COMPUTATIONAL MODELLING OF VISUAL ATTENTION

Laurent Itti and Christof Koch‡*

Five important trends have emerged from recent work on computational models of focal visual attention that emphasize the bottom-up, image-based control of attentional deployment. First, the perceptual saliency of stimuli critically depends on the surrounding context. Second, a unique 'saliency map' that topographically encodes for stimulus conspicuity over the visual scene has proved to be an efficient and plausible bottom-up control strategy. Third, inhibition of return, the process by which the currently attended location is prevented from being attended again, is a crucial element of attentional deployment. Fourth, attention and eye movements tightly interplay, posing computational challenges with respect to the coordinate system used to control attention. And last, scene understanding and object recognition strongly constrain the selection of attended locations. Insights from these five key areas provide a framework for a computational and neurobiological understanding of visual attention.

CENTRE–SURROUND MECHANISMS
These involve neurons that respond to image differences between a small central region and a broader concentric antagonistic surround region.

**Hedco Neuroscience Building, University of Southern California, 3641 Watt Way, Los Angeles, California 90089-2520, USA. ‡Computation and Neural Systems Program, Caltech, Pasadena, California 91125, USA. Correspondence to L.I. e-mail: itti@usc.edu*

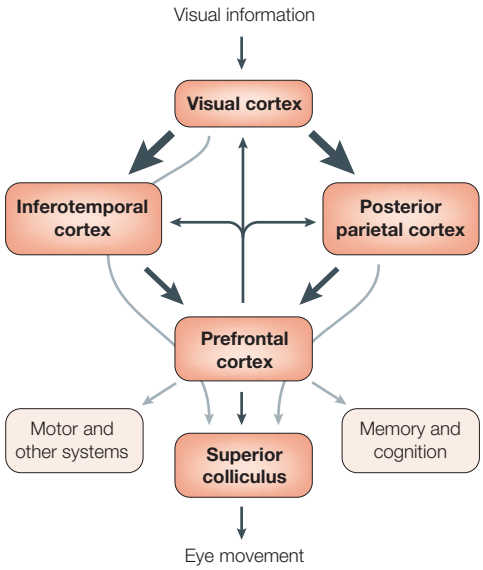
The most important function of selective visual attention is to direct our gaze rapidly towards objects of interest in our visual environment[1–9]. This ability to orientate rapidly towards salient objects in a cluttered visual scene has evolutionary significance because it allows the organism to detect quickly possible prey, mates or predators in the visual world. A two-component framework for attentional deployment has recently emerged, although the idea dates back to William James[1], the father of American psychology. This framework suggests that subjects selectively direct attention to objects in a scene using both bottom-up, image-based saliency cues and top-down, task-dependent cues.

Some stimuli are intrinsically conspicuous or salient in a given context. For example, a red dinner jacket among black tuxedos at a sombre state affair, or a flickering light in an otherwise static scene, automatically and involuntarily attract attention. Saliency is independent of the nature of the particular task, operates very rapidly and is primarily driven in a bottom-up manner, although it can be influenced by contextual, figure–ground effects. If a stimulus is sufficiently salient, it will pop out of a visual scene. This suggests that saliency is computed in a pre-attentive manner across the entire visual field, most probably in terms of hierarchical CENTRE–SURROUND MECHANISMS. The speed of this saliency-based form of attention is on the order of 25 to 50 ms per item.

The second form of attention is a more deliberate and powerful one that has variable selection criteria, depending on the task at hand (for example, 'look for the red, horizontal target'). The expression of this top-down attention is most probably controlled from higher areas, including the frontal lobes, which connect back into visual cortex and early visual areas. Such volitional deployment of attention has a price, because the amount of time that it takes — 200 ms or more — rivals that needed to move the eyes. So, whereas certain features in the visual world automatically attract attention and are experienced as 'visually salient', directing attention to other locations or objects requires voluntary 'effort'. Both mechanisms can operate in parallel.

Attention implements an information-processing bottleneck that allows only a small part of the incoming sensory information to reach short-term memory and

Box 1 | **Neuronal mechanisms for the control of attention**



The brain regions that participate in the deployment of visual attention include most of the early visual processing area. A simplified overview of the main brain areas involved is shown in the figure. Visual information enters the primary visual cortex via the lateral geniculate nucleus (not shown), although smaller pathways, for example, to the superior colliculus (SC), also exist. From there, visual information progresses along two parallel hierarchical streams. Cortical areas along the 'dorsal stream' (including the posterior parietal cortex; PPC) are primarily concerned with spatial localization and directing attention and gaze towards objects of interest in the scene. The control of attentional deployment is consequently believed to mostly take place in the dorsal stream. Cortical areas along the 'ventral stream' (including the inferotemporal cortex; IT) are mainly concerned with the recognition and identification of visual stimuli. Although probably not directly concerned with the control of attention, these ventral-stream areas have indeed been shown to receive attentional feedback modulation, and are involved in the representation of attended locations and objects (that is, in what passes through the attentional bottleneck). In addition, several higher-function areas are thought to contribute to attentional guidance, in that lesions in those areas can cause a condition of 'neglect' in which patients seem unaware of parts of their visual environment (see REF. 111 for an overview of the regions involved).

From a computational viewpoint, the dorsal and ventral streams must interact, as scene understanding involves both recognition and spatial deployment of attention. One region where such interaction has been extensively studied is the prefrontal cortex (PFC). Areas within the PFC are bidirectionally connected to both the PPC and the IT (see REF. 15). So, in addition to being responsible for planning action (such as the execution of eye movements through the SC), the PFC also has an important role in modulating, via feedback, the dorsal and ventral processing streams.

visual awareness[10,11]. So, instead of attempting to fully process the massive sensory input (estimated to be on the order of 10^7–10^8 bits per second at the optic nerve) in parallel, a serial strategy has evolved that achieves near real-time performance despite limited computational capacity. Attention allows us to break down the problem of understanding a visual scene into a rapid series of computationally less demanding, localized visual analysis problems. In addition to these orientating and scene analysis functions, attention is also characterized by a feedback modulation of neural activity for the visual attributes and at the location of desired or selected targets. This feedback is believed to be essential for binding the different visual attributes of an object, such as colour and form, into a unitary percept[2,12,13]. By this account, attention not only serves to select a location of interest but also enhances the cortical representation of objects at that location. As such, focal visual attention has been compared to a 'stagelight', successively illuminating different players as they take centre stage[14]. Finally, attention is involved in triggering behaviour, and is consequently intimately related to recognition, planning and motor control[15].

Developing computational models that describe how attention is deployed within a given visual scene has been an important challenge for computational neuroscience. The potential application of these architectures in artificial vision for tasks such as surveillance, automatic target detection, navigational aids and robotic control provides additional motivation. Here, we focus on biologically plausible computational modelling of a saliency-based form of focal bottom-up attention. Much less is known about the neural instantiation of the top-down, volitional component of attention[16,17]. As this aspect of attention has not been modelled in such detail, it is not our primary focus here.

The control of focal visual attention involves an intricate network of brain areas (BOX 1). In a first approximation, selecting where to attend next is primarily controlled by the DORSAL STREAM of visual processing[18], although object recognition in the VENTRAL STREAM can bias the next attentional shift through top-down control. The basis of most computational models are the experimental results obtained using the visual search paradigm of Treisman and colleagues, in particular the distinction between pop-out and conjunctive searches developed in the early 1980s[2].

The first explicit, neurally plausible computational architecture for controlling visual attention was proposed by Koch and Ullman[19] in 1985 (FIG. 1) (for an earlier related model of vision and eye movements, see Didday and Arbib[20]). Koch and Ullman's model was centred around a 'saliency map', that is, an explicit two-dimensional topographical map that encodes stimulus conspicuity, or saliency, at every location in the visual scene. The saliency map receives inputs from early visual processing, and provides an efficient control strategy in which the focus of attention simply scans the saliency map in order of decreasing saliency. Following this basic architecture, we now illustrate five essential com-

DORSAL STREAM
Visual brain areas involved in the localization of objects and mostly found in the posterior/superior part of the brain.

VENTRAL STREAM
Visual brain areas involved in the identification of objects and mostly found in the posterior/inferior part of the brain.

OVERT ATTENTION
Expression of attention involving eye movements.

COVERT ATTENTION
Expression of attention without eye movements, typically thought of as a virtual 'spotlight'.

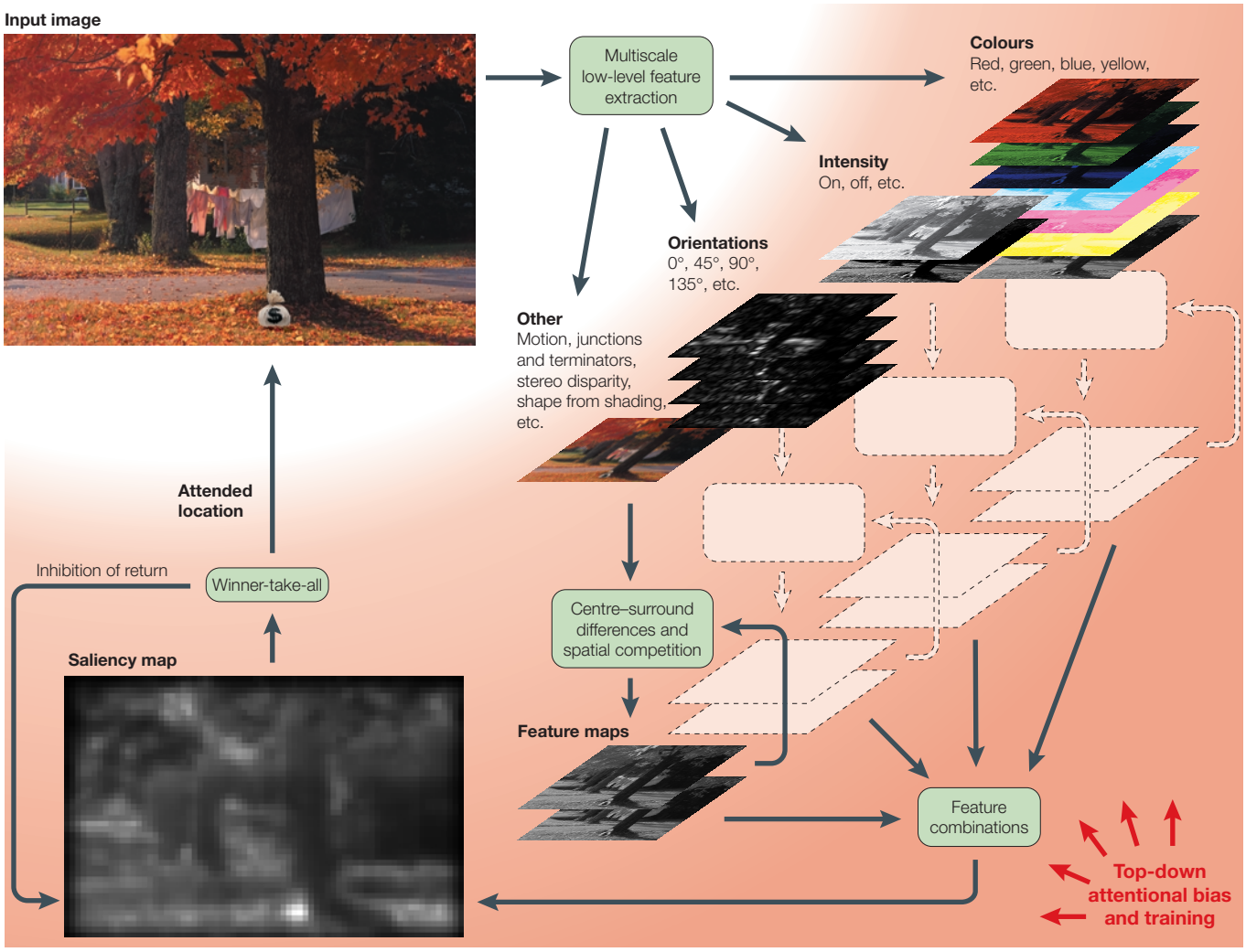


Figure 1 | **Flow diagram of a typical model for the control of bottom-up attention**. This diagram is based on Koch and Ullman's[19] hypothesis that a centralized two-dimensional saliency map can provide an efficient control strategy for the deployment of attention on the basis of bottom-up cues. The input image is decomposed through several pre-attentive feature detection mechanisms (sensitive to colour, intensity and so on), which operate in parallel over the entire visual scene. Neurons in the feature maps then encode for spatial contrast in each of those feature channels. In addition, neurons in each feature map spatially compete for salience, through long-range connections that extend far beyond the spatial range of the classical receptive field of each neuron (here shown for one channel; the others are similar). After competition, the feature maps are combined into a unique saliency map, which topographically encodes for saliency irrespective of the feature channel in which stimuli appeared salient. The saliency map is sequentially scanned by attention through the interplay between a winner-take-all network (which detects the point of highest saliency at any given time) and inhibition of return (which suppresses the last attended location from the saliency map, so that attention can focus onto the next most salient location). Top-down attentional bias and training can modulate most stages of this bottom-up model (red shading and arrows).

INTENSITY CONTRAST
Spatial difference (for example, detected by centre–surround mechanisms) in light intensity (luminance) in a visual scene.

COLOUR OPPONENCY
Spatial difference in colours, computed in the brain using red/green and blue/yellow centre–surround mechanisms.

NEURONAL TUNING
Property of visual neurons to only respond to certain classes of stimuli (for example, vertically orientated bars).

ponents of any model of bottom-up attention. These are the pre-attentive computation of early visual features across the entire visual scene, their integration to yield a single attentional control command, the generation of attentional scanpaths, the interaction between COVERT and OVERT attentional deployment (that is, eye movements) and the interplay between attention and scene understanding.

Pre-attentive computation of visual features

The first processing stage in any model of bottom-up attention is the computation of early visual features. In biological vision, visual features are computed in the retina, superior colliculus, lateral geniculate nucleus and early visual cortical areas[21]. Neurons at the earliest stages are tuned to simple visual attributes such as INTENSITY CONTRAST, COLOUR OPPONENCY, orientation, direction and velocity of motion, or stereo disparity at several spatial scales. NEURONAL TUNING becomes increasingly more specialized with the progression from low-level to high-level visual areas, such that higher-level visual areas include neurons that respond only to corners or junctions[22], shape-from-shading cues[23,24] or views of specific real-world objects[25–28].

Early visual features are computed pre-attentively in a massively parallel manner across the entire visual field (note, however, that we do not imply here that such computation is purely feedforward, as object recognition and attention can influence it[29]). Indeed, neurons fire vigorously in these early areas even if the

animal is attending away from the receptive field at the site of recording[30], or is anaesthetized[31]. In addition, several psychophysical studies, as well as introspection, indicate that we are not blind to the world outside the focus of attention. So we can make simple judgments on objects to which we are not attending[32], although those judgments are limited and less accurate than those made in the presence of attention[2,12,13,33–36]. Thus, although attention does not seem to be mandatory for early vision, it has recently become clear that attention can vigorously modulate, in a top-down manner, early visual processing, both in a spatially defined and in a non-spatial but feature-specific manner[37–39]. This modulatory effect of attention has been described as enhanced gain[30], biased[40,41] or intensified[33] competition, or enhanced spatial resolution[34], or as modulated background activity[42], effective stimulus strength[43] or noise[44]. That attention can modulate early visual processing in a manner equivalent to an increase of stimulus strength[43] is computationally an important finding, which directly supports the metaphor of attention as a stagelight. Of particular interest from a computational perspective is a recent study by Lee *et al.*[33] that measured PSYCHOPHYSICAL THRESHOLDS for three simple pattern-discrimination tasks (contrast, orientation and spatial-frequency discriminations) and two spatial-masking tasks (32 thresholds in total). A dual-task paradigm was used to measure thresholds either when attention was fully available to the task of interest, or when it was less available because it was engaged elsewhere by a concurrent attention-demanding task. The mixed pattern of attentional modulation observed in the thresholds (up to threefold improvement in orientation discrimination with attention, but only 20% improvement in contrast discrimination) can be quantitatively accounted for by a computational model. This model predicts that attention activates a winner-take-all competition among neurons tuned to different orientations and spatial frequencies within one cortical HYPERCOLUMN[33,45], a proposition that has recently received further experimental support[46]. Because feedback modulation influences the computation of bottom-up features, models of bottom-up attention need to take this into account. An example of a mixed bottom-up and top-down model in which attention enhances spatial resolution[47] is discussed later.

Computational models may or may not include explicit details about early visual feature extraction. Models that do not are restricted to images for which the responses of feature detectors can reasonably be guessed. Models that do have the widest applicability to any visual stimulus, including natural scenes. Computer implementations of early visual processes are often motivated by an imitation of biological properties. For example, the response of a neuron tuned to intensity centre–surround contrast can be computed by convolving the luminance channel of the input image by a DIFFERENCE-OF-GAUSSIANS (Mexican hat) filter. Similarly, the responses of orientation-selective neurons are usually obtained through convolution by GABOR WAVELETS, which resemble biological IMPULSE RESPONSE FUNCTIONS[48,49]. Another interesting approach consists of implementing detectors that respond best to those features that are present at the locations visited by observers while free-viewing images[50,51]. For instance, Zetzsche *et al.*[50,52] showed using an eye-tracking device how the eyes preferentially fixate regions with multiple superimposed orientations such as corners, and derived nonlinear operators that specifically detect those regions.

Irrespective of the method used for early feature detection, several fundamental computational principles have emerged from both experimental and modelling studies. First, different features contribute with different strengths to perceptual saliency[53], and this relative feature weighting can be influenced according to the demands of the task through top-down modulation[38,54] and through training[55–58]. Second, at a given visual location, there is little evidence for strong interactions across different visual modalities, such as colour and orientation[53]. This is not too surprising from a computational viewpoint, as one would otherwise expect these interactions to also be subject to training and top-down modulation, and this would result in the ability to learn to detect conjunctive targets efficiently, which we lack[2,59]. Within a given broad feature dimension, however, strong local interactions between filters sensitive to different properties of that feature (for example, between different orientations within the broad orientation feature) have been precisely characterized, both in physiology[60] and psychophysics[45]. Less evidence exists for within-feature competition across different spatial scales[45].

Last and most importantly, what seems to matter in guiding bottom-up attention is feature contrast rather than local absolute feature strength[61]. Indeed, not only are most early visual neurons tuned to some type of local spatial contrast (such as centre–surround or orientated edges), but neuronal responses are also strongly modulated by context, in a manner that extends far beyond the range of the classical receptive field (cRF)[62]. In a first approximation, the computational consequences of non-classical surround modulation are twofold. First, a broad inhibitory effect is observed when a neuron is excited with its preferred stimulus but that stimulus extends beyond the neuron's cRF, compared with when the stimulus is restricted to the cRF and the surrounding visual space is either empty or contains non-preferred stimuli[63–65]. Second, long-range excitatory connections in V1 seem to enhance the responses of orientation-selective neurons when stimuli extend to form a contour[66,67]. These interactions are thought to be crucial in perceptual grouping[68,69]. The net result is that activity in early cortical areas is surprisingly sparse when monkeys are free-viewing natural scenes[70], compared with the vigorous responses that can be elicited by small laboratory stimuli presented in isolation.

So, the computation of early visual features entails more than localized operations limited to the cRF of visual neurons, as local responses crucially depend on longer-range contextual influences. To explicitly demonstrate this idea with a computer model, Itti *et al.*[71] compared purely local spatial frequency 'richness'

PSYCHOPHYSICAL THRESHOLDS
Smallest difference between two visual stimuli (for example, vertical versus tilted bar) than can reliably (that is, with a given probability of error) be reported by an observer.

HYPERCOLUMN
A patch of cortex including neurons responding to all orientations and many spatial scales, all for a single location in the visual field.

DIFFERENCE-OF-GAUSSIANS
A filter obtained by taking the difference between a narrow Gaussian distribution (the excitatory centre) and a broader Gaussian distribution with the same mean (the inhibitory surround).

GABOR WAVELET
Product of a sinusoidal grating and a two-dimensional Gaussian function.

IMPULSE RESPONSE FUNCTION
The response of a filter to a single pulse (Dirac) stimulus.

(as measured by computing local Fourier components with a magnitude above a certain threshold) with a saliency measure that included broad non-classical surround inhibition. They designed images with uniformly rich spatial frequency content (using colour speckle noise), but which contained a perceptually salient target. Although the target was undifferentiated from its surround in terms of spatial frequency content, it was correctly detected by the mechanism that included contextual competition.

Pre-attentive mechanisms that extract early visual features across the entire visual scene should not be overlooked in future modelling efforts. Indeed, it has recently become clear that early vision is far from being a passive and highly prototyped image-processing front-end that can be accurately modelled by linear filtering operations. Perceptually, whether a given stimulus is salient or not cannot be decided without knowledge of the context in which the stimulus is presented. So, computationally, one must also account for nonlinear interactions across distant spatial locations, which mediate contextual modulation of neuronal responses.

Saliency

We have seen how the early stages of visual processing decompose the incoming visual input through an ensemble of feature-selective filtering processes endowed with contextual modulatory effects. The question that arises next is how to control a single attentional focus based on multiple neuronal networks that encode the incoming sensory signals using multiple representations. To solve this problem, most models of bottom-up attention follow Koch and Ullman[19] and hypothesize that the various feature maps feed into a unique 'saliency' or 'master' map[2,19]. The saliency map is a scalar, two-dimensional map whose activity topographically represents visual saliency, irrespective of the feature dimension that makes the location salient. That is, an active location in the saliency map encodes the fact that this location is salient, no matter whether it corresponds to a red object in a field of green objects, or to a stimulus moving towards the right while others move towards the left. On the basis of this scalar topographical representation, biasing attention to focus onto the most salient location is reduced to drawing attention towards the locus of highest activity in the saliency map.

Computationally, an explicit representation of saliency in a dedicated map reinforces the idea that some amount of spatial selection should be performed during pre-attentive feature detection. Otherwise, the divergence from retinal input to many feature maps could not be followed by a convergence into a saliency map without ending up with a representation in the saliency map that is as complex, cluttered and difficult to interpret as the original image. On the basis of this divergence, selection and convergence process, a location is defined as salient if it wins the spatial competition in one or more feature dimensions at one or more spatial scales. The saliency map then encodes an aggregate measure of saliency that is not tied to any particular feature dimension, thereby providing an efficient control strategy for focusing attention to salient locations without consideration of the detailed feature responses that made those locations salient.

Not surprisingly, many successful models of the bottom-up control of attention are built around a saliency map. What differentiates the models, then, is the strategy used to prune the incoming sensory input and extract saliency. In an influential model largely aimed at explaining visual search experiments, Wolfe[54] 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. Saliency is then computed in this model as the likelihood that a target will be present at a given location, based both on bottom-up feature contrast and top-down feature weight. This view has recently received experimental support from the many studies of top-down attentional modulation mentioned earlier.

Tsotsos and colleagues[72] implemented attentional selection using a combination of a feedforward bottom-up feature extraction hierarchy and a feedback selective tuning of these feature extraction mechanisms. In this model, the target for attention is selected at the top level of the processing hierarchy (the equivalent of a saliency map), on the basis of feedforward activation and possibly 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 saliency is thus refined at each level of processing, as the feedforward paths that do not contribute to the winning location are pruned (resulting in the feedback propagation of an 'inhibitory beam' around the selected target).

Milanese and colleagues[73] used a relaxation process to optimize an energy measure consisting of four terms: first, minimizing inter-feature incoherence favours those regions that excite several feature maps; second, minimizing intra-feature incoherence favours grouping of initially spread activity into small numbers of clusters; third, minimizing total activity in each map enforces intra-map spatial competition for saliency; and last, maximizing the dynamic range of each map ensures that the process does not converge towards uniform maps at some average value. Although the biological plausibility of this process remains to be tested, it has yielded a rare example of a model that can be applied to natural colour images.

Itti *et al.*[71,74] consider a purely bottom-up model, in which spatial competition for saliency is directly modelled after non-classical surround modulation effects. The model uses an iterative spatial competition 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, a nonlinear process that ensures that the locations losing the competition are entirely eliminated. The net effect of this competitive process is similar to a winner-take-all process with limited

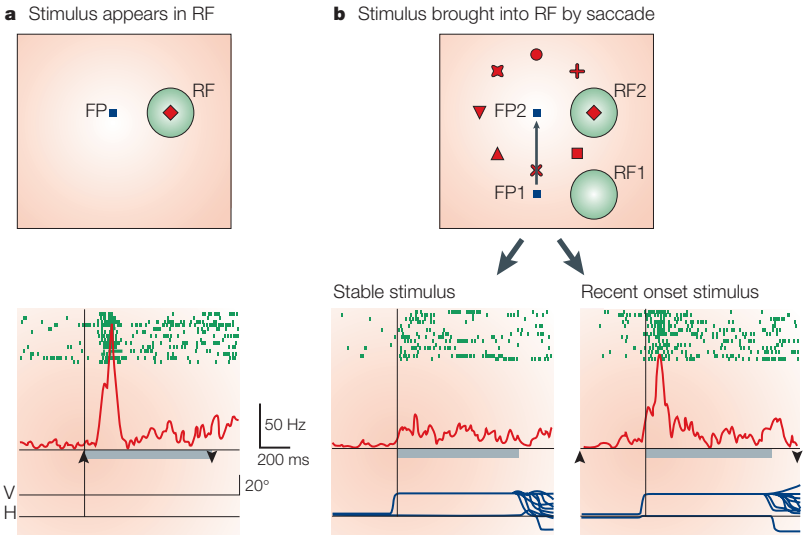


Figure 2 | **Recording saliency.** Once a purely computational hypothesis, the idea that saliency might be explicitly encoded by specific neurons in the cortex has recently received experimental support from many electrophysiological studies[77–82]. How can one design an experiment that specifically tests whether a neuron responds to the saliency of a stimulus, rather than to the mere presence of that stimulus in the visual environment? In a particularly interesting experiment, Gottlieb and colleagues[80], recording from the lateral intraparietal sulcus of the awake monkey, found neurons that responded to visual stimuli only when those stimuli were made salient (by rapidly flashing them on a computer screen), but not otherwise. Their experiment cleverly used the retinotopic nature of the receptive fields of these neurons to bring a stimulus into their receptive field (RF) through a SACCADIC EYE MOVEMENT. **a** | In the control condition, with the eyes stable at the fixation point (FP), a stimulus is presented in the RF of the neuron being recorded from. The response elicited by this stimulus could be simply visual, or indicate the saliency of this stimulus suddenly appearing in the visual field. **b** | To differentiate between these possibilities, two additional experiments were designed to be identical for the neuron of interest: a stimulus entered the RF through a saccade. However, a vigorous response was observed only when the stimulus had been made salient shortly before the beginning of the trial (by flashing it on and off while it was still outside the RF of the neuron; 'recent onset' condition). (Adapted with permission from REF. 80 © (1998) Macmillan Magazines Ltd.)

inhibitory spread, and allows only a sparse population of locations to remain active. After competition, all feature maps are simply summed to yield the scalar saliency map at the core of the model. Because it includes a complete front-end, this model has been widely applied to the analysis of natural colour scenes. Experimental results include the reproduction by the model of human behaviour in classical visual search tasks (pop-out versus conjunctive search, and search asymmetries[2,74]), a demonstration of very robust saliency computation with respect to image noise[71], the automatic detection of traffic signs and other salient objects in natural environments[58] and the detection of pedestrians in natural scenes. Finally, the performance of the model at detecting military vehicles in the high-resolution Search2 NATO database of colour rural scenes[75] exceeded human performance, in terms of the estimated number of locations that need to be visited by the attentional searchlight before the target is located[74].

In view of the numerous models based on a saliency map, it is important to note that postulating centralized control based on such a map is not the only computational alternative for the bottom-up guidance of attention. In particular, Desimone and Duncan[10] argued that saliency is not explicitly represented by specific neurons and by a saliency map, but instead is implicitly coded in a distributed modulatory manner across the various feature maps. Attentional selection is then performed on the basis of top-down enhancement of the feature maps relevant to a target of interest and extinction of those that are distracting, but without an explicit computation of salience. At least one model successfully applied this strategy to synthetic stimuli[76]; note, however, that such top-down biasing (also used in Wolfe's Guided Search model to select the weights of various feature contributions to the saliency map) requires that a specific search task be performed for the model to yield useful predictions.

Although originally a theoretical construct supported by sparse experimental evidence, the idea of a unique, centralized saliency map seems today to be challenged by the apparent existence of multiple areas that encode stimulus saliency in the visual system of the monkey. These regions include areas in the lateral intraparietal sulcus of the posterior parietal cortex (FIG. 2), the frontal eye fields, the inferior and lateral subdivisions of the pulvinar and the superior colliculus[77–82].

One possible explanation for this multiplicity could be that some of the neurons in these areas are indeed concerned with the explicit computation of saliency, but are located at different stages along the sensorimotor processing stream. For example, other functions have also been assigned to the posterior parietal cortex, such as that of mapping retinotopic to head-centred coordinate systems and of memorizing targets for eye or arm movements[83,84]. So more detailed experimental studies are needed to reveal subtle differences in the functions and representations found in these brain areas. Most probably, the main difference between these brain regions is the balance between their role in perception and action[15,82]. Meanwhile, it is worth noting that, in addition to the physiological findings just mentioned, recent psychophysical results also support the idea of an explicit encoding of saliency in the brain[85].

Attentional selection and inhibition of return

The saliency map guides where the attentional stagelight or spotlight[86] is to be deployed, that is, to the most salient location in the scene. One plausible neural architecture to detect the most salient location is that of a winner-take-all network, which implements a neurally distributed maximum detector[19,87]. Using this mechanism, however, raises another computational problem: how can we prevent attention from permanently focusing onto the most active (winner) location in the saliency map? One efficient computational strategy, which has received experimental support, consists of transiently inhibiting neurons in the saliency map at the currently attended location. After the currently attended location is thus suppressed, the winner-take-all network naturally converges towards the next most salient location, and repeating this process generates attentional scanpaths[19,71].

Such inhibitory tagging of recently attended locations has been widely observed in human psychophysics as a phenomenon called 'inhibition of return' (IOR)[88,89]. A typical psychophysical experiment to

SACCADIC EYE MOVEMENT
Very rapid, ballistic eye movement (with speeds up to 800 degrees per second).

Box 2 | **Attention and eye movements**

Most of the models and experiments reviewed in this article are concerned with covert attention, that is, shifts of the focus of attention in the absence of eye movements. In normal situations, however, we move our eyes 3–5 times per second (that is, 150,000 to 250,000 times every day), to align locations of interest with our foveas. Overt and covert attention are closely related, as revealed by psychophysical[112–115] physiological[79,81,83,116] and imaging[117,118] studies. The neuronal structures involved 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. An example of overlapping functionality in humans is the study by Hoffman and Subramaniam[114]. They designed an experiment in which subjects performed a saccade just preceded by a target detection task; the greater accuracy found when the target appeared at the endpoint of the saccade suggests that covert attention had been deployed to that endpoint in preparation for the saccade.

For models, the addition of eye movements poses several additional computational challenges. Of particular interest is the need for compensatory mechanisms to shift the saliency map (typically in retinotopic coordinates) as eye movements occur. Dominey and Arbib[119] proposed a biologically plausible computational architecture that could perform such dynamic remapping in posterior parietal cortex (PPC). They noted that eye velocity signals have not been found in PPC; however, cells modulated by eye position have been reported[83]. They thus devised an iterative scheme to shift the contents of the saliency map according to the difference between current eye position and a temporally damped eye position signal. Their algorithm builds a convolution kernel from the difference between current and damped eye positions, which, when applied to the saliency map, translates it in the direction opposite to that difference. A related approach was proposed by Pouget and Sejnowski[84], in which the observed modulation of neuronal responses in PPC by retinal location and eye position ('gain field'[83]) is modelled by a set of basis functions, then used to transform from retinotopic to head-centred coordinates.

The interaction between overt and covert attention is particularly important for models concerned with visual search[120–122]. Further modelling of such interactions promises a better understanding of many mechanisms, including saccadic suppression, dynamic remapping of the saliency map and inhibition of return, covert pre-selection of targets for overt saccades, and the online understanding of complex visual scenes.

evaluate IOR consists of performing speeded local pattern discriminations at various locations in the visual field; when a discrimination is performed at a location to which the observer has been previously cued, reaction times are slightly, but significantly, higher than at locations not previously visited[90]. These results indicate that visual processing at recently attended locations might be slower, possibly owing to some inhibitory tagging at attended locations. Several authors have specifically isolated an attentional component of IOR in addition to a motor (response delay) component[91–93].

Computationally, IOR implements a short-term memory of the previously visited locations and allows the attentional selection mechanism to focus instead on new locations. The simplest implementation of IOR consists of triggering transient inhibitory conductances in the saliency map at the currently attended location[74]. However, this only represents a coarse approximation of biological IOR, which has been shown to be object-bound, so that it should track and follow moving objects, and compensate for a moving observer as well[94–97]. The frame of reference in which IOR is expressed is an important issue when the eyes and the body move (BOX 2). This frame-of-reference problem should be accounted for in computational models of attention. Note that the idea of IOR is not necessarily contradicted by the recent findings of Horowitz and Wolfe[98] that visual search seems memoryless: when elements of a search array were randomly reorganized at short time intervals while subjects were searching for a specific target, search efficiency was not degraded compared with when the search array remained stationary. Although these results preclude perfect memorization of all previously attended locations (otherwise, search on a stable array should be more efficient than on a constantly changing array), they do not preclude an explanation in which the positions of the last few visited items were remembered, in accordance with the limited lifespan reported for IOR[90].

Although simple in principle, IOR is computationally a very important component of attention, in that it allows us — or a model — to rapidly shift the attentional focus over different locations with decreasing saliency, rather than being bound to attend only to the location of maximal saliency at any given time. The role of IOR in active vision and overt attention poses challenges that will need to be addressed in more detail by future models (BOX 2).

Attention and recognition

So far, we have reviewed computational modelling and supporting experimental evidence for a basic architecture concerned with the bottom-up control of attention: early visual features are computed in a set of topographical feature maps; spatial competition for saliency prunes the feature responses to only preserve a handful of active locations; all feature maps are then combined into a unique scalar saliency map; and, finally, the saliency map is scanned by the focus of attention through the interplay between winner-take-all and IOR. Although such a simple computational architecture might accurately describe how attention is deployed within the first few hundreds of milliseconds after the presentation of a new scene, it is obvious that a more complete model of attentional control must include top-down, volitional biasing influences as well. The computational challenge, then, lies in the integration of bottom-up and top-down cues, such as to provide coherent control signals for the focus of attention, and in the interplay between attentional orientating and scene or object recognition.

One of the earliest models that combines object recognition and attention is MORSEL[99], in which attentional selection was shown to be necessary for object recognition. This model is applied to the recognition of words processed through a recognition hierarchy. Without attentional selection, the representations of several words in a scene would conflict and confuse that recognition hierarchy, yielding multiple superimposed representations at the top level. The addition of a top-down attentional selection process allowed the model to disambiguate recognition by focusing on one word at a time. Another early model that is worth mentioning here is described in REF. 100.

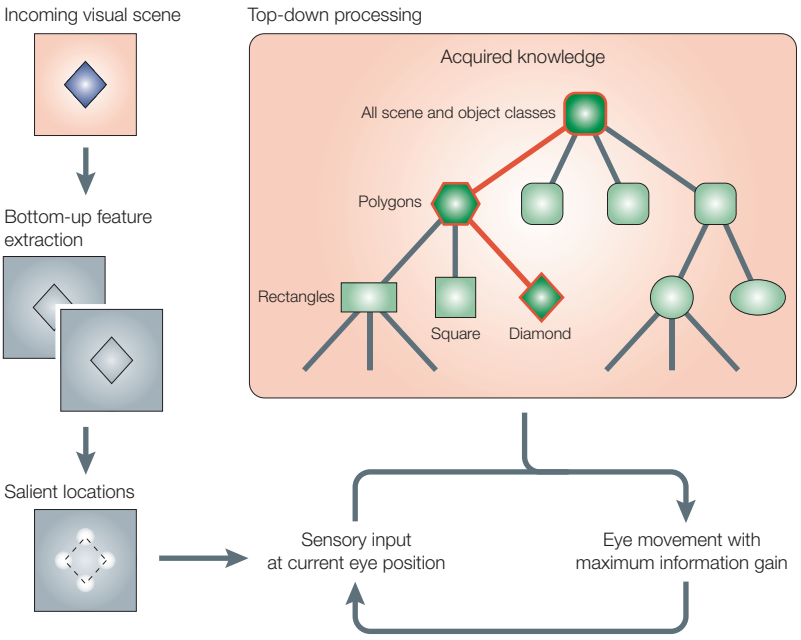


Figure 3 | **Combined model of attentional selection and object recognition.** Attention scans the scene to gather as much information as possible that can help discriminate between several recognition hypotheses. The model has two main components. First, a bottom-up feature extraction pathway extracts informative image regions from an incoming visual scene (for example, the corners of the diamond in the present illustration). Second, a trained knowledge base hierarchically represents object classes and encodes for both expected visual features at a set of critical points on these objects, and motor commands to move the eyes from one critical point to another. Recognition is then achieved by choosing for the next eye movement the movement that maximizes information gain, that is, that best prunes the tree of known objects. In the hypothetical example shown to illustrate this idea, the first eye movement might thus go to the top corner of the object; finding sharp edges there would then suggest that this is a polygon. The knowledge base would then direct gaze to the most salient point directly below the currently fixated point, as that eye movement would best discriminate between the several known polygons; looking at the orientations of the features there, it becomes clear that the object is a diamond rather than a square or one of several possible rectangles. (Adapted with permission from REF. 101.)

A very interesting model that uses spatial shifts of attention during recognition was recently provided by Schill *et al.*[101]. Their model performs scene (or object) recognition, using attention (or eye movements) to focus on those parts of the scene that are most informative for disambiguating identity. To this end, a hierarchical knowledge tree is built through training. Its leaves represent identified objects, intermediary nodes represent more general object classes and links between nodes contain sensorimotor information used for discrimination between possible objects (that is, bottom-up feature response to be expected for particular points in the object and eye movements targeted at those points). During the iterative recognition of an object, the system programs its next fixation towards the location that will maximize the gain of information about the object. This permits the model to discriminate between the current candidate object classes (FIG. 3).

Rybak *et al.*[102] proposed a related model, 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. When presented with a new image, the model starts by selecting candidate scanpaths by matching bottom-up features in the image to those stored in the 'what' memory. For each candidate scanpath, the model deploys attention according to the directives in the 'where' memory and compares the local contents of the 'what' memory at each fixation with the local image features. This model can recognize complex greyscale scenes and faces in a translation-, rotation- and scale-independent manner.

Deco and Zihl have recently proposed another model that combines attentional selection and object recognition[47]. Their model starts by selecting candidate object locations in a bottom-up manner through a coarse-scale analysis of the image. An attentional mechanism scans the candidate locations in a serial fashion and performs object recognition at progressively finer scales until a sufficient recognition score is obtained for an object stored in memory. This model has been successfully applied to psychophysical experiments that show attentional enhancement of spatial resolution (see also REFS 34,103 for related experiments and modelling).

A more extreme view is expressed by the 'scanpath theory' of Stark[104], in which the control of eye movements is almost exclusively under top-down control. The theory proposes that what we see is only remotely related to the patterns of activation in our retinas. This is suggested by our permanent illusion of vivid perception over the entire field of view, although only the central two degrees of our foveal vision provide crisp sampling of the visual world. Rather, the scanpath theory argues that a cognitive model of what we expect to see is the basis for our percept; the sequence of eye movements that we make to analyse a scene, then, is mostly controlled by our cognitive model of that scene. This theory has had several 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[105].

One important challenge for combined models of attention and recognition is finding suitable neuronal correlates for the various components. Despite the biological inspiration in these architectures, the models reviewed here do not relate in much detail to biological correlates of object recognition. Although several biologically plausible models have been proposed for object recognition in the ventral 'what' stream (in particular, REFS 106,107), their integration with neurobiological models concerned with attentional control in the dorsal 'where' stream remains an open issue. This integration will, in particular, have to account for the increasing experimental support for an object-based spatial focus of attention[108–110].

Summary

We have discussed recent advances in the study of biologically plausible computational models of attention, with a particular emphasis on bottom-up control of attentional deployment. Throughout this review, we have stressed five important computational trends that

have emerged from the literature. First, saliency is derived from low-level visual features but, more than absolute feature strength or other detailed characteristics of the features, what seems to be important for the computation of saliency is feature contrast with respect to the contextual surround. Second, saliency increasingly seems to be a quantity that is coded explicitly in cortex and separate from the visual features. This reinforces the once hypothetical concept of an explicit saliency map. Furthermore, several models have demonstrated the computational usefulness and plausibility of such an explicit map by successfully reproducing the behaviour of humans and monkeys in search tasks. Meanwhile, neural analogues of the saliency map are being found at multiple locations in the visual system of the macaque. This poses a new challenge of how to integrate these many maps to yield unitary behaviour. Third, attention will not shift unless the currently attended (most salient) location is somehow disabled (otherwise, any model looking for saliency will keep coming back to the most salient location). Inhibition of return is consequently an essential computational component of attention and, indeed, it has been recently described as a complex, object-based and dynamically adaptive process that needs to be better modelled. Fourth, covert attention and eye movements are increasingly believed to share a common neuronal substrate. This poses serious computational problems with respect to the frame of reference in which saliency and IOR are computed. Recent evidence for world-centred and object-centred frames of reference need to be integrated into models. Last, the control of attentional deployment is intimately related to scene understanding and object recognition. Although several computer vision models have been proposed that integrate both attentional orientating and object identification, many exciting research challenges still lie in attempting to provide a more complete account of the interactions between the dorsal and ventral processing streams in primate brains.

Controlling where attention should be deployed is not an autonomous feedforward process. Important future directions for modelling work include modelling of interactions between task demands and top-down cues, bottom-up cues, mechanistic constraints (for example, when eye and body movements are executed) and neuroanatomical constraints such as feedback modulation.

Links

FURTHER INFORMATION Laurent Itti's lab | Christof Koch's lab | Supplementary material for Figure 2

1. James, W. *The Principles of Psychology* (Harvard Univ. Press, Cambridge, Massachusetts, 1980/1981).
2. Treisman, A. M. & Gelade, G. A feature-integration theory of attention. *Cogn. Psychol.* **12**, 97–136 (1980).
An influential theory of attention and visual search.
3. Bergen, J. R. & Julesz, B. Parallel versus serial processing in rapid pattern discrimination. *Nature* **303**, 696–698 (1983).
4. Treisman, A. Features and objects: The fourteenth Bartlett memorial lecture. *Q. J. Exp. Psychol. A* **40**, 201–237 (1988).
5. Nakayama, K. & Mackeben, M. Sustained and transient components of focal visual attention. *Vision Res.* **29**, 1631–1647 (1989).
6. Braun, J. & Sagi, D. Vision outside the focus of attention. *Percept. Psychophys.* **48**, 45–58 (1990).
7. Hikosaka, O., Miyauchi, S. & Shimojo, S. Orienting a spatial attention — its reflexive, compensatory, and voluntary mechanisms. *Brain Res. Cogn. Brain Res.* **5**, 1–9 (1996).
8. Braun, J. & Julesz, B. Withdrawing attention at little or no cost: detection and discrimination tasks. *Percept. Psychophys.* **60**, 1–23 (1998).
9. Braun, J., Itti, L., Lee, D. K., Zenger, B. & Koch, C. in *Visual Attention and Neural Circuits* (eds Braun, J., Koch, C. & Davis, J.) (MIT, Cambridge, Massachusetts, in the press).
10. Desimone, R. & Duncan, J. Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* **18**, 193–222 (1995).
A complete review on attention.
11. Crick, F. & Koch, C. Constraints on cortical and thalamic projections: the no-strong-loops hypothesis. *Nature* **391**, 245–250 (1998).
12. Hummel, J. E. & Biederman, I. Dynamic binding in a neural network for shape recognition. *Psychol. Rev.* **99**, 480–517 (1992).
13. Reynolds, J. H. & Desimone, R. The role of neural mechanisms of attention in solving the binding problem. *Neuron* **24**, 19–29 (1999).
14. Weichselgartner, E. & Sperling, G. Dynamics of automatic and controlled visual attention. *Science* **238**, 778–780 (1987).
15. Miller, E. K. The prefrontal cortex and cognitive control. *Nature Rev. Neurosci.* **1**, 59–65 (2000).
16. Hopfinger, J. B., Buonocore, M. H. & Mangun, G. R. The neural mechanisms of top-down attentional control. *Nature Neurosci.* **3**, 284–291 (2000).
17. Corbetta, M., Kincade, J. M., Ollinger, J. M., McAvoy, M. P. & Shulman, G. L. Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nature Neurosci.* **3**, 292–297 (2000); erratum **3**, 521 (2000).
18. Ungerleider, L. G. & Mishkin, M. in *Analysis of Visual Behavior* (eds Ingle, D. J., Goodale, M. A. & Mansfield, R. J. W.) 549–586 (MIT, Cambridge, Massachusetts, 1982).
19. Koch, C. & Ullman, S. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.* **4**, 219–227 (1985).
One of the first explicit computational models of bottom-up attention, at the origin of the idea of a 'saliency map'.
20. Didday, R. L. & Arbib, M. A. Eye movements and visual perception: A "two visual system" model. *Int. J. Man–Machine Studies* **7**, 547–569 (1975).
21. Suder, K. & Worgotter, F. The control of low-level information flow in the visual system. *Rev. Neurosci.* **11**, 127–146 (2000).
22. Pasupathy, A. & Connor, C. E. Responses to contour features in macaque area v4. *J. Neurophysiol.* **82**, 2490–2502 (1999).
23. Braun, J. Shape-from-shading is independent of visual attention and may be a 'texton'. *Spat. Vis.* **7**, 311–322 (1993).
24. Sun, J. & Perona, P. Early computation of shape and reflectance in the visual system. *Nature* **379**, 165–168 (1996).
25. Logothetis, N. K., Pauls, J. & Poggio, T. Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* **5**, 552–563 (1995).
26. Bar, M. & Biederman, I. Localizing the cortical region mediating visual awareness of object identity. *Proc. Natl Acad. Sci. USA* **96**, 1790–1793 (1999).
27. Vogels, R., Biederman, I., Bar, M. & Lorincz, A. Inferior temporal neurons show greater sensitivity to nonaccidental than metric differences. *J. Cogn. Neurosci.* (in the press).
28. Kreiman, G., Koch, C. & Fried, I. Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neurosci.* **3**, 946–953 (2000).
29. He, Z. J. & Nakayama, K. Perceiving textures: beyond filtering. *Vision Res.* **34**, 151–162 (1994).
30. Treue, S. & Maunsell, J. H. Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature* **382**, 539–541 (1996).
31. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol. (Lond.)* **160**, 106–154 (1962).
32. DeSchepper, B. & Treisman, A. Visual memory for novel shapes: implicit coding without attention. *J. Exp. Psychol. Learn. Mem. Cogn.* **22**, 27–47 (1996).
33. Lee, D. K., Itti, L., Koch, C. & Braun, J. Attention activates winner-take-all competition among visual filters. *Nature Neurosci.* **2**, 375–381 (1999).
A detailed neural model is used to quantitatively predict attentional modulation of psychophysical pattern discrimination performance in terms of intensified competition between visual neurons.
34. Yeshurun, Y. & Carrasco, M. Attention improves or impairs visual performance by enhancing spatial resolution. *Nature* **396**, 72–75 (1998).
35. Mack, A., Tang, B., Tuma, R., Kahn, S. & Rock, I. Perceptual organization and attention. *Cogn. Psychol.* **24**, 475–501 (1992).
36. Moore, C. M. & Egeth, H. Perception without attention: evidence of grouping under conditions of inattention. *J. Exp. Psychol. Hum. Percept. Perform.* **23**, 339–352 (1997).
37. Motter, B. C. Neural correlates of attentive selection for color or luminance in extrastriate area V4. *J. Neurosci.* **14**, 2178–2189 (1994).
38. Treue, S. & Trujillo, J. C. M. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* **399**, 575–579 (1999).
Investigates two types of feedback attentional modulation: spatial-based, and non-spatial but feature-based.
39. Barcelo, F., Suwazono, S. & Knight, R. T. Prefrontal modulation of visual processing in humans. *Nature Neurosci.* **3**, 399–403 (2000).
40. Moran, J. & Desimone, R. Selective attention gates visual processing in the extrastriate cortex. *Science* **229**, 782–784 (1985).
41. Niebur, E., Koch, C. & Rosin, C. An oscillation-based model for the neuronal basis of attention. *Vision Res.* **33**, 2789–2802 (1993).
42. Chawla, D., Rees, G. & Friston, K. J. The physiological basis of attentional modulation in extrastriate visual areas. *Nature Neurosci.* **2**, 671–676 (1999).
43. Reynolds, J. H., Pasternak, T. & Desimone, R. Attention increases sensitivity of V4 neurons. *Neuron* **26**, 703–714 (2000).
44. Dosher, B. A. & Lu, Z. L. Mechanisms of perceptual attention in precuing of location. *Vision Res.* **40**, 1269–1292 (2000).
45. Itti, L., Koch, C. & Braun, J. Revisiting spatial vision: towards a unifying model. *J. Opt. Soc. Am. A* **17**, 1899–1917 (2000).
46. Carrasco, M., Penpeci-Talgar, C. & Eckstein, M. Spatial covert attention increases contrast sensitivity across the CSF: support for signal enhancement. *Vision Res.* **40**, 1203–1215 (2000).

47. Deco, G. & Zihl, J. A neurodynamical model of visual attention: Feedback enhancement of spatial resolution in a hierarchical system. *J. Comp. Neurosci.* (in the press).
48. Daugman, J. G. Spatial visual channels in the Fourier plane. *Vision Res.* **24**, 891–910 (1984).
49. Palmer, L. A., Jones, J. P. & Stepnoski, R. A. in *The Neural Basis of Visual Function* (ed. Leventhal, A. G.) 246–265 (CRC, Boca Raton, Florida, 1991).
50. Zetzsche, C. *et al*. Investigation of a sensorimotor system for saccadic scene analysis: an integrated approach. *Proc. 5th Int. Conf. Simulation Adaptive Behav*. **5**, 120–126 (1998).
51. Reinagel, P. & Zador, A. M. Natural scene statistics at the centre of gaze. *Network Comp. Neural Syst.* **10**, 341–350 (1999).
52. Barth, E., Zetzsche, C. & Rentschler, I. Intrinsic two-dimensional features as textons. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* **15**, 1723–1732 (1998).
53. Nothdurft, H. Salience from feature contrast: additivity across dimensions. *Vision Res.* **40**, 1183–1201 (2000).
Psychophysical study of how orientation, motion, luminance and colour contrast cues combine to yield the saliency of visual stimuli.
54. Wolfe, J. M. Visual search in continuous, naturalistic stimuli. *Vision Res.* **34**, 1187–1195 (1994).
55. Braun, J. Vision and attention: the role of training. *Nature* **393**, 424–425 (1998).
56. Ahissar, M. & Hochstein, S. The spread of attention and learning in feature search: effects of target distribution and task difficulty. *Vision Res.* **40**, 1349–1364 (2000).
57. Sigman, M. & Gilbert, C. D. Learning to find a shape. *Nature Neurosci.* **3**, 264–269 (2000).
58. Itti, L. & Koch, C. Feature combination strategies for saliency-based visual attention systems. *J. Electronic Imaging* (in the press).
59. Wolfe, J. in *Attention* (ed. Pashler, H.) 13–74 (University College London, London, 1996).
60. Carandini, M. & Heeger, D. J. Summation and division by neurons in primate visual cortex. *Science* **264**, 1333–1336 (1994).
61. Nothdurft, H. C. Texture discrimination by cells in the cat lateral geniculate nucleus. *Exp. Brain Res.* **82**, 48–66 (1990).
62. Allman, J., Miezin, F. & McGuinness, E. Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local–global comparisons in visual neurons. *Annu. Rev. Neurosci.* **8**, 407–430 (1985).
One of the first reports that activity of a visual neuron can be modulated by the presence of distant stimuli, far outside the neuron's receptive field.
63. Cannon, M. W. & Fullenkamp, S. C. Spatial interactions in apparent contrast: inhibitory effects among grating patterns of different spatial frequencies, spatial positions and orientations. *Vision Res.* **31**, 1985–1998 (1991).
64. Sillito, A. M., Grieve, K. L., Jones, H. E., Cudeiro, J. & Davis, J. Visual cortical mechanisms detecting focal orientation discontinuities. *Nature* **378**, 492–496 (1995).
65. Levitt, J. B. & Lund, J. S. Contrast dependence of contextual effects in primate visual cortex. *Nature* **387**, 73–76 (1997).
66. Gilbert, C. D. & Wiesel, T. N. Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *J. Neurosci.* **9**, 2432–2442 (1989).
67. Gilbert, C., Ito, M., Kapadia, M. & Westheimer, G. Interactions between attention, context and learning in primary visual cortex. *Vision Res.* **40**, 1217–1226 (2000).
68. Ben-Av, M. B., Sagi, D. & Braun, J. Visual attention and perceptual grouping. *Percept. Psychophys.* **52**, 277–294 (1992).
69. Grossberg, S. & Raizada, R. D. Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex. *Vision Res.* **40**, 1413–1432 (2000).
70. Vinje, W. E. & Gallant, J. L. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* **287**, 1273–1276 (2000).
71. Itti, L., Koch, C. & Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Patt. Anal. Mach. Intell.* **20**, 1254–1259 (1998).
72. Tsotsos, J. K. *et al*. Modeling visual-attention via selective tuning. *Artif. Intell.* **78**, 507–545 (1995).
73. Milanese, R., Gil, S. & Pun, T. Attentive mechanisms for dynamic and static scene analysis. *Opt. Eng.* **34**, 2428–2434 (1995).
74. Itti, L. & Koch, C. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res.* **40**, 1489–1506 (2000).
75. Toet, A., Bijl, P., Kooi, F. L. & Valeton, J. M. *A High-Resolution Image Dataset for Testing Search and Detection Models (TNO-TM-98–A020)* (TNO Human Factors Research Institute, Soesterberg, The Netherlands, 1998).
76. Hamker, F. H. in *Proc. 5th Neural Comp. Psychol. Workshop (NCPW'98)* (eds von Heinke, D., Humphreys, G. W. & Olson, A.) 252–261 (Springer Verlag, London, 1999).
77. Laberge, D. & Buchsbaum, M. S. Positron emission tomographic measurements of pulvinar activity during an attention task. *J. Neurosci.* **10**, 613–619 (1990).
78. Robinson, D. L. & Petersen, S. E. The pulvinar and visual salience. *Trends Neurosci.* **15**, 127–132 (1992).
79. Kustov, A. A. & Robinson, D. L. Shared neural control of attentional shifts and eye movements. *Nature* **384**, 74–77 (1996).
80. Gottlieb, J. P., Kusunoki, M. & Goldberg, M. E. The representation of visual salience in monkey parietal cortex. *Nature* **391**, 481–484 (1998).
Electrophysiological experiments in the awake monkey indicating that some neurons explicitly encode for saliency in the posterior parietal cortex.
81. Colby, C. L. & Goldberg, M. E. Space and attention in parietal cortex. *Annu. Rev. Neurosci.* **22**, 319–349 (1999).
82. Thompson, K. G. & Schall, J. D. Antecedents and correlates of visual detection and awareness in macaque prefrontal cortex. *Vision Res.* **40**, 1523–1538 (2000).
83. Andersen, R. A., Bracewell, R. M., Barash, S., Gnadt, J. W. & Fogassi, L. Eye position effects on visual, memory, and saccade-related activity in areas lip and 7a of macaque. *J. Neurosci.* **10**, 1176–1196 (1990).
84. Pouget, A. & Sejnowski, T. J. Spatial transformations in the parietal cortex using basis functions. *J. Cogn. Neurosci.* **9**, 222–237 (1997).
85. Blaser, E., Sperling, G. & Lu, Z. L. Measuring the amplification of attention. *Proc. Natl Acad. Sci. USA* **96**, 11681–11686 (1999).
86. Brefczynski, J. A. & DeYoe, E. A. A physiological correlate of the 'spotlight' of visual attention. *Nature Neurosci.* **2**, 370–374 (1999).
87. Amari, S. & Arbib, M. A. in *Systems Neuroscience* (ed. Metzler, J.) 119–165 (Academic, New York, 1977).
88. Posner, M. I. & Cohen, Y. in *Attention and Performance* Vol. X (eds Bouma, H. & Bouwhuis, D.) 531–556 (Erlbaum, Hillsdale, New Jersey, 1984).
89. Kwak, H. W. & Egeth, H. Consequences of allocating attention to locations and to other attributes. *Percept. Psychophys.* **51**, 455–464 (1992).
90. Klein, R. M. Inhibition of return. *Trends Cogn. Sci.* **4**, 138–147 (2000).
A complete review of inhibition of return.
91. Shimojo, S., Tanaka, Y. & Watanabe, K. Stimulus-driven facilitation and inhibition of visual information processing in environmental and retinotopic representations of space. *Brain Res. Cogn. Brain Res.* **5**, 11–21 (1996).
92. Kingstone, A. & Pratt, J. Inhibition of return is composed of attentional and oculomotor processes. *Percept. Psychophys.* **61**, 1046–1054 (1999).
93. Taylor, T. L. & Klein, R. M. Visual and motor effects in inhibition of return. *J. Exp. Psychol. Hum. Percept. Perform.* **26**, 1639–1656 (2000).
94. Tipper, S. P., Driver, J. & Weaver, B. Object-centred inhibition of return of visual attention. *Q. J. Exp. Psychol. A* **43**, 289–298 (1991).
95. Gibson, B. S. & Egeth, H. Inhibition of return to object-based and environment-based locations. *Percept. Psychophys.* **55**, 323–339 (1994).
96. Ro, T. & Rafal, R. D. Components of reflexive visual orienting to moving objects. *Percept. Psychophys.* **61**, 826–836 (1999).
97. Becker, L. & Egeth, H. Mixed reference frames for dynamic inhibition of return. *J. Exp. Psychol. Hum. Percept. Perform.* **26**, 1167–1177 (2000).
98. Horowitz, T. S. & Wolfe, J. M. Visual search has no memory. *Nature* **394**, 575–577 (1998).
99. Mozer, M & Sitton, S. in *Attention* (ed. Pashler, H.) 341–393 (University College London, London, 1996)
100. Guigon, E., Grandguillaume, P., Otto, I., Boutkhil, L. & Burnod, Y. Neural network models of cortical functions based on the computational properties of the cerebral cortex. *J. Physiol. (Paris)* **88**, 291–308 (1994).
101. Schill, K., Umkehrer, E., Beinlich, S., Krieger, G. & Zetzsche, C. Scene analysis with saccadic eye movements: top-down and bottom-up modeling. *J. Electronic Imaging* (in the press).
102. Rybak, I. A., Gusakova, V. I., Golovan, A. V., Podladchikova, L. N. & Shevtsova, N. A. A model of attention-guided visual perception and recognition. *Vision Res.* **38**, 2387–2400 (1998).
103. Deco, G. & Schurmann, B. A hierarchical neural system with attentional top-down enhancement of the spatial resolution for object recognition. *Vision Res.* **40**, 2845–2859 (2000).
104. Stark, L. W. & Choi, Y. S. in *Visual Attention and Cognition* (eds Zangemeister, W. H., Stiehl, H. S. & Freska, C.) 3–69 (Elsevier Science B. V., Amsterdam, 1996).
105. Stark, L. W. *et al*. Representation of human vision in the brain: how does human perception recognize images? *J. Electronic Imaging* (in the press).
106. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nature Neurosci.* **2**, 1019–1025 (1999).
107. Riesenhuber, M. & Poggio, T. Models of object recognition. *Nature Neurosci.* **S3**, 1199–1204 (2000).
108. O'Craven, K. M., Downing, P. E. & Kanwisher, N. fmri evidence for objects as the units of attentional selection. *Nature* **401**, 584–587 (1999).
109. Roelfsema, P. R., Lamme, V. A. & Spekreijse, H. Object-based attention in the primary visual cortex of the macaque monkey. *Nature* **395**, 376–381 (1998).
110. Abrams, R. A. & Law, M. B. Object-based visual attention with endogenous orienting. *Percept. Psychophys.* **62**, 818–833 (2000).
111. Webster, M. J. & Ungerleider, L. G. in *The Attentive Brain* (ed. Parasuraman, R.) 19–34 (MIT, Cambridge, Massachusetts, 1998).
112. Shepherd, M., Findlay, J. M. & Hockey, R. J. The relationship between eye movements and spatial attention. *Q. J. Exp. Psychol.* **38**, 475–491 (1986).
113. Sheliga, B. M., Riggio, L. & Rizzolatti, G. Orienting of attention and eye movements. *Exp. Brain Res.* **98**, 507–522 (1994).
114. Hoffman, J. E. & Subramaniam, B. The role of visual attention in saccadic eye movements. *Percept. Psychophys.* **57**, 787–795 (1995).
115. Kowler, E., Anderson, E., Dosher, B. & Blaser, E. The role of attention in the programming of saccades. *Vision Res.* **35**, 1897–1916 (1995).
116. Schall, J. D., Hanes, D. P. & Taylor, T. L. Neural control of behavior: countermanding eye movements. *Psychol. Res.* **63**, 299–307 (2000).
117. Corbetta, M. Frontoparietal cortical networks for directing attention and the eye to visual locations: identical, independent, or overlapping neural systems? *Proc. Natl Acad. Sci. USA* **95**, 831–838 (1998).
118. Nobre, A. C., Gitelman, D. R., Dias, E. C. & Mesulam, M. M. Covert visual spatial orienting and saccades: overlapping neural systems. *Neuroimage* **11**, 210–216 (2000).
119. Dominey, P. F. & Arbib, M. A. A cortico-subcortical model for generation of spatially accurate sequential saccades. *Cereb. Cortex* **2**, 153–175 (1992).
120. Motter, B. C. & Belky, E. J. The guidance of eye movements during active visual search. *Vision Res.* **38**, 1805–1815 (1998).
121. Gilchrist, I. D., Heywood, C. A. & Findlay, J. M. Saccade selection in visual search: evidence for spatial frequency specific between-item interactions. *Vision Res.* **39**, 1373–1383 (1999).
122. Wolfe, J. M. & Gancarz, G. in *Basic and Clinical Applications of Vision Science* (ed. Lakshminarayanan, V.) 189–192 (Kluwer Academic, Dordrecht, The Netherlands, 1996).

Acknowledgements

The research carried out in the laboratories of the authors on visual attention is supported by the National Science Foundation, the National Institute of Mental Health and the Office of Naval Research. We thank Alex Pouget for excellent comments and suggestions.

Bios

Laurent Itti received his M.S. in Image Processing from the Ecole Nationale Superieure des Telecommunications, Paris, in 1994, and Ph.D. in Computation and Neural Systems from Caltech, Pasadena, California, in 2000. In September 2000, he became assistant professor of Computer Science at the University of Southern California, Los Angeles, California. His primary research interest is in biologically plausible computational brain modelling, and in the comparison of model simulations to empirical measurements from living systems. Of particular interest in his laboratory is the development of computational models of biological vision with applications to machine vision.

Christof Koch is the Lois and Victor Troendle Professor of Cognitive and Behavioural Biology at the California Institute of Technology. He obtained his Ph.D. in Physics from the University of Tübingen, Germany. His research focuses on understanding the biophysical mechanisms underlying information processing in individual nerve cells as well as the neuronal operations underlying spatial vision, motion, shape perception and visual attention in the primate visual system, using electrophysiological, brain imaging, psychophysical and computational tools. Together with Dr Francis Crick, he works on the neuronal basis of visual consciousness.

Summary

- We review recent work on computational models of focal visual attention, with emphasis on the bottom-up, saliency- or image-based control of attentional deployment. We highlight five important trends that have emerged from the computational literature:
- First, the perceptual saliency of stimuli critically depends on surrounding context; that is, the same object may or may not appear salient depending on the nature and arrangement of other objects in the scene. Computationally, this means that contextual influences, such as non-classical surround interactions, must be included in models.
- Second, a unique 'saliency map' topographically encoding for stimulus conspicuity over the visual scene has proved to be an efficient and plausible bottom-up control strategy. Many successful models are based on such architecture, and electrophysiological as well as psychophysical studies have recently supported the idea that saliency is explicitly encoded in the brain.
- Third, inhibition-of-return (IOR), the process by which the currently attended location is transiently inhibited, is a critical element of attentional deployment. Without IOR, attention would endlessly be attracted towards the most salient stimulus. IOR thus implements a memory of recently visited locations, and allows attention to thoroughly scan our visual environment.
- Fourth, attention and eye movements tightly interplay, posing computational challenges with respect to the coordinate system used to control attention. Understanding the interaction between overt and covert attention is particularly important for models concerned with visual search.
- Last, scene understanding and object recognition strongly constrain the selection of attended locations. Although several models have approached, in an information-theoretical sense, the problem of optimally deploying attention to analyse a scene, biologically plausible implementations of such a computational strategy remain to be developed.

Links

Laurent Itti's lab
http://ilab.usc.edu/

Christof Koch's lab
http://www.klab.caltech.edu/

Supplementary information for Figure 2
http://ilab.usc.edu/itti/nrn/