network will always converge to a unique solution within at most $2log_m n$ time steps. It can be built with no more than $2n\ m/(m-1)$ units. This can be immediately established by noticing that the total number of units required is always smaller or equal to the infinite geometric series. Thus, for all integers $m$ the WTA network can always be built with less than $4n$ units.

Assuming that the optic nerve contains approximately $10^6$ fibers and that $m = 10$ neurons can compare their activity simultaneously, a WTA network covering the entire retinal image would require no more than $2.2 \cdot 10^6$ neurons, a small fraction of all visual neurons. Moreover, a solution will always be found after at most 12 time steps. Since time-constants for neurons are in the ms range, this number seems broadly compatible with the estimated 30–50 ms required to shift visual attention to a new location (Bergen and Julesz 1983). If the Y-system, with its associated short delay, high movement sensitivity, large receptive fields and transient temporal response, provided the major input to the WTA network, this number would drop substantially (Lennie 1980; Sherman 1985). In the cat, about 4% of all ganglion cells are of the Y-type. If this percentage carries over to primates and man, a WTA network for the entire visual field could be built with just $10^5$ neurons (Lennie 1980; Sherman 1985). Interestingly, the computational architecture of the WTA network is reminiscent of the K- and P-pyramids proposed by Minsky for his K-line theory of memory (Minsky 1979). In the following, we will only consider the second implementation of a WTA network.

One cautionary note here. Since neurons rarely show all-or-none behavior, we do not expect all units in the saliency map to be completely inhibited while only the unit corresponding to the selected location fires. Rather, this unit may have an enhanced response while all other units are depressed.

*Mapping the selected location into the central representation.* Once the most conspicuous point has been localized in the saliency map, its properties, i.e. the information contained within the early representation, must be copied into the central representation. The routing of this information can be achieved by removing some tonic inhibitory influence or by increasing the amount of excitation at the selected location in the early representation. We will not suggest here specific mechanisms for the mapping operation. The crucial point is that the WTA network directs the "copy" operation to a single selected location. Note, that the selection system itself is *not* responsible for the information processing relevant to the visual task but simply selects which area of visual space should be inspected (Posner et al. 1980). It can be likened to a spotlight illuminating some portion of the visual field.

## Shifting the processing focus

Until now we have only considered the initial selection of an "interesting" location. But how does the selection process move from one location to the next, i.e. how can selective attention shift across the visual field (Shulman et al. 1979)? From psychophysical experiments it is known that it takes some measurable time to shift the focus of attention from one location to another (Eriksen and Schultz 1977; Tsal 1983). There is some evidence that this time increases with the distance between these locations (Shulman et al. 1979; Tsal 1983; see however, Remington and Pierce 1984).

A simple way to introduce such dynamics into our model is to let the conspicuity of the maximal active unit in the saliency map decay, even if constant visual stimuli are present. This decay may be implemented either locally or centrally (or by some combination of the two methods). By "local" we mean that an active location in the saliency map adapts and decays after a while. By "central" we mean that once the information from the early representation has been relayed to the central representation a signal is sent back, inhibiting the most active unit in the saliency map, i.e. its conspicuity fades. The WTA network responds to the new input configuration by shifting away from the presently selected location and towards the next most conspicuous

[3] Hierarchical, pyramid-like computer architectures have been proposed for image processing and analysis. For an overview of their use see Rosenfeld (1984)

[4] The computational structure is similar to the Wimbledon tennis tournament where players drop out if they lose a single match (a so-called knock-out competition)

location. The convergence time, i.e. the time taken by the WTA network to converge to the newly selected location, depends primarily on the distance between the two locations. In the worst case it will take $2log_m n$ time steps for the new maximum to propagate up, and subsequently down, the $log_m n$ layers (Fig. 4), assuming that the comparison of $m$ units can be done in one time step. Shorter convergence time can be achieved if the two locations are close to each other. Note, that the dependency of the convergence time on distance follows naturally from the computational architecture of the WTA network and does not have to be artificially imposed.

The local scheme is similar, except that the most active unit is locally inhibited, for instance some fixed time interval after the WTA mechanism has converged. These schemes are non-exclusive; in fact, it seems likely that some local, automatic mechanism might always be in operation. The central mechanism may be invoked when a voluntary shift of attention is desired (Posner 1980). The basis for both mechanisms is a long-lasting inhibition of the selected unit in the saliency map preventing, for a given time period, that the attentional focus will revisit this location. A temporary inhibition, lasting more than 500 ms, has been reported by Posner et al. (1982) after attentional shifts away from a cued location.

Parallel and serial search can now be quite simply explained in terms of our mechanism. When searching an array of objects, among which at least one object has a salient property differentiating it from its neighbours, then that particular location will be quite conspicuous in the corresponding feature and saliency maps. If no other distracting objects exist within some neighbourhood, the WTA will immediately converge to this location and the object will be detected, independent (within limits) of the total numbers of surrounding objects. In other words, the red line "immediately" pops out. When searching for an object defined by the conjunction of two different features, the situation is more complicated. The saliency map will have numerous local peaks, in the worst case as many as there are objects displayed. The WTA mechanism must shift to each one individually, until the correct target has been identified. If no further search strategy were used, then on the average $n/2$ objects must be scanned before the search can be successfully terminated. On this view parallel "pop-out" and serial search are not fundamentally different: an element pops out since, due to its saliency, it is the first item to be visited. In the next section we will discuss two schemes for accelerating the search and for which there is some psychophysical support.

The above scenario also suggests a simple explanation for the fact that the presence of a particular object can be masked by other objects (Treisman 1982). Two different camouflage strategies are possible. One can either reduce the conspicuity of the to-be-hidden object at the level of the saliency map by blending this object with its background (one of the functions of combat fatigue) or one can place the to-be-hidden object among a background of very conspicuous objects, distracting "attention". In both cases, the activity at the corresponding location in the saliency map is reduced in relation to the activity of its neighbours, thereby making it less likely that the WTA network will shift to this unit and thus camouflaging the object at that particular location. Finally, shifts may possibly be directed under voluntary control (Posner 1980), although we consider in this paper only involuntary, automatic aspects of selective attention.

## Two rules for shifting the processing focus

Should there be any systematic relationship between the current location and the next location to be selected? If no such relationship is encouraged, it would seem difficult to visually inspect areas of the visual field without constantly shifting to conspicuous, but distant, locations. Thus, it would seem desirable for the visual system to be able to select potential targets according to some useful criteria.

Objects tend to occupy a compact region in space with similar properties (color, motion, etc.). If the shifting apparatus is to scan automatically different parts of a given object, it is useful to introduce a bias based on both spatial proximity and similarity. Searching for an "interesting" target around a selected location would profit from a selection mechanism biased to nearby locations (what we call *proximity preference*). Scanning the visual field for objects with a common identifying feature, for instance the color red, would be likewise facilitated if locations with similar features to the presently selected location are preferrentially selected (*similarity preference*). Both mechanisms are related to phenomena in perceptual grouping and "Gestalt effects" which occur as a function of object similarity and spatial proximity (Wertheimer 1923; Beck 1967). The next two sections discuss these rules in more detail.

### *Proximity preference*

It would seem advantageous from a computational point of view, if the selection process shifts preferentially to conspicuous locations in the neighborhood of the presently selected location, rather than to the global maximum independent of any locality considerations. The simplest way of implementing such a proximity preference within the framework of the WTA mechanism is to enhance all units in the neighborhood of the currently selected unit in the saliency map. Such a preference can be incorporated in a straightforward manner into the network described earlier. More specifically, we assume that the output of the WTA mechanism associated with the presently attended location enhances the conspicuity of nearby units in the saliency map by a factor depending on the distance between the location and its neighbors, thereby facilitating shifts of the processing focus to nearby locations. This is equivalent to postulating the existence of an attractive potential around every selected location. Some experimental evidence for this type of interaction is provided by Engel (1971, 1974). His results indicate that the probability of detecting a target depends on the proximity of the location being attended to (see also Sagi and Julesz 1984).

### *Similarity preference*

On similar computational grounds one can justify the existence of a similarity preference. We postulate therefore the existence of an interaction between similar, elementary features: the processing focus will preferentially shift to

a location with the same or similar elementary features as the presently selected location. Such a mechanism assumes interactions within individual elementary feature maps, but not between them, and therefore it does not require precise topographic mappings between the different elementary feature maps. The interaction will be activated by the output of the WTA network. This output ($y_k$ in Fig. 2) increases the excitability, viz. the conspicuity, of all units in a neighborhood of the selected location within those elementary feature maps where the corresponding features have been detected. If the currently selected location contains for instance a red, horizontal line, then neighboring units in the feature map for horizontal and red will be facilitated. The processing focus will now preferentially shift to either red and/or horizontal targets. The effect of the similarity preference is opposite to the initial bias towards conspicuous locations. Locations with similar properties initially inhibit each other. After a location has been selected, it tends to facilitate the conspicuity of nearby locations with similar properties. Although the two processes have opposite effects, they can both be implemented without causing undesirable contradiction or interference. The first occurs early on within the individual maps and is implemented by local inhibition within the maps. Finally, it would be expedient if the similarity preference for individual features could be switched on or off voluntarily (look for red objects i.e. facilitate the red feature map), but it is unclear to what degree such a control actually exists.

A partial experimental support for this type of interactions comes from a recent study by Geiger and Lettvin (Geiger 1984) who investigated the influence of the attended location on lateral masking. If the subjects fixate a central point, while a string of three letters is flashed onto the screen at some distance from the central point, the subjects are usually unable to name the central latter. However, if a copy of the interior letter is flashed at the fixation point, the letter in the periphery transiently stands out against its neighbors in the string.

## Selective visual attention and the fusion of information

Since all the properties of the selected location are mapped together into the central representation, selective attention can serve as a "glue", integrating the initially separate features into unitary objects. According to this view, the fusion of the information contained in the different feature maps at a single location only occurs once these features have been loaded into the central representation, that is once attention has shifted to this location. This role of selective attention was suggested by Treisman and Gelade (1980). Thus, a line which is both red and horizontal will only be considered as red horizontal line if attention has focused onto it. This assumption predicts that if "attention" is unable for some reason to focus correctly onto an unknown object, its composite features will not be correctly identified and conjunctions of these features will be formed on a random basis. Treisman calls this phenomenon *illusory conjunctions*. Such illusonary conjunctions have been experienced among stimuli varying in color, shape and size (Treisman and Schmidt 1982; Treisman and Paterson 1984).

Although this integrative role of selective attention was only tested for color, orientation of line segments and certain shape parameters, it is possible that attention provides the necessary mechanism to fuse information provided by the different early vision processes like stereo, motion, shape-from-shading, edge detection etc. (Marr 1982; Terzopoulos 1985). In other words, selective visual attention might operate early on in the visual system, possibly at the level of Marr's $2\ ^1/_2$D sketch. Of particular interest is the issue of whether motion and stereo are automatically combined across the whole visual scene or whether they are combined via selective attention in order to perceive both. In the later case, at least two predictions can be made. (1) The search time for a feature defined by both motion and stereo attributes should increase linearly with the number of distracting items and (2) illusory conjunctions should occur between these modalities, i.e. between motion and stereo.

## Biological considerations

Until now we have restricted ourselves to the computational and the "algorithmic" side of selective visual attention, without considering the question of the implementation into neuronal hardware (Marr and Poggio 1977). We will now briefly discuss the possible anatomical substratum of visual attention.

Our scheme for selective visual attention (Fig. 5) may be implemented in at least two different ways. First, information can flow from the retina to the early representation, where simple properties are extracted and represented in parallel. Next, these maps feed into the saliency map which

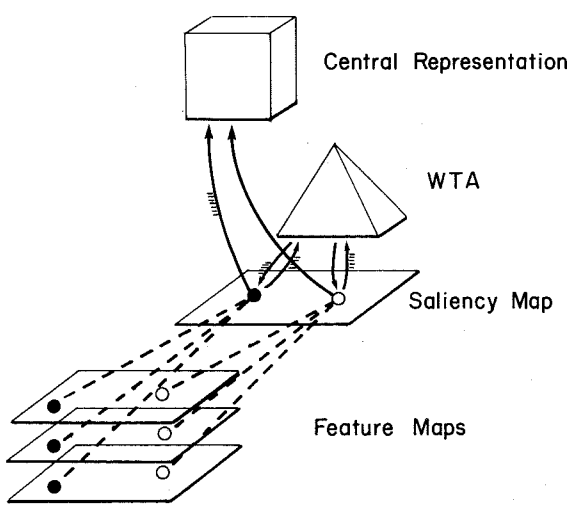


**Fig. 5.** A schematic drawing summarizing the workings of our selection process. Every location in the visual scene is analyzed in terms of elementary features, such as color, orientation of line segments, motion disparity etc. and is represented in different feature maps. Lateral inhibition within the feature maps enhances the local conspicuity. The output of these maps is combined in the saliency map, encoding salient features in the visual scene. The WTA network subsequently selects the most conspicuous location, routing the corresponding properties to the central representation. After the selection process, the central representation contains the properties of a single, the selected location. Note, that the visual input from the visual scene can either terminate on the saliency map or on the feature maps

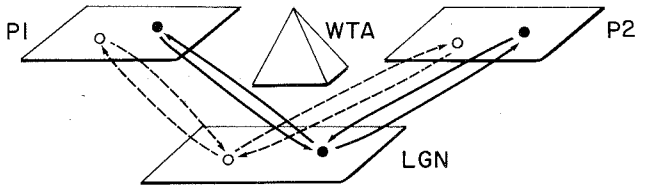


**Fig. 6.** A possible biological implementation for the selection process. The saliency map may be localized either within the lateral geniculate nucleus (LGN) or within the striate cortex (V1). The backprojection from the different cortical maps for different properties (for instance P1 and P2) solve the spatial register problem. The WTA network selects the most active unit in the saliency map, subsequently routing the information corresponding to this *selected* location into the central representation. A simple alternative (not shown) is that each feature map (P1 or P2) has its own WTA instead of having one global WTA for the saliency map, simplifying the problem of mapping the properties of the selected location into the central representation considerably. Interestingly, Crick proposed recently (1984; see also Yingling and Skinner 1977) that the attentional searchlight is controlled by the thalamic reticular nucleus, a layered structure surrounding the thalamus. It receives extensive feedback from the visual cortex and projects onto the principal relay cells in the LGN

in turns provides the input to the WTA network and to the central representation. Since the maps for the different elementary features are most likely localized in areas within and beyond striate cortex, such as MT and MST for motion, and perhaps V4 for color (Van Essen and Maunsell 1983), this implies that the saliency map must be located beyond striate cortex. An intriguing alternative is that the saliency map, in fact, preceeds the early representation. In this case, visual information is first represented in the saliency map and subsequently routed to the individual feature maps.

The second implementation seems to suggest that the saliency map resides early on in the visual system, either at the level of the lateral geniculate nucleus (LGN) or in the striate cortex, V1 (Fig. 6). The geniculate in the cat and in the striate cortex in primates represent the last major station along the retino – geniculo – cortico pathway before the visiual information is dispersed to different regions. The Y or magnocellular pathway projects from the LGN to the striate and extrastriate cortex (V1, V2 and V3) in cat, but predominately to V1 in the monkey. The X or parvocellular pathway behaves similar in both cat and monkey, projecting predominantly to V1 (Lennie 1980, Fries 1981; Sherman 1985).

One puzzling feature about the LGN is the existence of an extensive reciprocal projection from layer VI pyramidal cells in cortex onto the LGN (Updyke 1975; for an anatomical summary see Macchi and Rinvik 1976). This connection observes the general principle that for every geniculo-cortical projection there is a corresponding cortico-thalamic pathway. Although little information on the number of fibers involved in this back projection is available, estimates suggest that it is at least as massive as the forward projection, and perhaps considerably stronger (Gilbert and Kelly 1975) estimate that about half of all cells in layer VI in the cat striate cortex send their axons to the LGN). Cross-correlation analysis between visual cortex and LGN neurons reveals an excitatory pathway if the receptive field center of both neurons are separated by less than 1.7° (Tsumoto et al. 1978). Inhibitory cortico-geniculate interactions were demonstrated in most cases if the receptive field centers of the cortical and the geniculate neuron were separated by more than 1.8°.

These strong reciprocal connections could be used to solve the spatial register problem in the manner suggested in Figure 6. The visual environment is encoded at the level of the LGN or V1 in neurons having circular-symmetric receptive fields. Subsequently, different properties such as color, motion, disparity etc. are processed, analyzed and represented in different regions of the cortex. These regions then project back to the LGN (via V1). If, for instance, in the area computing color a single location stands out among all others, this location will enhance the corresponding location in the LGN. Similarly, the different visual maps all feed back into the saliency map, providing it with a measure of the strength and importance of the different features. The WTA network now localizes the most active unit in the saliency map. This arrangement provides a mechanism for spatial register, since all the information pertaining to the selected location is transmitted together to the central representation. A notable limitation of this mechanism is that spatial register is obtained for one location at a time, a property that is consistent with psychophysical evidence (Treisman and Gelade 1980). Finally, although this arrangement presupposes a precise topographic cortico-geniculate projection, it places no such demands on the interconnections among the different visual maps.

## Summary

We have suggested that selective visual attention requires three different stages (Fig. 5). First, a set of elementary features is computed in parallel across the visual field and is represented in a set of cortical topographic maps. Locations in visual space that differ from their surround with respect to an elementary feature such as orientation, color or motion are singled out in the corresponding map. These maps are combined into the saliency map, encoding the relative conspicuity of the visual scene. Second, the WTA mechanism, operating on this map, singles out the most conspicuous location. Thirdly, the properties of this selected location are routed to the central representation. The WTA network then shifts automatically to the next most conspicuous location. The shift can be biased by proximity and similarity preferences. The visual system processes a scene in a sequential and automatic way by selectively inspecting the information present in conspicuous locations. The mechanism sketched here might of course not only be used for the shift of the attentional focus but also for such visual routines as tracking of contours, counting objects or marking a specific location (Ullman 1984).

*Acknowledgements.* We would like to thank John Maunsell and Tomaso Poggio for discussions and for critically reading the manuscript. We thank K. P. Hadeler for suggesting Eq. (2). C. K was supported by a fellowship from the Fritz Thyssen Stiftung and is presently being supported by a grant from the Office of Naval Research, Engineering Psychology Division. Support for the Center of Biological Information Processing is provided in part by a grant from the Sloan foundation and in part by Whitaker College at MIT.

## References


Barlow HB (1981) Critical limiting factors in the design of the eye and visual cortex. Proc Roy Soc (Lond) B212:1–35

Bashinski HS, Bacharach VR (1980) Enhancement of perceptual sensitivity as the results of selectively attending to spatial locations. Percep Psychophys 28:241–248

Beck J (1967) Perceptual grouping produced by line figures. Percept Psychophys 2:491–495

Bergen JR, Julesz B (1983) Focal attention in rapid pattern discrimination. Nature 303:696–698

Bushnell C, Goldberg ME, Robinson DL (1981) Behavioral enhancement of visual responses in monkey cerebral cortex. I. Modulation in posterior parietal cortex related to selective visual attention. J Neurophysiol 4:755–772

Campbell FWC, Robson J (1968) Application of Fourier analysis of the visibility of gratings. J Physiol 197:551–566

Crick F (1984) The function of the thalamic reticular complex: the searchlight hypothesis. Proc Natl Acad Sci USA 81:4586–4590

Engel FL (1971) Visual conspicuity, directed attention and retinal locus. Vision Res 11:563–576

Engel FL (1974) Visual conspicuity and selective background interference in eccentric vision. Vision Res 14:459–471

Eriksen CW, Hoffman JE (1972) Temporal and spatial characteristics of selective encoding from visual displays. Percept Psychophys 12(2B):201–204

Eriksen CW, Schultz DW (1977) Retinal locus and acuity in visual information processing. Bull Psychonomic Soc 9:81–84

Feldman JA (1982) Dynamic connections in neural networks. Biol Cybern 46:27–39

Feldman JA, Ballard DH (1982) Connectionist models and their properties. Cognit Sci 6:205–254

Fries W (1981) The projection from the lateral geniculate nucleus to the prestriate cortex of the macaque monkey. Proc Roy Soc (Lond) B213:73–80

Geiger G (1984) Eccentric enhancement of form perception. Proc IEEE Int Conf Syst Man Cybern

Gilbert CD, Kelly JP (1975) The projections of cells in different layers of the cat's visual cortex. J Comp Neur 163:81–106

Goldberg ME, Wurtz RH (1972) Activity of superior colliculus in behaving monkey. II. Effect of attention and neural responses. J Neurophysiol 35:560–574

Hadeler KP (1974) On the theory of lateral inhibition. Kybernetik 14:161–165

Haenny P, Maunsell J, Schiller P (1984) Cells in prelunate cortex alter response to visual stimuli of different behavioral significance. Perception 13:A7

Julesz B (1984) A brief outline of the texton theory of human vision. Trends Neurosci 7:41–48

Julesz B, Bergen JR (1983) Textons, the fundamental elements in preattentive vision and perception of textures. Bell Syst Tech J 62:1619–1645

Koch C, Ullman S (1984) Selecting one among the many: a simple network implementing shifts in selective visual attention. Artificial Intelligence Lab Memo No 770. MIT, Cambridge

Land EH (1983) Recent advances in retinex theory and some implications for cortical computations: color vision and the natural image. Proc Natl Acad Sci USA 80:5163–5169

Land EH, Hubel DH, Livingstone MS, Perry SH, Burns MM (1983) Colour-generating interactions across the corpus callosum. Nature 303:616–618

Lennie P (1980) Parallel visual pathways: a review. Vision Res 20: 561–594

Macchi G, Rinvik E (1976) Thalamo-telencephalic circuits: a neuroanatomical survey. In: Creutzfeldt O (ed) Handbook of Electroencephalography and Clinical Neurophysiology, Vol 2(A). Elsevier, Amsterdam

Marr D (1982) Vision. Freeman Co, San Francisco

Marr D, Poggio T (1977) From understanding computation to understanding neural circuitry. Neurosci Res Prog Bull 15:470–488

McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 5:115–133

Minsky M (1979) K-lines: a theory of memory. Artificial Intelligence Lab Memo No 516. MIT, Cambridge, Mass

Minsky M, Papert S (1969) Perceptrons. MIT Press, Cambridge, Mass

Moran J, Desimone R (1985) Selective attention gates visual processing in the extrastriate cortex. Science 229:782–784

Mountcastle VB, Andersen RA, Motter BC (1981) The influence of attentive fixation upon the excitability of the light-sensitive neurons on the posterior parietal cortex. J Neurosci 1:1218–1235

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

Poggio T (1984) Routing thoughts. Artificial Intelligence Lab Working paper No 258. MIT, Cambridge, Mass

Posner MI (1980) Orienting of attention. Quart J Exp Psychol 32: 3–25

Posner MI, Cohen Y, Rafal RD (1982) Neural systems control of spatial orienting. Phil Trans R Soc (Lond) B298:187–198

Posner MI, Snyder CRR, Davidson BJ (1980) Attention and the detection of signals. J Exp Psychol Gen 109:160–174

Remington R, Pierce L (1984) Moving attention: evidence for time-invariant shifts of visual selective attention. Percept Psychophys 35:393–399

Rosenfeld A (ed) (1984) Multiresolution image processing and analysis. Springer, Berlin Heidelberg New York

Sagi D, Julesz B (1984) Probing the mind's "fovea" around a peripheral target with a small test light. Perception 13:A23

Schneider W, Shiffrin RM (1977) Controlled and automatic human information processing. I. Detection, search and attention. Psychol Rev 84:1–57

Sherman SM (1985) Functional organization of the W-, X-, and Y-cell pathways in the cat: a review and hypothesis. In: Sprague JM, Epstein AN (eds) Progress in psychobiology and physiological psychology. Academic Press, New York

Shulman GL, Remington RW, McLean JP (1979) Moving attention through visual space. J Exp Psychol Human Percept 5:522–526

Terzopoulos D (1985) Integrating visual information from multiple sources for the cooperative computation of surface shape. In: Pentland A (ed) From pixels to predicates: recent advances in computational and robotic vision. Ablex

Treisman A (1982) Perceptual grouping and attention in visual search for features and for objects. J Exp Psychol Human Percept 8: 194–214

Treisman A (1983) The role of attention in object perception. In: Braddick OJ, Sleigh AC (eds) Physical and biological processing of images. Springer, Berlin Heidelberg New York

Treisman A, Gelade G (1980) A feature-integration theory of attention. Cog Psychol 12:97–136

Treisman A, Paterson R (1984) Emergent features, attention and object perception. J Exp Psychol Human Percept 10:12–31

Treisman A, Schmidt H (1982) Illusory conjunctions in the perception of objects. Cogn Psychol 14:107–141

Tsal Y (1983) Movements of attention across the visual field. J Exp Psychol Human Percept 9:523–530

Tsumoto T, Creutzfeldt OD, Legendy CR (1978) Functional organization of the corticofugal system from visual cortex to lateral geniculate nucleus in the cat. Exp Brain Res 32:345–364

Ullman S (1984) Visual Routines. Cognition 18:97–159

Updyke BV (1975) The patterns of projection of cortical areas 17, 18 and 19 onto the laminae of the dorsal lateral geniculate nucleus in the cat. J Comp Neurol 163:377–395

Van Essen DC, Maunsell J (1983) Hierarchical organization and functional streams in the visual cortex. Trends Neurosci 6:370–375

Wertheimer M (1923) Untersuchungen zur Lehre von der Gestalt II. Psychol Forsch 4:301–350

Wilson HR, Bergen JR (1979) A four mechanism model for threshold spatial vision. Vis Res 19:19–32

Yingling CD, Skinner JE (1977) Gating of thalamic input to cerebral cortec by nucleus reticularis thalami. In: Desmedt JE (ed) Attention, voluntary contraction and event-related cerebral potentials. Prog Clin Neurophysiol bf 1. Karger, Basel

Zeki SM (1978) Functional specialization in the visual cortex of the rhesus monkey. Nature 274:423–428