Chapter 1

Modeling Primate Visual Attention

1.1 Introduction

Selective visual attention is the mechanism by which we can rapidly direct our gaze towards objects of interest in our visual environment [26, 52, 3, 51, 34, 6, 18, 5]. From an evolutionary viewpoint, this rapid orienting capability is critical in allowing

living systems to quickly become aware of possible preys, mates or predators in their cluttered visual world. It has become clear that attention guides where to look next based on both bottom-up (image-based) and top-down (task-dependent) cues [26]. As such, attention implements an information processing bottleneck, only allowing a small part of the incoming sensory information to reach short-term memory and visual awareness [15, 12]. That is, instead of attempting to fully process the massive sensory input in parallel, nature has devised a serial strategy to achieve near real-time performance despite limited computational capacity: Attention allows us to break down the problem of scene understanding into rapid series of computationally less demanding, localized visual analysis problems.

These orienting and scene analysis functions of attention are complemented by a feedback modulation of neural activity at the location and for the visual attributes of the desired or selected targets. This feedback is believed to be essential for binding the different visual attributes of an object, such as color and form, into a unitary percept [52, 22, 41]. That is, attention not only serves to select a location of interest, but also enhances the cortical representation at that location. As such, focal visual attention is often compared to a rapidly shiftable spotlight [57, 7], which scans our visual environment both overtly (with accompanying eye movements) or covertly (with the eyes fixed).

Finally, attention is involved in triggering behavior, and consequently is intimately related to recognition, planning and motor control [31]. Of course, not all of vision is attentional, as we can derive coarse understanding from presentations of visual scenes that are too brief for attention to explore the scene. Vision thus relies 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 several critical aspects of selective visual attention: First, the brain area involved in its control and deployment; second, the mechanisms by which attention is attracted in a bottom-up or image-based manner towards conspicuous or 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.

1.2 Brain Areas

The control of focal visual attention involves an intricate network of brain areas, spanning from primary visual cortex to prefrontal cortex. In a first approximation, selecting where to attend next is carried out, to a large extent, 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 [16]: 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 [56]. It is important to note, however, that object recognition in the ventral stream can bias the next attentional shift, for example via top-down control when an object is recognized that suggests where the next interesting object may be located. Similarly, we will see how attention strongly modulates activity in the object recognition system.

Among the brain regions participating to the deployment of visual attention include most of the early visual processing areas and the dorsal processing stream **(Figure 1)**. These include the lateral geniculate nucleus of the thalamus (LGN) and cortical areas V1 (primary visual cortex) through the parietal cortex along the dorsal stream [50]. In addition, overt and covert attention have been shown to be closely related, as revealed by psychophysical [47, 46, 19, 28], physiological [1, 29, 9, 44], and imaging [10, 36] studies. Directing covert attention thus involves a number of subcortical structures that are also instrumental in producing directed eye movements. These include the deeper parts of the superior colliculus; parts of the pulvinar; the frontal eye fields in the macaque and its homologue in humans; the precentral gyrus; and areas in the intraparietal sulcus in the macaque and around the intraparietal and postcentral sulci and adjacent gyri in humans.

1.3 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 can easily be studied using simple visual search tasks as described below. Based on these experimental results, several computational theories and models have been developed for how attention may be attracted towards a particular object in the scene rather than another.

1.3.1 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 [52]. 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 extensive scanning and inspection of the various stimuli in the display was necessary before the target stimulus could be located **(Figure 2)**. The pop-out cases suggest that the target can be effortlessly located by relying on preattentive vi-

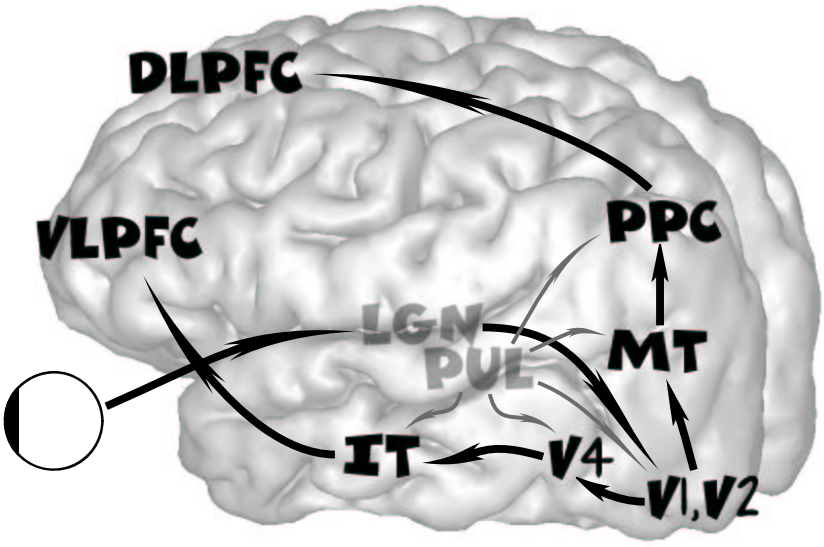

Figure 1

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.

sual 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 extensively scanned until the target becomes the object of attentional selection.

Further experimentation has revealed that the original dichotomy between the fast, parallel search observed with pop-out displays and slower, serial search observed with conjunctive displays represent the two extremes of a continuum of search difficulty [58]. 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 potentially very strong component of attentional control, guiding attention from the bottom of the visual processing hierarchy up.

1.3.2 Computational Models and the Saliency Map

The feature integration theory of Treisman and colleagues [52] 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 incoming visual scene, in early visual processing areas including primary visual cortex. Attention is then necessary to bind those early features into a more sophisticated object representation, and the selected bound representation is (to a first approximation) the only part of the visual world which passes though the attentional bottleneck for further processing.

The first explicit neurally-plausible computational architecture of a system for the bottom-up guidance of attention was proposed by Koch and Ullman [27], and is closely related to the feature integration theory. Their 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** [24]. 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. In order to control a single attentional focus based on this multiplicity in the representation of the incoming sensory signals, it is assumed that all feature maps provide input to the saliency map, which topographically represents visual salience, irrespectively of the feature dimension by which a given location was salient. 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. In order to prevent attention from permanently focusing onto the most active (winner) location in the saliency map, the currently attended location is transiently inhibited in the saliency map by an inhibition-of-return mechanism. After the most salient location is thus suppressed, the winner-take-all network naturally converges towards the next most salient location, and repeating this process generates attentional scanpaths [27, 24].

Many successful models for the bottom-up control of attention are architectured

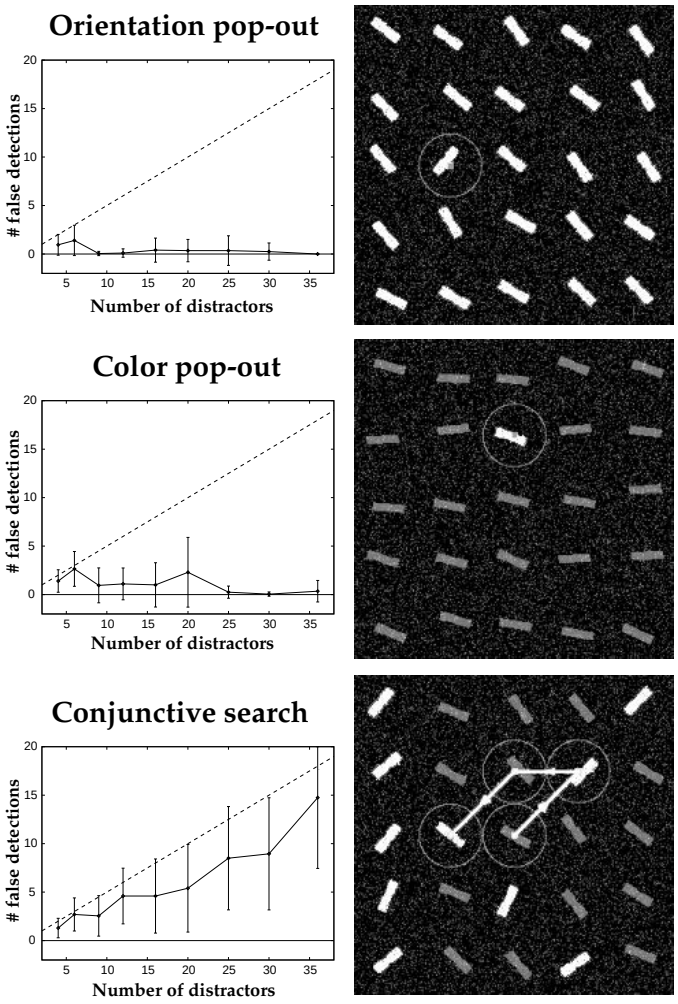

Figure 2

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 dark *and* oriented like the brighter elements); in this case, a serial search is initiated, which will require more time as the number of elements in the display is increased.

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

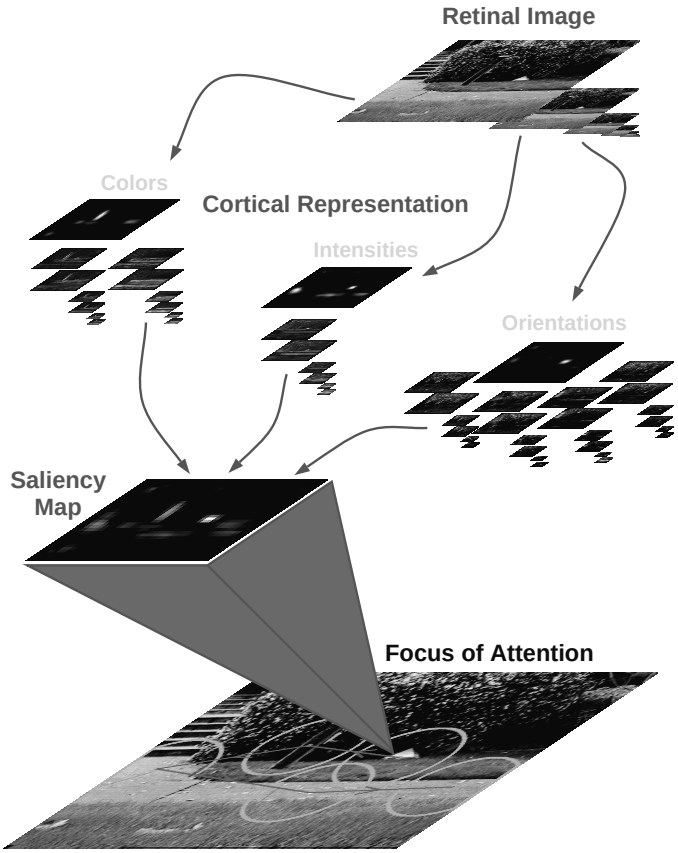

Figure 3

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 aximum activity in the saliency map is the next attended location. Transient inhibition of this location in the saliency map allows the system to shift towards the next most salient location.

model mostly aimed at explaining visual search experiments, Wolfe [59] 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 [55] implemented attentional selection using a combina-

tion of a feedforward bottom-up feature extraction hierarchy and a feedback selective tuning of these feature extraction mechanisms. In this model, 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.* [25, 23, 24] 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 [23]. 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. In view of the affluence of models based on a saliency map, it is important to note that postulating centralized control based on such map is not the only computational alternative for the bottom-up guidance of attention. In particular, Desimone and Duncan [15] 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 [59]) requires that a specific search task be performed for the model to yield useful predictions.

1.4 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?

1.4.1 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 [38, 40] 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 in which a jet airplane is present and one of its reactors has been erased from one of the two photographs to be compared). Although the alteration is obvious when one directly attends to it, it takes naive observers several tens of seconds to locate it. Not unexpectedly, instances of this experiment which are the most difficult for observers involve a change at a location that is of little interest in terms of understanding and interpreting the scene (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 below, this does necessarily mean that there is no vision other than through the attention bottleneck.

1.4.2 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 [54], or is anesthetized [21]. Behaviorally, we can also perform fairly specific spatial judgments on objects not being attended to [6, 14], though those judgments are less accurate than in the presence of attention [30, 61]. 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 [30].

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 [33, 53, 2]. This modulatory effect of attention has been described as enhanced gain [54], biased [32, 35] or intensified [30] competition, enhanced spatial resolution [61], or as modulated background activity [8], effective stimulus strength [42] or noise [17].

Of particular interest in a computational perspective, a recent study by Lee *et*

al. [30] measured psychophysical thresholds for five simple pattern discrimination tasks (contrast, orientation and spatial frequency discriminations, and two spatial masking tasks; 32 thresholds in total). They employed a dual-task paradigm to measure thresholds either when attention was fully available to the task of interest, or when it was 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 in orientation discrimination with attention, 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 [30], 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.

1.5 Top-Down Deployment of Attention

The precise mechanisms by which voluntary shifts of attention are elicited remain elusive, although several studies have narrowed down the brain areas primarily involved [20, 11, 24]. 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 (so that none of the stimuli is more salient than the others), and, second, how eye movements recorded from observers inspecting a visual scene with the goal of answering a question about that scene are dramatically influenced by the question being answered.

1.5.1 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 presented on a computer screen. The cue indicates to the observer where to focus on, but only at a high cognitive level (e.g., verbal cue), so that nothing in the display would directly attract attention bottom-up towards the desired stimulus. Detection or discrimination of the stimulus at the 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 that is 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 not only where to attend to, but also the specific features of a stimulus to be attended. These results are closely related to and consistent with the spatial and featural nature of attentional modulation mentioned in the previous section.

1.5.2 Influence of Task

Recording eye movements from human observers while they inspect a visual scene has revealed a profound influence of task demands on the pattern of eye movements generated by the observers [60]. 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 an identical visual stimulus, the patterns of eye movements recorded differ dramatically depending on the question being addressed by each observer. These experiments clearly demonstrate that task demands play a critical role in determining where attention is to be focused next.

Building in part on eye tracking experiments, Stark and colleagues [37] 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 percept. 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 sequences to a small number of circumscribed regions important for a given task.

1.6 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 show how already from a very brief presentation of a scene we are able to extract its gist, basic layout, and a number of other characteristics. This suggests that another part of our visual system, which operates 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.

1.6.1 Is scene understanding purely attentional?

Psychophysical experiments pioneered by Biederman and colleagues [4] 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.

1.6.2 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.* [45]. Their

model aims at performing scene (or object) recognition, using attention (or eye movements) to focus on those parts of the scene being analyzed which are most informative in disambiguating its identity. To this end, 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 [43, 48, 49, 13, 24], 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).

While these models provide very interesting examples of cooperation between a fast attentional cueing system and a slower localized feature analysis system, their relationship to biology has not been emphasized beyond the general architectural level. Teasing apart the brain mechanisms by which attention, localized recognition, and rapid computation of scene gist and layout collaborate in normal vision remains one of the most exciting challenges for modern visual neuroscience [39].

1.6.3 Attention as a component of vision

In this section, 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 **(Figure 4)**.

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

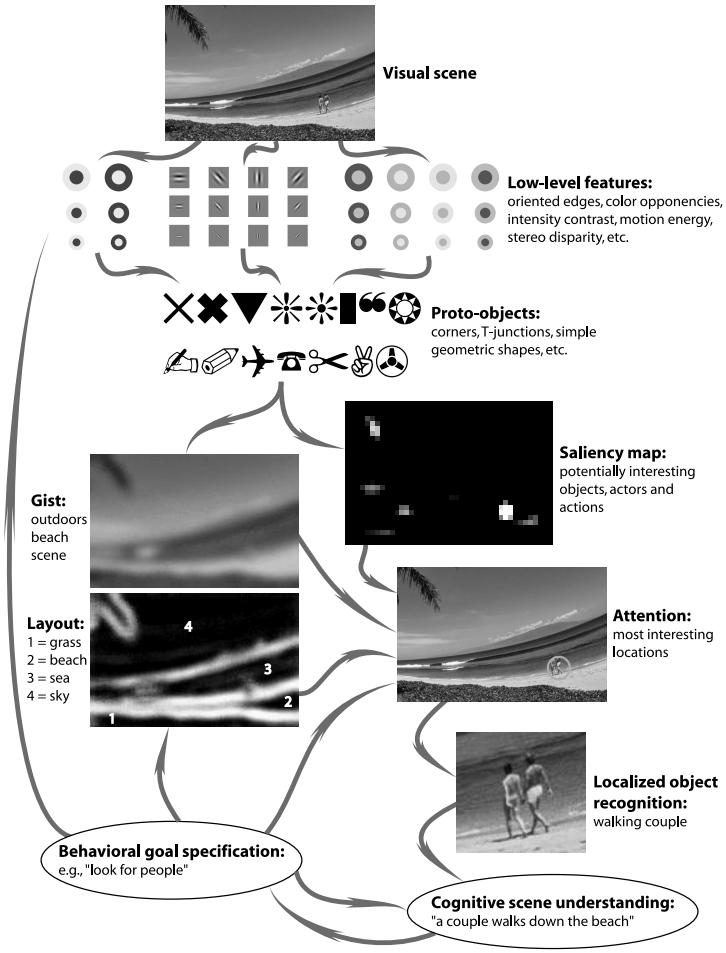

Figure 4

Simplified architecture for the understanding of visual scenes, extended from Rensink's (2000) triadic model. The incoming visual scene is analyzed by low-level visual processes (top) in a massively-parallel, full-field and pre-attentive manner up to a fairly simple "proto-object" representation. Building on this representation, gist and layout of the scene are computed in a fast, probably feedforward and non-iterative manner (left). Also building on this representation, the saliency map describes potentially interesting locations in the scene (right). Guided by saliency, gist, layout, and behavioral goal specifications, focal attention selects a region of the scene to be analyzed in further details. The result of this localized object recognition is used to incrementally refine the cognitive understanding of the contents of the scene. This understanding as well as the goal specification bias the low-level vision through feedback pathways.

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.

1.7 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 accumulated so far on the brain areas involved with the control of attention has revealed a complex interconnected network, which spans from the earliest stages of visual processing up to prefrontal cortical areas. To a large extent, this network serves not only the function of guiding attention, but is 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, which include bottom-up signals derived from the visual input, and top-down contraints derived from task priority and scene understanding. The bottom-up control of attention is clearly evidenced by simple visual search experiments, in which our attention is automatically drawn towards targets that pop-out from surrounding distractors. This bottom-up guidance is certainly the best understood component of attention, and many computational models have been proposed which replicate some of the human performance at exploring visual search stimuli. Most models have embraced the idea that a single topographic saliency map may be an efficient centralized representation for guiding attention. Several of these models have been applied to photographs of natural scenes, yielding remarkably plausible results. One of the important theoretical results derived from bottom-up modeling is the critical role of cortical interactions in pruning the massive sensory input such as to extract only those elements of the scene that are conspicuous.

In part guided by bottom-up cues, attention thus 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 image details outside the focus of attention, there is still substantial residual vision of unattended objects. That is, the attentional bottleneck is not strict, and some elements in the visual scene may reach our conscious perception if they are sufficiently salient, even though attention might be engaged elsewhere in the visual environment.

In addition, attentional selection appears to be a two-way process, in which not only selected scene elements are propagated up the visual hiererchy, but the representation of these elements is also enhanced down to the earliest levels of the hierarchy

through feedback signals. Thus attention not only serves the function of selecting a subset of the current scene, but also profoundly alters the cortical representation of this subset. Computationally, one mechanism for this enhancement which enjoys broad validity across a variety of visual discrimination tasks is that attention may activate a winner-take-all competition among visual neurons representing different aspects of a same visual location, thus making more explicit what the dominant characteristic of that location is. Top-down attentional modulation can be triggered not only on the basis of location, but also towards specific visual features.

Introspection easily makes evident that attention is not exclusively controlled bottom-up. Indeed, we can with little effort focus attention onto any region of our visual field, no matter how inconspicuous. Volitional shifts of attention are further evidenced by psychophysical experiments in which improved performance is observed when subjects know in advance where or what to look for, and hence presumably use a volitional shift of attention (across space or feature dimensions) in preparation for performing a visual judgement. The exact mechanisms by which volitional attention shifts are elicited remain rather elusive, but it has been widely demonstrated that high-level task specifications, such as a question asked about a visual scene, have dramatic effects on the deployment of attention and eye movements onto the scene.

Finally, it is important to consider attention not as a visual subsystem of its own that would have little interaction with other aspects of vision. Indeed, we have seen that it is highly unlikely, or impossible under consitions of very brief presentation, that we analyze and understand complex scenes only through attentional scanning. Rather, attention, object recognition, and rapid machanisms for the extraction of scene gist and layout cooperate in a remarkable multi-threaded analysis which exploits different time scales and levels of details within interacting processing streams. Although tremendous progress has been made over the past century of the scientific study of attention, starting with William James, 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

[1] R A Andersen, R M Bracewell, S Barash, J W Gnadt, and L Fogassi. Eye position effects on visual, memory, and saccade-related activity in areas lip and 7a of macaque. *J Neurosci*, 10(4):1176–96, Apr 1990.

[2] F Barcelo, S Suwazono, and R T Knight. Prefrontal modulation of visual processing in humans. *Nat Neurosci*, 3(4):399–403, Apr 2000.

[3] J Bergen and B Julesz. Parallel versus serial processing in rapid pattern discrimination. *Nature (London)*, 303:696–698, 1983.

[4] I. Biederman. Perceiving real-world scenes. *Science*, 177(43):77–80, Jul 1972.

[5] J Braun and B Julesz. Withdrawing attention at little or no cost: detection and discrimination tasks. *Percept Psychophys*, 60(1):1–23, Jan 1998.

[6] J Braun and D Sagi. Vision outside the focus of attention. *Percept Psychophys*, 48(1):45–58, Jul 1990.

[7] J A Brefczynski and E A DeYoe. A physiological correlate of the 'spotlight' of visual attention. *Nat Neurosci*, 2(4):370–374, Apr 1999.

[8] D Chawla, G Rees, and K J Friston. The physiological basis of attentional modulation in extrastriate visual areas. *Nat Neurosci*, 2(7):671–676, Jul 1999.

[9] C L Colby and M E Goldberg. Space and attention in parietal cortex. *Annu Rev Neurosci*, 22:319–49, 1999.

[10] M Corbetta. Frontoparietal cortical networks for directing attention and the eye to visual locations: identical, independent, or overlapping neural systems? *Proc Natl Acad Sci U S A*, 95(3):831–8, Feb 1998.

[11] M Corbetta, J M Kincade, J M Ollinger, M P McAvoy, and G L Shulman. Voluntary orienting is dissociated from target detection in human posterior parietal cortex [published erratum appears in nat neurosci 2000 may;3(5):521]. *Nat Neurosci*, 3(3):292–297, Mar 2000.

[12] F Crick and C Koch. Constraints on cortical and thalamic projections: the no-strong-loops hypothesis. *Nature*, 391(6664):245–50, Jan 1998.

[13] G Deco and J Zihl. A neurodynamical model of visual attention: Feedback enhancement of spatial resolution in a hierarchical system. *Journal of Computational Neuroscience*, page in press, 2001.

[14] B DeSchepper and A Treisman. Visual memory for novel shapes: implicit coding without attention. *J Exp Psychol Learn Mem Cogn*, 22(1):27–47, Jan 1996.

[15] R Desimone and J Duncan. Neural mechanisms of selective visual attention. *Annu Rev Neurosci*, 18:193–222, 1995.

[16] R L Didday and M A Arbib. Eye movements and visual perception: A "two visual system" model. *Int J Man-Machine Studies*, 7:547–569, 1975.

[17] B A Dosher and Z L Lu. Mechanisms of perceptual attention in precuing of location. *Vision Res*, 40(10-12):1269–1292, 2000.

[18] O Hikosaka, S Miyauchi, and S Shimojo. Orienting a spatial attention-its reflexive, compensatory, and voluntary mechanisms. *Brain Res Cogn Brain Res*, 5(1-2):1–9, Dec 1996.

[19] J E Hoffman and B Subramaniam. The role of visual attention in saccadic eye movements. *Percept Psychophys*, 57(6):787–795, Aug 1995.

[20] J B Hopfinger, M H Buonocore, and G R Mangun. The neural mechanisms of top-down attentional control. *Nat Neurosci*, 3(3):284–291, Mar 2000.

[21] D H Hubel and T N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol (London)*, 160:106–54, 1962.

[22] J E Hummel and I Biederman. Dynamic binding in a neural network for shape recognition. *Psychol Rev*, 99(3):480–517, Jul 1992.

[23] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506, May 2000.

[24] L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, Mar 2001.

[25] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, Nov 1998.

[26] W James. *The Principles of Psychology*. Harvard University Press, Cambridge, MA, 1890/1981.

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

[28] E Kowler, E Anderson, B Dosher, and E Blaser. The role of attention in the programming of saccades. *Vision Res*, 35(13):1897–916, Jul 1995.

[29] A A Kustov and D L Robinson. Shared neural control of attentional shifts and eye movements. *Nature*, 384(6604):74–7, Nov 1996.

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

[31] E K Miller. The prefrontal cortex and cognitive control. *Nat Reviews Neurosci*, 1(1):59–65, 2000.

[32] J Moran and R Desimone. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229(4715):782–4, Aug 1985.

[33] B C Motter. Neural correlates of attentive selection for color or luminance in extrastriate area v4. *J Neurosci*, 14(4):2178–89, Apr 1994.

[34] K Nakayama and M Mackeben. Sustained and transient components of focal visual attention. *Vision Research*, 29:1631–1647, 1989.

[35] E Niebur, C Koch, and C Rosin. An oscillation-based model for the neuronal basis of attention. *Vision Res*, 33(18):2789–802, Dec 1993.

[36] A C Nobre, D R Gitelman, E C Dias, and M M Mesulam. Covert visual spatial orienting and saccades: overlapping neural systems. *Neuroimage*, 11(3):210–216, Mar 2000.

[37] D Noton and L Stark. Scanpaths in eye movements during pattern perception. *Science*, 171(968):308–11, Jan 1971.

[38] J K O'Regan, R A Rensink, and J J Clark. Change-blindness as a result of 'mudsplashes'. *Nature*, 398(6722):34, Mar 1999.

[39] R. A. Rensink. The dynamic representation of scenes. *Vis Cogn*, 7:17–42, 2000.

[40] R. A. Rensink. Change detection. *Annu Rev Psychol*, 53:245–277, 2002.

[41] J H Reynolds and R Desimone. The role of neural mechanisms of attention in solving the binding problem. *Neuron*, 24(1):19–29, 111–25, Sep 1999.

[42] J H Reynolds, T Pasternak, and R Desimone. Attention increases sensitivity of v4 neurons [see comments]. *Neuron*, 26(3):703–714, Jun 2000. Comment in: Neuron 2000 Jun;26(3):548-50.

[43] I A Rybak, V I Gusakova, A V Golovan, L N Podladchikova, and N A Shevtsova. A model of attention-guided visual perception and recognition. *Vision Res*, 38(15-16):2387–2400, Aug 1998.

[44] J D Schall, D P Hanes, and T L Taylor. Neural control of behavior: countermanding eye movements. *Psychol Res*, 63(3-4):299–307, 2000.

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

[46] B M Sheliga, L Riggio, and G Rizzolatti. Orienting of attention and eye movements. *Exp Brain Res*, 98(3):507–22, 1994.

[47] M Shepherd, J M Findlay, and R J Hockey. The relationship between eye movements and spatial attention. *Q J Exp Psychol*, 38:475–491, 1986.

[48] L W Stark and Y S Choi. Experimental methaphysics: The scanpath as an epistemological mechanism. In W H Zangemeister, H S Stiehl, and C Freska, editors, *Visual Attention and Cognition*, pages 3–69. Elsevier Science B.V., 1996.

[49] L W Stark, C M Privitera, H Yang, M Azzariti, Y F Ho, T Blackmon, and D Chernyak. Representation of human vision in the brain: how does human perception recognize images? *Journal of Electronic Imaging*, 10(1), 2001.

[50] K Suder and F Worgotter. The control of low-level information flow in the visual system. *Rev Neurosci*, 11(2-3):127–146, 2000.

[51] A Treisman. Features and objects: the fourteenth bartlett memorial lecture. *Q J Exp Psychol [A]*, 40(2):201–37, May 1988.

[52] A M Treisman and G Gelade. A feature-integration theory of attention. *Cognit Psychol*, 12(1):97–136, Jan 1980.

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

[54] S Treue and J H Maunsell. Attentional modulation of visual motion processing in cortical areas mt and mst. *Nature*, 382(6591):539–41, Aug 1996.

[55] J K Tsotsos, S M Culhane, W Y K Wai, Y H Lai, N Davis, and F Nuflo. Modeling visual-attention via selective tuning. *Artificial Intelligence*, 78(1-2):507–45, 1995.

[56] L G Ungerleider and M Mishkin. Two cortical visual systems. In D G Ingle, M A A Goodale, and R J W Mansfield, editors, *Analysis of visual behavior*, pages 549–586. MIT Press, Cambridge, MA, 1982.

[57] E Weichselgartner and G Sperling. Dynamics of automatic and controlled visual attention. *Science*, 238(4828):778–780, Nov 1987.

[58] J Wolfe. Visual search: a review. In H Pashler, editor, *Attention*. London, UK: University College London Press, 1996.

[59] J M Wolfe. Visual search in continuous, naturalistic stimuli. *Vision Res*, 34(9):1187–95, May 1994.

[60] A Yarbus. *Eye Movements and Vision*. Plenum Press, New York, 1967.

[61] Y Yeshurun and M Carrasco. Attention improves or impairs visual performance by enhancing spatial resolution. *Nature*, 396(6706):72–75, Nov 1998.