Visual Attention

Laurent Itti, University of Southern California,
Hedco Neuroscience Building HNB-30A, Los Angeles, CA 90089-2520
itti@usc.edu, Tel: +1(213)740-3527, Fax: +1(213)740-5678

INTRODUCTION

Selective visual attention is the mechanism by which we can rapidly direct our gaze towards objects of interest in our visual environment. It can be bottom-up (image-based) or top-down (task-dependent). Attention only allows a small part of the incoming sensory information to reach short-term memory and visual awareness, allowing us to break down the problem of scene understanding into rapid series of computationally less demanding, localized visual analysis problems. Moreover attention mechanisms can provide feedback modulation of neural activity at the location and for the visual attributes of the desired or selected targets. This feedback may be essential for binding the different visual attributes of an object, such as color and form, into a unitary percept. As such, focal visual attention is often compared to a rapidly shiftable SPOTLIGHT, which scans our visual environment both overtly (with accompanying eye movements) or covertly (with the eyes fixed). Of course, not all of vision is attentional, as we can derive coarse understanding from presentations of visual scenes that are so brief that they do not leave time for attention to explore the scene. Vision thus appears to rely on sophisticated interactions between coarse, massively parallel, full-field pre-attentive analysis systems and the more detailed, circumscribed and sequential attentional analysis system.

In what follows, we focus on, first, the brain areas involved in the control and deployment of attention; second, the mechanisms by which attention is attracted in a bottom-up or image-based manner towards salient locations in our visual environment; third, the mechanisms by which attention modulates the early sensory representation of attended stimuli; fourth the mechanisms for top-down or voluntary deployment of attention; and fifth the interaction between attention, object recognition and scene understanding.

BRAIN AREAS

To a first approximation, selecting where to attend next is carried out by distinct brain structures from recognizing what is being attended to. This suggests that a cooperation between "two visual systems" is used by normal vision (Didday & Arbib, 1975): Selecting where to attend next is primarily controlled by the DORSAL VISUAL PROCESSING STREAM (or "where/how" stream) which comprises cortical areas in posterior parietal cortex, whereas the VENTRAL VISUAL PROCESSING STREAM (or "what" stream), comprising cortical areas in inferotemporal cortex, is primarily concerned with localized object recognition (Ungerleider & Mishkin, 1982). It is important to note, however, that object recognition in the ventral stream can bias the next attentional shift by suggesting where the next interesting object may be located. Similarly, attention strongly modulates activity in the object recognition system.

Brain regions participating in the deployment of visual attention **(Figure 1)** include the lateral geniculate nucleus of the thalamus (LGN) and cortical areas V1 (primary visual cortex) through the parietal cortex along the dorsal stream as well as subcortical structures instrumental in producing directed eye movements. These include the deeper parts of the superior colliculus; parts of the pulvinar; the frontal eye fields; the precentral gyrus; and areas in the intraparietal sulcus in the macaque and around the intraparietal and postcentral sulci and adjacent gyri in humans.

BOTTOM-UP CONTROL

One important mode of operation of attention is largely unconscious and driven by the specific attributes of the stimuli present in our visual environment. This so-called BOTTOM-UP CONTROL OF VISUAL ATTENTION may be studied using simple VISUAL SEARCH tasks. Based on these experiments, computational models have been developed for how attention is attracted towards salient scene locations.

Visual Search and Pop-Out

One of the most effective demonstrations of bottom-up attentional guidance uses simple visual search experiments, in which an odd target stimulus to be located by the observer is embedded within an array of distracting visual stimuli (Treisman & Gelade, 1980). Originally, these experiments suggested a dichotomy between situations where the target stimulus would visually pop-out from the array and be found immediately, and situations where locating the target required extensive scanning of the array **(Figure 2)**. The pop-out cases suggest that the target can be effortlessly located by relying on preattentive visual processing over the entire visual scene. In contrast, the conjunctive search cases suggest that attending to the target is a necessary precondition to being able to identify it as being the unique target, thus requiring that the search array be scanned until the target becomes the object of attentional selection.

Further experimentation has revealed that the original dichotomy between fast, parallel pop-out and slower, serial search represent the two extremes of a continuum of search difficulty (Wolfe, 1996). Nevertheless, these experiments clearly demonstrate that if a target differs significantly from its surround (in ways which can be characterized in terms of visual attributes of the target and distractors), it will immediately draw attention towards itself. Thus, these experiments evidence how the composition of the visual scene alone is a strong component of attentional control.

Computational Models and the Saliency Map

The FEATURE INTEGRATION THEORY of Treisman and colleagues (Treisman & Gelade, 1980) that was derived from visual search experiments has served as a basis for many computational models of bottom-up attentional deployment. This theory proposed that only fairly simple visual features are computed in a massively parallel manner over the entire visual scene, in early visual processing areas including primary visual cortex. Attention is then necessary to bind those early features into more sophisticated object representations, and the selected bound representations are (to a first approximation) the only part of the visual world which pass though the ATTENTIONAL BOTTLENECK for further processing.

The first explicit neurally-plausible bottom-up computational architecture was proposed by Koch and Ullman (Koch & Ullman, 1985), and is closely related to the feature integration theory. The model is centered around a SALIENCY MAP, that is, an explicit two-dimensional topographic map that encodes for stimulus conspicuity, or SALIENCE, at every location in the visual scene. The saliency map receives inputs from early visual processing, and provides an efficient control strategy by which the focus of attention simply scans the saliency map in order of decreasing saliency.

This general architecture has been further developed and implemented, yielding the computational model depicted in **Figure 3** (Itti & Koch, 2001). In this model, the early stages of visual processing decompose the incoming visual input through an ensemble of feature-selective filtering processes endowed with contextual modulatory effects. To control a single attentional focus based on this multiplicity of features, all feature maps provide input to the saliency map, which encodes visual salience irrespectively feature dimensions. Biasing attention to focus onto the most salient location is then reduced to drawing attention towards the locus of highest activity in the saliency map. This is achieved using a WINNER-TAKE-ALL neural network, which implements a neurally distributed maximum detector. To allow attention to shift to the next most salient location, each attended location is transiently inhibited in the saliency map by an INHIBITION-OF-RETURN mechanism, such that the winner-take-all network naturally converges towards the next most salient location (Koch & Ullman, 1985; Itti & Koch, 2001).

Many successful models for the bottom-up control of attention are architectured around a saliency map. What differentiates the models, then, is the strategy employed to prune the incoming sensory input and extract salience. In an influential model mostly aimed at explaining visual search experiments, Wolfe (Wolfe, 1996) hypothesized that the selection of relevant features for a given search task could be performed top-down, through spatially-defined and feature-dependent weighting of the various feature maps. Although limited to cases where attributes of the target are known in advance, this view has recently received experimental support from studies of top-down attentional modulation (see below).

Tsotsos and colleagues (Tsotsos *et al.*, 1995) implemented attentional selection using a combination of a feedforward bottom-up feature extraction hierarchy and a feedback selective tuning of these feature extraction mechanisms. The target of attention is selected at the top level of the processing hierarchy (the equivalent of a saliency map), based on feedforward activation and on possible additional top-down biasing for certain locations or features. That location is then propagated back through the feature extraction hierarchy, through the activation of a cascade of winner-take-all networks embedded within the bottom-up processing pyramid. Spatial competition for salience is thus refined at each level of processing, as the feedforward paths not contributing to the winning location are pruned (resulting in the feedback propagation of an "inhibitory beam" around the selected target).

Itti *et al.* (Itti & Koch, 2001) recently proposed a purely bottom-up model, in which spatial competition for salience is directly modeled after non-classical surround modulation effects. The model employs an iterative scheme with early termination. At each iteration, a feature map receives additional inputs from the convolution of itself by a large difference-of-Gaussians filter. The result is half-wave rectified, with a net effect similar to a winner-take-all with limited inhibitory spread, which allows only a sparse population of locations to remain active. After competition, all feature maps are simply summed to yield the scalar saliency map. Because it includes a complete biological front-end, this model has been widely applied to the analysis of natural color scenes (Itti & Koch, 2001). The non-linear interactions implemented in this model strongly illustrate how, perceptually, whether a given stimulus is salient or not cannot be decided without knowledge of the context within which the stimulus is presented.

Many other models have been proposed, which typically share some of the components of the three models just described. It is thus important to note that postulating centralized control based on such map is not the only computational alternative. In particular, Desimone and Duncan (Desimone & Duncan, 1995) argued that salience is not explicitly represented by specific neurons, but instead is implicitly coded in a distributed modulatory manner across the various feature maps. Attentional selection is then performed based on top-down weighting of the bottom-up feature maps that are relevant to a target of interest. This top-down biasing (also used in Wolfe's Guided Search model (Wolfe, 1996)) requires that a specific search task be performed for the model to yield useful predictions.

TOP-DOWN MODULATION OF EARLY VISION

The general architecture for the bottom-up control of attention presented above opens two important questions on the nature of the attentional bottleneck. First, is it the only means through which incoming visual information may reach higher levels of processing? Second, does it only involve one-way processing of information from the bottom-up, or is attention a two-way process, also feeding back from higher centers to early processing stages?

Are we blind outside of the focus of attention?

Recent experiments have shown how fairly dramatic changes applied to a visual scene being inspected may go unnoticed by human observers, unless those changes occur at the location currently being attended to. These CHANGE BLINDNESS experiments (O'Regan *et al.*, 1999) can take several forms, yielding essentially the same conclusions. One implementation consists of alternatively flashing two versions of a same scene separated by a blank screen, with the two versions differing very obviously at one location (for example, a scene containing a jet airplane, and one of its reactors has been erased from one of the photographs). Although the alteration is obvious when directly attended to, naive observers typically require several tens of seconds to locate it. Not unexpectedly, the most difficulty instances of this experiment involve a change at a location of little interest in terms of scene understanding (for example, the aforementioned scene with an airplane also contains many people, who tend to be inspected in priority).

These experiments demonstrate the crucial role of attention in conscious vision: unless we attend to an object, we are unlikely to consciously perceive it in any detail and detect when it is altered. However, as we will see, this does necessarily mean that there is no vision other than through the attention bottleneck.

Attentional Modulation of Early Vision

A number of psychophysical end electrophysiological studies indicate that we are not entirely blind outside the focus of attention. At the early stages of processing, responses are still observed even if the animal is attending away from the receptive field at the site of recording, or is anesthetized. Behaviorally, we can also perform fairly specific spatial judgments on objects not being attended to, though those judgments are less accurate than in the presence of attention. This is in particular demonstrated by DUAL-TASK PSYCHOPHYSICAL EXPERIMENTS in which observers are able to simultaneously discriminate two visual stimuli presented at two distant locations in the visual field (Lee *et al.*, 1999).

While attention thus appears not to be mandatory for early vision, it has recently become clear that it can vigorously modulate, top-down, early visual processing, both in a spatially-defined and in a non-spatial but feature-specific manner (Treue & Martinez Trujillo, 1999). This modulatory effect of attention has been described as enhanced gain, biased or intensified competition, enhanced spatial resolution, or modulated background activity, effective stimulus strength or noise (Itti & Koch, 2001).

Of particular interest in a computational perspective, a recent study by Lee *et al.* (Lee *et al.*, 1999) measured psychophysical thresholds for five pattern discrimination tasks (contrast, orientation and spatial frequency discriminations, and two spatial masking tasks). They employed a dual-task paradigm to measure thresholds either with attention fully available to the task of interest, or poorly available because engaged elsewhere by a concurrent attention-demanding task. The mixed pattern of attentional modulation observed in the thresholds (up to 3-fold improvement with attention in orientation discrimination, but only 20% improvement in contrast discrimination) was quantitatively accounted for by a computational model. It predicted that attention strengthens a winner-take-all competition among neurons tuned to different orientations and spatial frequencies within one cortical hypercolumn (Lee *et al.*, 1999), a proposition which has recently received additional experimental support.

These results indicate that attention does not implement a feed-forward, bottom-up information processing bottleneck. Rather, attention also enhances, through feedback, early visual processing for both the location and visual features being attended to.

TOP-DOWN DEPLOYMENT

The precise mechanisms by which voluntary shifts of attention are elicited remain elusive, although several studies have narrowed down the brain areas primarily involved (see (Itti & Koch, 2001) for a review). Here we focus on two types of experiments that clearly demonstrate how, first, attention may be deployed on a purely voluntary basis onto one of several identical stimuli, and, second, how eye movements over a scene may dramatically differ dependent on task demands.

Attentional Facilitation and Cueing

Introspection easily reveals that we are able to voluntarily shift attention towards any location in our visual field, no matter how inconspicuous that location may be. More formally, psychophysical experiments may be used to demonstrate top-down shifts of attention. A typical experiment involves cueing an observer towards one of several possible identical stimuli, but only at a high cognitive level (e.g., verbal cue), so that nothing in the visual display distinguishes the target from distractors. Detection or discrimination of the stimulus at the cued (and presumably attended) location are significantly better (e.g., lower reaction time or lower psychophysical thresholds) than at uncued locations. These experiments hence suggest that voluntarily shifting attention towards a stimulus improves the perception of that stimulus.

Similarly, experiments involving decision uncertainty demonstrate that if a stimulus is to be discriminated by a specific attribute known in advance (e.g., discriminate the spatial frequency of a grating), performance is significantly improved compared to situations where one randomly chosen of several possible stimulus attributes are to be discriminated (e.g., discriminate the spatial frequency, contrast or orientation of a grating). Thus, we appear to also be able to voluntarily select the specific features of a stimulus. These results are closely related to and consistent with the spatial and featural nature of attentional modulation mentioned in the previous section.

Influence of Task

Recording eye movements from human observers while they inspect a visual scene has revealed a profound influence of task demands on eye movements (Yarbus, 1967). In a typical experiment, different observers examine a same photograph while their eye movements are being tracked, but are asked to answer different questions about the scene (for example, estimate the age of the people in the scene, or determine the country in which the photograph was taken). Although all observers are presented with the same visual stimulus, the patterns of eye movements recorded differ dramatically depending on the question being addressed by each observer.

Building in part on eye tracking experiments, Stark and colleagues (Noton & Stark, 1971) have proposed the SCANPATH THEORY of attention, according to which eye movements are generated almost exclusively under top-down control. The theory proposes that what we see is only remotely related to the patterns of activation of our retinas; rather, a cognitive model of what we expect to see is at the basis of our percepts. The sequence of eye movements which we make to analyze a scene, then, is mostly controlled top-down by our cognitive model, and serves the goal of obtaining specific details about the particular scene instance being observed, to embelish the more generic internal model. This theory has had a number of successful applications to robotics control, in which an internal model of a robot's working environment was used to restrict the analysis of incoming video to a small number of task-dependent circumscribed regions.

ATTENTION AND SCENE UNDERSTANDING

We have seen how attention is deployed onto our visual environment through a cooperation between bottom-up and top-down driving influences. One difficulty which then arises is the generation of proper top-down biasing signals when exploring a novel scene; indeed, if the scene has not been analyzed and understood yet using thorough attentional scanning, how can it be used to direct attention top-down? Below we explore two dimensions of this problem: First, we describe how already from a very brief presentation of a scene we are able to extract its gist, basic layout, and other characteristics. This suggests that another part of our visual system, operating much faster than attention, might be responsible for this coarse analysis. The results of this analysis may then be used to guide attention top-down. Second, we explore how several computer vision models have used a collaboration between the where and what subsystems to yield sophisticated scene recognition algorithms. Finally, we cast these results into a more global view of our visual system and the function of attention in vision.

Is scene understanding purely attentional?

Psychophysical experiments pioneered by Biederman and colleagues (Biederman, 1972) have demonstrated how we can derive coarse understanding of a visual scene from a single presentation that is so brief (80 ms or less) that it precludes any attentional scanning or eye movement. A particularly striking example of such experiments consists of presenting to an observer a rapid succession of unrelated photographs of natural scenes at a high frame rate (over 10 scenes/s). After presentation of the stimuli for several tens of seconds, observers are asked whether a particular scene, for example an outdoors market scene, was present among the several hundred photographs shown. Although the observers are not made aware in advance of the question, they are able to provide a correct answer with an overall performance well over chance (Biederman, personal communication). Furthermore, observers are able to recall a number of coarse details about the scene of interest, such as whether it contained human persons, or whether it was highly colorful or rather dull.

These and many related experiments clearly demonstrate that scene understanding does not exclusively rely on attentional analysis. Rather, a very fast visual subsystem which operates in parallel with attention allows us to rapidly derive the gist and coarse layout of a novel visual scene. This rapid subsystem certainly is one of the key components by which attention may be guided top-down towards specific visual locations.

Cooperation between Where and What

Several computer vision models have been proposed for extended object and scene analysis that rely on a cooperation between an attentional (where) and localized recognition (what) subsystems.

A very interesting instance was recently provided by Schill *et al.* (Schill *et al.*, in press). Their model aims at performing object recognition, using eye movements to focus on those parts of the scene that are most informative in disambiguating its identity. A hierarchical knowledge tree is trained into the model, in which leaves represent identified objects, intermediary nodes represent more general object classes, and links between nodes contain sensorimotor information used for discrimination between possible objects (i.e., bottom-up feature responses to be expected for particular points in the object, and eye movement vectors targeted at those points). During the iterative recognition of an object, the system programs its next fixation towards the location which will maximally increase information gain about the object being recognized, and thus will best allow the model to discriminate between the various candidate object classes.

Several related models have been proposed (see (Itti & Koch, 2001)), in which scanpaths (containing motor control directives stored in a "where" memory and locally expected bottom-up features stored in a "what" memory) are learned for each scene or object to be recognized. The difference between the various models comes from the algorithm used to match the sequences of where/what information to the visual scene. These include using a deterministic matching algorithm (i.e., focus next onto the next location stored in the sequence being tested against the new scene), hidden Markov models (where sequences are stored as transition probabilities between locations augmented by the visual features expected at those locations), or evidential reasoning (similar to the model of Schill and colleagues). These models typically demonstrate strong ability to recognize complex grayscale scenes and faces, in a translation, rotation and scale independent manner, but cannot account for non-linear image transformations (e.g., three-dimensional viewpoint change).

Attention as a component of vision

We have seen how vision relies not only on the attentional subsystem, but more broadly on a cooperation between crude preattentive subsystems for the computation of gist, layout and for bottom-up attentional control, and the attentive subsystem coupled with the localized object recognition subsystem to obtain fine details at various locations in the scene.

This view on the visual system raises a number of questions which remain fairly controversial. These are issues of the internal representation of scenes and objects (e.g., view-based versus three-dimensional models, or a cooperation between both), and of the level of detail with which scenes are stored in memory for later recall and comparison to new scenes (e.g., snapshots versus crude structural models). Many of these issues extend well beyond the scope of the present discussion of selective visual attention. Nevertheless, it is important to think of attention within the broader framework of vision and scene understanding, as this allows us to delegate some of the visual functions to non-attentional subsystems.

DISCUSSION

We have reviewed some of the key aspects of selective visual attention, and how these contribute more broadly to our visual experience and unique ability to rapidly comprehend complex visual scenes.

Looking at the evidence on the brain areas involved with the control of attention has revealed a complex interconnected network, shared with other subsystems, including the guidance of eye movements, the computation of early visual features, the recognition of objects and the planning of actions.

Attention is guided towards particular locations in our visual world under a combination of competing constraints, including bottom-up from the visual input and top-down from task priorities and scene understanding. Bottom-up control is evidenced by visual search experiments, in which attention is automatically drawn towards targets that pop-out from surrounding distractors. It is the best understood component of attention, and computational models exist which replicate some of the human search performance. Most models have embraced the idea that a single topographic saliency map efficiently guide attention. An important theoretical result is the critical role of cortical interactions in pruning the massive sensory input, to isolate conspicuous elements from the scene.

Attention implements an information processing bottleneck, which allows only select elements in the scene to reach higher levels of processing. But not all vision is attentional, and even though we may easily appear blind to unattended image details, there is still substantial residual vision outside the focus of attention. That is, the attentional bottleneck is not strict. In addition, it is a two-way process, such that not only are selected stimuli propagated up the visual hierarchy, but are also enhanced through top-down feedback. Computationally, a proposed unifying mechanism is the activation by attention of a winner-take-all competition among visual neurons representing different aspects of stimulus, making its dominant characteristics more explicit. Top-down attentional modulation appears both location- and feature-specific.

Introspection evidences how attention is not exclusively controlled bottom-up. Improved performance when subjects know in advance where or what to look for provides further support for top-down attentional guidance. The exact mechanisms by which volitional attention shifts remain rather elusive. Nevertheless, high-level task specifications, e.g., a question asked about a visual scene, dramatically influence the deployment of attention and gaze.

Finally, it is important to consider attention not as an autonomous visual subsystem, concerned only with stimulus selection. Indeed, it is highly unlikely, or impossible under consitions of very brief presentation, that we analyze complex scenes only through attention. Rather, attention, object recognition, and rapid evaluation of scene gist and layout, cooperate in a remarkable multi-threaded analysis that exploits multiple time scales and levels of details within interacting processing streams. Despite tremendous recent progress, many of the key components of this complex interacting system remain poorly understood and elusive, thus posing ever renewed challenges for future neuroscience research.

References

Biederman, I. 1972. Perceiving real-world scenes. *Science*, **177**(43), 77–80.

Desimone, R, & Duncan, J. 1995. Neural mechanisms of selective visual attention. *Annu Rev Neurosci*, **18**, 193–222.

Didday, R L, & Arbib, M A. 1975. Eye Movements and Visual Perception: A "Two Visual System" Model. *Int J Man-Machine Studies*, **7**, 547–569.

Itti, L., & Koch, C. 2001. Computational Modeling of Visual Attention. *Nature Reviews Neuroscience*, **2**(3), 194–203.

Koch, C, & Ullman, S. 1985. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, **4**(4), 219–27.

Lee, D. K., Koch, C., & Braun, J. 1999. Attentional capacity is undifferentiated: concurrent discrimination of form, color, and motion. *Percept Psychophys*, **61**(7), 1241–1255.

Noton, D, & Stark, L. 1971. Scanpaths in eye movements during pattern perception. *Science*, **171**(968), 308–11.

O'Regan, J K, Rensink, R A, & Clark, J J. 1999. Change-blindness as a result of 'mudsplashes'. *Nature*, **398**(6722), 34.

Schill, K, Umkehrer, E, Beinlich, S, Krieger, G, & Zetzsche, C. in press. Scene analysis with saccadic eye movements: top-down and bottom-up modeling. *J Electronic Imaging*.

Treisman, A M, & Gelade, G. 1980. A feature-integration theory of attention. *Cognit Psychol*, **12**(1), 97–136.

Treue, S, & Martinez Trujillo, J C. 1999. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, **399**(6736), 575–579.

Tsotsos, J K, Culhane, S M, Wai, W Y K, Lai, Y H, Davis, N, & Nuflo, F. 1995. Modeling Visual-Attention via Selective Tuning. *Artificial Intelligence*, **78**(1-2), 507–45.

Ungerleider, L G, & Mishkin, M. 1982. Two cortical visual systems. *Pages 549–586 of:* Ingle, D G, Goodale, M A A, & Mansfield, R J W (eds), *Analysis of visual behavior*. Cambridge, MA: MIT Press.

Wolfe, J. 1996. Visual search: a review. *In:* Pashler, H (ed), *Attention*. London, UK: University College London Press.

Yarbus, A. 1967. *Eye Movements and Vision.* New York: Plenum Press.

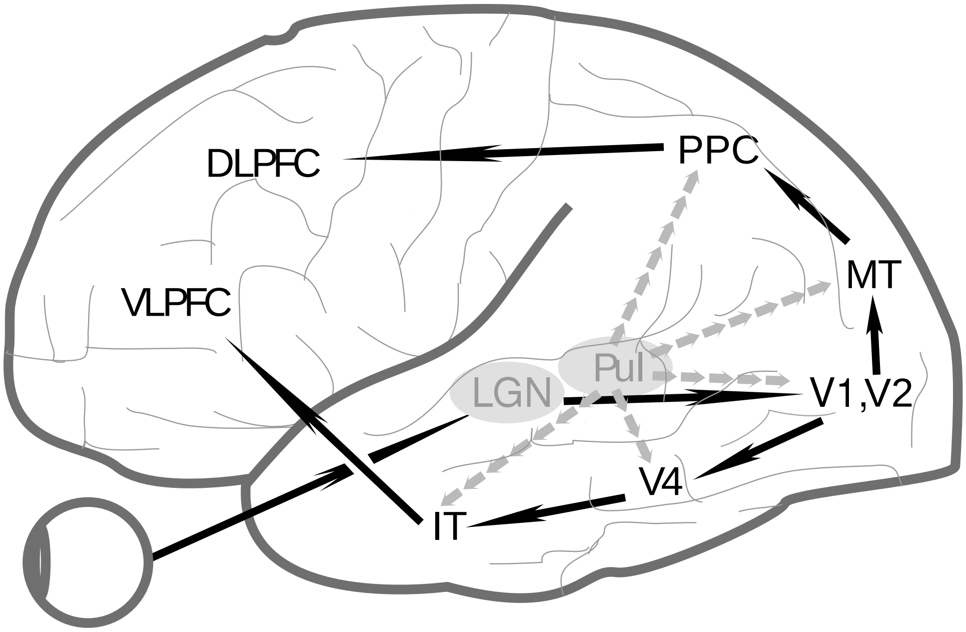

Figure 1: Major brain areas involved in the deployment of selective visual attention. Although single-ended arrows are shown to suggest global information flow (from the eyes to prefrontal cortex), anatomical studies suggest reciprocal connections, with the number of feedback fibers often exceeding that of feedforward fibers (except between retina and LGN). Cortical areas may be grouped into two main visual pathways: the dorsal "where/how" pathway (from V1 to DLPFC via PPC) is mostly concerned with spatial deployment of attention and localization of attended stimuli, while the ventral "what" pathway (from V1 to VLPFC via IT) is mostly concerned with pattern recognition and identification of the attended stimuli. In addition to these cortical areas, several subcortical areas including LGN and Pul play important roles in controlling where attention is to be deployed. **Key to abbreviations:** LGN: lateral geniculate nucleus; Pul: Pulvinar nucleus; V1, V2, V4: early cortical visual areas; MT: Medial temporal area; PPC: posterior parietal cortex; DLPFC: dorsolateral prefrontal cortex; IT: inferotemporal cortex; VLPFC: ventrolateral prefrontal cortex.

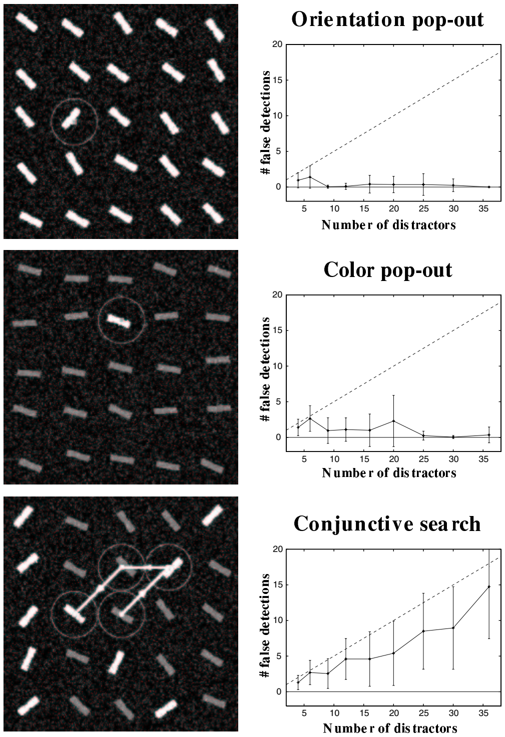

Figure 2: Search array experiments of the type pioneered by Treisman and colleagues. The top two panels are examples of pop-out cases where search time (here shown as the number of locations fixated before the target if found) is small and independent of the number of elements in the display. The bottom panel demonstrates a conjunctive search (the target is the only element that is bright *and* oriented like the darker elements); in this case, a serial search is initiated, which will require more time as the number of elements in the display is increased.

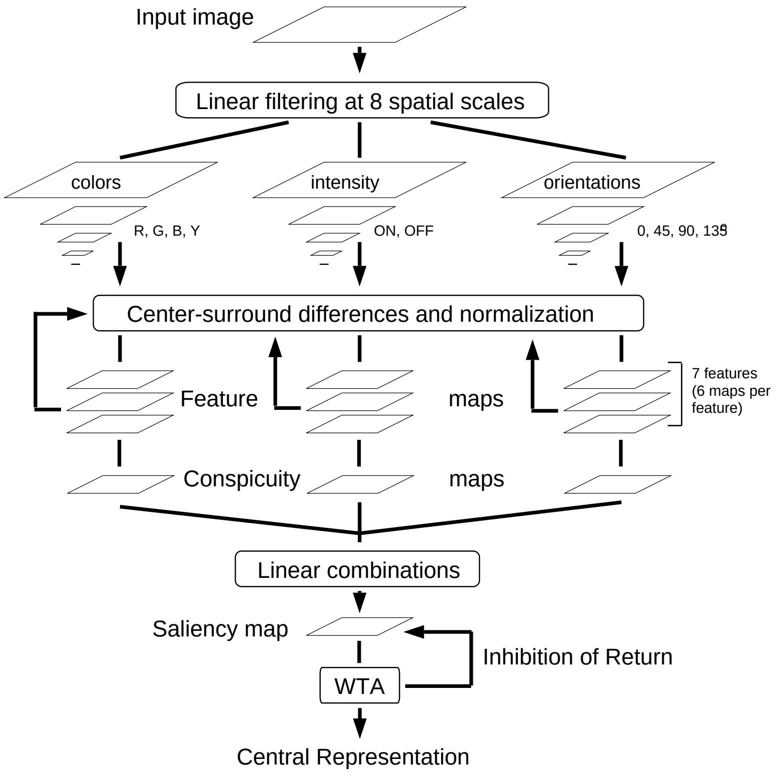

Figure 3: Typical architecture for a model of bottom-up visual attention based on a saliency map. The input image is analyzed by a number of early visual filters, sensitive to stimulus properties such as color, intensity and orientation, at several spatial scales. After spatial competition for salience within each of the resulting feature maps, input is provided to a single saliency map from all of the feature maps. The maximum activity in the saliency map (detected by a winner-take-all network; WTA) is the next attended location. Transient inhibition-of-return at this location in the saliency map allows the system to shift towards the next most salient location.